İNSAN VE TOPLUM BİLİMLERİ ARAŞTIRMALARI DERGİSİ
Cilt / Vol: 6, Sayı/Issue: 2, 2017
Sayfa:777-798
Received/Geliş: Accepted/Kabul:
[06-03-2017] – [04-04-2017]

Çok Değişkenlik Kaynaklı Rasch Ölçme Modeli ve Hiyerarşik Puanlayıcı Modeli İle Kestirilen Puanlayıcı Parametrelerinin Karşılaştırılması[1]

Müge ULUMAN
Öğr. Gör. Dr., Marmara Üniversitesi Atatürk Eğitim Fakültesi, Eğitim Bilimleri Böl.
RA. ,Marmara Unv. Ataturk Faculty of Education
mugeulumann@gmail.com

Ezel TAVŞANCIL
Prof. Dr., Ankara Üniversitesi Eğitim Bilimleri Fak.
Prof. Ankara University Faculty of Education
etavsancil@gmail.com

Öz

Bu araştırmada, açık uçlu maddelere ilişkin, aynı sınananlar tarafından verilen yanıtların, birden fazla puanlayıcı tarafından puanlanması durumunda, çok değişkenlik kaynaklı Rasch ölçme modeli (ÇDKRÖM) ve hiyerarşik puanlayıcı modeli (HPM) ile puanlayıcı katılık/cömertlik ve değişkenlik parametrelerinin kestirilmesi ve her iki modele ilişkin parametrelerin birlikte değerlendirilmesi amaçlanmıştır. Temel araştırma modelindeki araştırmanın verileri, 2012-2013 eğitim-öğretim yılı ikinci döneminde Ankara ili Çankaya ilçesinde yer alan, 10 okulda öğrenim gören, 15 yaş grubu 380 öğrencinin sekiz açık uçlu maddeye verdikleri yanıtlara beş ortaöğretim matematik öğretmeni tarafından atanmış puanlardan oluşmaktadır. Araştırma sonucunda, ÇDKRÖM ve HPM puanlayıcı parametre sonuçlarının genel olarak benzer olduğu saptanmıştır. Her iki modele ait sapma bilgi kriteri değerlerine göre; HPM'nin ÇDKRÖM'e göre araştırma verilerine daha iyi uyum sağladığı, tek bir maddenin tek bir yanıtına ilişkin atanan çoklu puanlara ait bir yapının HPM'yle daha iyi yansıtıldığı sonucuna ulaşılmıştır.

Anahtar Kelimeler: Rasch Ölçme Modeli, Hiyerarşik Puanlayıcı Model, Puanlayıcı Parametreler, Model, Parametre.

Comparing Parametersof Many Facetrasch Measurement Model And Hierarchial Rater Model

Abstract

This study aims at estimating the parameters with many facet Rasch measurement model (MFRMM) and hierarchical rater model (HRM) and evaluating together the rater severty/leniency and parameters obtained from both models if responses given by the same examinees for open-ended items are scored by multiple raters. In the scope of collecting study data, the scores assigned by five secondary school mathematics teachers for responses to eight open-ended items by 380 students, aged 15, from 10 schools in Çankaya District of Ankara province were used during the 2nd semester of the 2012-2013 academic year. The study revealed that rater parameters of MFRMM and HRM were similar in general. According to the deviation in formation criteria for both models; it was concluded that HRM provides better fit the data than MFRMM and the structure of assigned multiple scores regarding one single response to one single item is reflected beter by theHRM.

Keywords: Rasch Measurement Model, Hierarchical Rater Model, Rater Parameters, Model, Parameters.

[1]Bu çalışma Müge ULUMAN'ın doktora tezinden uyarlanmıştır.

Giriş

Eğitimde açık uçlu maddelerin kullanımı son yıllarda gerçekleştirilen birçok geniş ölçekli uluslararası değerlendirme çalışmaları (National Assessment of Educational Progress-NEAP, Graduate Record Examination-GRE, Programme for International Student Assessment-PISA) ile artmıştır (Kim, 2009; Mariano, 2002). Öğrencilerden, verilen bilgileri açıklamaları, düzenlemeleri, tanımlamaları, bir senteze ulaşmaları ve özgün fikirler üretmeleri isteniyorsa açık uçlu maddelerin kullanımı çok uygundur (Roid ve Haladyna, 1982). Bu madde tipi yeterliliğin doğrudan (direct) ve otantik (authentic) değerlendirilmesine olanak sunarak, eğitime olumlu yönde katkı sağlar (Messick, 1994). Fakat açık uçlu maddelerin puanlanma biçimlerine ilişkin zorluklarla karşılaşılabilir. Çünkü açık uçlu maddeler net ve tek bir doğru yanıta sahip değildir. Bu maddelerin puanlanma sürecinde, birden fazla puanlayıcı yer almakta ve dereceli puanlama anahtarı kullanılmaktadır. Puanlama puanlayıcıların kararları doğrultusunda yapılmaktadır.

Açık uçlu maddelerin puanlanmasındaki puanlayıcı etkileri modellenmezse, madde parametreleri ve sınanan yeterliliği kestiriminin duyarlılığı zedelenebilir. Bu nedenle açık uçlu maddeler kullanıldığında puanlayıcı etkilerinin varlığının dikkate alınması çok önemlidir (Kim, 2009; Linacre, 1994).

Diğer kuramlara nazaran çok daha uzun süredir kullanılagelmiş klasik test kuramına dayalı; puanlayıcı davranış ve etkilerini ortaya koyan, puanlayıcılar arası güvenirliğin belirlenmesinde kullanılan tekniklerin (basit yüzde tekniği, kesin uyum yüzdesi, Cohen kappa katsayısı, Fleiss kappa katsayısı vb.) ortak sınırlılığı; tek bir hata kaynağını dikkate alarak sonuç vermeleri biçiminde ifade edilebilir. Bu sınırlılığın üzerine Cronbach, Gieser, Nanda ve Rajarantnam (1972) tarafından geliştirilmiş olan genellenebilirlik kuramının (Cardinet, Tourneur ve Allal, 1981) kökleri klasik test teorisine ve varyans analizine (ANOVA) dayanmaktadır (Brennan, 1992; Brennan, 2010). Bu kuram araştırmacıya, herhangi bir ölçme durumuna ilişkin tüm potansiyel hata kaynaklarını (puanlayıcı, madde, zaman, vb.) birlikte değerlendiren kavramsal bir çerçeve sunar (Cardinet ve diğerleri, 1981).

Genellenebilirlik kuramı, puanlayıcılar koşulu üzerine kurulmamıştır. Dolayısıyla puanlayıcıların bireysel olarak doğrudan değerlendirilmesi mümkün değildir ve genellikle ek analizlere ihtiyaç duyulmaktadır (Patz, Junker, Johnson ve Mariano, 2002). Benzer bir zorluk, sınanan yeterliliği için de gözlenebilir (Mariano, 2002). Ayrıca bu kuram, test ham puanlarının doğrusal olmayan dönüşümlerini içeren uygulamalar için yeterince geliştirilmemiştir (Brennan, 1997; Mariano, 2002; Patz ve diğerleri, 2002). Bu yönüyle de madde, puanlayıcı ve sınanan arasındaki ilişkilerin niceliğini belirlemek için sınırlı yeteneğe sahiptir (Patz ve diğerleri, 2002).

John M. Linacre tarafından geliştirilmiş olan çok değişkenlik kaynaklı Rasch ölçme modeli (ÇDKRÖM), derecelenmiş ölçek modeline (Andrich, 1978) puanlayıcı parametresinin eklendiği, Rasch yaklaşımının bir uzantısı olarak tanımlanabilir (Linacre 1989; Linacre, 1994). ÇDKRÖM ölçme sürecine puanlayıcı parametresinin de dâhil edilmesiyle sadece sınanana ait yetenek düzeyi ve maddeye ait güçlük düzeyinin değil puanlayıcıya ait katılık düzeyinin de eş zamanlı olarak kestirilmesi yönünden çok kullanışlıdır (Linacre, Wright ve Lunz, 1990).

Atanan her bir puan, birbirini etkileyen dört bileşenin olasılıksal sonuçları olarak nitelendirilebilir. ÇDKRÖM eşitliği aşağıda sunulmuştur (Linacre, 1994).

$$\log(P_{nijk} / P_{nijk}\text{-}1) = B_n - D_i - C_j - F_k$$

P_{nijk}: Sınanan "n"in "i" maddesinde gösterdiği performansın "j" puanlayıcısı tarafından "k" kategorisinde puanlanma olasılığıdır.

P_{nijk}-1: Sınanan "n"in "i" maddesinde gösterdiği performansın "j" puanlayıcısı tarafından "k-1" kategorisinde puanlanma olasılığıdır.

B_n: Sınanan "n"in yetenek düzeyi

D_i: Madde "i"nin güçlük düzeyi

C_j: Puanlayıcı "j"nin katılık düzeyi

F_k: Kategori "k-1"den kategori "k"ya geçişin güçlük düzeyi

Modelin temel eşitliği doğrultusunda sınanan, puanlayıcı ve maddeler değişkenlik kaynaklarıdır. Değişkenlik kaynaklarını oluşturan her bir eleman parametrelerle temsil edilmektedir.

Literatürde ÇDKRÖM kullanılarak gerçekleştirilmiş birçok çalışmanın (Akın ve Baştürk 2012; Atılgan, 2005; Engelhard, 1994; Engelhard ve Myford, 2003; Iramaneerart, Myford, Yudkowsky ve Lowenstein, 2009; Nakamura, 2000) yanı sıra ÇDKRÖM üzerine de birçok çalışma yapılmıştır (Casabianca ve Junker, 2013; DeCarlo 2010; DeCarlo ve diğerleri, 2011; Lynch ve McNamara, 1998; Mariano 2002; Patz ve diğerleri 2002; Sudweeks, Reeve ve Bradshaw, 2004; Verhelst ve Verstralen, 2001; Wilson ve Hoskens 2001). Bu çalışmaların büyük bir kısmı ÇDKRÖM'in, sınanan, madde ve puanlayıcı parametrelerini veren her bir puanın birbirinden bağımsız olduğu ve tüm puanlayıcıların eşit güvenirliğe sahip olduğu önermelerinin sınırlılıklarına yöneliktir (Mariano, 2002).

ÇDKRÖM'ne alternatif olarak geliştirilen modellemelerden biri Patz ve diğerleri (2002) tarafından ortaya konmuş, genellenebilirlik teorisi yapısına madde tepki kuramı modelini dâhil eden ve hiyerarşik bir bayes modeli olan, hiyerarşik puanlayıcı modelidir (HPM) (Mariano ve Junker, 2007).

HPM, sınananın verdiği aynı yanıtın çoklu puanlarının birbirinden bağımsız olduğu varsayımının aksine; çoklu puanlar arasındaki, yapısı gereği var olan bağımlılığı tanımlama amacıyla tasarlanmıştır (Mariano, 2002). Bu bağlamda, puanlayıcıların verdiği puanlar; doğrudan sınanan yeterliliğinin bir göstergesi olmaktan ziyade dereceli puanlama anahtarı kullanılarak elde edilen maddenin ait olduğu kategorinin göstergesidir (DeCarlo ve diğerleri, 2011).

HPM, değişkenlik kaynaklarının özünde bulunan hiyerarşik yapıyı kullanır ve örtük özelliklerin dağılımını, öğrencilerin örtük özelliklerini veren sınanan yanıtlarının dağılımını ve yanıtların niteliğini veren puanların dağılımını modelleyerek avantaj sağlar (DeCarlo ve diğerleri, 2011; Mariano ve Junker, 2007). Dolayısıyla, birden fazla puanlayıcının yer aldığı desenlerde görülen, bireysel olarak puanlayıcı etkilerinin izlenmesine imkân tanır. Ayrıca sınananın doğru yanıt verme yeterliliğini modellediği gibi puanlayıcıların doğru puan atama yeterliliğini de modeller (Patz ve diğerleri, 2002).

HPM, üç düzeyli bir hiyerarşiden oluşmaktadır. Bu hiyerarşi iki kademeli bir işlemle bağlantılıdır. Modelin ilk düzeyi, yanıtların niteliğini veren atanan puanların (X_{ijr}) dağılımıdır. İkinci düzey ise sınananlara ait örtük özelliği veren, sınanan yanıtlarının ideal puan (ξ) dağılımıdır. Son olarak üçüncü düzey, örtük özelliklerin (θ_i) dağılımıdır. Hiyerarşinin bağlantılı olduğu iki kademe, j kadar maddenin sınananlar tarafından yanıtlanması ve bu yanıtların r kadar puanlayıcı tarafından değerlendirilmesinden oluşmaktadır (Casabianca, Junker ve Patz, 2014; Mariano ve Junker, 2007).

$\theta_i \sim N(\mu, \sigma^2)$, i = 1, ..., N,

ξ_{ij}~ Çok Sonuçlu Madde Tepki Kuramı Modelleri, j = 1, …, J, her bir i için

$\mathbf{X}_{ijr}$~ Sinyal Tespit Modeli, r = 1, …R, her bir i ve j için

Bu hiyerarşik yapı, puanlayıcı performansının tanımlanması esnasında; puanlayıcılar arası, puanlayıcılar içi ve puanlayıcı performansının ortak değişkenliğinin modellenmesi noktasında da esneklik sağlar (Mariano, 2002).

İlk düzeyde, puanlayıcıların verdiği puanlar, maddenin ait olduğu gerçek kategorinin sıralı göstergeleriyken; ikinci düzeyde örtük kategoriler (latent categories) sınanan yeterliliğinin sıralı göstergeleridir. HPM'nin ilk düzeyinde sinyal tespit modelinden faydalanılır ki bu puanlayıcı modeli olarak adlandırılır. İkinci düzeyi için madde tepki kuramının uygun olan bir modeli kullanılır ve bu da madde modeli olarak adlandırılır (DeCarlo ve diğerleri, 2011).

ÇDKRÖM ve HPM puanlama sürecinde birden fazla puanlayıcının yer aldığı durumlarda güvenirliği belirlemek amacıyla kullanılan ve puanlayıcı davranışlarına da yer veren modellerdir. Eğitim alanının yanı sıra puan

atanmasına ihtiyaç duyulan tüm alanlar için puanlayıcı davranışlarının ek parametreleri barındıran modellerle daha derinlemesine incelenmesi, çalışma sonuçlarının güvenirliği ve geçerliliği açısından önemlidir. Bu nedenle puanlayıcı davranışlarının farklı modeller aracılığıyla incelendiği, kullanılan modellerin tanıtıldığı ve elde edilen sonuçların karşılaştırılarak tartışıldığı araştırmalara ihtiyaç duyulduğu düşünülmektedir.

Bu araştırmanın amacı, çok değişkenlik kaynaklı Rasch ölçme modeli ve hiyerarşik puanlayıcı modeli kullanılarak puanlayıcı parametrelerinin kestirilmesi ve her iki modele ait parametrelerin birlikte değerlendirilmesidir. Belirtilen amaç doğrultusunda aşağıdaki sorulara yanıtlar aranmıştır.

1-Çok değişkenlik kaynaklı Rasch ölçme modeli'nin;

a- Model veri uyumu nasıldır?

b- Puanlayıcı katılık/cömertlikleri ve uygunluk istatistikleri nasıldır?

2-Hiyerarşik puanlayıcı modeli'nin puanlayıcıların katılık/cömertlik ve değişkenlik istatistikleri nasıldır?

3-Çok değişkenlik kaynaklı Rasch ölçme modeli ile hiyerarşik puanlayıcı modeli'ne ait sapma bilgi kriteri değerleri nasıldır?

Yöntem

Bu araştırma, açık uçlu maddelere verilen yanıtların birden fazla puanlayıcı tarafından puanlanması ile elde edilen gerçek veri setinin, aynı amaç doğrultusunda geliştirilmiş çok değişkenlik kaynaklı Rasch ölçme modeli ile hiyerarşik puanlayıcı modeli uygulamaları üzerinden gerçekleştirilmiştir. Her iki model için gerçekleştirilen uygulama neticesinde elde edilen sonuçların; birbirlerine göre benzerlik ve farklılıkları, kullanışlılık açısından avantaj ve dezavantajlarının neler olduğu, hangi modelin daha fazla bilgi sağladığı incelenmiştir. Bu bağlamda araştırma "temel araştırma" niteliği taşımaktadır.

Çalışma Grubu

Çalışma grubunu: ulaşım kolaylığı nedeniyle 2012-2013 eğitim-öğretim yılı ikinci döneminde, Ankara ili Çankaya ilçesinde yer alan 10 okulda öğrenim gören, 15 yaş grubu öğrencileri oluşturmaktadır. Araştırmada yer alan öğrenci sayısı, her iki model için tamamen çaprazlanmış desende (Fully-Crossed Design) gerçekleştirilmiş çalışmalara (Atılgan, 2005; Baştürk, 2010; Patz ve diğerleri, 2000; Patz ve diğerleri, 2002; Turner, 2003) ait öğrenci, puanlayıcı ve madde sayılarıyla birlikte öğrenci, puanlayıcı ve madde sayılarının çarpımı sonucu elde edilen toplam veri sayısı dikkate alınarak belirlenmiştir. Bu bağlamda öğrencilerden isteyenlerin katıldığı, ders esnasında gerçekleştirilmiş olan uygulama sonucu toplam 380 öğrenciye ulaşılmıştır. Öğrencilerden 350'sinin yanıtları analize dâhil edilmiş, açık uçlu

maddelere verdikleri yanıtlar doğrultusunda her okuldan başarılı, başarısız ve orta düzeyde başarılı olduğu düşünülen 3, toplam 30 öğrencinin verdiği yanıtlar ise bütünsel dereceli puanlama anahtarının hazırlanması sürecinde kullanılmak üzere analiz dışında tutulmuştur. Öğrencilerin açık uçlu maddelere verdikleri yanıtlara puan atayacak olan puanlayıcılar, gönüllülük esası doğrultusunda çalışmaya katılmak istemiş olup; 5 ortaöğretim matematik öğretmeninden oluşmaktadır.

Veri Toplama Araçları

Araştırma kapsamında maddelerin yazılması yerine hâlihazırda uzman bir grup tarafından geliştirilmiş ve uygulanmış maddelerin kullanımı tercih edilmiştir. Var olan açık uçlu maddelerinin niceliği ve araştırma kapsamında kullanılabilirliği doğrultusunda PISA ikinci dönem uygulamasında yer almış ve açıklanmış olan, çok kategorili puanlanan 10 açık uçlu madde kullanılmak istenmiştir. İlgili maddeler OECD tarafından gerçekleştirilen uygulamada puanlayıcılar tarafından üç kategoride puanlanmıştır. Araştırma amaç ve alt amaçları doğrultusunda puanlayıcıların, puan atama sürecinde kendi içlerinde ve aralarında sergileyecekleri benzerlik ve farklılıkların, daha açık bir biçimde ortaya konulması ve araştırmada kullanılan her iki modelin bu benzerlik ve farklılıkları yansıtma derecelerinin belirlenmesi önemsenmektedir. Bu nedenle maddelerin puanlayıcılar tarafından beş kategoride puanlanması uygun bulunmuştur. Kullanılması amaçlanan 10 madde farklı zamanlarda iki matematik ve bir ölçme ve değerlendirme alan uzmanının görüşüne sunulmuştur. Alan uzmanlarının ortak görüşleri doğrultusunda iki madde, beş kategoride puanlanmasının mümkün olmaması nedeniyle araştırma kapsamı dışında tutulmuştur.

Araştırmada, ürünün ya da performansın bir bütün olarak ve daha hızlı puanlanmasına olanak sağlaması (Mertler, 2001); maddelerle ölçülmesi amaçlanan performansın alt bileşenlerinin ölçülmesine ihtiyaç duyulmadığı durumlarda uygulanabilir olması (Jonsson ve Svingby, 2007; Quinlan, 2011); araştırma grubunda sayıca oldukça fazla bireyin bulunması (Lund ve Veal, 2013; Quinlan, 2011) gerekçeleri dikkate alınarak bütünsel dereceli puanlama anahtarının kullanılması tercih edilmiştir. Bütünsel dereceli puanlama anahtarının hazırlanması sürecinde litaratürde konuyla ilgili kaynaklarda (Airasian, 2001; Mertler, 2001; Popham, 1997; Stevens ve Levi, 2005) yer alan adımlardan, araştırmaya katılmaya gönüllü olmuş, daha önceden görüşü alınmayan, bir ortaöğretim matematik öğretmeni ile bir matematik alan uzmanından yardım alınmıştır. Bununla birlikte, maddelere yönelik öğrenci performans göstergelerinin listelenebilmesi ve böylece kategorilere ait performans tanımlarının oluşturulmasına kolaylık sağlaması bakımından (Airasian, 2001; Stevens ve Levi, 2005) öğrenci yanıtlarına ihtiyaç duyulmuştur. Bu nedenle açık uçlu maddelere verdikleri yanıtlar doğrultusunda, her okulda öğrenciler başarılı, başarısız ve orta düzeyde

başarılı şeklinde gruplanmış ve her bir grupta yer alan öğrencilerden biri seçkisiz bir biçimde belirlenmiştir. Toplam 30 öğrencinin yanıtları bütünsel dereceli puanlama anahtarının hazırlanması sürecinde kullanılmak üzere analizler dışında tutulmuştur. Son olarak da hazırlanmış olan dereceli puanlama anahtarları daha önce görüşüne başvurulmuş bir ve daha önce görüşü alınmamış iki olmak üzere toplam üç matematik alan uzmanı ile daha önce görüşüne başvurulmamış iki ölçme ve değerlendirme uzmanına verilmiştir. Alan uzmanlarından gelen görüşler doğrultusunda dereceli puanlama anahtarları yeniden düzenlenmiş ve uygulamaya hazır hâle getirilmiştir.

İşlem

Maddeleri içeren soru kitapçıkları öğrencilerin okumakta zorlanmayacakları ve yanıtlarını rahatlıkla yazabilecekleri biçimde hazırlanmıştır. Soru kitapçıkları çoğaltıldıktan sonra araştırma grubunda yer alan okullara gidilmiş ve uygulama zamanı için okul rehber öğretmeni aracılığıyla ders öğretmenlerinden randevular alınmıştır. Uygulama yapılacak sınıflarda öncelikle araştırmaya ilişkin açıklamalar yapılmış ve öğrenciler tarafından sorulan sorular cevaplandırılmıştır. Daha sonra, araştırmaya katılmaya gönüllü olan öğrencilerle uygulama gerçekleştirilmiştir. Soru kitapçığında öğrencilere dağıtıldığında gidilen okulu ya da öğrenciyi işaret eden herhangi bir numaralandırma ya da simgeleme konulmamış ve öğrencilerden soru kitapçıklarına isimlerini yazmaları istenmemiştir. Soru kitapçıkları öğrencilerden alındıktan sonra her bir puanlayıcının aynı sıralamayla puanlama yapabilmesi açısından kitapçıklara sıra numaraları verilmiştir.

Uygulamalar tamamlandıktan sonra araştırmaya katılmakta gönüllü olan puanlayıcılara gerekli açıklamalar yapılmış ve her bir puanlayıcının isteği doğrultusunda puanlama yapabileceği bir zaman dilimi için randevu alınmıştır. Puanlayıcıların atadıkları puanları üzerine yazacakları, öğrencileri ve öğrencilerin maddelere verdikleri yanıtları temsil eden bir çizelge oluşturulmuştur. Kitapçıklar, dereceli puanlama anahtarları ve çizelge puanlayıcılara kendilerinin oluşturduğu takvim doğrultusunda gönderilmiş ve süreç sonunda geri alınmıştır. Böylelikle, 350 öğrencinin 8 açık uçlu maddeye verdikleri yanıtlara ait 5 puanlayıcının puan ataması sonucunda toplam 14000 veri elde edilmiştir.

Verilerin Analizi

Araştırma amacı doğrultusunda, verilerin analizinde ÇDKRÖM ve HPM için tüm öğrenciler tüm maddeleri yanıtlamış ve tüm yanıtlar tüm puanlayıcılar tarafından puanlanmıştır. Bu bağlamda "Tamamen Çaprazlanmış Desen"den faydalanılmıştır.

Bu çalışmada, model kurulumu ve model uyumu bakımından örtük değişkenlere ait parametrelerin kestirilmesine yapısı gereği uygun olması

(Patz ve Junker, 1999a, 1999b; Patz ve diğerleri, 2002); sınır değer problemiyle (boundary value problems) başa çıkma noktasında kullanışlılığı (DeCarlo ve diğerleri, 2011; Gelman, Carlin, Stern ve Rubin, 1995) ve literatürde yer alan HPM uygulamalarında büyük çoğunlukla Bayes kestirimi kullanılmış olması (Casabianca ve Junker, 2013; Casabianca ve diğerleri, 2014; Mariano, 2002; Patz ve Junker, 1999a, 1999b; Patz ve diğerleri, 2002) nedeniyle HPM uygulamaları Bayes kestirimi kullanılarak gerçekleştirilmiştir.

Bayes kestirimi için kullanılan bilgilendirici olmayan önsel dağılımlar Tablo 1'de özetlenmiştir.

Tablo 1. Önsel Dağılımlar

Parametreler	Bilgilendirici Olmayan Önsel Dağılımlar
β_j	N (0,10) dan Bağımsız Özdeşçe Dağılmış
γ_{jk}	N (0,10) dan Bağımsız Özdeşçe Dağılmış
ϕ_r	N (0,10)
ψ_r	log ψ_r) ~ N(0,10)
μ	N (0,10)
σ^2	1/ σ^2 ~ Gamma (α,η), α=η=1

HPM analizleri, 5000 burn-in periodu, 30,000 iterasyon sayısı, üç Marcov zinciri ve 10 seyreltme kullanılarak, OpenBUGS programında gerçekleştirilmiştir. Analiz, yaklaşık olarak toplam 135 saatte tamamlanmıştır.

Bir modelleme çalışması olan HPM, diğer modelleme çalışmalarında olduğu gibi analiz öncesi test edilmesi gereken varsayımlara sahip değildir. Elde edilen sonuçların raporlaştırılabilir olup olmadığı, analizde kullanılan zincirlerin denge dağılımını yakınsamasına bağlıdır. Bu yakınsamanın sağlanıp sağlanmadığına, modelden elde edilen her bir parametre için Brooks-Gelman-Rubin (BGR) tanısalı (diagnostic) ve zaman serileri (history) diyagramlarının incelenmesi ile karar verilebilir (Kéry, 2010; Spiegelhalter, Thomas, Best ve Lunn, 2003).

Bu çalışmada da HPM'nin model veri uyumunun ölçülmesi ve HPM ile ÇDKRÖM sonuçlarının karşılaştırılması için Sapma Bilgi Kriterinden faydalanılmıştır.

Bulgular

Çok Değişkenlik Kaynaklı Rasch Ölçme Modeli Model-Veri Uyumu

Model-veri uyumu; model varsayımı olarak verilen, beklenmeyen (unexpected) yanıtların incelenmesiyle elde edilebilir (Eckes, 2005). Verinin modele uygun olabilmesi için standartlaştırılmış artık değerlerin (standardized residuals) yaklaşık olarak %1'inden azı -/+3'den ve %5'inden azı -/+2'den büyük olmalıdır (Linacre, 1994; Linacre, 2003). Analizde yer alan toplam 14000 verinin standartlaştırılmış artık değerlerinin -/+3'den büyük olanlarının sayısı 119 (%0,85), -/+2'den büyük olanlarının sayısı ise 348'dir (%2,49). Elde edilen bu bulgular doğrultusunda araştırma verilerinin, ÇDKRÖM analizi kapsamında kullanılan modele uygun olduğu ifade edilebilir.

Çok Değişkenlik Kaynaklı Rasch Ölçme Modeli Puanlayıcı Katılık/Cömertlik ve Uygunluk İstatistikleri

Puanlayıcılara ait ÇDKRÖM analizi sonuçları Tablo 2'de verilmiştir. Tablo 2'de yer alan bulgular doğrultusunda yorumlamalar yapılmıştır.

Tablo 2. ÇDKRÖM analizi puanlayıcı ölçümü sonuçları

			Puanlayıcı Katılığı		Uygunluk İçi		Uygunluk Dışı	
Puanlayıcı No	Puanlayıcı Ort.	Puanlayıcı Toplam r	Logit Ölçüsü	S.H.	Kareler Ort.	Z Std.	Kareler Ort.	Z Std.
2	3.0	3.10	.14	.01	1.0	1.0	1.0	0.0
5	3.1	3.19	.10	.01	1.0	0.0	1.0	-1.0
3	3.2	3.38	.01	.02	1.1	2.0	1.0	0.0
4	3.4	3.57	-.09	.02	0.9	-2.0	0.9	-2.0
1	3.5	3.70	-.16	.02	1.0	0.0	0.9	-2.0
Ortalama	3.3	3.39	.00	.02	1.0	0.1	1.0	-1.5
SS	0.2	0.23	.11	.00	0.0	1.8	0.0	1.3

RMSE (Model) = .02 SS = .11 Ayırma İndeksi = 7.25 Güvenirlik = .98

Tamamı Aynı Ki-Kare = 266.0 Sd = 4 p =.00

Tablo 2'de puanlayıcılar 0.14 ile -0.16 aralığında yer alan logit değerlerince en katı olan puanlayıcıdan en cömert olan puanlayıcıya doğru sıralanmıştır. Bu değerlere göre en katı puanlayıcı, ikinci puanlayıcı olurken birinci puanlayıcı, en cömert puanlayıcıdır. Puanlayıcıların logit değerleri ortalaması 0.00 ve standart sapması ise 0.11'dir. Elde edilen puanlayıcı katılık ve cömertlik değerlerine ait standart hata (0.02) oldukça düşük bulunmuştur. Düzeltilmiş standart hata değerinin 0.11 de kritik değerin (1.0) altında olduğu tespit edilmiştir. Puanlayıcı ayırma indeksi (7.25) istenen düzeyin (0.00 ve 0.00'a yakın) üstünde bir değer bulunmuştur. Bu değer,

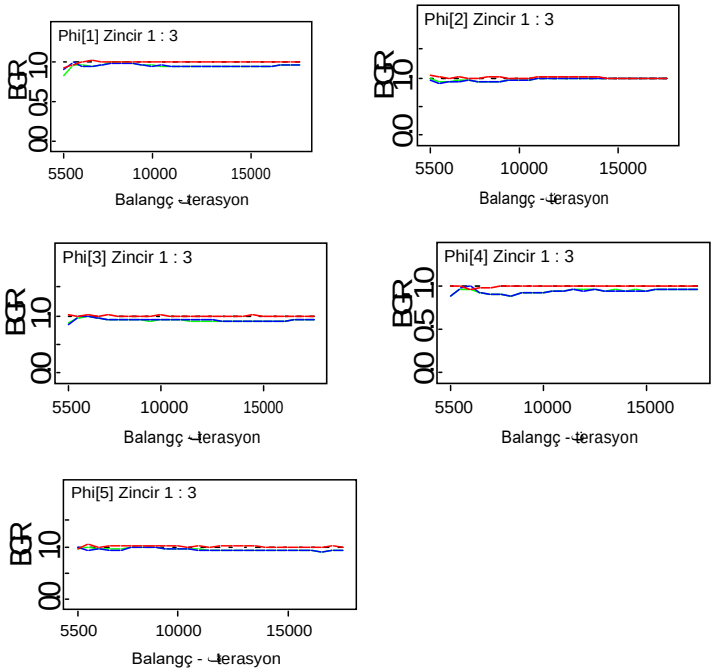
puanlayıcılar arası puan atama noktasında farklılıkların olduğunu, puanlayıcıların katılık ve cömertlik düzeylerine göre farklılaştığını ve puanlayıcıların atadıkları puanlarda cömertlik/katılık hatasının yer aldığını yansıtmaktadır. Puanlayıcı ayırma indeksi güvenirliği 0.98'dir. Bu, puanlayıcılar arası istenmeyen varyansın göstergesi olarak yorumlanabilir. Her iki değer göz önünde bulundurulduğunda, puanlayıcıların birbirleri yerine geçmeleri durumunun sakıncalı olabileceği ve bir noktaya kadar maddelere atanan puanların sadece maddenin niteliğine bağlı olmaksızın, hangi puanlayıcının puanladığına da bağlı olduğu belirtilebilir (Sudweeks ve diğerleri, 2004). Ayırma indeksi ve güvenirlik test edilmiş ve sabit etki (fixed effect) hipotezi Ki-kare testiyle (X^2= 266, sd= 4, p= 0.00) reddedilmiştir. Yukarıda elde edilen bulgularla paralel biçimde, beş puanlayıcının istatistiksel olarak katılık ve cömertlik düzeyleri arasında farklılıkların olabileceği ifade edilebilir. Bu bulgularla birlikte, puanlayıcıların standart Z puanlarının birbirine oldukça yakın ve katılık/cömertlik bakımından, birbirleri arasında bir logit birimin biraz üzerinde fark olduğu görülmektedir. Bu nedenle, puanlayıcılar arasında görülen farklılıkların kabul edilebilir düzeyde olduğu, katılık ve cömertlik bakımından aynı davranışı sergilemeseler de birbirlerine yakın hareket ettikleri ifade edilebilir (Lee ve Kantor, 2003). Her bir puanlayıcı için katılık ve cömertlik parametresi kestiriminin kararlılığını gösteren, model standart hata sütunu incelendiğinde elde edilen değerlerin oldukça küçük olduğu ve bu doğrultuda modelin kararlı olduğu ifade edilebilir.

Her bir puanlayıcı için uygunluk içi ($\overline{X}$ =1, s.s.=0.0) ve uygunluk dışı ($\overline{X}$ =1, s.s.=0.0) değerleri incelendiğinde tüm değerlerin istenen aralıkta (0.8-1.2) (Linacre, 1989) olduğu saptanmıştır. Modelde yer alan değişkenlik kaynaklarına ilişkin kesin yorumlar yapılmadan önce tüm değerlerin incelenmesi ve birlikte değerlendirilmesi tavsiye edilmektedir (Linacre, 1994). Bununla birlikte puanlayıcıların, puan atama noktasında birbirleriyle ve kendi içlerinde tutarlılık gösterdikleri ifade edilebilir.

Hiyerarşik Puanlayıcı Modeli'nin Puanlayıcı Katılık/Cömertlik ve Değişkenlik İstatistikleri

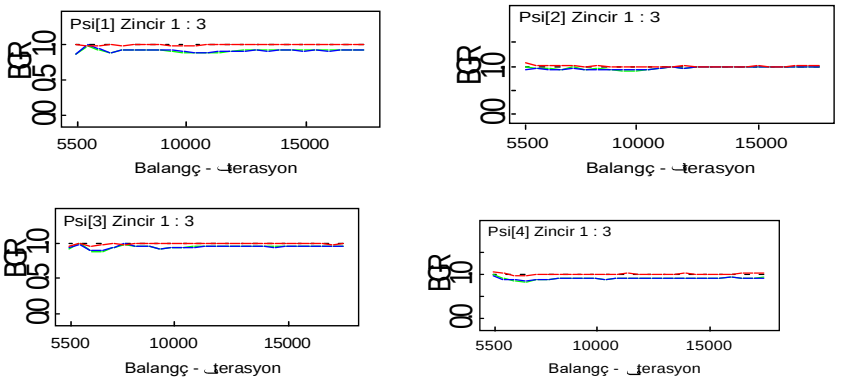
Analizde yer alan parametreler rapor edilmeden önce ilgili parametreler için analizin gerçekleştirildiği üç zincirin yakınsama durumu kontrol edilmelidir. Sonsal dağılımdan geçerli örneklem elde edildiğine dair kanıtlara ulaşıldıktan sonra gözlemlere dayanan parametrelere ilişkin çıkarımlarda bulunulabilir (Christensen, Johnson, Branscum, ve Hanson, 2011). Bu nedenle, Markov zincirlerinin kararlı denge dağılımına erişip erişmediği, başka bir ifadeyle zincirlerin yakınsayıp yakınsamadığı incelenmelidir. Bu amaç doğrultusunda, litaretürde oldukça yaygın olarak kullanılan Brooks-Gelman-Rubin (BGR) tanısalı (diagnostic) ve zaman serileri diyagramından faydalanılmıştır (Kéry, 2010).

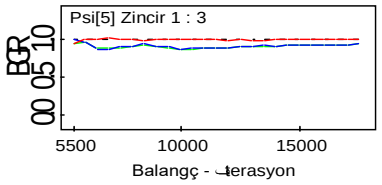
Grafik 1. Puanlayıcı Katılık Parametrelerine İlişkin BGR Grafikleri



Puanlayıcı katılık parametreleri, BGR grafikleri incelendiğinde, özellikle ikinci puanlayıcının neredeyse bir, diğer puanlayıcıların da bire çok yakın değerler aldığı görülmektedir. Sadece dördüncü puanlayıcıya ait katılık parametresi, BGR grafiğinin bir değerinden çok az da olsa uzaklaştığı gözlenmiş ve BGR değeri incelenmiştir. Son iterasyon aralığı için BGR değerinin, 1.001 olduğu ve bu değerin 1.0 ile 1.1 aralığında yer aldığı tespit edilmiştir. Elde edilen grafikler ve değerler doğrultusunda puanlayıcı katılık parametreleri rapor edilebilir niteliğe sahiptir. Puanlayıcı değişkenlik parametrelerine ilişkin BGR grafikleri ise aşağıda verilmiştir.

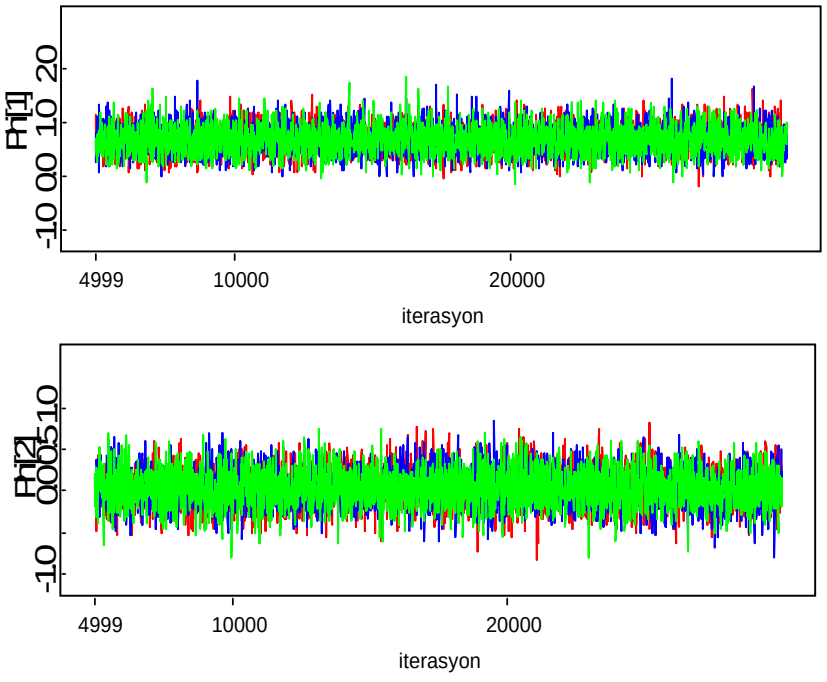
Grafik 2. Puanlayıcı Değişkenlik Parametrelerine İlişkin BGR Grafikleri

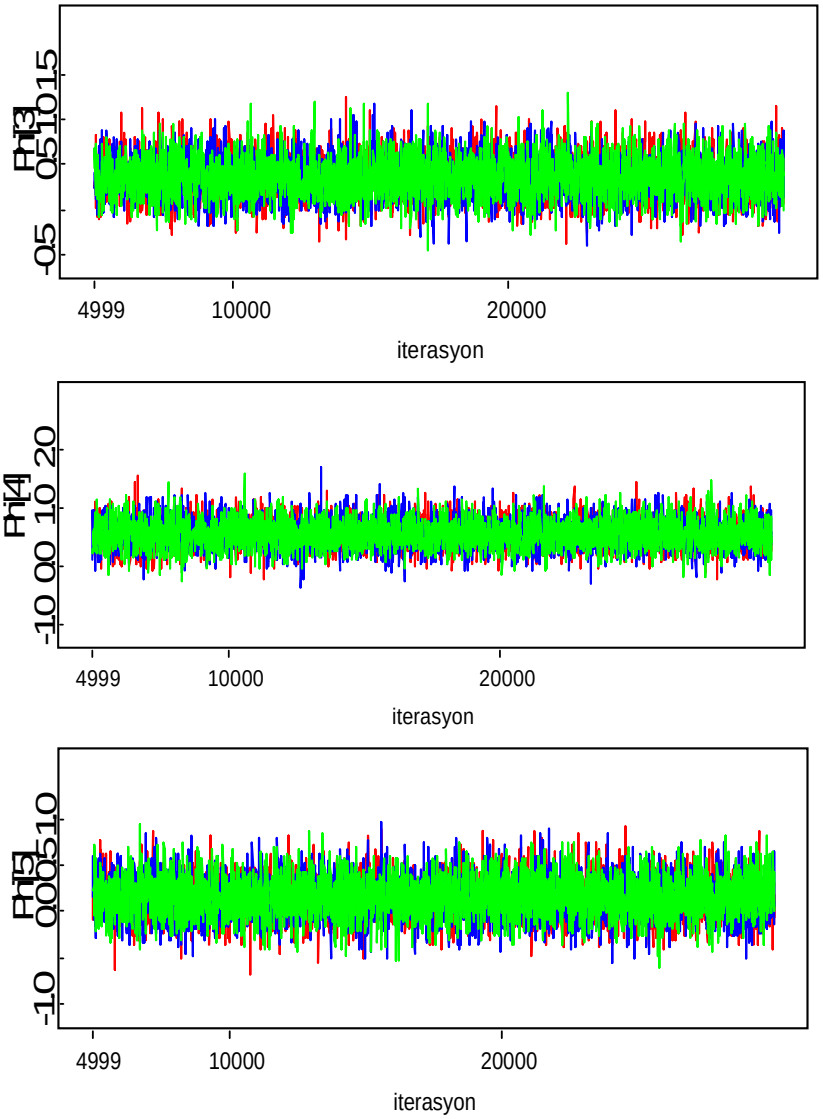




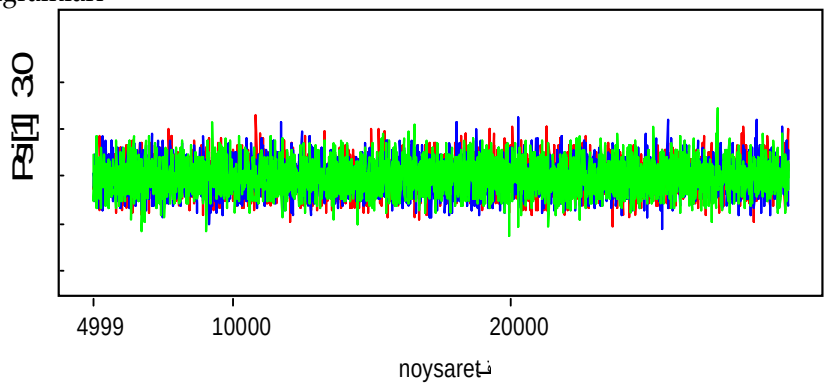
Puanlayıcı katılık parametrelerine benzer olarak, puanlayıcı değişkenlik parametreleri, BGR grafikleri için ikinci puanlayıcının neredeyse bir, diğer puanlayıcıların da bire oldukça yakın değerler aldığı görülmektedir. Fakat puanlayıcı bir ve puanlayıcı beşin diğer puanlayıcılara göre bir değerinden biraz daha fazla uzaklaştığı görülmektedir. Bu nedenle, ilgili puanlayıcılar için BGR değerleri incelenmiş ve birinci puanlayıcının son iterasyon aralığı için BGR değerinin, 1.002 olduğu ve beşinci puanlayıcı için bu değerin 1.004 olduğu görülmüştür. Her iki değerin, 1.0 ile 1.1 aralığında yer alması dolayısıyla kabul edilebilir değerlere sahip oldukları ifade edilebilir. İncelenmesi gereken diğer bir gösterge; zaman serileri diyagramları, puanlayıcı katılık ve değişkenlik parametreleri için aşağıda verilmiştir.

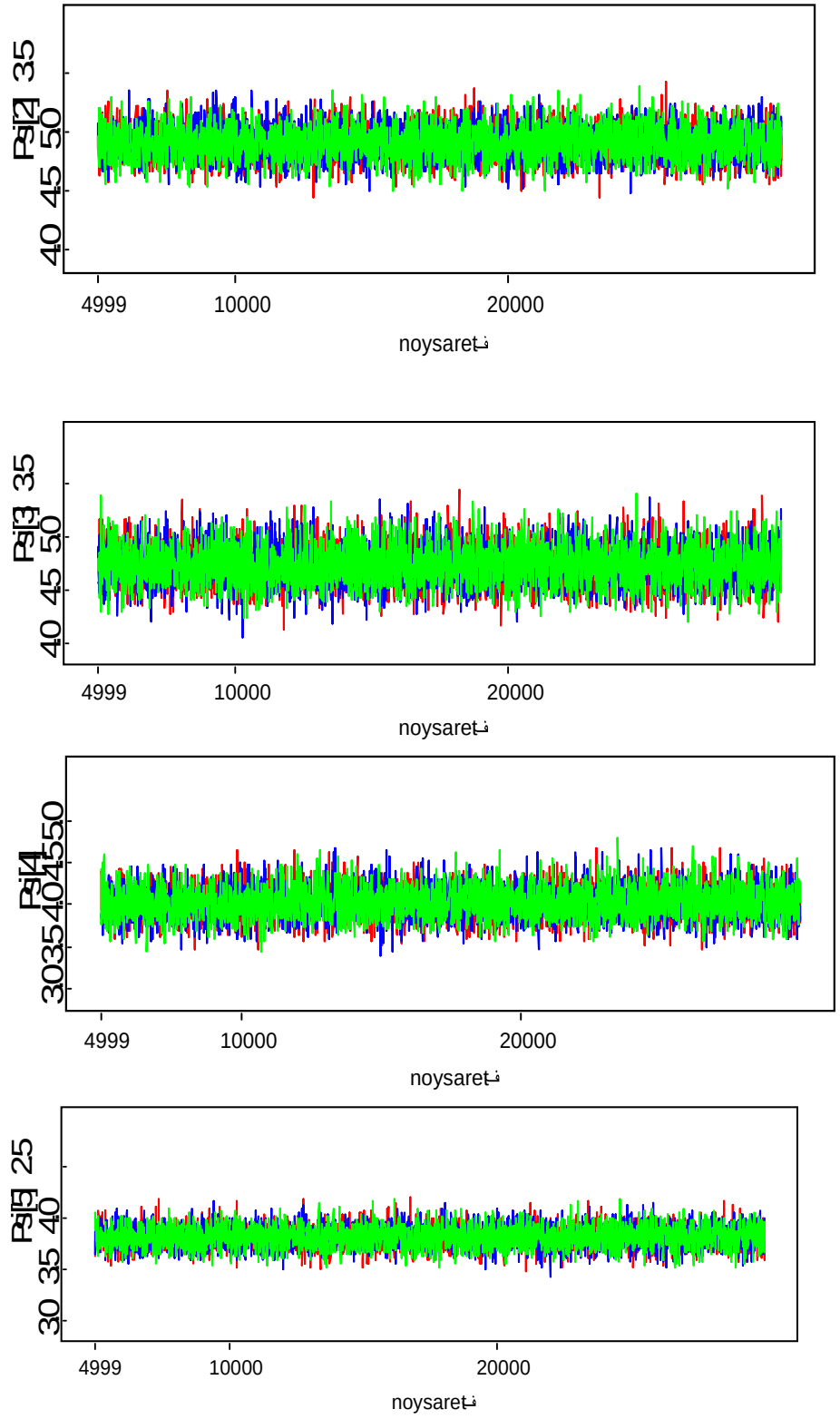
Diyagram 1. Puanlayıcı Katılık Parametrelerine İlişkin Zaman Serileri Diyagramları





Diyagram 2. Puanlayıcı Değişkenlik Parametrelerine İlişkin Zaman Serileri Diyagramları




Psi2
Psi3
Psi4
Psi5
4999
10000
20000
noysaretف

Puanlayıcı parametrelerine ait diyagramlar incelendiğinde, örneklenen değerlerin, ortak bir ortalama değeri etrafında seçkisiz olarak hareketli olduğu ve üç Markov zincirinin binişik ya da bir noktada birleşmiş olduğu görülmektedir. Bu bağlamda zaman serileri diyagramlarının istenen şekle sahip olduğu, başka bir ifadeyle üç zincirin yakınsadığı söylenebilir. BGR grafikleri ve zaman serileri diyagramlarına dayanılarak, her bir puanlayıcı için iki parametrenin de rapor edilebilir niteliğe sahip olduğu ifade edilebilir. Puanlayıcı parametrelerinin, sonsal dağılımı yakınsadığı belirlendikten sonra Tablo 3'de, puanlayıcıların katılık ve değişmezlik değerlerine ait ortalama, standart sapma, MC hatası, medyan ve güven aralığı değerleri verilmiştir.

Tablo 3. HPM Puanlayıcı Katılığı MCMC Kestirimi Sonsal Değerleri

Parametreler	Ortalama	Standart Sapma	MC Hatası	Median	Güven Aralığı Alt sınır - Üst Sınır
Phi (ϕ_1)	0.6873	0.12705	0.001649	0.6872	0.7124 – 0.6621
Psi (ψ_1)	0.5008	0.16780	0.001789	0.5208	0.5218 –0.4801
Phi (ϕ_2)	0.0346	0.14505	0.001580	0.0347	0.0628 – 0.0049
Psi (ψ_2)	0.4735	0.13510	0.001659	0.4766	0.4911 – 0.4624
Phi (ϕ_3)	0.3558	0.12500	0.001433	0.3559	0.3799 – 0.3313
Psi (ψ_3)	0.4865	0.18051	0.002178	0.4869	0.5071 – 0.4659
Phi (ϕ_4)	0.5476	0.11760	0.002310	0.5476	0.5707 – 0.5245
Psi (ψ_4)	0.4988	0.18590	0.002142	0.4991	0.5216 – 0.4771
Phi (ϕ_5)	0.1400	0.14470	0.001530	0.1401	0.1686 – 0.1111
Psi (ψ_5)	0.5493	0.10311	0.001226	0.5493	0.5671 – 0.5334

Phi (ϕ_r) parametresi bireysel olarak puanlayıcı r'nin katılık (severity) ya da yanlılık (bias) olarak da adlandırılan değerlerinin ölçülmesine olanak sağlar. Bu parametre, 0.0 a eşit olduğunda (ϕ_r =0) puanlayıcı r'nin çoğunlukla ideal puan kategorisinde, puan atadığı (ϕ_r =ξ); -0.5'ten küçük olduğunda (ϕ_r<-0.5) puanlayıcı r'nin çoğunlukla ideal puan kategorisinden daha düşük kategorilerde puan atadığı (ϕ<ξ), başka bir ifadeyle ideal kategori göz önünde bulundurulduğunda daha katı bir puanlayıcı davranışı sergilediği söylenebilir. Bu değer 0.5'ten büyük olduğunda (ϕ>0.5) ise puanlayıcı r'nin çoğunlukla ideal puan (ϕ_r>ξ) kategorisinden daha yüksek kategorilerde puan atadığı ya da ideal kategori göz önünde bulundurulduğunda daha

cömert bir puanlayıcı davranışı sergilediği ifade edilebilir. Psi (ψ_r) parametresi ise bireysel olarak puanlayıcı r'nin güvenirlik noktasında yetersizliğini (lack of reliability) yansıtmaktadır. Phi (ψ_r) parametresinin 0'a yakın değerler alması yüksek tutarlılığın ya da atanan puanların güvenirliğinin göstergesiyken, yüksek değerler atanan puanlardaki zayıf tutarlılığın göstergesidir (Casabianca ve diğerleri, 2014). Daha genel olarak, tüm puanlayıcıların ϕr ve ψr parametrelerinden 0.0'a yakın bir değer alması istenen bir durumun göstergesidir. Çünkü ancak bu durumda, güvenilir biçimde birbirleriyle fikir birliği oluşturmuş bir grup puanlayıcının varlığından söz edilebilir (Patz ve diğerleri, 2002).

İdeal puan kategorisi, atanan tüm puanlar dikkate alınarak HPM tarafından elde edilir. Temelde puanlayıcıların üzerinde fikir birliğine ulaştıkları puanlardır (consensus rating). Bu nedenle sıfıra yakın değerler alan puanlayıcı parametreleri, her bir öğrenci yanıtlarının puanlanması noktasında, puanlayıcıların fikir birliğine ulaştıklarının bir göstergesidir.

Tablo 3'de yer alan değerler incelendiğinde; 2., 3. ve 5. puanlayıcıların puanlayıcı katılık parametrelerinin mutlak değerce 0.5'ten küçük ($|\phi_r|<0.5$) olduğu görülmektedir. İlgili puanlayıcıların diğer kategorilerden ziyade ideal puan kategorisinde ya da bu kategoriye yakın bir kategoride puanlama yaptığı söylenebilir. Ayrıca bu üç puanlayıcının maddelere verilen öğrenci yanıtlarına puan atama noktasında birbirleriyle uyum içinde oldukları da ifade edilebilir. Bu puanlayıcılar arasında 2. puanlayıcının en az yanlılık gösteren puanlayıcı olduğu ve ideal puan kategorisine en yakın puanları atadığı söylenebilir. Puanlayıcı katılık parametreleri 0.5'ten büyük olan iki puanlayıcı (1.ve 4.) vardır. Her iki puanlayıcının parametre değerleri (sırasıyla; 0.6872, 0.5476) doğrultusunda ideal puanlardan daha yüksek kategorilerde puanlama yaptıkları başka bir ifadeyle daha cömert puanlayıcılar oldukları ve pozitif yanlılık gösterdikleri söylenebilir. Fakat 4. puanlayıcının 1. puanlayıcıya nazaran ideal puan kategorisine daha yakın puanlar atadığı ve 0.5 değerine oldukça yakın bir değer aldığı da dikkate alınmalıdır. Bu doğrultuda, puanlayıcılar arasında en yanlı davranan puanlayıcının 1. puanlayıcı olduğu ifade edilebilir. Puanlayıcılardan hiç birinin negatif yanlılık göstermediği ya da öğrencilere ideal puan değerinden daha düşük değerler verme eğiliminde olmadığı da saptanmıştır.

Puanlayıcı katılık parametrelerinin aksine, puanlayıcı değişkenlik parametrelerinin birbirine oldukça yakın değerler aldığı tablo 3'de görülmektedir. Puanlayıcı değişkenlik parametresi değeri 0'a en yakın olan puanlayıcı 0.4766 değeri ile 2. puanlayıcıdır. Başka bir ifadeyle, 2. puanlayıcının en güvenilir puanlayıcı olduğu söylenebilir. Ayrıca 2. puanlayıcının tablo 3'de yer alan her iki parametre değeri incelendiğinde diğer puanlayıcılara nispeten hem daha güvenilir hem de daha az yanlı puanlar atadığı da ifade edilebilir. Puanlayıcı katılık parametresi dikkate

alındığında ideal kategoriye oldukça yakın puan atayan 5. puanlayıcı; değişkenlik parametresi dikkate alındığında en yüksek değere sahip ve dolayısıyla en az güvenilir olan puanlayıcıdır. Başka bir ifadeyle, 5. puanlayıcının aynı niteliğe sahip öğrenci yanıtlarına puan atama noktasında daha az tutarlı olduğu söylenebilir. Elde edilen bu bulgunun, sadece katılık parametresi dikkate alınarak puanlayıcılara dair yargıya varmanın zaman zaman yanıltıcı olabileceğinin göstergesi olması bakımından önemli olduğu düşünülmektedir.

Çok Değişkenlik Kaynaklı Rasch Ölçme Modeli ve Hiyerarşik Puanlayıcı Modeline ait Sapma Bilgi Kriteri Değerleri

HPM model veri uyumuna ilişkin bilgi edinilebilmesi ve ÇDKRÖM ile HPM sonuçlarının karşılaştırılabilmesi açısından SBK değerlerinden faydalanılmıştır. Tablo 4'de SBK ve SBK'nın dayandığı değerlere yer verilmiştir.

Tablo 4. ÇDKRÖM ve HPM'ye ait Sapma Bilgi Kriteri Değerleri

	pD	D	SBK
ÇDKRÖM	36	3657	3693
HPM	44	3641	3685

Tablo 4'de, her iki model için de pD, D ve $\overline{SBK}$ değerleri yer almaktadır. Model karşılaştırmalarında Literatürde yer alan çalışmaların bir kısmında SBK'nın, doğru modele değil de var olan modellerden veriye en iyi uyum gösteren modele işaret ettiği vurgulanmakla birlikte; daha küçük SBK değerine sahip olan modeli, doğru model olarak nitelendiren çalışmalar bulunmaktadır (Liddle, 2007; Mason, Richardson ve Best, 2012; Spiegelhalter, Best, Carlin ve Van der Linde, 1998; Spiegelhalter ve diğerleri, 2002). Ayrıca, SBK için önemli sayılabilecek farklılığın ne olduğuna dair net bir bilgi literatürde yer almamaktadır. Genel olarak, farklılaşan 10 ve üzeri SBK değerleri için yüksek SBK'ya sahip olan modelin çıkarılması ya da reddedilmesi önerilmiştir. Bununla birlikte 5-10 birim arasında gözlenen farklılığın önemli olarak değerlendirilmesi ve düşük SBK değerine sahip olan modelin tercih edilmesi belirtilmiştir. Farklılaşan beş ve altı SBK değeri için de modellerin çok farklı sonuçlara sahip olduğu ve sadece düşük SBK değerinin raporlaştırılmasının yanıltıcı olacağı ifade edilmiştir (MRC Biostatistics Unit, 2014). Bu bilgiler ışığında, literatürde farklı değerler ve yorumlamalar bulunmakla birlikte hangi ya da hangilerinin dikkate alınacağının araştırmacıya ait bir karar olduğu ve tüm çalışmalarda kabul gören ölçütün, daha küçük SBK değerine sahip olan modelin ele alınan modeller arasında en iyisi olarak düşünülebileceği ifade edilebilir. Son

olarak da küçük SBK değerine sahip olan model için diğer modellere göre SBK'yı oluşturan unsurlardan $\overline{D}$'in küçük pD'nin ise büyük olması beklenen durum olduğu ifade edilebilir (Berg ve diğerleri, 2004). HPM için elde edilen SBK değeri, 3685 iken ÇDKRÖM için 3693'dür. İki modelin SBK değeri farkı, sekiz olarak bulunmuştur. Başka bir ifadeyle, HPM SBK değeri, ÇDKRÖM SBK değerinden daha küçük bir değer olarak bulgulanmıştır. Her iki modele ait $\overline{D}$ ve pD değerleri incelendiğinde SBK değerleriyle paralel ve beklenen yönde olduğu görülmektedir. HPM için elde edilen $\overline{D}$ değeri (3641), ÇDKRÖM için elde edilen $\overline{D}$ değerinden (3657) daha küçükken, HPM için elde edilen pD değeri (44) ÇDKRÖM için elde edilen pD değerinden (36) daha büyüktür. Bu bağlamda, karşılaştırılan iki model içinden HPM'nin, çalışmada kullanılan veriler açısından daha iyi bir model olduğu ve SBK değerleri arasında gözlenen farklılığın önemli olarak nitelendirilebileceği belirtilebilir. Başka bir deyişle, HPM'nin ÇDKRÖM'ne nazaran verilere önemli derecede daha iyi uyum gösterdiği ve model dâhilinde ulaşılan parametreler için daha doğru kestirimler yaptığı ifade edilebilir. Elde edilen bu bulguların, araştırma verileri kapsamında, her iki modele ilişkin literatürde yer alan kısıtlı sayıdaki çalışmalarla (Mariano, 2002; Patz ve diğerleri, 2000; Patz ve diğerleri, 2002) paralellik taşıdığı ifade edilebilir.

Tartışma

Öğrenci yanıtlarına, birbirinden bağımsız beş puanlayıcının atadığı puanların, öğrenciler için uygun olduğu tespit edilmiştir. Ayırma indeksi ve güvenirliği dikkate alındığında, bu puanlayıcıların katılık/cömertlik düzeylerinin farklılaştığı belirlenmiştir. Bu farklılığa karşın, puanlayıcıların katılık/cömertlik düzeylerinin birbirine yakın olduğu ve her bir puanlayıcının, kendi içinde öğrencilere atadıkları puanların tutarlı olduğu sonucuna varılmıştır.

HPM'de, puanlayıcılar için analizde faydalanılan üç zincirin, yakınsadığı kanıtlarına erişilmiştir. Puanlayıcı katılık parametresi doğrultusunda, puan atayan toplam beş puanlayıcının, en az en çok yanlılık gösteren puanlayıcılar olduğu belirlenirken, üç puanlayıcının birbirleriyle oldukça uyum içinde ve ideal şekilde puan atadıkları tespit edilmiştir. Özellikle birinci puanlayıcının, diğer puanlayıcılardan farklı olarak pozitif yanlılık gösterdiği ve negatif yanlılık gösteren bir puanlayıcının olmadığı sonucuna varılmıştır. Puanlayıcı davranışlarının daha detaylı incelenmesine olanak sağlayan, puanlayıcı değişkenlik parametresi göz önünde bulundurulduğunda, genel olarak puanlayıcıların birbirine yakın güvenirlikte puan atadıkları sonucu elde edilmiştir. Puanlayıcı katılık parametresine göre uygun puan atayan beşinci puanlayıcının, benzer öğrenci yanıtlarına daha az tutarlı (diğer puanlayıcılara göre daha az güvenilir olması) puan ataması ise araştırma kapsamında ulaşılan, dikkat çekici bir sonuçtur.

İki modele ilişkin, sadece karşılaştırabilir değerleri doğrultusunda ve sıralama düzeyinde elde edilen bilgilerin yetersiz olması nedeniyle her iki modele yönelik doğrudan bilgi elde edilmesi amacıyla sapma bilgi kriterinden faydalanılmıştır. Bu kriter doğrultusunda, HPM'nin ÇDKRÖM'e göre araştırma verilerine daha iyi uyum sağladığı ve tek bir maddenin, tek bir yanıtına ilişkin atanan çoklu puanlara ait bir yapının, hiyerarşik puanlayıcı modelince daha iyi yansıtıldığı sonucuna ulaşılmıştır.

Puanlayıcı davranışlarının incelenmesinin önemli görüldüğü durumlarda, HPM'nin kullanımının daha detaylı bilgi sunması bakımından yararlı olabileceği ve kullanılabileceği önerilebilir.

Analizin gerçekleştirildiği programa verilerin tanıtılması, analizin gerçekleştirilebilmesi için harcanan emek ve analiz çıktılarının elde edilme süresi gibi koşullar dikkate alındığında, ÇDKRÖM'nin HPM'ye göre oldukça kullanışlı olması, araştırmanın amacına uygun modelin seçimi noktasında göz önünde bulundurulmalıdır.

Kaynakça

Airasian, P.W. (2001). Classroom assessment: Concepts and applications. Boston: McGraw-Hill.

Akın, Ö. & Baştürk, R. (2012). Keman Eğitiminde Temel Becerilerin Rasch Ölçme Modeli İle Değerlendirilmesi. Pamukkale Üniversitesi Eğitim Fakültesi Dergisi, 31 (31), 175-187.

Andrich, D. (1978). A rating formulation for ordered response categories. Psychometrika, 43(4), 561-573.

Atılgan, H. (2005b). Müzik öğretmenliği özel yetenek seçme sınavının çok-yüzeyli rasch modeli ile analizi (İnönü üniversitesi örneği). Eurasian Journal of Educational Measurement, 20, 62 – 73.

Brennan, R.L. (1992). Generalizability theory. Educational Measurement: Issues and Practice, 11(4), 27-34.

Brennan, R.L. (1997). A Perspective on the history of generabability theory. Educational Measurement: Issues and Practice, 16(4), 14-20.

Brennan, R.L. (2010). Generalizability theory and classical test theory. Applied Measurement in Education. 24(1), 1-21.

Cardinet, J., Tourneur, Y. & Allal, L. (1981). Extension of generalizability theory and its applications in educational measurement. Journal of Educational Measurement, 18(4), 183-204.

Casabianca, J.M. & Junker, B. (2013). Hierarchical rater models for longitudinal assessments. Annual Meeting of the National Council for Measurement in Education'da sunulan bildiri. San Francisco, California.

Casabianca, J.M. & Junker, B. (2014). The hierarchical rater model for evaluating changes in traits over time. 121st Annual Convention of the American Psychological Association, Division 5: Evaluation, Measurement and Statistics'te sunulan bildiri. Washington D.C.

Christensen, R., Johnson, W., Branscum, A. & Hanson, T.E. (2011). Bayesian ideas and data analysis: An introduction for scientists and statisticians. CRC Press, USA.

DeCarlo, L.T. (2005). A model of rater behavior in essay grading based on signal detection theory. Journal of Educational Measurement, 42(1), 53-76.

DeCarlo, L.T. (2010). Studies of a latent class signal detection model for constructed response scoring II: Incomplete and hierarchical designs. ETS Research Report Series, (08). Princeton, NJ: Educational Testing Service.

DeCarlo, L.T., Kim, Y.K. & Johnson, M.S. (2011). A hierarchical rater model for constructed responses, with a signal detection rater model. Journal of Educational Measurement, 48(3), 333-356.

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.

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. & 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, (01). Princeton, NJ: Educational Testing Service.

Gelman, A., Carlin, J.B., Stern, H.S. & Rubin, D.B. (1995). Bayesian data analysis. New York, NY: Chapman & Hall.

Iramaneerat, C., Myford, C.M., Yudkowsky, R. & Lowenstein, T. (2009). Evaluating the effectiveness of rating instruments for a communication skills assessment of medical residents. Advances İn Health Sciences Education,14(4), 575-594.

Jonsson, A. & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. Educational Research Review, 2(2), 130-144.

Kastner, M. & Stangla, B. (2011). Multiple choice and constructed response tests: Do test format and scoring matter? Procedia-Social and Behavioral Sciences, 12, 263-273.

Kéry, M. (2010). Introduction to WinBUGS for ecologists: Bayesian approach to regression, ANOVA, mixed models and related analyses. USA: Academic Press.

Kim, Y.K. (2009). Combining constructed response items and multiple choice items using a hierarchical rater model (Doktora Tezi). Teachers College, Columbia University.

Liddle, A.R. (2007). Information criteria for astrophysical model selection. Monthly Notices of the Royal Astronomical Society: Letters, 377(1), 74-78.

Linacre, J.M. (1989). Many facet rasch measurement (Doktora tezi). University of Chicago, Chicago.

Linacre, J.M., Wright B.D. & Lunz M.E. (1990). A Facets Model of Judgmental Scoring. Memo 61. MESA Psychometric Laboratory. University of Chicago. www.rasch.org/memo61.html.

Linacre, J.M. (1994). Many-facet Rasch measurement. Chicago: Mesa Press.

Linacre, J.M. (2003). The hierarchical rater model from a Rasch perspective. Rasch Measurement Transactions (Transactions of the Rasch Measurement SIG American Educational Research Association), 17(2), 928.

Lund, J.L. & Veal, M.L. (2013). Assessment-driven ınstruction in physical education with web resource: A standards-based approach to promoting and documenting learning. Human Kinetics.

Lynch, B.K. & McNamara, T.F. (1998). Using G-theory and many-facet rasch measurement in the development of performance assessments of the ESL speaking skills of ımigrants. Language Testing, 15(2), 158-180.

Mariano, L.T. (2002). Information accumulation, model selection and rater behavior in constructed response student assessments (Doktora Tezi). Carnegie Mellon University, Pennsylvania.

Mariano, L.T. & Junker, B.W. (2007). Covariates of the rating process in hierarchical models for multiple ratings of test items. Journal of Educational and Behavioral Statistics, 32, 287–314.

Mertler, C.A. (2001). Designing scoring rubrics for your classroom. Practical Asssessment Reaserch and Evaluation, 7(25), 1-10.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. Educational Researcher, 23(2), 13-23.

Nakamura, Y. (2000). Many facet rasch based analsis of communıcative language testing results. Journal of Communication Students, 12, 3-13.

Patz, R. J. & Junker, B. W. (1999a). The hierarchical rater model for rated test items and its application to large-scale assessment data. Annual meeting of

the American Educational Research Association'nda sunulan bildiri. Montreal, Quebec, Canada.

Patz, R.J. & Junker, B.W. (1999b). A straightforward approach to Markov chain Monte Carlo methods for item response models. Journal of Educational and Behavioral Statistics, 24(2), 146-178.

Patz R.J., Junker B.W. & Johnson M.S. (2000) The Hierarchical Rater Model for Rated Test Items and its Application to Large-Scale Educational Assessment Data. Revised AERA Paper.

Patz, R.J., Junker, B.W., Johnson, M.S. & Mariano, L.T. (2002). The hierarchical rater model for rated test items and its application to large-scale educational assessment data. Journal of Educational and Behavioral Statistics, 27(4), 341384.

Popham, W.J. (1997). What's wrong-and what's right-with rubrics. Educational Leadership, 55, 72-75

Popham, W.J. (2008). Classroom assessment what teachers need to know. USA: Pearson Education.

Quinlan, A.M. (2011). A complete guide to rubrics: assessment made easy for teachers, kd college. R&L Education.

Roid, G.H. & Haladyna T.M. (1982). A technology for test-ıtem writing. New York: Academic Pres.

Rodriquez, M. C. (2002). Choosing An Item Format. Tindal, G. ve Haladyna, T.M. (Ed.). Large-Scale Assessment Programs For All Students (213-231). New Jersey: Lawrence Erlbaum Associates Publishers.

Spiegelhalter, D., Thomas, A., Best, N. & Lunn, D. (2003). WinBUGS user manual.

Stevens, D. & Levi, A. (2005). Introduction to rubrics. Sterling, Va.: Stylus Pub.

Sudweeks, R.R., Reeve, S. & Bradshaw, W.S. (2004). A comparison of generalizability theory and many-facet Rasch measurement in an analysis of college sophomore writing. Assessing Writing, 9(3), 239-261.

Turner, J. (2003). Examining on art portfolio assessment using a many facet rasch measurement model (yayınlanmamış doktora tezi). Boston College, Boston.

Verhelst, N. & Verstralen, H. (2001). IRT models for multiple raters. A. Boomsma, T. Snijders, and M. van Duijn, (Ed.). In essays in ıtem response modeling. New York: Springer-Verlag.

Wilson, M. & Hoskens, M. (2001). The rater bundle model. Journal of Educational and Behavioral Statistics, 26, 283–306.