ANADOLU ÜNİVERSİTESİ

Açıköğretim Uygulamaları ve Araştırmaları Dergisi
AUAd

auad.anadolu.edu.tr

Açıköğretim sistemindeki açık uçlu soruların Çok Yüzeyli Rasch Modeli analizi yöntemiyle puanlanması

Araş. Gör. Kübra KARAKAYA ÖZYER[a]

[a]Eskişehir Osmangazi Üniversitesi

Özet

Açık uçlu soruların oluşturduğu sınavların puanlanmasında, bir puanlayıcıdan beklenen soruların zorluk derecesinden ve değerlendirme yöntemlerinden bağımsız olarak değerlendirme aşamasını nesnel ve adil olarak gerçekleştirmesidir. 2016-2017 eğitim öğretim yılı itibariyle Açıköğretim Fakültesindeki bazı sınavlarda açık uçlu sorular kullanılmaya başlanmış ve bu denli nesnel yargılamaya açık sınavların değerlendirilmesinde yeni yöntemler arayışı içine girilmiştir. Mevcut yöntemlerden en etkililerinden biri de farklı yüzeyleri analize dâhil edebilen Çok Yüzeyli Rasch Modelidir. Bu çalışmada Rasch modelinin geliştirilmiş formu olan çok yüzeyli Rasch analizinin açık uçlu soruların değerlendirilmesi açısından açıköğretim sistemine sağlayacağı katkı açıklanmaktadır.

Anahtar Sözcükler: Açık uçlu sorular, Çok yüzeyli Rasch modeli, Madde tepki kuramı

Abstract

In order to obtain a score for an open-ended question, a judge should assess examinee's performance objectively regardless of item difficulty and grading method. Open education system at Anadolu University started to use constructed-response items in their assessment process. Since there are some measurement problems for judgement of open-ended questions, it is important to apply objective methods to improve reliability and validity. One of the effective method is Multi-facet Rasch Model which includes facets (variables) into analysis simultenously. This article presents an extention of the Rasch model, Multi-facet Rasch model, and its contribution of the open education system at Anadolu University.

Keywords: Constructed-response items, Multi-facet Rasch Model, Item response theory

Kaynak Gösterme

Karakaya Özyer, K. (2018). Açıköğretim sistemindeki açık uçlu soruların Çok Yüzeyli Rasch Modeli analizi yöntemiyle puanlanması. *AUAd, 4*(1), 61-77.

Giriş

Geniş ölçekli sınavların ölçme ve değerlendirme aşamasında hatadan arınık olmak ve geçerli sonuçlar elde etmek önemli ölçütlerdir. Tüm aşamaların dikkatle planlanması ve izlenmesine karşın hataların ölçme sonuçlarına etki etmesi kaçınılmazdır (Güler, 2012). Çoktan seçmeli sınavlarda öğrencilerin bireysel başarı düzeylerinden kaynaklanan hatalar ön plana çıkarken yazılı-yoklama sınavlarında cevapları değerlendirenlerin bireysel eğilimlerinden kaynaklanan hatalar dikkate alınması gereken önemli bir konu olarak karşımıza çıkmaktadır (Goodwin, 2001).

Açık Uçlu Soruların Değerlendirilmesi

Eğitimde sıklıkla kullanılan açık uçlu soru tipinden oluşan sınavlarda nesnel puanlandırma oldukça önemlidir. Açık uçlu maddeler, cevaplayıcılara kendilerini özgür bir biçimde ifade etme olanağı sunmakta ve üst düzey bilişsel becerilerin ölçülmesine olanak sağlamaktadır (Tan ve Erdoğan, 2004; Turgut ve Baykul, 2012). Tüm bu getirilerine karşın açık uçlu soruların en belirgin sakıncası değerlendirmede nesnelliğin sağlanamaması ve puanlayıcı yanlılığıdır (Atılgan, Kan ve Doğan, 2009). Puanlayıcılardan, maddelerin zorluklarına ve öğrencilerin özelliklerine bakmaksızın yanıtları değerlendirmeleri beklenmektedir (Lunz, Wright ve Linacre, 1990). Bu tür maddeleri içeren sınavlarda, puanlayıcıların yüksek puanları üst düzey performans gösteren cevaplayıcılara vermesi beklenir. Ancak birçok çalışma gösteriyor ki puanlayıcılar klasik tür sınavları değerlendirmekte hatalar yapmakta ve aynı performansı gösteren kişilere farklı puanlar verme eğilimi göstermektedir (Eckes, 2005; Engelhard, 1992; Lunz vd., 1990). Puanlayıcılara verilen eğitimlerin ölçme sonuçlarını olumsuz yönde etki edebilecek hataları en aza indirgemesi beklense de bu eğitimlerin etkinliği konusunda farklı sonuçlara ulaşılmıştır (Lunz vd., 1990). Sonuç olarak, hedeflenen ve hatadan arındırılmış puanları elde etmek ve istenmeyen varyasyonun (hatanın) etkisini azaltmak için etkili yöntemlere gereksinim duyulmuştur.

Açıköğretim Fakültesinin Yeni Uygulamaları

Anadolu Üniversitesi Açıköğretim sistemi, uzaktan eğitim almak isteyen bireylere olanak sunmaktadır. Anadolu Üniversitesi Açıköğretim Fakültesi 2016-2017 eğitim öğretim yılında çoktan seçmeli soruların yanı sıra açık uçlu ve kısa cevaplı soruları da sınav sistemlerine dâhil etmiştir. Öncelik olarak dört farklı programda (Tarih, Edebiyat, Sosyoloji

ve Felsefe) uygulanan bu yeni sistemde yanıtlayıcılar çoktan seçmeli sorulara ek olarak 3 adet açık uçlu soruyu da yanıtlamak durumundadır. İlk iki soru kısa yanıt gerektiren sorular olmakla birlikte üçüncü soru daha kapsamlı cümleler kurulması beklenen açık-uçlu soru olarak tasarlanmıştır. Bu son sorunun sistematik olarak değerlendirilebilmesi için açık yönerge hazırlanmıştır.

Açıköğretim sınavlarındaki açık uçlu (klasik) sorular öğrencinin bilgi düzeyine dair kapsamlı bilgi verseler de bu tür soruların sonuçlarının geçerliliği ile ilgili sorunlar ortaya çıkabilmektedir. Böylesine geniş ölçekli ve yüksek risk içeren sınavlarda, açık uçlu maddelerin kullanılmasıyla ilgili önemli sorunlardan biri öznel puanlayıcı etkisidir (Reynolds, Livingston ve Willson, 2009). Birçok araştırmanın sonuçlarına göre kapsamlı ve iyi yapılandırılmış eğitim seminerlerine karşın puanlayıcılar arasındaki farklılıklar göz ardı edilemez boyutlara ulaşmıştır (Lunz vd., 1990). Şekil 3. Açıköğretim fakültesinde yeni uygulamaya konan açık uçlu soruların değerlendirilmesine dair kavram haritasını ifade etmektedir.

Açık Uçlu Soruların Değerlendirilmesinde Puanlayıcı Etkisi

Açık öğretim fakültesinin sınavlarındaki açık uçlu soruların değerlendirilmesi için 100'e yakın uzman görevlendirilmektedir ve sınava katılan her bir bireyin cevapları en az iki puanlayıcıya gönderilmektedir. Tüm yanıtlar puanlayıcılardan gelen notların ortalamaları alınarak puana dönüştürülmektedir. Sistemde puanlayıcılar arası tutarsızlığı gidermeye dair belirli adımlar atılmış olsa da halo etkisi, yanlılık, kendi içindeki tutarsızlık, ranj sınırlaması ve merkeze yönelme gibi puanlayıcı etkilerini belirlemekte iyileştirmeye gereksinim duyulmaktadır. Halo etkisi, puanlayıcıların farklı özelliklere sahip bir bireylerin cevaplarına benzer puanları verme eğilimi göstermesidir (Eckes, 2009; Engelhard, 2002). Merkeze yönelme etkisi ise puanlayıcıların en yüksek ve en düşük puanları vermekten kaçınması ve yanıtlayıcı hangi performans düzeyinde olursa olsun merkez yakın olan puanları vermek istemeleridir (Eckes, 2009). Ranj sınırlaması etkisi merkeze yönelme etkisine benzemekle birlikte puanlayıcıların sadece ölçeğin orta noktasını kullanmasını değil herhangi bir noktasını tutarlı bir biçimde kullanmasını ifade etmektedir (İlhan, 2015). Diğer bir ifade ile merkeze yönelme etkisi ranj sınırlaması etkisinin özel bir halidir. Tutarsızlık etkisi ise herhangi bir puanlayıcının kullanılan ölçeği diğer değerlendiricilerden farklı bir mantıkla kullanması ve tesadüfi hatalara yol açılmasıdır (Eckes, 2015). Adından da anlaşılacağı üzere, yanlılık etkisi puanlayıcıların belirli özellikteki öğrencilere daha katı veya daha cömert

davranmasıdır (Eckes, 2015). Bu özellikler kimi zaman yaş, etnik köken, cinsiyet kimi zaman da öğrencinin diğer sorulara verdiği cevap olabilmektedir. Hangi etki olursa olsun tüm etki türleri istenmeyen varyasyona sebep olduğu için ölçme işleminin güvenirliği etkilemektedir.

Puanlayıcı Etkisini Azaltmak İçin Yöntemler

Puanlayıcılar arasındaki ve puanlayıcıların kendi içindeki bu tutarsızlıkları belirleyecek bazı yaklaşımlar mevcuttur. Bu yaklaşımlar klasik test kuramına dayalı yöntemler, genellenebilirlik kuramına dayalı yöntemler ve son olarak madde tepki kuramına dayalı yöntemler olmak üzere üç sınıfa ayrılmıştır (Swanlund, 2016). Klasik test kuramına dayalı yöntemlerden puanlayıcılar arası güvenirlik katsayısı bulma (inter-rater reliability) ve Tek yönlü Varyans analizi (ANOVA) açık uçlu soruların analiz edilmesinde en çok kullanılan yöntemlerdendir. Ek olarak, genellenebilirlik kuramı çalışmaları grup düzeyindeki puanlayıcıların etkisini ortaya koymada yararlı veriler sunmaktadır. Ne yazık ki, tüm bu yöntemler öğrencilerin becerilerinin ham puanlarını temel alarak analiz yapmaktadır. Madde tepki kuramına (MTK) dayalı yöntemler ise bireysel puanlayıcı etkisini belirlemekte ve daha detaylı sonuçlar vermektedir. Madde zorluklarını temel alan Rasch modeli, Madde tepki kuramı modelinin özel bir formu olarak bilinir (Chapman, Letourneau ve Sheidow, 2013). Linacre (1989), Rasch modelinin geliştirilmiş bir modelini farklı yüzeyleri analize dahil ederek geliştirmiş ve bu yeni model Çok Yüzeyli Rasch Modeli (ÇYRM, Multi-facet Rasch Model-MFRM) olarak adlandırılmıştır (Chapman vd., 2013). Bu modelin en önemli getirisi ise diğer yöntemlerin çözümleyemediği ölçme sorunlarına çözüm getirmesidir (Schumacker, 1996). Tablo 1'de Klasik test kuramı, Genellenebilirlik teorisi ve Çok Yüzeyli Rasch Modelinin açık uçlu soruların değerlendirilmesinde sunduğu imkânları göstermektedir.

Tablo 1

Klasik Test Kuramı, Genellenebilirlik Teorisi ve Çok Yüzeyli Rasch Modeli Karşılaştırılması (Mulqueen, Baker ve Dismukes, 2002)

Yüzeyler	Puanlayıcılar arası güvenirlik	Genellenebilirlik Teorisi	Çok yüzeyli Rasch Modeli
Puanlayıcılar	X	X	X
Görevler		X	X
Kriterler		X	X
Etkileşim		X	X
Ölçmenin odağı			
Grup düzeyi	X	X	
Bireysel düzey			X

Madde Tepki Kuramı ve Çok Yüzeyli Rasch Modeli

Çok yüzeyli Rasch modeli (ÇYRM, MFRM), birden fazla değişkenin diğer bir adıyla yüzeyin çözümleme sürecine dâhil edilmesiyle elde edilen bir yöntemdir. Bu modelde söz edilen yüzey kavramı, ölçme sürecinde değerlendirme sonuçlarına etkisi olabilecek bir boyut olarak tanımlanmıştır (Eckes, 2015). Örneğin, puanlayıcılar, öğrencilerin özellikleri ya da öğrencilerin yapması beklenen etkinlikler ÇYRM yönteminde yüzey olarak adlandırılmaktadır. Her yüzey diğer yüzeylerden tamamıyla bağımsız olmalıdır. Diğer yandan, analize dâhil edilen tüm yüzeyler birbirleri ile karşılaştırabilmeleri için aynı ölçek üzerinde birleştirilebilmelidir (Schumacker, 1999).

Genellenebilirlik teorinin aksine, çok yüzeyli Rasch analizi puanlayıcıların bireysel etkilerini daha yakından inceleme olanağı sunmaktadır (Engelhard ve Myford, 2003; Linacre, 1993). Özellikle, ÇYRM puanlayıcıların yanıtları değerlendirirken yaptıkları hataları belirlemede etkilidir. Dahası, ÇYRM puanlayıcılardan bağımsız olarak öğrencilerin gerçek performanslarını elde edebilmektedir (Eckes, 2015).

Çok yüzeyli Rasch modeli yöntemi olağan değişim aralığında bulunmayan gözlemleri belirlemede etkili bir yaklaşımdır. Dahası, ÇYRM her yüzey için grup-düzeyinde ve bireysel-düzeydeki etkileri analiz edebilmektedir. Bu yöntem puanlayıcıların puanlama örüntüsünü ortaya çıkartma olanağı da sunmaktadır. Ek olarak, ÇYRM yaklaşımı her bir yüzeyin ölçme ortamına katkısını ayrıştırarak bu yüzeylerin istenilen düzeydeki katkılarını araştırmaya

olanak sağlamaktadır (Myford ve Wolfe, 2003). Örneğin, ÇYRM analizi puanlayıcıların öğrencilerin yanıtlarını değerlendirirken puanlama katılık derecesinde farklılık olup olmadığını ortaya koyabilmektedir. Hatta farklılık tespit edilen durumlardaki puanlayıcıların kaç farklı katılık düzeyinde değerlendirme yaptıklarını da göstermektedir. Diğer bir deyişle çok yüzeyli Rasch analizi hangi puanlayıcıların daha katı hangilerinin daha cömert puanlama yaptığını ortaya koymaya yardımcı olmaktadır (Wolfe ve Chiu, 1997). Bu yaklaşım puanlayıcıların tutarsızlıklarının yanı sıra halo etkisi gibi durumlara maruz kalan öğrencileri de ortaya çıkartabilmektedir (Eckes, 2015; Myford ve Wolfe, 2003). Sonuç olarak, ÇYRM analizi puanlayıcıların öğrencilerin yanıtlarını değerlendirirken yaptıkları hataları belirleyip öğrencilerin aldıkları puanlarda uyarlama yapabilmektedir.

Çok yüzeyli Rasch analizi lojistik regresyona benzemekle birlikte bu analiz öğrencilerin ve puanlayıcıların pozisyonlarını aynı anda gösteren bir harita (kalibrasyon haritası, değişken haritası, logit ölçeği) oluşturmaktadır. ÇYRM yaklaşımında gözlenen değerlerin lojistik transformasyonları bağımlı değişken ve yüzeyler (örneğin puanlayıcılar) bağımsız değişkenler olarak kabul edilir (Behizadeh ve Engelhard, 2014). Logit ölçek (kalibrasyon haritası, değişken haritası) gerçek aralıklı bir ölçek olduğu için (Kondo-Brown, 2002) tüm yüzeylerin ve yüzeylere ait elemanların eşzamanlı karşılaştırılmasını görselleştirmeye olanak sağlamaktadır (Güler, 2014).

Diğer madde tepki kuramı modelleri gibi Rasch modeli için de model-veri uyumu, tek boyutluluk ve yerel bağımsızlık varsayımlarının sağlanması gerekmektedir (Eckes, 2015). ÇYRM aşağıdaki gibi ifade edilmektedir:

Eşitlik 1.1

$$ln = \left[\frac{p_{nijk}}{p_{nijk-1}}\right] = \theta_n - \beta_i - \alpha_j - \tau_k$$

p_{nijk} = j. puanlayıcı tarafından n. Öğrencinin i. görevine k kriterini verilme olasılığı

p_{nijk-1} =j. puanlayıcı tarafından n. öğrencinini. görevine k-1 kriterini verilme olasılığı

θ_n = n. öğrencinin yeterlilik düzeyi,

β_i = i. etkinliğin zorluk derecesi,

α_j = j. puanlayıcının katılık derecesi,

τ_k = k-1 kriteri yerine k kriterini alma zorluğu.

1.1'deki eşitlik basit bir doğrusal eklemeli model olmakla birlikte logit ölçek negatif sonsuzdan pozitif sonsuza kadar gitmektedir. Logit ölçekteki değerler her yüzey için farklı yorumlanmaktadır. Örneğin, puanlayıcı yüzeyindeki pozitif log-odd'lar katı puanlamayı temsil ederken, negatif log-odda sahip puanlayıcılar daha cömert puan veren olarak tanımlanmaktadır. Öte yandan, öğrenci performanslarına dair pozitif log-odd değerleri yüksek beceriye sahip öğrencileri, negatif log-odd değerleri ise düşük beceriye sahip öğrencileri ifade etmektedir.

Öncelikle, verinin model ile genel uyumu standardize edilmiş artıklar aracılığıyla incelenmelidir (Engelhard ve Myford, 2003; İlhan, 2015). Büyük standartlaştırılmış artık değerler beklenmeyen sonuçları veya uç değerleri ifade etmektedir. Linacre'e (2004) göre kabul edilebilir model uyumunu göstermenin iki gereksinimi vardır. İlk olarak, 2 ve 2'den fazla olan standartlaştırılmış artık değerlerin tüm verilerde en fazla %5 düzeyinde olmalı ve ek olarak 3 ve 3'ten büyük olan standardize artıklar en fazla %1 düzeyinde olmalıdır (Eckes, 2005; Linacre, 2004).

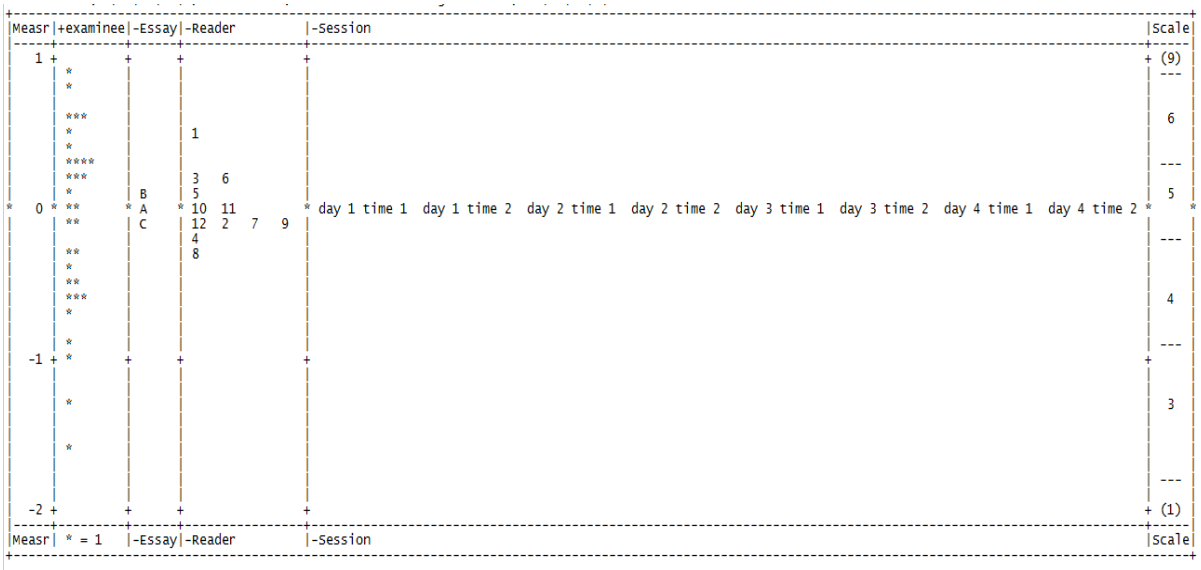
Açık uçlu soruların değerlendirilmesi ile ilgili kavram haritası Şekil 3'te ayrıntılı bir biçimde belirtilmiştir.

FACETS Programı ve Kullanımı

Açık uçlu soruların analizini çok yüzeyli Rasch modeli yöntemi ile analiz etmek için Linacre (1993) FACETS programını geliştirmiştir. Bu programın çalışma ilkesi Linacre'nin (2008) yayınladığı kılavuz niteliğindeki kitabında mevcuttur. 2017 yılı itibari ile FACETS programının 3.80.3 versiyonu piyasaya sürülmüştür (ayrıntılar için bakınız www.winsteps.com). FACETS açık uçlu sorulardan elde edilen parametre tahminlerini ve grup düzeyindeki ve bireysel düzeydeki hataları belirleyerek ölçümlerin güvenirliği ve geçerliliğine dair veri sağlayan uygunluk istatistikleri sunmaktadır.

FACETS programı ilk olarak belirli yinelemeler (iterasyonlar, iteration) sonucunda verilerin model ile uyumunu göstermektedir. Sonrasında değişkenlerin diğer bir ifadeyle yüzeylerin bütüncül olarak değerlendirilebildiği ve birbiriyle karşılaştırılabildiği bir değişken haritası üretmektedir(İlhan, 2015). Bu harita logit ölçek olarak da adlandırılmakta ve sütun kısmındaki değerler yüzeyleri temsil etmektedir. Diğer bir deyiş ile değişken haritasında analize dâhil edilen yüzey kadar sütun bulunmaktadır (Eckes, 2015). Örneğin, Şekil 1 dört yüzeyli bir Rasch analizi sonuç çıktısındaki değişken haritasını göstermektedir. Örnek analiz

için yüzeyler, öğrenciler, komposizyon soruları, puanlayıcılar ve sınav zamanları olarak belirlenmiştir.



Şekil 1. Örnek Veri Kalibrasyon haritası

Şekil 1'deki örnek harita (ölçek) incelendiğinde puanlayıcıların katılık-cömertlik anlamında birbirinden farklılaştığı gözlenmektedir. Ayrıca sınavın yapıldığı zamanların aynı satırda yan yana sıralanması istatistiksel anlamda birbirlerinden farklı olmadıkları sonucunu göstermektedir.

FACETS programı yardımıyla yapılan ÇYRM analizi sonucunda bireysel düzeyde fit indeksleri elde edilmektedir. Bu indeksler ölçüm raporları tablosunda gösterilmektedir (Şekil 2). Ölçüm raporlarında puanlayıcıların öğrencilere verdiği puanların toplamı (total score), her bir puanlayıcının kaç adet puanlama yaptığı (total count), öğrencilerin gözlenen ortalama puanları (ObsvdAverage) ve öğrencilerin ÇYRM sonucunda elde edilmiş diğer bir ifadeyle düzeltilmiş puanları (Fairaverage) listelenmiştir. Düzeltilmiş öğrenci puanları bireyin beceri kestirimi ve puanlayıcıların katılık-cömertlik dereceleri dikkate alınarak elde edilmiştir. + Ölçüm (+ measure) sütunundaki değerler mevcut puanlayıcıların logit ölçeği üzerindeki konumunu göstermektedir. Kalibrasyon haritasından da anlaşıldığı gibi bu sütundaki değerlere bakılarak puanlayıcılar katılık ve cömertlik derecelerine göre kategorize edilebilmektedir.

Total Score	Total Count	Obsvd Average	Fair(M) Average	Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Correlation PtMea	Correlation PtExp	Exact Agree. Obs %	Exact Agree. Exp %	Nu	Reader
508	96	5.29	5.26	-.30	.08	1.23	1.6	1.21	1.4	.75	.60	.62	20.8	20.4	8	8
485	96	5.05	5.00	-.16	.08	.52	-4.2	.53	-4.1	1.48	.67	.62	21.2	21.7	4	4
484	96	5.04	4.99	-.15	.08	1.02	.1	1.01	.0	.97	.64	.62	24.1	21.6	9	9
479	96	4.99	4.93	-.12	.08	1.13	.9	1.13	.9	.83	.55	.62	28.8	21.7	7	7
473	96	4.93	4.86	-.08	.08	1.06	.5	1.06	.4	.93	.56	.62	20.8	21.7	2	2
470	96	4.90	4.83	-.06	.08	1.40	2.6	1.37	2.4	.63	.65	.62	27.8	22.0	12	12
466	96	4.85	4.79	-.04	.08	1.14	.9	1.11	.8	.81	.62	.61	30.6	21.9	11	11
461	96	4.80	4.73	.00	.08	.71	-2.3	.71	-2.2	1.31	.68	.61	42.4	21.9	10	10
444	96	4.63	4.55	.11	.08	.85	-1.1	.84	-1.1	1.14	.60	.61	36.1	22.0	5	5
434	96	4.52	4.44	.17	.08	1.04	.3	1.06	.4	.93	.65	.60	38.9	21.9	6	6
433	96	4.51	4.43	.18	.08	1.06	.4	1.03	.2	.99	.64	.60	27.8	21.8	3	3
392	96	4.08	4.00	.45	.08	.79	-1.5	.79	-1.5	1.23	.48	.59	19.7	19.8	1	1
460.8	96.0	4.80	4.73	.00	.08	1.00	-.1	.99	-.2		.61				Mean (Count: 12)	
29.5	.0	.31	.32	.19	.00	.23	1.8	.22	1.7		.05				S.D. (Population)	
30.8	.0	.32	.33	.20	.00	.24	1.9	.23	1.8		.06				S.D. (Sample)	

Model, Populn: RMSE .08 Adj (True) S.D. .17 Separation 2.17 Strata 3.22 Reliability (not inter-rater) .82
Model, Sample: RMSE .08 Adj (True) S.D. .18 Separation 2.28 Strata 3.38 Reliability (not inter-rater) .84
Model, Fixed (all same) chi-square: 66.3 d.f.: 11 significance (probability): .00
Model, Random (normal) chi-square: 9.4 d.f.: 10 significance (probability): .49
Inter-Rater agreement opportunities: 384 Exact agreements: 108 = 28.1% Expected: 82.6 = 21.5%

Şekil 2. Örnek Puanlayıcı Ölçüm Raporları

Ayırma güvenirliği - ayırma indeksi. Global model uyumunun yanı sıra, ayırma indeksi (seperation index), ayırma güvenirliği (seperation reliability) ve kikare homojenlik test istatistikleri (chi-square) de FACETS programından elde edilmektedir. Ayırma indeksi (oranı) ve ayırma güvenirliği temelde aynı sonuçları göstermekle beraber farklı metrik düzeyde ifade edilmektedir (Çetin ve İlhan, 2017). Ayırma indeksi yüzey bileşenlerinin kaç farklı boyutta yer aldığını ifade etmektedir (Eckes, 2015; Engelhard ve Myford, 2003). Ki-kare homojenlik test istatistiği ise ayırma indeksi ve ayırma güvenirliğinin anlamlılığını göstermektedir. Bu uyum değerleri her yüzey (değişken) için farklı anlam ifade etmekte ve farklı biçimde yorumlanmaktadır.

Öğrenci ve etkinlik (soru, madde gibi) yüzeyleri için ayırma güvenirliği ve ayırma indekslerinin yüksek olması beklenen durumlardır. Yüksek güvenirlik indeksleri farklı yetenek düzeyindeki öğrencilerin birbirinden ayırt edilebildiğini gösterir (Sudweeks, Reeveb ve Bradshawc, 2004; Aktaran İlhan, 2015). Düşük güvenirlik indeksi ve ayırma indeksi değerleri özellikle ranj sınırlaması etkisi olasılığını işaret etmektedir. Bu durumda puanlayıcılar farklı yetenek düzeyine sahip öğrencilere ölçeğin alt sınırındaki, üst sınırındaki veya orta noktasındaki puanları verme eğilimi gösterirler (İlhan, 2015). Bu yüzey için ayırma güvenirliği klasik anlamdaki Cronbach alpha ve KR-20 güvenirlik değerlerine benzer biçimde yorumlanabilmektedir (Baştürk, 2010; Engelhard ve Myford, 2003). Diğer bir ifade ile ayırma güvenirliği indeksi 0 ile 1 değerleri arasında değişmektedir. Ayırma güvenirliği 1'e yaklaştıkça gözlenen puan değişimi ile gerçek puan değişimi arasındaki fark azaltmakta ve ölçme hata payı azalmaktadır. Güvenirlik değeri 0.5'in altına düştükçe varyasyonun büyük oranda ölçme hatasından kaynaklandığı söylenebilir (Bahrouni, 2016; Engelhard vd., 2003). Bu durum özellikle büyük ölçekli açık öğretim sistemi sınavlarında istenmeyen sonuçlar

doğurabilmektedir. Ayırma indeksi güvenirlik değerinden farklı olarak 0 ile sonsuz arasında herhangi bir değer alabilmektedir.

Görev yüzeyi için elde edilen güvenirlik ve ayırma indeksleri ise farklı yapıları ölçen etkinliklerin birbirinden ne derecede ayırt edilebildiğini göstermektedir. Etkinlik yüzeyine ait ölçüm tablosundaki yüksek güvenirlik indeksi ve ayırma indeksi puanlayıcıların farklı performansları ölçen etkinlikleri birbirinden ayırt edebildiklerini anlamına gelmektedir.

Puanlayıcılara ait ölçüm tablosundaki düşük ayırma güvenirliği diğer yüzeydekilerin tersine istendik bir durumdur. Yüksek ayırma indeksi (> 2.0) ve güvenirlik indeksi (> 0.5) puanlayıcılardan kaynaklanan hataların olduğu ve yapılan puanlamaların katılık-cömertlik düzeyinde tutarsız olduğu sonucunu göstermektedir. Başka bir deyiş ile puanlayıcılar öğrencilerin cevaplarını değerlendirmede farklı katılık ve cömertlik eğilimi göstermektedir.

Puanlayıcıların katılık-cömertlik karşılaştırılması. Puanlayıcıların katılık-cömertlik derecelerini belirlemek için genel ve bireysel düzeyde farklı indekslere bakmak gerekmektedir. Öncelikle, ayırma indeksi, ayırma güvenirliği değerlerini inceleyerek genel düzeyde puanlayıcılar arasında katılık-cömertlik konusunda farklılık olup olmadığı belirlenir. Sonrasında ortaya çıkan bu farklılığın hangi puanlayıcılar tarafından gerçekleştiğini ortaya çıkartmak için FACETS programı sonuç çıktılarında ki-kare testi, "iç uyum (infit)" ve "dış uyum (outfit)" olmak üzere üç farklı fit istatistiği sunmaktadır. İlk olarak ki-kare testi "puanlayıcıların katılıkları veya cömertlikleri arasında istatistiksel anlamda bir fark var mıdır?" sorusunun yanıtını verir. Bu test için verilen p değeri 0.05 değerinden küçük ise öğrencilerin yanıtlarını değerlendiren puanlayıcıların katılığı-cömerliği arasında anlamlı bir fark vardır yorumu yapılabilir. Puanlayıcılar arasında İstatistiksel anlamda bir fark bulunduğu takdirde Tablo 2'den ya da değişken haritasından (logit ölçek) hangi puanlayıcıların katı hangilerinin cömertçe not verdiği belirlenebilmektedir.

İç uyum indeksi, ağırlandırılmış standart artıkların karesi istatistiklerine göre, dış uyum indeksi ise ağırlandırılmamış standart artıkların karesi istatistiklerine göre hesaplanmıştır. Bu istatistikler standartlaştırılmış artıklara göre örneklem büyüklüğünden daha az etkilenmektedir (İlhan, 2015). Beklenen değer ile gözlenen değer arasındaki farkı en az hatayla gösterdiği için iç uyum ve dış uyum indekslerinin kullanılması önerilmektedir (Linacre, 2010). Dış uyum indeksi beklenmeyen aşırı değerlere karşı daha duyarlı olduğu için iç uyum değerleri puanlayıcı tutarsızlıklarını göstermede daha önemli veriler sağlamaktadır (Engelhard, 1994; Yan, 2014). Dış uyum ve iç uyum istatistiklerinin beklenen değerleri 1.0 (standart hatası 0)ise mükemmel model-veri uyumunu göstermektedir. Ancak Liancre

(2002a) 1.5'ten yüksek ve 0.5'ten küçük değerlere sahip uyum indekslerinin beklenmeyen ve tutarsız sonuçlara işaret ettiğini ifade etmektedir. Uyum indekslerindeki kabul edilebilir sınırlar konusunda tam bir uzlaşma sağlanmış değildir. Wright ve Linacre 1994 yılında yaptıkları çalışmada içuyum ve dış uyum değerlerinin 0.6 ve 1.4 aralığında olmasını önermişlerdir. Ancak Myford ve Wolfe (2003) yapılan sınavın amacına göre bu sınırların genişletilip daraltabileceğini iddia etmişlerdir. Açık öğretim fakültesi sınavları gibi büyük ölçekli sınavlarda daha dar bir aralık olan 0.8 ve 1.2 değerlerinin kabul edilmesinin daha doğru sonuçlar vereceği belirtilmektedir (Myford ve Wolfe, 2003).

Sonuç olarak, bu indeksler model tarafından kestirilen puanlama sonuçları ile kişilerin gözlenen puanlama sonuçlarının farklarına göre bulunmaktadır (Eckes, 2009). Puanlayıcı yüzeyindeki veriler için uyum indeksleri puanlayıcı değerlendirmelerindeki tutarsızlıkları ortaya koymaktadır (Eckes, 2009). Özellikle bu indeksler, puanlayıcı etki türlerinden olan katılık-cömertlik, halo etkisi, ranj sınırlaması ve merkeze yönelme etkisi gibi durumları belirlemek için önemli veriler sunmaktadır (Eckes, 2015; Engelhard, 2002).

Yanlılık analizleri. FACETS programı ÇYRM analiz sürecine yüzey etkileşimlerini de dâhil ederek yanlılık olup olmadığını araştırmaya imkân tanımaktadır. Programda analiz edilmesi istenen yüzeyler ve etkileşimleri belirtilerek yanlılık puanları elde edilir. Kalibrasyon haritasında belirtilen yanlılık puanlarından 2 ve daha fazla standart z puanına sahip olanlar anlamlı etkileşimi göstermektedir (Mulqueen vd., 2000). Puanlayıcıların öğrencilerin bir kısmına daha katı veya daha cömert davrandığını belirlemek amacıyla puanlayıcı X öğrenci yüzeylerinin etkileşimi incelenir. Eğer puanlayıcıların bazı etkinlikleri (soruları, maddeleri) diğerlerinden daha farklı bir biçimde yorumladığına dair şüpheler var ise o zaman puanlayıcı X etkinlik yüzeylerinin etkileşimi analizi dikkate alınır. Son olarak, puanlayıcı X öğrenci X etkinlik etkileşim etkisi analiz edilerek puanlayıcıların belirli öğrencilerin bazı etkinliklerine daha düşük puan verip bazı öğrencilerin belirli etkinliklerine yüksek puan verdikleri tespit edilebilir.

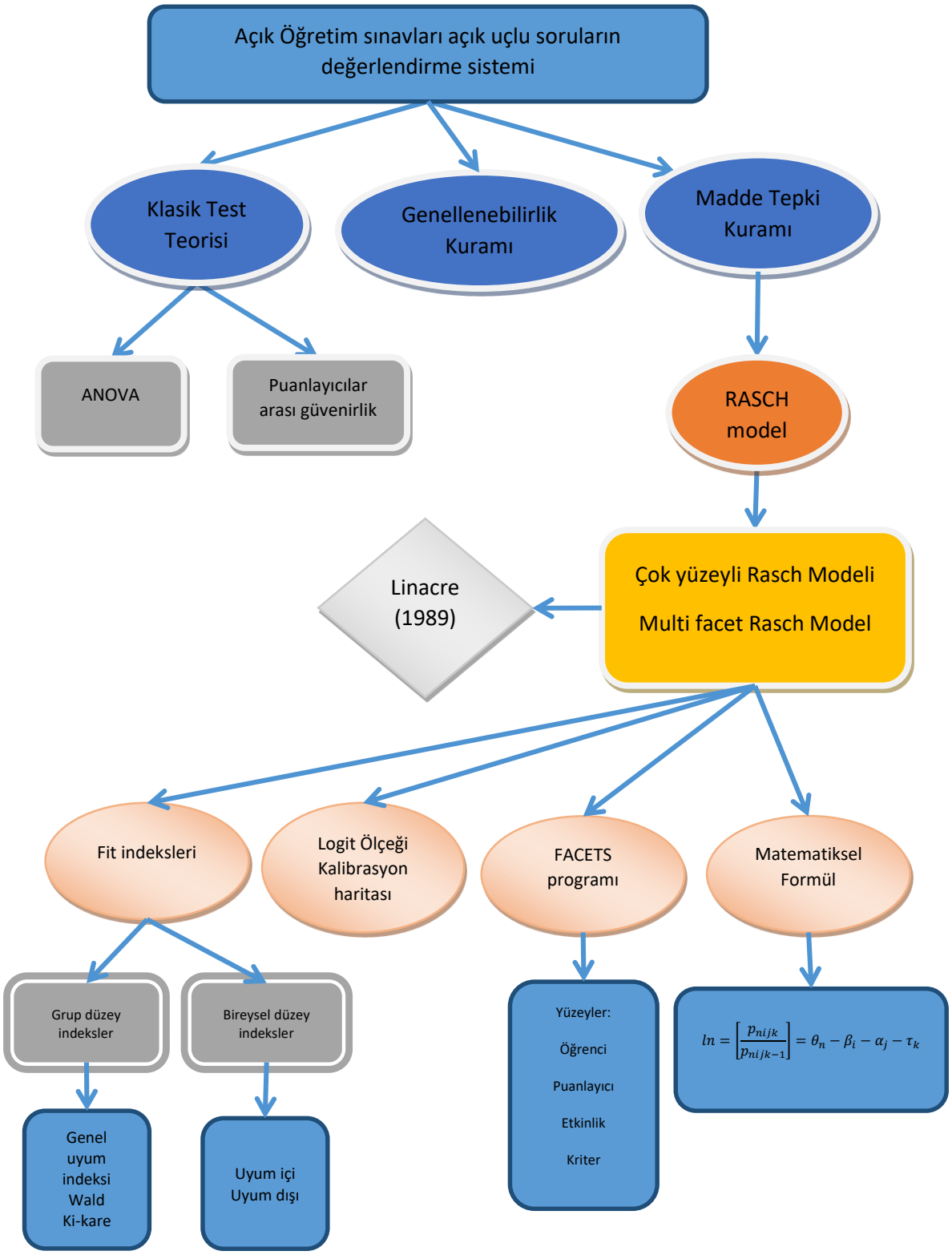
Sonuçlar ve Öneriler

ÇYRM puanlayıcılardan elde edilen veriyi değerlendirmek ve düzenlemek için kullanılan bir istatistiksel yöntemdir (Myford ve Wolfe, 2003). Bu yaklaşım puanlayıcı etkisi gibi boyutlardan kaynaklanan potansiyel ölçme hatalarını analiz etmede önemli yararlar sağlamaktadır. Linacre (1989) birden fazla yüzeyi aynı anda kalibre ederek tüm yüzeyleri tek bir ölçek üzerinde incelemeye yardımcı olan ÇYRM/MFRM yöntemini geliştirmiştir.

Açıköğretim fakültelerinin yapmakta olduğu çok sayıda öğrencinin katıldığı sınavlarda kullanılan açık uçlu maddelerin değerlendirilmesinde en etkili yöntem olan ÇYRM, puanlayıcıların denetlenmesine önemli katkı sağlamaktadır. Bu yöntem ile genel olarak puanlayıcıların katılık-cömertlik konusunda farklılık gösterip göstermediği belirlenebildiği gibi bireysel anlamda hangi puanlayıcının daha az puan verme eğilimi gösterdiği de kolayca bulunabilmektedir. Böylece sisteme dâhil olan puanlayıcılar katılık-cömertlik derecelerine göre sınıflandırılabilir ve öğrencilerin puanları onları değerlendiren kişilerden bağımsız olarak düzeltilebilir. Çok yüzeyli Rasch modeli puanlayıcı havuzu oluşturma konusunda da yararlı bilgiler sunmaktadır. Sisteme veri sağlayan her puanlayıcı kendi içindeki tutarlılık düzeyine göre sınıflandırılarak sonraki sınavlarda bu puanlayıcıların yer alıp almayacağına karar verilebilir. Şekil 3'te genel anlamda açık uçlu soruların değerlendirilmesine dair kavram haritası bulunmaktadır.

Tüm bu katkıların yanında, farklı katılık-cömerlik düzeyinde puan verme eğilimi gösteren puanlayıcıların etkisi azaltılarak yeni düzenlenmiş puanlar elde etmek mümkündür. Bu sayede açıköğretim sınavlarına giren öğrencilerin yanıtları puanlayıcı yanlılığı ve tutarsızlığı gibi istenmeyen etkilerden arındırılarak değerlendirilebilmektedir.

Kısaca özetlenirse, çok yüzeyli Rasch modeli yaklaşımı açık uçlu soruların sorulduğu sınavların değerlendirmesinde önemli katkılar sağlamaktadır. Ek olarak, farklı değişkenlerin analize katılması sonucunda ortaya çıkabilecek ölçme hatalarıyla başa çıkılmasına olanak tanımaktadır. Bu model, birden fazla değişkenin (öğrenci, puanlayıcı, etkinlikler gibi) aynı anda kalibre edilmesini ve tek bir ölçek üzerinde temsil edilmesini sağlamaktadır. Açık öğretim sınavlarındaki açık uçlu soruların yanıtlarının çok yüzeyli Rasch analizi yöntemiyle analiz edilmesi olası puanlayıcı tutarsızlıklarının belirlenmesine ve öğrencilerin performanslarının daha güvenilir bir biçimde ölçülmesine olanak sağlayacaktır.



Şekil 3. Açık uçlu soruların değerlendirilmesi kavram haritası

Kaynakça

Atılgan, H., Kan, A., & Doğan, N. (2009). *Eğitimde ölçme ve değerlendirme*. Anı Yayıncılık.

Bahrouni, F. (2016). usıng multı-facet rasch model (mfrm) ın rater-mediated assessment. *Journal of Teaching English for Specific and Academic Purposes*, *4*(1), 1, 95-212.

Behizadeh, N., & Engelhard, G. (2014). Development and validation of a scale to measure perceived authenticity in writing. *Assessing Writing*, *21*, 18-36.

Baştürk, R. (2010). Bilimsel araştırma ödevlerinin çok yüzeyli Rasch ölçme modeli ile değerlendirilmesi. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, *1*(1).

Chapman, J. E., McCart, M. R., Letourneau, E. J., & Sheidow, A. J. (2013). Comparison of youth, caregiver, therapist, trained, and treatment expert raters of therapist adherence to a substance abuse treatment protocol. *Journal of Consulting and Clinical Psychology*, *81*(4), 674.

Çetin, B., & İlhan, M. (2017). An analysis of rater severity and leniency in open-ended mathematic questions rated through standard rubrics and rubrics based on the SOLO taxonomy. *Eğitim ve Bilim*, *42*(189).

Eckes, T. (2005). Examining rater effects in TestDaF writing and speaking performance assessments: A many-facet Rasch analysis. *Language Assessment Quarterly: An International Journal*, *2*(3), 197-221.

Eckes, T. (2009). Many-facet Rasch measurement. *Reference supplement to the manual for relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment*, 1-52.

Engelhard Jr, G. (1992). The measurement of writing ability with a many-faceted Rasch model. *Applied Measurement in Education*, *5*(3), 171-191.

Engelhard, G. (1994). Examining rater errors in the assessment of written composition with a many-faceted Rasch model. *Journal of Educational Measurement*, *31*(2), 93-112.

Engelhard, G. (2002). Monitoring raters in performance assessments. *Large-scale assessment programs for all students: Validity, technical adequacy, and implementation*, 261-287.

Engelhard, G., & Myford, C. M. (2003). Monitoring faculty consultant performance in the advanced placement English Literature and composition program with a many-faceted Rasch model. *ETS Research Report Series*, *2003*(1).

Goodwin, L. D. (2001). Interrater agreement and reliability. *Measurement in Physical Education and Exercise Science*, *5*(1), 13-34.

Güler, N. (2014). Analysis of Open-Ended Statistics Questions with Many Facet Rasch Model. *Eurasian Journal of Educational Research*, *55*, 73-90.

İlhan, M. (2015). *Standart ve SOLO taksonomisine dayalı rubrikler ile puanlanan açık uçlu matematik sorularında puanlayıcı etkilerinin çok yüzeyli Rasch modeli ile incelenmesi* Doctoral dissertation, Doctoral dissertation, Gaziantep University, Gaziantep, Turkey. Retrieved from https://tez. yok. gov. tr/UlusalTezMerkezi.

Kondo-Brown, K. (2002). A FACETS analysis of rater bias in measuring Japanese second language writing performance. *Language Testing*, *19*(1), 3-31.

Linacre, J. M. (1989). Multi-faceted Rasch measurement.

Linacre, J. M. (2003). Size vs. significance: Standardized chi-square fit statistic. *Rasch Measurement Transactions*, *17*(1), 918.

Linacre, J. M. (2004). Rasch model estimation: Further topics. *Journal of Applied Measurement*, *5*(1), 95-110.

Linacre, J. M., & Wright, B. D. (1994). Chi-square fit statistics. *Rasch Measurement Transactions*, *8*(2), 350.

Lunz, M. E., Wright, B. D., & Linacre, J. M. (1990). Measuring the impact of judge severity on examination scores. *Applied Measurement in Education*, *3*(4), 331-345.

Mulqueen, C., Baker, D. P., & Dismukes, R. K. (2002). Pilot instructor rater training: The utility of the multifacet item response theory model. *The International Journal of Aviation Psychology*, *12*(3), 287-303.

Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many facet Rasch measurement: Part I. *Journal of Applied Measurement*, *4*(4), 386-422.

Reynolds, C. R., Livingston, R. B., & Willson, V. (2009). *Measurement and assessment in education.* Columbus, OH: Merrill.

Schumacker, R. E. (1996). Many-Facet Rasch Model Selection Criteria: Examining Residuals and More.

Swanlund, A. P. (2016). *Correcting for Rater Bias in the Presence of Non-Ignorable Missing Ratings* (Doctoral dissertation).

Tan, Ş., & Erdoğan, A. (2005). *Öğretimi planlama ve değerlendirme: Öğretim yönetim ve teknikleri ölçme ve değerlendirme KPSS el kitabı*. Pagem yayınları.

Turgut, M. F., & Baykul, Y. (2015). *Eğitimde ölçme ve değerlendirme* (Yedinci Baskı). Pegem A yayıncılık, Ankara.

Wolfe, E. W., & Chiu, C. W. (1997). Detecting Rater Effects with a Multi-Faceted Rating Scale Model.

Yan, X. (2014). An examination of rater performance on a local oral English proficiency test: A mixed-methods approach. *Language Testing*, *31*(4), 501-527.

Yazar Hakkında

Kübra KARAKAYA ÖZYER

Kübra Karakaya Özyer, 2015 yılından itibaren Eskişehir Osmangazi Üniversitesi Eğitimde Ölçme ve Değerlendirme anabilim dalında araştırma görevlisi olarak hizmet vermektedir. Karakaya Özyer, Yüksek Lisans eğitimini Amerika Birleşik Devletlerindeki kuzey Karolina Üniversitesi - Chapel Hill eğitim psikolojisi, ölçme ve değerlendirme alanında tamamlamıştır. Yazar, halen Eskişehir Osmangazi Üniversitesi Eğitim Bilimleri Enstitüsü Eğitimde Araştırma Yöntemleri ve İstatistik bütünleşik doktora programında öğrenimine devam etmektedir. Yazarın eğitim istatistiği, uluslar arası öğrenciler ve eğitim teknolojileri alanında çeşitli yayınları bulunmaktadır.

Posta adresi: Eskişehir Osmangazi Üniversitesi, Eğitim Fakültesi, Ölçme ve Değerlendirme A.B.D.
Tel (İş): +90 222 229 31 23- 1664
Eposta: kozyer@ogu.edu.tr
URL: http://egitim.ogu.edu.tr/files/kubra_karakaya.pdf