ANKARA ÜNİVERSİTESİ
EĞİTİM BİLİMLERİ ENSTİTÜSÜ

ÖLÇME VE DEĞERLENDİRME ANABİLİM DALI

ÇOK DEĞİŞKENLİK KAYNAKLI RASCH ÖLÇME MODELİ VE HİYERARŞİK PUANLAYICI MODELİ İLE KESTİRİLEN PARAMETRELERİN KARŞILAŞTIRILMASI

DOKTORA TEZİ

MÜGE ULUMAN

ANKARA, TEMMUZ, 2015

ANKARA ÜNİVERSİTESİ
EĞİTİM BİLİMLERİ ENSTİTÜSÜ
ÖLÇME VE DEĞERLENDİRME ANABİLİM DALI

ÇOK DEĞİŞKENLİK KAYNAKLI RASCH ÖLÇME MODELİ VE HİYERARŞİK PUANLAYICI MODELİ İLE KESTİRİLEN PARAMETRELERİN KARŞILAŞTIRILMASI

DOKTORA TEZİ

Müge ULUMAN

Tez Danışmanı
Prof. Dr. Ezel TAVŞANCIL

Ankara
Temmuz, 2015

Eğitim Bilimleri Enstitüsü Müdürlüğü'ne

Müge ULUMAN'ın hazırladığı "Çok Değişkenlik Kaynaklı Rasch Ölçme Modeli ve Hiyerarşik Puanlayıcı Modeli ile Kestirilen Parametrelerin Karşılaştırılması" başlıklı bu çalışma jürimiz tarafından Ölçme ve Değerlendirme Anabilim Dalı/Ölçme ve Değerlendirme Programı'nda Doktora Tezi olarak kabul edilmiştir.

		İmza
Başkan	Prof. Dr. Nizamettin KOÇ	.....
Üye	Prof. Dr. Ezel TAVŞANCIL (Danışman)	.....
Üye	Prof. Dr. Mehtap ÇAKAN	.....
Üye	Doç. Dr. Ömay ÇOKLUK	.....
Üye	Doç. Dr. Duygu ANIL	.....

ONAY

Bu tez Ankara Üniversitesi Lisansüstü Eğitim-Öğretim ve Sınav Yönetmeliği'nin ilgili maddeleri uyarınca yukarıdaki jüri üyeleri tarafından ./..../20.. tarihinde uygun görülmüş ve Enstitü Yönetim Kurulunca .../.../20.. tarihinde kabul edilmiştir.

Prof. Dr. İsmail GÜVEN

Eğitim Bilimleri Enstitüsü 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.

Müge ULUMAN

ÖZET

ÇOK DEĞİŞKENLİK KAYNAKLI RASCH ÖLÇME MODELİ VE HİYERARŞİK PUANLAYICI MODELİ İLE KESTİRİLEN PARAMETRELERİN KARŞILAŞTIRILMASI

Uluman, Müge
Doktora, Ölçme ve Değerlendirme Anabilim Dalı
Tez Danışmanı: Prof. Dr. Ezel TAVŞANCIL
Temmuz 2015, xii + 111 sayfa

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 parametrelerin kestirilmesi ve her iki modele ilişkin parametrelerin birlikte değerlendirilmesi amaçlanmıştır.

Temel araştırma modelindeki araştırmanın verileri, 2012-2013 eğitim-öğretim yılı II. 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. Öğrencilerin 30'undan elde edilen yanıtlar, öğretmenlere puan atama sürecinde rehber olması bakımında bütünsel dereceli puanlama anahtarlarının geliştirilmesi aşamasında kullanılmak üzere analiz dışında tutulmuştur. Araştırma kapsamında tamamen çaprazlanmış desen kullanılmış, ÇDKRÖM ve HPM analizlerinin gerçekleştirilebilmesi için OpenBUGS ve FACET programlarından faydalanılmıştır.

Araştırma sonucunda, ÇDKRÖM ve HPM öğrenci, puanlayıcı ve madde değişkenlik kaynaklarına ait parametre sonuçlarının genel olarak benzer olduğu saptanmıştır. Değişkenlik kaynakları arasında en yüksek ilişkinin puanlayıcı katılık/cömertlik parametrelerinde görüldüğü ve her iki model için elde edilen, puanlayıcı katılık/cömertlik sıralamalarının aynı olduğu tespit edilmiştir. Madde değişkenlik kaynağı için korelasyon katsayısının yüksek düzeyde ilişkiye işaret ettiği belirlenmiş ve değerleri birbirine yakın olmakla birlikte iki madde dışında sıralamanın benzer olduğu; öğrenci değişkenlik kaynağına göre de yüksek düzeyde ilişkinin görüldüğü sonuçları elde edilmiştir. 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.

SUMMARY

COMPARING PARAMETERS OF MANY FACET RASCH MEASUREMENT MODEL AND HIERARCHIAL RATER MODEL

Uluman, Müge

Ph.D.,Department of Measurementand Evaluation

Supervisor: Prof. Dr. Ezel TAVŞANCIL

July 2015, xii + 111 pages

This study aims at estimating the parameters with many facet Rasch measurement model (MFRMM) and hierarchical rater model (HRM) and evaluating together the parameters obtained from both models if responses given by the same examinees for open-ended items are scored by multiple raters.

The study was designed with basic research. 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. Responses obtained from 30 points were excluded from the analysis to be used in development of the holistic scoring rubrics in order to guide teachers during the score assignment process. In this research, fully crossed design was used and OpenBUGS and FACET programs were utilized for realization of the MFRMM and HRM analyses.

The study revealed that parameters of MFRMM and HRM examinee, rater and item variability sources were similar in general. The highest relationship between sources of variability was found in rater severity/leniency parameters; also the order of rater severity/leniency obtained for both models was found to be thesame. For sources of item variability, it was determined that the correlation coefficient indicated a high level of relationship. Though values are close, the order was found similar except for two items and a high level of relationship was seen by source of student variability. 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 the HRM.

ÖNSÖZ

Eğitim alanındaki gelişmeler doğrultusunda, öğrenci başarısının belirlenmesi sürecinde çoktan seçmeli maddelerin yanı sıra açık uçlu maddelere de ihtiyaç duyulmuştur. Açık uçlu maddelerin kullanımı, birden fazla puanlayıcının işe koşulması gerekliliğini doğurarak, ölçme ve değerlendirme sürecini daha karmaşık hale getirmiştir. Bu durum, analizlerin gerçekleştirilebilmesi için farklı modellemelerin geliştirilmesini sağlamıştır. Literatürde bu modellemelere ait bilgiler bulunmakla birlikte modellemelerin sahip oldukları avantaj ve dezavantajlarına, kullanım durumlarına ilişkin yeterli bilgi bulunmamaktadır. Bu araştırmada da, modellemeler arasında yer alan çok değişkenlik kaynaklı Rasch ölçme modeli ve hiyerarşik puanlayıcı modellerine ait gerçekleştirilen uygulamaların, hem araştırmacılara hem de uygulayıcılara katkı getirmesi umulmaktadır.

Bu araştırmanın gerçekleştirilmesinde desteğini esirgemeyen, değerli öneri ve yorumlarıyla araştırmanın tamamlanmasına büyük katkı sağlayan danışmanım ve sevgili hocam Prof. Dr. Ezel TAVŞANCIL'a; tez izleme komitemde yer alan, değerli fikirleriyle tezime katkıda bulunan hocalarım Doç. Dr. Ömay ÇOKLUK ve Doç. Dr. Duygu ANIL'a teşekkürlerimi ve saygılarımı borç bilirim.

Doktora eğitimim boyunca ölçme ve değerlendirme alanında bilgilerini benimle paylaşan, bilim insanı olma yolunda büyük emeği olan değerli hocalarım Prof. Dr. Nizamettin KOÇ'a, Prof. Dr. Nükhet DEMİRTAŞLI'ya çok teşekkür ederim.

Ülkemden kilometrelerce uzaktayken, konu ne olursa olsun her tökezlediğimde beni destekleyen Doç. Dr. Asuman TÜRKMEN'e; araştırma konusunun görmekte zorlandığım temel noktalarının sabırla, tek tek aydınlanmasını sağlayan, beni cesaretlendiren Prof. Dr. Paul DEBOECK'a; görüş ve önerileriyle çalışmaya yeni yollar açan, sorularımla bunalttığım sevgili Yrd. Doç. Dr. Dylan MOLENAAR'a; çalışma boyunca gerek birebir gerekse kilometrelerce öteden desteğini hiç esirgemeyen, çok sevgili arkadaşım Dr. Haijin CHEN'e; Ohio Devlet Üniversitesi'nde araştırmacı olarak bulunma fırsatı sunan, doğru kaynaklara

ulaşmamı sağlayan ve konu ne olursa olsun desteğini hissettiğim sevgili Yrd. Doç. Dr. C. Paul GUGİU'ya teşekkürü borç bilirim.

Araştırma boyunca en kötü anlarımda nefes almamı sağlayan ve sabırla beni dinleyen, bilim insanı olma yolunda destek olan, sevgili arkadaşlarım, Fatih KEZER, Sabri DOĞAN, Emel BAYDAN, Rümeysa ŞAHİN, Ela FURAT, Gonca USTA ve Alper BÜYÜKÇAKIR'a çok teşekkür ederim.

Doktora öğrenimi süresince yurtiçi doktora burs imkânı sağlayan TÜBİTAK'a teşekkür ederim.

Hayatımın en önemli ve değerli paydaşları olan ailem: ilk öğretmenim-babam, Yaşar ULUMAN; kıymetlim- annem Nilüfer ULUMAN ve diğer yarım-kardeşim Çağla ULUMAN'a her şey için teşekkürü borç bilirim.

Müge ULUMAN

İÇİNDEKİLER

Sayfa

ÇİZELGELER DİZİNİ

Sayfa

ŞEKİLLER DİZİNİ

Sayfa

BÖLÜM I

GİRİŞ

Bu bölümde, madde türlerine, puanlayıcı davranışlarına değinilmiş ve çok değişkenlik kaynaklı Rasch ölçme modeli ile hiyerarşik puanlayıcı modeli hakkında genel bilgiler sunulmuştur. Bununla birlikte, araştırmanın problemi, amaç, önem, sayıltılar, sınırlılıklar ve tanımlar verilmiştir.

Problem

Ölçme, belli bir nesnenin ya da nesnelerin belli bir özelliğe sahip olup olmadığının, sahipse sahip oluş derecesinin gözlenip gözlem sonuçlarının özellikle sayı sembolleriyle ifade edilmesidir (Tekin, 1987). Öğrenci davranışlarının ölçülebilmesi için ya davranışların, ya o davranışlar sonunda ortaya çıkan ürünün ya da her ikisinin gözlenip nicelendirilmesi gerekir. Ölçülecek davranışların her türünü gözlemlemek özellikle gerçek hayat koşullarında pek de mümkün değildir. Bu nedenle, eğitimde bilgiler, zihinsel beceriler, tutumlar, tercihler doğrudan değil, davranışın dışa vurulup sergileneceği ölçme durumları aracılığıyla ölçülür. Ölçme araçları ve özellikle araçlardaki maddeler davranışları ortaya çıkaracak uyarıcılardır (Turgut ve Baykul, 2012).

Öğrenci başarısının belirlenmesi, izlenmesi, ölçme ve değerlendirme sürecinin niteliği, bu süreçte kullanılan ölçme araçlarının niteliği ile doğrudan ilgilidir. Bu araçların niteliği, bir başka deyişle amaca uygun ve olabildiğince hatadan arınık ölçme yapabilme gücü ise maddelerin niteliği ile belirlenmektedir. Eğitim sürecinde farklı düzeylerde gerçekleşen öğrenmeleri yoklamak üzere farklı madde türleri geliştirilmiştir (Çıkrıkçı, 2010). Madde türlerine dair farklı sınıflamalar bulunsa da bilişsel alana ilişkin öğrencinin yanıt verme biçimine göre yapılan sınıflamalar daha yaygın kullanılmaktadır. Bu sınıflama içerisinde öğrencinin yanıtını seçerek verdiği çoktan seçmeli maddeler ve öğrencinin yanıtını kendisinin yapılandırdığı kurgu tepki maddeleri yer almaktadır (constructed response item) (Crocker ve Algina, 1986; Roid ve Haladyna, 1982).

Madde türü seçimine yönelik literatürde birçok tartışma gerçekleştirilmiştir. Herhangi bir madde türünün incelenen tüm özellikler bakımından diğerine göre üstünlüğünün mümkün olmadığı vurgulanmış ve ölçmek istenilen özelliğe hangi madde türü daha çok hizmet ediyorsa o madde türünün kullanılması önerilmiştir (Kastner ve Stangla, 2011; Popham, 2008; Rodriquez, 2002; Roid ve Haladyna, 1982). Martinez (1999) ise hiçbir madde türünün yalnız başına tüm eğitim amaçlarına uygun olmadığını belirterek, farklı madde formatlarının farklı boyutlarda (bilişsel özellikler, madde, test karakteristiği ve ekonomik koşullar) daha güçlü olduğunu ileri sürmüştür.

Eğitimde en sık kullanılan madde türü: öğrencinin yanıtını seçerek verdiği maddeler arasında yer alan çoktan seçmeli maddelerdir. Çoktan seçmeli maddeler ağırlıklı olarak bilginin ve zihinsel becerilerin ölçülmesine yönelik tasarlanmıştır. Çoktan seçmeli maddeler, problemin yer aldığı madde kökü ile probleme ilişkin doğru ve yanlış yanıtları içeren seçeneklerden oluşmaktadır. Yanlış yanıtları içeren seçenekler çeldiriciler olarak adlandırılır. Çoktan seçmeli maddeler öğrencinin doğru seçeneği seçmesine dayanır. Haladyna (1997), çoktan seçmeli maddelerin tercih edilme nedenlerini; daha iyi bir kapsam geçerliliği sunması, yüksek güvenirliğe sahip olması, kullanışlı ve nesnel olması ve optik okuyucular aracılığıyla puanlanabilir olması biçiminde sıralamıştır. Çoktan seçmeli maddelerin sınırlılıklarını ise öğrenme üzerindeki etkisi, belli amaçlara (yaratıcı düşünme, yazma vb.) hizmet edecek nitelikte madde yazılamaması ve yazma becerisinden yoksunluk biçiminde ifade etmiştir. Doğru yanıtın bulunması noktasında ölçülen bilgiye sahip olunmaksızın test becerisiyle hareket etmeye yönlendirmesi ve şans başarısı da çoktan seçmeli maddelerin sınırlılıkları arasında gösterilebilir.

Son yıllarda Ulusal Eğitimsel İlerlemeyi Değerlendirme (National Assessment of Educational Progress-NEAP), Eğitim Yetenek Testi (Scholastic Aptitude Test-SAT), İleri Düzey Yerleştirme (Advanced Placement-AP) Programı, Lisans Giriş Sınavı (Graduate Record Examination-GRE), Uluslararası Matematik ve Fen Eğilimleri Çalışması (Trends in International Mathematics and Science Study-TIMMS), Uluslararası Öğrenci Değerlendirme Programı (Programme for International Student Assessment-PISA) gibi birçok geniş ölçekli ulusal ve uluslararası değerlendirme çalışmaları, hem çoktan seçmeli hem de açık uçlu maddeleri içermektedir (DeCarlo, Kim, ve Johnson, 2011; Kim, 2009; Mariano, 2002). Bir sınananın kompozisyon oluşturması, matematik problemine çözüm

üretmesi ya da performansını ortaya koyması gibi durumlarda işe koşulan tüm maddeler kurgu tepki maddeleri olarak adlandırılabilir. Bunun yanı sıra açık uçlu maddeler de kurgu tepki maddeleri altında yer almaktadır (Kim, 2009; Mariano 2002). Çoktan seçmeli maddelerin aksine açık uçlu maddeler; seçenekler arasından doğru yanıtı seçmek yerine sınananların kendi yanıtlarını yapılandırmalarını gerektirir (Kastner ve Stangla, 2011; Messick, 1994; Rodriquez, 2002; Roid ve Haladyna, 1982). Öğretim, kolayca seçilecek yanıtın aksine öğrencinin yöneltilen soruya vereceği yanıtı planlayarak yapılandırmasını gerektiriyorsa açık uçlu maddeler çok kullanışlıdır (Haladyna, 1997). Açık uçlu maddelerin kullanımıyla, sınanan bireydeki bilgiye ilişkin daha derin ölçmeler yapılabilir ve çoktan seçmeli maddelerde sıklıkla karşılaşılan sınav deneyimi probleminden kaçınılabilir (Pollack, Rock ve Jenkins, 1992; Rodriquez, 2002). Bunun yanı sıra tam olarak yerleşmemiş bilgiler de saptanabilir (Cooper, 1984). Bu maddeler sadece yüksek yeterlilik seviyesine sahip öğrenciler hakkında değil, çok düşük yeterlilik seviyesine sahip öğrenciler hakkında da bilgi edinilmesine olanak sağlar (Ercikan ve diğerleri, 1998).

Açık uçlu maddelerin kullanımı; yeterliliğin doğrudan (direct) ve otantik (authentic) değerlendirilmesine olanak sunarak, eğitim için olumlu yönde katkı sağlamaktadır (Messick, 1994). Öğrencilerden, verilen bilgileri açıklamaları, düzenlemeleri, tanımlamaları, bir senteze ulaşmaları ve özgün fikirler üretmeleri isteniyorsa açık uçlu maddelerin kullanımı çok uygundur (Roid ve Haladyna, 1982). Açık uçlu maddelerin kullanılma nedenleri arasında; uygulanan testlerde sadece çoktan seçmeli madde kullanımının yetersizliğinin belirtilmesi ve bu dayanak alınarak gerçekleştirilen okul reformları, politik faktörler, üst düzey zihinsel becerilerin ölçülme ihtiyacı, bilişsel psikoloji, hesap verebilirlik ve testlerin kötüye kullanımı, çoktan seçmeli maddelere dayalı öğretimin yaygınlaşması gösterilebilir (Haladyna, 1997). Açık uçlu maddelerle, öğrencilerin gerçek hayatta sergilemeleri gereken davranış biçimlerine yakın öğrenci yanıtları elde edilebilir (Popham, 2008). Açık uçlu maddelerin sınırlılıkları arasında; maddelerin yanıtlanma sürelerinin uzun olması ve testte daha az sayıda madde bulunması nedeniyle kapsam geçerliliğinin zarar görmesi gösterilebilir. Haladyna (1997)a göre maliyetinin fazla olması, yanlılık ve puanlamadaki tutarsızlık da dikkate alınmalıdır.

Çoktan seçmeli maddeler ile açık uçlu maddeleri birbirinden ayıran önemli farklılıklardan birinin puanlanma biçimleri olduğu ifade edilebilir. Nesnel bir biçimde; doğru ya da yanlış olarak puanlanabilen çoktan seçmeli maddelerin aksine

açık uçlu maddelerin puanlanması çok daha güçtür (DeCarlo, 2005, 2010; DeCarlo ve diğerleri, 2011; Linacre, 2003; Popham, 2008; Wang, 2012). Aynı çoktan seçmeli test maddesine iki farklı puanlayıcının vereceği puan aynı olurken, aynı açık uçlu test maddesine iki farklı puanlayıcının aynı puanı ataması her zaman mümkün olmayabilir (Haladyna, 1997). Çünkü açık uçlu maddeler çoktan seçmeli maddelerde olduğu gibi 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.

Mariano (2002) ise her bir sınanan yanıtı için birden çok puanlayıcı kullanımının hem puanlayıcılar arası, hem de puanlayıcıların kendi kararları arasındaki (puanlayıcı içi) mevcut durumu doğrudan ve doğru biçimde modelleyebilmeyi sağlarken, yetenek kestirimindeki kesinliği arttırabilmek için de bir fırsat sunacağını ifade etmiştir. Donoghue ve Hombo'ya (2000) göre puanlayıcı sayısının birden fazla olması; sınanan yeterliliğinin belirlenmesindeki kesinliği arttırırken, bazı problemleri (puanlayıcı eğilimlerinin, puanlayıcılar arası ortak yönlerin belirlenmesi vb. noktasında karşılaşılan zorluklar) de beraberinde getirir.

Test sürecindeki amaç, her bir sınanan tarafından sergilenen performansa puanlayıcılar tarafından atanan puanların, olabildiğince doğru (accurate), adaletli (fair) ve yararlı (useful) bir biçimde saptanabilmesidir (Linacre, 1994). Puanlama sürecinde puanlayıcıların yer alması birçok potansiyel ölçme hatasının (puanlayıcılar tarafından durumların hatalı ya da tutarsız bir biçimde tanımlanması ve çeşitli puanlayıcı cömertliği, tutarsızlığı ve yanlılığı gibi puanlayıcı tipi hatalar) ortaya çıkmasına neden olur (Myford ve Wolfe, 2003). Potansiyel ölçme hataları sonucu elde edilen puan değişimleri, ölçmedeki uygun olmayan varyansın doğmasına ve puanların anlamlı bir biçimde yorumlanması bakımından geçerliliğin tehdidine yol açar. Bu bağlamda, bu ölçme hatalarını yansıtan niteliğin kontrol edilmesi, elde edilen puanların geçerli bir biçimde yorumlanması yönünden kritik öneme sahiptir (Downing ve Haladyna, 2004; Iramaneerat, Yudkowsky, Myford ve Downing 2008). Özellikle puanlayıcıların atadıkları puanlar arasında farklılıklar olduğunda genellikle çözülmesi zor olan bir durumla karşılaşılır ve ölçme modelinde puanlayıcıların nasıl farklılaştığının, bu farklılıkların nasıl ele alınacağının ve kontrol edileceğinin tanımlanmasına ihtiyaç duyulmaktadır (Linacre 1990).

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). Puanlayıcı etkilerinin varlığının tespitine yönelik farklı kuramlar altında literatürde kullanılan, birçok teknik bulunmaktadır.

Diğer kuramlara nazaran çok daha uzun süredir kullanılagelmiş klasik test kuramına dayalı; puanlayıcı davranış ve etkilerini ortaya koyan, puanlayıcılar arası güvenirliğin belirlenmesinde kullanılan teknikler aşağıda sıralanmıştır.

- Basit Yüzde Tekniği,
- Kesin Uyum Yüzdesi,
- Cohen Kappa Katsayısı,
- Cohen Ağırlıklandırılmış Kappa Katsayısı,
- Fleiss Kappa Katsayısı,
- Kendall Uyum Katsayısı,
- Pearson Çarpım Moment Korelasyon Katsayısı,
- Eşleştirilmiş Gruplar t Testi,
- Tekrarlı Ölçmeler için ANOVA

Bu tekniklerin her birinin kendine ait avantaj ve dezavantajları, farklı kullanım durumları olmakla birlikte hepsinde var olan ortak sınırlılık; tek bir hata kaynağını dikkate alarak sonuç vermeleri biçiminde ifade edilebilir. Oysa ölçme durumlarında, farklı hata kaynakları ve bu hata kaynakları etkileşimi sonucu ortaya çıkan yeni hata kaynakları da yer alabilmektedir. Tüm bu hata kaynaklarının birlikte değerlendirilmesine olanak sağlayan analiz ise genellenebilirlik kuramıdır.

Cronbach, Gieser, Nanda ve Rajarantnam (1972) tarafından ortaya atılmış 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 (Brennan, 1992; Cardinet ve diğerleri, 1981; Shavelson, Webb ve Rowley, 1989). Literatürde yaygın olarak kullanılan bu kurama dair, Türkiye'de ve yurt dışında, gerek kuramı tanıtmak ve geliştirmek adına, gerekse farklı alanlarda kuram aracılığıyla gerçekleştirilmiş birçok çalışma (Anıl ve Büyükkıdık, 2012; Atılgan, 2005a; Brennan, 1992, 1997, 1998, 2000a, 2000b, 2010; Cardinet, Tourneur ve Allal, 1976, 1981; Güler, 2009, 2011; Güler ve Gelbal, 2010;

Mushquash ve O'Connor, 2006; Shavelson ve diğerleri, 1989; Yelboga ve Tavsancil, 2010; Yılmaz Nalbantoğlu ve Gelbal, 2011; Yılmaz Nalbantoğlu ve Tavsancil, 2014) yer almaktadır.

Genellenebilirlik kuramı, test sürecinin her bir aşamasındaki değişkenliğin bu sürece katkısını inceler (Mariano, 2002). Güvenirliğe ilişkin daha kapsamlı bilgi sunması en önemli avantajlarından biri olan bu kuramın, her kuram gibi dezavantajları da bulunmaktadır. 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). Bu bağlamda ek analizler olmadan, standardın altında performans sergileyen puanlayıcıların belirlenmesini zorlaştırır. 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).

Araştırmacının amacı doğrultusunda, açık uçlu maddelerin kullanıldığı bir teste uygun olan yaklaşım sınanan yeterliliklerinin, puanlayıcı performanslarının ve maddelerin güçlük düzeylerinin eş zamanlı olarak kestirilmesini sağlamalı ve doğrusal olmayan dönüşümlerini içeren uygulamalar için de kullanışlı olmalıdır. Bu niteliklere sahip olması yönüyle öne çıkan yaklaşımlardan biri de çok değişkenlik kaynaklı Rasch ölçme modelidir.

Çok Değişkenlik Kaynaklı Rasch Ölçme Modeli

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). Böylece sınanan ve madde ölçümlerinde bulunan, puanlayıcı kaynaklı yanlılıklar elimine edilir (Linacre,

1989; Sudweeks, Reeve ve Bradshaw, 2004). ÇDKRÖM, atanan puanlar doğrultusunda elde edilen verileri analiz etmek ve temel etkilere sahip faktörler kapsamında tüm puan örüntülerini özetlemek için istatistiksel bir çerçeve oluşturmak amacıyla kullanılır. Ayrıca bu yaklaşımın kullanımı araştırmacılara, her bir faktör için modelde gözlenen elemanlara ait bireysel etkilerin incelenmesine de olanak sağlar (Linacre ve diğerleri, 1990). Genellikle dört faktör, çalışmanın odak noktasını oluşturur ve sınananın performansını veren puanlar üzerinde etkilidir. Bu faktörler: sınanan yeterliliği, madde ya da görev güçlüğü, puanlayıcı katılığı ve derecelendirme ölçeğidir (Linacre ve diğerleri, 1990; Linacre ve Wright, 2004). ÇDKRÖM kapsamında bu faktörler değişkenlik kaynakları (facets) olarak tanımlanmaktadır. Modele istendiği kadar değişkenlik kaynağı eklenebilir fakat eklenen değişkenlik kaynaklarının test geliştiricileri ve kullanıcıları için anlamlı olması gerekmektedir (Linacre, 1989).

Her bir değişkenlik kaynağının kendi içinde elemanları dikkate alınarak (örn; her bir sınananın kendisine özgü bir yetenek düzeyi; her bir puanlayıcının kendisine özgü katılık düzeyi ve her bir maddenin kendisine özgü bir güçlük düzeyine sahip olduğu temelinde ele alınması) ve bir bütün olarak değerlendirilmesi puanlama sürecin kapsamlı olarak incelenmesi bakımından önemlidir (Iramaneerat, Myford, Yudkowsky ve Lowenstein, 2009; Linacre, 1994).

Sınanan yeterlilik düzeyinin göstergesi, değişkenlik kaynaklarından ilkidir. Bu değişkenlik kaynağı, gözlemlenen değişkenlik kaynaklarına ait her bir elementin katılımıyla, dayanak alınan yapı içinde diğer değişkenlik kaynaklarından bağımsız olarak kestirilir (Linacre, 1989; Linacre, 1994).

İkinci değişkenlik kaynağı ise puanlayıcılardır. Puanlayıcı katılık (severity) ya da cömertliği (leniency); herhangi bir puanlayıcının diğer puanlayıcılar tarafından atanan ortalama puanlardan, sistematik olarak daha yüksek ya da daha düşük puan atamasıdır. Bu durum puanlayıcı etkisi (rater effect) veya puanlayıcı hatası (rater error) olarak da tanımlanmaktadır (Engelhard ve Myford, 2003). Literatürde kendisinden önce geliştirilmiş modellerle kıyaslandığında, ÇDKRÖM'nin en önemli avantajlarından biri puanlayıcı katılık düzeylerine yer vermesidir. Geleneksel analizler, her bir maddeye ilişkin gösterilmiş performanslar için atanan puanlar boyunca puanlayıcıların görüş birliğinde olmasına önem vermişlerdir. Rasch model atanan puanların sayısal değerlerinden ziyade bu puanların ne ifade ettiğinin belirleyici olduğunu savunur (Linacre, 1989). Başka bir deyişle temel amaç; her bir

maddede sınanan performansına ilişkin ideal puanlayıcıların görüş birliğine vardığı gerçek puanların tanımlanması değil, her bir puanlayıcının gösterdiği, sınananın örtük yetenek seviyelerinin kestirilmesidir (Linacre, 1994). Bu bağlamda ölçmedeki kesinlik; ideal bir puanlayıcının bulunmasından ziyade her bir puanlayıcının atadığı puanlarda gözlenen, tekrar eden davranışları boyunca gerçek puanlama maksadının kavranmasına dayanır (Linacre, 1989). Etkili bir puanlayıcı ise bireysel olarak aynı katılık seviyesinde puanlama yapabilen ve diğer puanlayıcılarla derecelendirme ölçeğine ilişkin ortak bir anlayışı paylaşandır.

Sınananlara verilen maddeler ise üçüncü değişkenlik kaynağını oluşturur. Bu değişkenlik kaynağı, maddelerin güçlük değerleri ile model içindeki etkilerine ilişkin bilgi sunmaktadır.

Sonuncu değişkenlik kaynağı ise, puanlama ölçeğidir. Puanlama ölçeğinde yer alan kategori sayısı, bu kategorilerin puanlayıcılar tarafından nasıl algılandığı ve nasıl kullanıldığı, atanan puanları doğrudan etkilemektedir. Bu nedenle puanlama ölçeği, puanlama süreci açısından önemlidir.

Atanan her bir puan, birbirini etkileyen dört bileşenin olasılıksal sonuçları olarak nitelendirilebilir. Diğer taraftan sınanan, puanlayıcı ve maddeler üzerinden elde edilen puan dağılımı, sınanan yeteneğinin, puanlayıcı katılığının, madde güçlüğünün ve derecelendirme ölçeğinin yapısının tanımlanmasını sağlar (Engelhard ve Myford, 2003). Bu tanımlamalar aşağıda yer alan ÇDKRÖM eşitliğinden faydalanılarak gerçekleştirilebilir (Linacre, 1994).

(1)

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

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.

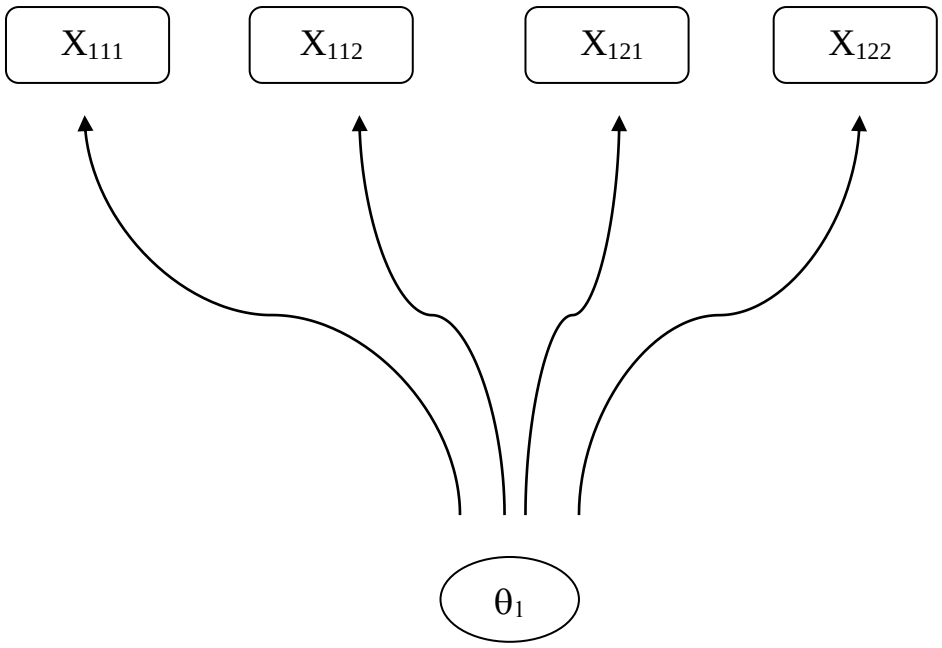
B_n: Sınanan "n"in yetenek düzeyi

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. ÇDKRÖM altında atanan puanlar; madde güçlüğü, puanlayıcı katılığı ve sınanan yeteneğinin etkileşimini tanımlamak amacıyla, her bir kategorinin bitişik alt kategorisiyle karşılaştırılmasının sonucu biçiminde ele alınmaktadır. ÇDKRÖM'in gösterimi Şekil 1'de verilmiştir.



Şekil 1. ÇDKRÖM'in Gösterimi (Patz ve diğerleri, 2002)

Şekil 1'de bir sınananın iki maddeye verdiği yanıtların, iki puanlayıcı tarafından puanlanması sürecinin ÇDKRÖM ile gösterimi sunulmuştur. Sınanan yanıtlarına puanlayıcılar tarafından atanan puanlar başka bir ifadeyle, gözlenen sıralı puanlar Şekil 1'de "X" (X_{111}; sınanın birinci maddeye verdiği yanıta birinci puanlayıcının atadığı puan) sembolüyle; sınanan yeterliliği ise "θ" sembolüyle temsil edilmektedir. X'den θ'ya giden kavisli oklar ise doğrusal olmayan bir ilişkiye işaret etmektedir.

ÇDKRÖM, gözlenen puanların yanıtlama olasılıklarının göstergesi olan log odds'a ya da logit ölçeğe dönüşümüne dayanan, toplamsal (additive) doğrusal bir modeldir (Engelhard ve Myford, 2003). Bu model, verilen kategorilere atanan puanların odd oranının logaritmik fonksiyonunu kullanır (Iramaneerat ve diğerleri,

2009). Toplamsallık, tüm elemanların ortak bir doğrusal ölçeği paylaşıyor olmalarının sonucudur (Linacre, 1994). Analiz esnasında, farklı değişkenlik kaynakları aynı anda analiz edilir fakat bu değişkenlik kaynakları istatistiksel olarak birbirinden bağımsızdır (Engelhard ve Myford, 2003; Iramaneerat ve diğerleri, 2009). Analiz sonucunda değişkenlik kaynakları içindeki her bir elemanın, doğrusal ve eşit aralıklı bir yapıya sahip olan ölçek üzerindeki yeri elde edilebilir. Sınanan, puanlayıcı ve madde parametreleri için logit ölçek üzerinde yüksek değere sahip olmak; daha yetenekli sınananların, daha katı puanlayıcıların ve daha zor maddelerin göstergesidir (Iramaneerat ve diğerleri, 2008). Buna paralel olarak sınanan, puanlayıcı ve madde parametreleri için logit ölçek üzerinde düşük değere sahip olmak; yetenek düzeyi daha düşük sınananların, daha cömert puanlayıcıların ve daha kolay maddelerin göstergesidir.

ÇDKRÖM için model veri uyumunun sağlanması durumunda, yerel bağımsızlığın bir aksiyomu olan, atanan puanların herhangi birinin diğerinden bağımsız olması; istatistiksel kestirimlerin madde, puanlayıcı ya da sınanandan olabildiğince bağımsız ölçümler elde edilmesini mümkün kılar (Linacre, 1989; Linacre, 1994). Başka bir ifadeyle, her bir sınananın yeteneğinin kestirimi; puanlama sürecinde yer alan herhangi bir puanlayıcının katılık düzeyinden, herhangi bir maddenin güçlük düzeyinden ve keyfi (arbitrary) olarak tanımlanmış kategorilerden bağımsızdır (Linacre ve Wright, 2004). Bu, puanlama durumuna ilişkin detayların ötesinde bir genellenebilirlik anlamını taşır.

Rasch (1968)a göre ölçme kestiriminin genellenebilirliği, nesnellik (objectivity) olarak adlandırılır (akt: Linacre ve diğerleri, 1990). Sınanan ölçümünün nesnelliği ise ilgili maddelerin evreninden hangi maddenin seçildiği ve ilgili puanlayıcıların evreninden hangi puanlayıcının seçildiği fark etmeksizin, sınanan için aynı ölçüm elde edilebilir olması durumunda sağlanabilir (Linacre ve diğerleri, 1990). Özellikle puanlayıcılar dikkate alındığında, hangi puanlayıcı hangi sınananı puanlarsa puanlasın, sınanan ölçümleri birinden diğerine her zaman aynı ilişkiyi taşımalıdır. Bu da nesnelliği işaret eder (Linacre, 1994).

Fisher (1925), nesnelliğin genel bir tanımının yapılmasının çok zor olduğunu ifade etmiştir. Roskam ve Jansen (1984) ise tartışmaya açık olan nesnelliğin, Schleibechner'in ortaya koyduğu önermeler (axiomatization) doğrultusunda sunulabileceğini belirtmiştir (akt: Linacre, 1994).

Yukarıda verilen önermelerden ilkine göre, puanlama sürecinin her bir elemanı, parametrelerle ifade edilmektedir. Her bir parametre, diğer parametrelerden bağımsızdır. Bu parametreler sınanan, puanlayıcı ve maddeleri temsil ederken, bağımsız parametreler seti derecelendirme ölçeğinin yapısını temsil eder. Bunların birleşimi belirli bir maddede bir sınananın performansına, bir puanlayıcı tarafından herhangi bir kategoriye atanan puanın görülme olasılığını verir (Thurstone, 1927, akt: Linacre, 1994). İkincisine göre, parametrelerin birleşimi yöntemi toplamsaldır. Derecelendirme ölçeği, aynı niteliğin miktarının artışını temsil ettiği için bu parametrelerin doğrusal ölçeği paylaştıkları biçiminde yorumlanabilir. Başka bir deyişle, test tek boyutludur. Üçüncüsüne göre, her bir parametrenin kestirimi sadece süreçte yer alan atanan tüm puanların yığılmasına (accumulation) dayanır. Puanların herhangi biri diğerleri tarafından etkilenmez. Başka bir ifadeyle, puanların toplamı parametrelerin kestirimi için yeterli bir istatistiktir (Fisher, 1925, akt: Linacre, 1994). Bu bağlamda herhangi bir toplam puanının elde edilmesi sürecinde, bir puanlayıcı tarafından atanan beklenmeyen yüksek puan, sınanan yeterliliğini arttırmaz. Benzer olarak, başka bir puanlayıcı tarafından atanan beklenmeyen düşük puan da sınanan yeterliliğini düşürmez. Fakat bu tip puanlar kestirimin geçerliliğini tehdit eder ve modelin veriye uyum sağlamamasına neden olabilir (Linacre, 1994). Bu üç önermenin bileşimi mümkün olabildiğince bağımsız parametre kestirimlerinin yapılabilmesine olanak sağlar.

ÇDKRÖM, tüm değişkenlik kaynaklarının tüm elemanları için bir ölçüm (analiz sonucu elde edilen logit kestirimler), bir standart hata (logit kestirimin kesinliğine ilişkin bilgi) ve uyum göstergelerine (verinin modele ne derece uyum sağladığına ilişkin bilgi) ilişkin değerler üretmektedir (Engelhard ve Myford, 2003). Modelin veri ile uyumunun istatistiksel olarak nicelendirilmesiyle parametre kestirimlerinin geçerlilik derecesi elde edilir (Wright ve Masters, 1982). Uyum istatistikleri, maddeler ve puanlayıcılar boyunca sınananın sahip olduğu yeterlilik düzeyinin belirlenmesine ilişkin tutarlılık derecesinin göstergesidir (Iramaneerat ve diğerleri 2008). Bu istatistikler, beklenen puanların gözlenen puanlarla ne derece eşleştiğini belirler (Engelhard ve Myford, 2003; Linacre, 1989; Wright ve Linacre 1994). Başka bir deyişle, verinin ölçme modeline ne derece uyum sağladığının bilgisini verir (Lee ve Kantor, 2003). Gözlenen ve beklenen puanlar arasındaki büyük farklılıklar (standartlaştırılmış artıklar olarak ifade edilen) şaşırtıcı ya da beklenmeyen sonuçların göstergesidir. Bu artık değerler, tipik olarak uyum dışı

(outfit) ve uyum içi (infit) olarak adlandırılan hata kareleri ortalaması istatistiği şeklinde özetlenebilir. Uyum dışı istatistikleri, beklenen ve gözlenen puan örüntüsü arasındaki artık değerlerin ağırlıklandırılmamış kareler ortalamasının indeksini verir. Puanlayıcılar için aşağıdaki formül aracığıyla hesaplanabilir (Engelhard, 1994).

(2)

$$u_j = \sum_{n=1}^{N} \sum_{i=1}^{I} z_{nij}^2 / (N + I)$$

Uyum dışı istatistikleri, beklenmeyen uç puanlara karşı oldukça hassastır. Uyum içi istatistikleri, ağırlıklandırılmış artık değerler kareler ortalamasıdır. Artık değer istatistiği olması dolayısıyla beklenmeyen uç puanlara karşı daha az hassastır (Engelhard ve Myford, 2003; Iramaneerat ve diğerleri, 2008). Bu istatistik uyum dışı istatistiği ile aynı dağılıma ve yorumlanmaya sahiptir (Linacre, 1994). Puanlayıcılar için aşağıdaki formül aracığıyla hesaplanabilir (Engelhard, 1994).

(3)

$$v_j = \sum_{n=1}^{N} \sum_{i=1}^{I} W_{nij} z_{nij}^2 / \sum_{n=1}^{N} \sum_{i=1}^{I} W_{nij}$$

Uyum dışı ve uyum içi eşitlikleri sınananlar (n) ve maddeler (i) için benzer şekilde yazılarak değerler hesaplanabilir.

Uyum istatistikleriyle, maddeler boyunca alınan puanlarda beklenmeyen örüntü sergileyen sınananlar (beklenmeyen biçimde iyi ya da beklenmeyen biçimde kötü) tanımlanabilir (Engelhard ve Myford, 2003). Puanlayıcılar bireysel olarak ölçeklediğinde, kaç tane puanın her bir puanlayıcı tarafından kaç tane kategoriye atandığının incelenmesi, uyum istatistiklerindeki değişimin açıklanabilmesini sağlar (Linacre ve diğerleri, 1990). Her iki uyum istatistiği 0 ile artı (+) sonsuz arasında değerler almaktadır. Ulaşılmak istenen değer 1.00'dır. Bu, verinin modele uyum sağladığının işaretidir. Uyum istatistikleri nitelik kontrol değerlerinin alt ve üst sınırlarının belirlenmesine dair kesin bir kural yoktur (Engelhard ve Myford, 2003). Var olan nitelik kontrol değerleri, farklı değişkenlik kaynakları ya da uyum dışı veya

uyum içi istatistiği olma durumu dikkate alındığında farklılaşmaktadır (Iramaneerat ve diğerleri 2008; Nakamura, 2000; Nakamura, 2002; Wright ve Linacre 1994). Nitelik kontrol değerleri kabul aralığı daraldıkça, araştırmanın amacı doğrultusunda analiz sonucu ulaşılacak verinin, model veri uyumu elbette daha yüksek olacaktır. Linacre (1989, 1994)a göre uyum istatistikleri nitelik kontrol değerleri kabul aralığı 0.8-1.2'dir. Bu aralıkta elde edilen değerler verimli olarak nitelendirilerek, veri model uyumunun sağlandığı sonucuna ulaşılabilir (Linacre, 1989). Uyum içi ve dışı istatistiklerinin, 0.8'den küçük olması; gözlenen çok az değişkenliğin (kullanılan kategorilere ilişkin olası ranj daralması) ve maddenin işlevsizliğinin (maddenin başka bir maddeye bağımlı olduğunun) göstergesidir (Engelhard, 2002; Linacre ve Wright 1994; Linacre ve diğerleri, 1990). Bu tür maddeler yeterince bilgi sağlamamakla birlikte, yanıtlanma örüntüleri diğer maddelerin yanıtlanma örüntülerince tahmin edilebilirdir. Daha kötüsü, diğer maddeler üzerinde bağımlılığın yapılandırılmasına neden olabilirler. Örneğin, madde yedi sadece madde altı doğru yanıtlandığında doğru yanıtlanıyorsa, madde yedi çok küçük bir değişkenliğe sahiptir. Bu maddeye ait değişkenlik bir önceki madde ile sınırlandırılmıştır. Madde yedi ölçülmek istenen özelliğe dair bağımsız bir katkı sağlamamaktadır (McNamara, 1996). Elde edilen değerin 1.2'den büyük olması ise çok fazla değişkenliğin, beklenmeyen puanların, maddelerdeki olası çok boyutluluğun ve kategorilerin kullanımı noktasında tutarsızlığın göstergesidir (Engelhard, 2002; Linacre ve diğerleri, 1990).

Model ve veri arasındaki uyumsuzluk; modelin başarısızlığının göstergesi olmaktan ziyade, verinin eşit aralıklı ölçüm yapısını desteklemediğinin ve modelin değişkenlik kaynaklarının tek bir boyut altında birleşmediğinin göstergesidir (Linacre, 1989). Eğer veri modele uyum sağlarsa, araştırmacılar değişkenlik kaynakları arasında kullanışlı, bilgilendirici karşılaştırmalar sunabilirler (Engelhard ve Myford, 2003).

ÇDKRÖM analizinde her bir değişkenlik kaynağı için iki farklı güvenirlik değeri bulunur. Bunlardan ilki ayırma oranı kullanılarak hesaplanan, ayırma indeksidir (Engelhard ve Myford, 2003; Sudweeks ve diğerleri, 2004). Bu indeks her bir değişkenlik kaynağına ait tüm elemanların birbirlerinden ne derece ayrıldığına dair bilgi sunmaktadır (Lee ve Kantor, 2003). Başka bir deyişle, değişkenlik kaynağının kesinliğine ilişkin bir yayılma (değişkenlik) ölçüsü vermektedir. Öncelikle gerçek standart sapmanın (SS) hataların ortalama karekökünе (RMSE) oranı, yani ayırma oranı (G) Eşitlik 4 ve gerçek SS Eşitlik 5 ile hesaplanmaktadır.

(4)

$$G = \text{Gerçek SS} + \text{RMSE}$$

(5)

$$(\text{Gerçek SS})^2 = (\text{Gözlenen SS})^2 - (\text{RMSE})^2$$

Ayırma indeksi ise Eşitlik 6 ile hesaplanabilmektedir.

(6)

$$\text{Ayırma İndeksi} = (4\ \text{Gerçek SS} + \text{RMSE}) / (3\ \text{RMSE}) = (4G + 1) / 3$$

Her bir değişkenlik kaynağı için bu indeksler yukarıda yer alan eşitliklerden faydalanılarak elde edilebilir. Daha anlaşılır ve kullanılabilir olması bakımından ayırma indeksi rapor edilmektedir (Engelhard ve Myford, 2003).

Analiz sonucunda ulaşılan ikinci güvenirlik değeri; ayırma indeksi güvenirliğidir (R). Bu indeks, bir değişkenlik kaynağındaki elemanların güvenilir bir şekilde değişkenlik kaynağını tanımlamak amacıyla, ne kadar iyi ayrılabildiğine dair bilgi sağlar. KR-20, Cronbach alfa gibi geleneksel güvenirlik istatistikleriyle benzerlik gösterir (Bond ve Fox, 2001; Engelhard ve Myford, 2003; Myford ve Wolfe, 2003; Sudweeks ve diğerleri, 2004). Eşitlik 7'den faydalanılarak hesaplanabilir.

(7)

$$R = (\text{Gerçek SS})^2 / (\text{Gözlenen SS})^2 = G^2 / (1 + G^2)$$

Her bir değişkenlik kaynağı için ayırma indeksi güvenirliği 0.0 ile 1.0; ayırma indeksi ise 1 ile sonsuz arasında değişmektedir (Sudweeks ve diğerleri, 2004). Ayırma indeksi güvenirliğinin 1.0 a yakın değerler alması yüksek düzey bir güvenirliğin göstergesi olup, istenen bir durumdur (Bond ve Fox, 2001). Ayırma indeksi ve ayırma indeksi güvenirliği farklı metriklerden rapor edilseler de, aynı bilgiye dayanılarak hesaplanır ve belirli bir değişkenlik kaynağı için benzer sonuçları verir. Fakat bu iki istatistiğin yorumlanması değişkenlik kaynağına göre farklılaşır (Sudweeks ve diğerleri, 2004). Sınanan değişkenlik kaynağı için ayırma indeksi ve ayırma indeksi güvenirliğinin yüksek bir değer alması istenen durumken diğer

değişkenlik kaynakları için düşük değer alması istenen durumdur. Çünkü diğer değişkenlik kaynaklarında yer alan elemanlar arasındaki değişkenlik puanlardaki istenmeyen varyansın göstergesidir (Engelhard ve Myford, 2003; Sudweeks ve diğerleri, 2004). Örneğin: sınanan ayırma indeksinin 3.0 olduğu bir çalışmada sınananların üç ayrık istatistiksel gruba ayrılabildiği sonucuna ulaşılır ve bu istenmeyen bir durum değildir. Sınanan ayırma indeksi 1.0 a eşit olduğunda ise herhangi bir sınananın yeterliliğinin başka bir sınananın yeterliliğinden ayırt etmek mümkün değildir (Sudweeks ve diğerleri, 2004). Puanlayıcılar dikkate alındığında, bu indeks puanlayıcı katılık düzeyleri yayılımını verir. İndeksin 1.0 olması, puanlayıcıların benzer katılık düzeyinde puanlama yaptıkları ve birbirlerinin yerine geçebileceklerinin göstergesi olarak kabul edilebilir ve istenen bir durumdur (Engelhard ve Myford, 2003). Sınanan değişkenlik kaynağı için bu iki değerin düşük olması, puanlarda merkezi eğilim hatasının (central tendency error) olduğunun göstergesi olabilir. Bu, puanlayıcıların sınananlara ait performansları birbirinden ayırma noktasında, yeterli olmadıkları şeklinde de yorumlanabilir (Myford ve Wolfe 2003, 2004). Diğer değişkenlik kaynakları için bu iki istatistiğin düşük değerler alması, ilgili değişkenlik kaynağına ait farklı elemanlar için elde edilen ölçümlerin yüksek derecede kararlılık (tutarsızlık bulunmaması) gösterdiği biçiminde yorumlanabilir (Sudweeks ve diğerleri, 2004).

ÇDKRÖM'nin verdiği bir diğer istatistikte ki-kare istatistiğidir. Ki-kare istatistiği değişkenlik kaynakları elemanları arasında manidar bir fark olup olmadığının hesaplanmasında kullanılmaktadır. Başka bir ifadeyle ki-kare testi; sınanan yetenek düzeyleri, puanlayıcıların katılık ve cömertlikleri ve maddelerin güçlükleri arasında manidar bir fark olmadığını belirten yokluk hipotezinin test edilmesi için kullanılır (sınanan yetenek düzeyleri arasında manidar bir fark yoktur, puanlayıcıların katılık düzeyleri arasında manidar bir fark yoktur ve hiçbir madde bir diğerinden daha kolay ya da zor değildir). Manidar bir ki-kare istatistiği ($p<0.05$) sınananın yetenek düzeyi, puanlayıcı katılık düzeyi ya da madde güçlük düzeyinden en az ikisi arasında gözlenen farklılığın olduğuna işaret eder (Myford ve Wolfe, 2004).

ÇDKRÖM, modelde yer alan tüm elemanlara ilişkin nitelik kontrol ve güvenirlik değerleri çerçevesinde bilgiler sunar. Böylece, model kapsamında ölçmenin hatasına neden olan elemanlar belirlenip, niteliğin geliştirilmesine olanak sağlanır.

ÇDKRÖM, doğrusal lojistik test modelleriyle aynı matematik forma sahip, puanlayıcı etkilerini modellemeyi amaçlayan popüler bir yaklaşımdır. Literatürde ÇDKRÖM kullanılarak gerçekleştirilmiş birçok çalışmanın (Akın ve Baştürk 2012; Atılgan, 2005b; Baştürk, 2010; Engelhard, 1994; Engelhard ve Myford, 2003; Iramaneerart ve diğerleri, 2009; Linacre ve diğerleri, 1990; Nakamura, 2000, 2002) yanı sıra ÇDKRÖM üzerine de birçok çalışma yapılmıştır (Casabianca ve Junker, 2013, 2014; DeCarlo 2010; DeCarlo ve diğerleri, 2011; Iramaneerat ve diğerleri, 2008; Kim 2009; Lynch ve McNamara, 1998; Mariano 2002; Patz, Junker ve Johnson 2000; Patz ve diğerleri 2002; Sudweeks ve diğerleri, 2004; Verhelst ve Verstralen, 2001; Wilson ve Hoskens 2001). Bu çalışmaların büyük bir kısmı ÇDKRÖM, sınanan, madde ve puanlayıcı parametrelerini veren her bir puanın birbirinden bağımsız olduğu ve tüm puanlayıcıların eşit güvenirliğe sahip olduğu önermelerine yöneliktir (Mariano, 2002).

Wilson ve Hoskens (2001) tekrarlayan puan problemi (repeated ratings problem) olarak adlandırdıkları ÇDKRÖM'e ilişkin karşılaşılan problemi şu şekilde örneklendirmişlerdir; a) bir sınanana ait bir yanıtın beş farklı puanlayıcı tarafından beş kez puanlandığı ve b) aynı sınanana ait beş yanıtın her birinin farklı puanlayıcı tarafından puanlandığı iki durumun var olduğunu düşünelim: a durumu; sınanan yanıtına ilişkin bize daha iyi bilgi sağlarken b durumu; sınanan yeterliliğine ilişkin bize daha iyi bir bilgi sağlar. ÇDKRÖM, her iki durum için atanan puanların sınanan yeterliliğine ilişkin eşit düzeyde bilgi sağladığını ifade eder. Verhelst ve Verstralen (2001) çalışmalarının sonucunda, herhangi bir sınananın her bir puanlayıcı için bağımsız yanıtlar atayamayacağını, tüm puanlayıcıların aynı yanıta ilişkin bir yargıya vardıklarını böylece ulaşılan yargıların ilişkili olduğunu belirtmişlerdir. Casabianca ve Junker (2013), DeCarlo (2010), DeCarlo ve diğerleri (2011), Kim (2009), Mariano (2002), Patz ve diğerleri (2000), Patz ve diğerleri (2002) ise bir sınanan yanıtı farklı puanlayıcılar tarafından puanlandığında, atanan puanlar aynı sınanan yanıtına dayandığı için ilişkili olduğunu vurgulamışlardır. Bunun yanı sıra, açık uçlu maddeler için çoklu puanlamaların kullanımı, deneysel desenlerdeki tekrarlanan ölçümlerle benzer yapıya sahip olduğunu ve eğer aynı maddeye ilişkin çoklu puanlamaların tekrar eden yapısı dikkate alınmazsa, sınanan yeterliliğinin standart hatasının önemli derecede yanlılık göstereceğini raporlaştırmışlardır. Benzer olarak, Junker ve Patz (1998) aynı madde için daha fazla atanan puan varlığı, her zaman yetenek kestiriminin kesinliğini yükselteceğini; hatta modele eklenen atanan

puanların yetenek kestirimi için yeni bir bilgi vermediğinde bile yetenek kestiriminin kesinliğini yükselteceği sonucuna ulaşmışlardır. Çalışmaları kapsamında sınanan yeterliliğine ilişkin daha kesin ölçümlerin, sınanana daha fazla madde vermek yerine daha fazla puanlayıcı kullanarak elde edilebilir olduğunu belirtmişlerdir. Junker ve Patz (1998) ile benzer olarak, DeCarlo ve diğerleri (2011), Donough ve Hombo (2000), Mariano (2002), Patz ve diğerleri, (2000), Patz ve diğerleri (2002) ise çok az sayıda (iki ya da üç) puanlayıcıyla, aynı sınanan yanıtına ait çoklu puanlamaların birbirinden bağımsız olduğu önermesi doğrultusunda ulaşılan sonucun sınanan yeterliliğine ilişkin var olandan daha iyimser olduğunu raporlaştırmışlardır.

Açık uçlu maddelerin kullanıldığı testlerde her yanıt için atanan çoklu puanlar, puanlayıcılardan kaynaklanan değişkenliğin azalmasında kullanışlı olabilir. Ancak, modelden faydalanırken çoklu puanlar arası bağımlılık dikkate alınmalıdır (Mariano, 2002).

Hiyerarşik Puanlayıcı Modeli

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, 2002; 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). Başka bir ifadeyle aynı yanıtın birbirinden bağımsız olarak çoklu puanlanması, yeterliliğe ilişkin kanıtın tek bir parçasına ait çoklu değerlendirmelerdir; yeterliliğe ilişkin kanıtın çoklu parçalarına ait değerlendirmeler değildir. HPM aynı yanıta atanan çoklu puanları, bu ikisi arasındaki ayrımı ortaya koyarak modellemektedir (Mariano, 2002).

Tek bir yanıt için atanan daha fazla puan, gerçek puan hakkındaki belirsizliği düşürebilir. Fakat uygun bir şekilde puan atanmış maddenin verdiği yeterlilikle ilgili belirsizliği düşürmeyecektir. Tek bir yanıt için atanan daha fazla puan, sınanan

yeterliliğine ilişkin biriken bilgi miktarını arttıracaktır. Fakat biriken bu bilgi, yanıtlara nesnel olarak puan atanırsa kullanışlı olacaktır (Mariano, 2002). HPM, sınanan yanıtına atanmış farklı puanlar arasındaki bağımlılığı dikkate alarak, birden fazla puanlayıcıdan elde edilen bilgileri bir araya getirmek için uygun bir yol sağlar. Böylece sınanan, puanlayıcı ve madde parametrelerine ilişkin daha güvenilir ve geçerli bilgiler elde edilebilir (Patz ve diğerleri, 2002).

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 ve kalibrasyonuna 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_{ij}) 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. Böylece puanlayıcılar tarafından sınanan yanıtlarına atanan puanlar, sınanan yanıtlarında koşullu olur. Başka bir ifadeyle, atanan puanlar sınanan yanıtlarına bağlı hale gelir (Casabianca, Junker ve Patz, 2014; Mariano ve Junker, 2007).

(8)

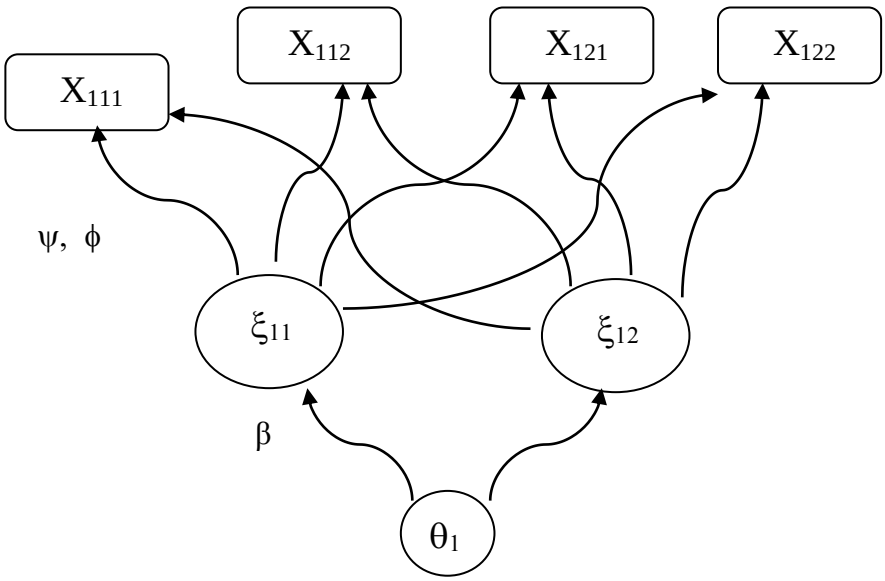
$\boldsymbol{\theta_i}$~ N (μ, σ^2), i = 1, …, N,

$\boldsymbol{\xi_{ij}}$~ 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).



Şekil 2. HPM'nin Gösterimi (Patz ve diğerleri, 2002)

Şekil 2'de bir sınananın iki maddeye verdiği yanıtların, iki puanlayıcı tarafından puanlanması sürecinin HPM ile gösterimi sunulmuştur. Sınanan yanıtlarına puanlayıcılar tarafından atanan puanlar, başka bir ifadeyle gözlenen sıralı puanlar Şekil 2'de "X" (X_{111}**;** sınananın birinci maddeye verdiği yanıta birinci puanlayıcının atadığı puan) sembolüyle; puanlayıcıların tespit etmeye çalıştığı örtük kategoriler ise " ξ" (ξ_{11}) sembolüyle temsil edilmektedir. Puanlayıcı katılığını simgeleyen "ϕ" ile puanlayıcı değişkenliğini simgeleyen, "ψ" ise HPM'nin birinci aşamaya ait olan parametreleridir. HPM'nin ikinci aşamasında sınanan yanıtlarının gerçek kategorisi olan ξ, sınanan yeterliliğini simgeleyen θ'nın sıralı göstergeleri olarak ele alınır. ξ'den X'e ve θ'dan ξ'ye giden kavisli oklar doğrusal olmayan bir ilişkiye işaret etmektedir. Son olarak "β" ise HPM'nin madde güçlüğünü simgeleyen ikinci aşama parametresidir.

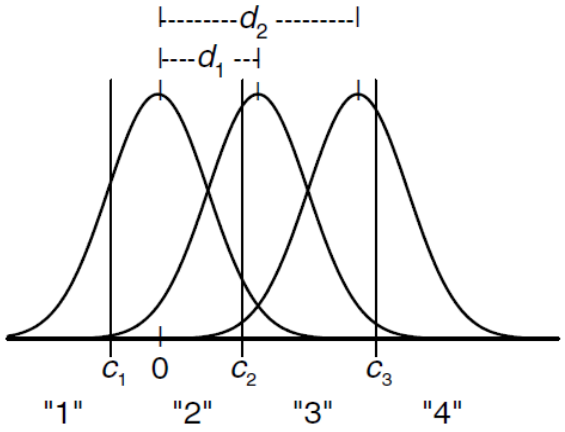
Daha detaylı incelenecek olursa; HPM'nin ilk aşamasında, gözlenemeyen örtük değişkenler olarak ele alınan, i sınananının j maddesindeki performansını açıklayan, ideal puanlar bulunmaktadır (Casabianca ve Junker, 2013; Patz ve diğerleri, 2002). İdeal puanlar, dereceli puanlama anahtarı kullanılarak tanımlanan sıralı örtük kategorilerdir (Kim, 2009).

Kavramsal olarak, bir madde için dereceli puanlama anahtarı kullanıldığında, olası sınanan tepkilerinin tümüne ait birtakım sıralı puan noktalarını içeren alanın haritası tanımlanır. İdeal puan, i sınananının j maddesine verdiği tepkileri yansıtan haritanın, ideal uygulama sonucunu veren nokta olarak görülebilir. İstatistiksel olarak, ideal puanlar sınanan yanıtı için atanmış, çoklu puanlamalar arasındaki bağımlılığı yansıtır (Patz ve diğerleri, 2002).

İlk olarak, çoklu puanlayıcılardan elde edilen puanlara dayanılarak maddeye ait ideal puanlar sinyal tespit teorisi benzeri modellenir (Casabianca ve Junker, 2013). Sinyal tespit modeli, sıralı kategorilerden oluşan eğitim ve psikolojide kullanılan yapıların ölçülmesi noktasında oldukça kullanışlıdır. HPM altındaki kullanım amacı ise puanlayıcıların maddeleri puanlarken ne yaptıklarına ilişkin bilgi sağlamaktır (DeCarlo, 2005).

Bir sınanan yanıtının ait olduğu gerçek kategori doğrudan gözlenemez ya da bilinemez, gerçek kategori örtüktür. Bu nedenle, puanlayıcıların görevi sinyal tespitidir (signal detection). Başka bir ifadeyle, verilen belirli bir yanıt için puanlayıcıların görevi, yanıtın ait olduğu kategoriyi tespit etmektir. Böylece, sinyal tespit modeli HPM'nin ilk aşaması olan, puanlayıcı modelin doğal bir parçası hâline gelir (DeCarlo ve diğerleri, 2011).

Sinyal tespit modeli, puanlayıcı yargılarını gerçek kategorilerin hatalı olabilen göstergeleri olarak ele alır. (Casabianca ve Junker, 2013, 2014; DeCarlo ve diğerleri, 2011). Puanlayıcı kararı, yanıtın (bütünsel bir puanlama için) tüm niteliğine ilişkin, bir dereceye kadar kendi algılayışına dayanır. Yanıtın niteliğine ilişkin algı, her bir örtük sınıfla ilişkilendirilmiş, farklı olasılık dağılımlarıyla birlikte, temelde sürekli bir olasılık dağılımının kavranışı biçiminde incelenebilir. Dağılımlar arasındaki uzaklık (d) esas ilgilenilendir. Çünkü puanlayıcının örtük sınıflar arasında ayrım yapabilme yeteneğini yansıtır. Bu Şekil 3'te gösterilmektedir (DeCarlo, 2005).



Şekil 3. Üç Örtük Sınıflı ve 1-4 Kategorili Sinyal Tespit Teorisi Gösterimi

Verilen bir yanıta puan atandığında, puanlayıcı yanıtın niteliğine ilişkin algısıyla dereceli puanlama anahtarında var olan yanıt kategorilerini karşılaştırarak bir karara vardığı varsayılır. Böylece, gözlenen yanıtlar hem algılayış hem de karara ait bakış açısını yansıtır. Yanıtın niteliğine ilişkin puanlayıcı algısını ve puanlayıcının yanıt kategorilerini kullanımını birbirinden ayırmak sinyal tespit modelinin oldukça önemli bir yönüdür (DeCarlo, 2005).

HPM'de temel olarak; dereceli puanlama anahtarıyla tanımlanan örtük sınıflar ideal puanlardır ve puanlayıcılar ideal puanları tespit edenler olarak incelenir. (Casabianca ve diğerleri, 2014; Casabianca ve Junker, 2013; DeCarlo, 2005; DeCarlo ve diğerleri, 2011). Başka bir deyişle puanlayıcılar, yanıtların örtük sınıfları arasında ayırım yapanlar olarak ele alınır (DeCarlo, 2005).

Sınanan performansını sınıflandırmak için dereceli puanlama anahtarının doğru kullanımıyla elde edilen sonuçlar ideal puanlardır. Gözlenen puanlar ise puanlayıcıların atadığı puanlar doğrultusunda ulaşılan, olası hatalı sınıflamalar biçiminde açıklanabilir (Mariano ve Junker, 2007; Patz ve diğerleri, 2002). Bu, dereceli puanlama anahtarına (az ya da oldukça fazla ayrıntıya sahip olması) ve atanan puanlara (puanlayıcının katı ya da hoşgörülü olması; bir ya da daha fazla kategorinin oldukça az kullanılması) ilişkin neyin iyi, neyin kötü olduğunu saptamamıza yardımcı olur. Daha genel olarak, puanlayıcıların gerçekte puanlama kategorilerini kullanımı (atanan puanlar tarafından yansıtılan) ile puanlama kategorilerinden istenilen kategorileri (ξ_{ij} tarafından yansıtılan) seçme durumlarının

karşılaştırılması, istenen (daha nitelikli bir dereceli puanlama anahtarı) ve istenmeyen durumların (bireysel puanlayıcı katılığının veya puanlayıcı değişkenliğinin artışı) tanımlanmasına yardımcı olur (Patz ve diğerleri, 2002).

Başka bir ifadeyle ideal puanlar, sınanan i'nin madde j için verdiği yanıtın, yanlı olmayan ve mükemmel tutarlılığa sahip bir puanlayıcı tarafından puanlanması sonucu elde edilen puanlarını yansıtır (Casabianca ve Junker 2013, 2014; Mariano, 2002). HPM'de atanan puanlar ile ideal puanlar arasındaki sapmalar, puan olasılıkları matrisini kullanarak, yanıtların niteliğini gösteren sinyal tespit teorisiyle modellenir. Puan olasılıkları matrisi, atanan ve ideal puanlar arasındaki ilişkiyi tanımlar. Çizelge 1'de yer alan olasılıklar puanlayıcı davranışları değişkenliğinin ve sınanan, puanlayıcı, madde arasındaki etkileşimin yansıtılması amacıyla oluşturulur (Casabianca ve diğerleri, 2014).

Çizelge 1. HPM'de Sinyal Tespit İşlemini Tanımlayan, Atanan Puan Olasılıkları Matrisi

	Gözlenen Puanlar (k)				
İdeal Puanlar (ξ)	0	1	2	3	4
0	P_{00r}	P_{01r}	P_{02r}	P_{03r}	P_{04r}
1	P_{10r}	P_{11r}	P_{12r}	P_{13r}	P_{14r}
2	P_{20r}	P_{21r}	P_{22r}	P_{23r}	P_{24r}
3	P_{30r}	P_{31r}	P_{32r}	P_{33r}	P_{34r}
4	P_{40r}	P_{41r}	P_{42r}	P_{43r}	P_{44r}

(Patz ve diğerleri, 2002)

Sinyal tespit teorisi, HPM içinde ideal puanları veren atanan puanların olasılığını, matrisin her bir dizisi için kesikli tek modlu bir dağılım kullanarak verir. Bu dağılımın modu, puanlayıcı yanlılığı ya da katılığını verirken, dağılımın yayılımı ise puanlayıcı değişkenliği ya da güvenilir olmama durumuna ilişkin bilgi verir (Casabianca ve diğerleri, 2014; Patz ve diğerleri, 2002). İdeal puanların kestirimi, j maddesine i sınananının verdiği yanıta dair puanlayıcılar arasında, üzerinde fikir birliğine varılmış atanan puanlar olarak görülebilir (Patz ve diğerleri, 2002). Elde edilen parametreler sınananlar, puanlayıcılar ve maddeler arasındaki olası etkileşimleri ele alır. Atanan puanlar, çeşitli puanlayıcı ortak değişkenlikleri de

(farklılaşan eğitim alt yapıları, geçmişleri) dikkate alınarak bu aşamada modellenir (Casabianca ve diğerleri, 2014; Mariano ve Junker, 2007; Patz ve diğerleri, 2002).

HPM'nin pratikte uygulanabilmesi için Eşitlik 9'da verilen, HPM hiyerarşisi içinde seçilen modellerin tanımlanmasına ihtiyaç duyulmaktadır. Eşitlik 9'un en alt düzeyinde, Çizelge 1'in her bir sırası için atanan puan olasılıkları parametrelerle ifade edilmektedir. Böylece model bireysel olarak her bir puanlayıcının katılık ve değişkenlik parametrelerine duyarlı hale gelecektir. Bu, ortalaması $\xi+\phi_r$ ve standart sapması ψ_r olan, atanan puanlarda normal yoğunlukla (Normal density) orantılı olarak, Çizelge 1'de yer alan her bir satırın olasılıklarının $P_{\xi kr} \equiv$ [puanlayıcı r atanan puan k | ideal puan ξ] elde edilmesiyle gerçekleştirilir. Olasılıkların elde edilmesinde faydalanılan, sinyal tespit teorisine dayalı olarak yazılmış eşitliğe aşağıda yer verilmiştir.

(9)

$$p_{\xi kr} = P[X_{ijr} = k \mid \xi_{ij} = \xi] \propto \exp\left\{-\frac{1}{2\psi_r^2}[k - (\xi + \phi_r)]^2\right\}$$

i = 1,…, N;

j = 1,…,J;

r = 1,…, R.

$P_{\xi kr}$: Atanan puan olasılıkları matrisinin her bir satırına ait kestirilen olasılıklar.

ψ_r: Puanlayıcı r'nin katılık düzeyi

ϕ_r: Puanlayıcı r'nin değişkenlik düzeyi

HPM'nin ikinci aşamasında, örtük kategoriler, sınanan yeterliliğinin sıralı göstergeleri olarak ele alınır. Başka bir ifadeyle, sınanan tarafından verilen yanıtın ait olduğu gerçek kategori, sınanan yeterliliğinin bir göstergesine dönüşür (DeCarlo ve diğerleri, 2011).

HPM, ideal puanlar için geleneksel madde tepki kuramı modellerinden uygun olan (aşamalı tepki modeli, dereceli tepki modeli, kısmi kredi modeli, genelleştirilmiş kısmı kredi modeli) birini kullanarak sınanan yeterliliklerini kestirir

(Casabianca ve Junker, 2013, 2014; Casabianca ve diğerleri, 2014; Mariano ve Junker, 2007). Bu modellerden hangisinin kullanılacağına ilişkin herhangi bir kural bulunmamaktadır. Hangi modelin kullanılacağı araştırmacının seçimine bağlıdır. Bu modellerden aşamalı tepki modeli, dereceli tepki modeli, kısmi kredi modeli ve genelleştirilmiş kısmi kredi modeline ait açıklamalara aşağıda yer verilmiştir.

Aşamalı tepki modeli

Samejima tarafından 1969 yılında geliştirilen aşamalı tepki modeli (graded response model-GRM), iki parametreli losjistik modelin bir uzantısıdır (Hays, Morales ve Reise, 2000; Samejima, 2008). Madde yanıtlarının, sıralanmış kategorik yanıtlar olarak değerlendirildiği durumlarda kullanışlıdır. Bu modelde her bir maddenin aynı yanıt kategorisine sahip olma gibi bir zorunluluğu yoktur (Hays ve diğerleri, 2000). Bu bağlamda, farklı sayıda yanıt kategorisine sahip maddelerin analizi için uygun olduğu ifade edilebilir. Aşamalı tepki modelinde (ATM) her bir madde, madde ayırıcılık parametresi ve maddenin sahip olduğu yanıt kategorilerinin bir eksiği kadar kategoriler arası eşik parametresi aracığıyla tanımlanır (Ferrando, 2009; Hays ve diğerleri, 2000). AŞM analizi gerçekleştirilirken, lojistik dağılım fonksiyonundan faydalanıldığında lojistik AŞM; normal dağılım fonksiyonundan faydalanıldığında ise normal ogiv AŞM olarak isimlendirilir (Forero ve Maydeu-Olivares, 2009).

Dereceli tepki modeli

Andrich'in 1978 yılında geliştirdiği dereceli tepki modeli (rating response model-RRM), iki kategoride puanlanan Rasch modelinin bir uzantısıdır (Chang ve Reeve, 2005). Bu modelde, sıralı kategorilerin yer aldığı ve tüm maddeler için eşik değerlerin eşit olduğu varsayımı bulunmaktadır (Andrich, 1978). Başka bir ifadeyle, yanıt kategorilerinin belirli bir sırasının olduğu ve eşik değerlerin zorluk düzeylerinin maddeden maddeye değişmediği varsayımlarına dayanır (Chang ve Reeve, 2005). Bunun yanı sıra dereceli tepki modelinde (DTM) her bir maddenin aynı miktarda bilgi sağladığı varsayımı da kabul edilmektedir. Ayrıca madde güçlüğü, belirli bir madde için kategori eşik değerlerine göre ortalama güçlüğü yansıtır ve tek bir parametreyle tanımlanır (Chang ve Reeve, 2005; Hays ve diğerleri, 2000). Rasch

ailesinin bir üyesi olan DTM için de diğer Rasch modellerinde olduğu gibi, tüm maddelerin madde ayırıcılık parametresi birbirine eşittir.

Kısmi kredi modeli

Masters tarafından 1982 yılında, Rasch'ın iki kategoride puanlanan (dichotomous) modeli kullanılarak geliştirilmiş olan kısmi kredi modeli (Muraki, 1992), örtük puanların tespiti için toplam puanları yeterli bir istatistik olarak kullanan çok kategoride puanlanan (polytomous) bir madde tepki kuramı modelidir (Ligtvoet, 2012).

Kısmi kredi modeli (partial credit model-PCM), çözüm sürecinde birden fazla adımı içeren test maddelerinin analizinde ve kısmi puan atamanın önemli olduğu durumlarda işe koşulur. Ayrıca tutum, kişilik ve inanç gibi, sınananların çok noktalı (multipoint) ölçekler üzerinden verdikleri yanıtların analizi için de oldukça uygundur (Masters, 1982).

Çok yaygın bir kullanım alanına sahip olan kısmi kredi modeli (KKM) (Fischer ve Molenaar, 1995; Rasch, 1960; Smith ve Smith, 2004; Wright ve Masters, 1982), diğer çok kategoride puanlanan modellerden (DYM, GKKM vb.) farklı olarak madde ayırıcılık parametresini içermez. Bu nedenle, göreceli olarak, madde ayırıcılık parametresini içeren modellere nazaran, daha küçük örneklem büyüklüklerine sahip araştırmalarda da kullanılabilir. Ayrıca KKM, Rasch ailesinin bir ögesidir ve dolayısıyla, Rasch modellerinin sahip olduğu tüm avantajlara sahiptir (Penfield, Myers ve Wolfe, 2008). Bu model, maddeler arası yanıt kategorilerinin göreceli zorluklarına ilişkin herhangi bir varsayıma sahip değildir (Chang ve Reeve, 2005). Başka bir ifadeyle, her bir madde için yanıt kategorileri arasındaki uzaklıklar farklılık gösterebilir.

Genelleştirilmiş kısmi kredi modeli

KKM'nin genelleştirilmiş hâli olan bu model, 1992 yılında Muraki tarafından geliştirilmiştir (Muraki, 1993). Çok kategoride puanlanan bir model olan genelleştirilmiş kısmi kredi modeli (generalized partial credit model-GPCM), iki parametreli lojistik modelin bir uzantısı şeklinde de ifade edilebilir (Davis, 2004).

Bu modelde madde adım parametresi eşik parametresi ve konum parametresine ayrılmıştır (Muraki, 1992). Dolayısıyla, genelleştirilmiş kısmi kredi modeli (GKKM) her bir maddenin ayırıcılık parametresinin sabit olduğu varsayımını taşımaz ve her bir madde için farklı ayırıcılık parametresi verebilir.

Bu çalışmada, yukarıda bahsedilmiş olan avantajları ve analiz için ihtiyaç duyulan kodların gözden geçirilip düzeltilmesindeki kolaylığı nedeniyle, HPM'nin ikinci aşamasında kullanılmak üzere KKM tercih edilmiştir. KKM için sınananın herhangi bir kategoride yanıt verme olasılığının hesaplanmasında kullanılan formüle aşağıda yer verilmiştir.

(10)

$$P[\xi_{ij} = \xi \mid \theta_i, \beta_j, \gamma_{j\xi}] = \frac{\exp\left\{\sum_{k=1}^{\xi}(\theta_i - \beta_j) - \gamma_{jk}\right\}}{\sum_{h=0}^{K-1}\exp\left\{\sum_{k=1}^{h}(\theta_i - \beta_j) - \gamma_{jk}\right\}}$$

i = 1,…, N;

j = 1,…,J;

θ_i: Sınanan yeterliliği (örtük özellik).

β_j: Madde güçlük parametresi.

γ_{jk}: Kategorilere ait adım güçlük parametresi.

K-1: Kategori eşik (parametre) sayısı

Alınan eğitimsel kararlarda çoktan seçmeli maddelerin yanı sıra açık uçlu maddelerin kullanımı da gereklilik hâline gelmiştir. Bu maddelerin kullanımı, puanlayıcıların rolünü ön plana çıkarmış; puanlamadaki güvenirliğin ve puanlayıcı etkilerinin belirlenme ihtiyacı ise çeşitli analiz ve modelleri doğurmuştur. 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.

Literatür incelemesi sonunda, HPM'nin 2002'den itibaren kullanılmaya başlandığı ve bu modele ilişkin kısıtlı sayıda çalışmanın var olduğu (Casabianca ve Junker, 2013, 2014; Casabianca ve diğerleri, 2014; DeCarlo ve diğerleri, 2011; Mariano, 2002; Patz ve diğerleri, 2000; Patz ve diğerleri, 2002) görülmektedir. ÇDKRÖM ve HPM'yi içinde barındıran oldukça az sayıda çalışma (Mariano, 2002; Patz ve diğerleri, 2002) bulunmakla birlikte Türkiye'de ise yapılmış herhangi bir çalışmaya rastlanmamıştır. Bu açıdan benzer amaçlar doğrultusunda geliştirilmiş ÇDKRÖM ve HPM'yi, aynı koşullar altında ve gerçek veri seti üzerinde karşılaştırarak, avantaj ve dezavantajlarının, bilgi sağlama güçlerinin, her iki modelden elde edilen bilgilerin benzerlik ve farklılıklarının ortaya konulduğu çalışmalara ihtiyaç duyulduğu ifade edilebilir. Bunun yanı sıra modeller üzerinde yürütülen uygulamalarda karşılaşılan problemler ya da eksikliklerin gözlenmesiyle, gerek var olan modellerin yeniden yapılandırılması gerekse yeni modellerin geliştirilmesi bakımından, her iki modelin birlikte kullanıldığı çalışmalara ihtiyaç duyulduğu düşünülmektedir. Bu bağlamda, araştırma kapsamında kullanılan açık uçlu maddelere ilişkin, aynı sınananlar tarafından verilen yanıtların, birden fazla puanlayıcı tarafından puanlanması durumunda, ÇDKRÖM ve HPM ile parametrelerin kestirilmesi ve her iki modele ait elde edilen parametrelerin birlikte değerlendirilmesi araştırmanın problemini oluşturmaktadır.

Amaç

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

1. Çok değişkenlik kaynaklı Rasch ölçme modeli'nin;
    a) Model veri uyumu nasıldır?
    b) Öğrenci yetenekleri ve uygunluk istatistikleri nasıldır?

c) Puanlayıcı katılık/cömertlikleri ve uygunluk istatistikleri nasıldır?
d) Madde güçlükleri ve uygunluk istatistikleri nasıldır?

2. Hiyerarşik puanlayıcı modeli'nin;
   a) Öğrenci yetenek istatistikler nasıldır?
   b) Puanlayıcıların katılık/cömertlik ve değişkenlik istatistikleri nasıldır?
   c) Madde güçlük istatistikleri nasıldır?
3. Çok değişkenlik kaynaklı Rasch ölçme modeli ile Hiyerarşik puanlayıcı modeli'ne ait;
   a) Öğrenci, puanlayıcı ve madde parametreleri korelasyon değerleri nasıldır?
   b) Sapma bilgi kriteri değerleri nasıldır?

Önem

Bu araştırmada, benzer amaçlara hizmet eden ÇDKRÖM ve HPM'nin gerçek veri seti üzerinden gerçekleştirilen uygulamalarıyla elde edilmiş sonuçları incelenmektedir. Araştırmanın, araştırma kapsamında kullanılan modellerin kullanışlılığına, sağladıkları bilgi miktarı ve bu bilgilerin örtüşen ve örtüşmeyen noktalarının neler olduğuna; uygulama esnasında karşılaşılabilen olası problemlere ve getirilebilecek çözüm önerilerine ilişkin fikir vermesi yönüyle araştırmacılara ve uygulayıcılara katkı sağlayacağı düşünülmektedir.

Araştırmanın, her iki model için sınanan yeterliliğine, puanlayıcılara, test maddelerine ve değişkenlik kaynakları arasındaki etkileşimlere ilişkin kullanışlı bilgiler sunması ve dolayısıyla modellerin kuramsal yapılarına dair edinilen bilgi düzeyini arttırması bakımından, literatüre ve ihtiyaç duyulabilecek yeni modellerin geliştirilmesine katkı getireceği düşünülmektedir.

Gerçek veri seti üzerinde her iki modelin incelenmesi, uygulayıcı ve araştırmacılara hangi modelin daha çok bilgi sağladığına, pratikte daha kullanışlı olduğuna ve amaçları doğrultusunda kendi çalışmaları için kullanılacak en etkili modeli seçmelerine ilişkin yararlı olacağı ifade edilebilir.

Sayıltılar

1. Puanlayıcılar, araştırma grubunda yer alan her bir sınananın verdiği yanıtlara birbirinden bağımsız olarak puan atamışlardır.

Sınırlılıklar

1. Araştırmada, PISA ikinci dönem uygulamasında yer almış ve açıklanmış olan 10 açık uçlu madde kullanılmak istenmiştir. İki madde uzman görüşü doğrultusunda beş kategoride puanlanamayacağı için uygulama dışında tutulmuş ve araştırma toplam 8 açık uçlu madde ile sınırlandırılmıştır.
2. ÇDKRÖM ve HPM için parametre kestirimine olanak veren belirli bir örneklem büyüklüğü sayısına ulaşılamamıştır. Literatürde modellerin kullanıldığı çalışmalar (Mariano, 2002; Patz ve diğerleri 2002) ve kestirim yapılacak parametre sayıları temel alınmak suretiyle, çalışma grubunu oluşturacak öğrenci ve puanlayıcı sayısı belirlenmiştir. Bu bağlamda, elde edilen veriler 2012-2013 eğitim-öğretim yılı Ankara ili Çankaya ilçesinde yer alan 10 okul, 350 öğrenci ve 5 ortaöğretim matematik öğretmeni ile sınırlı tutulmuştur.
3. HPM'nin parametre kestirimleri, Bayes kestirim yönteminin kullanımıyla sınırlandırılmıştır.

Tanımlar

Sınanan: Ölçme işlemine tabi tutulan bireydir. Araştırmanın giriş başlığı altında, daha genel bir anlam taşıması nedeniyle sınanan kavramının kullanılması tercih edilmiştir. Fakat araştırmada ölçme işlemine tabi tutulan sınananların, öğrenciler olması nedeniyle bulgular ve yorumlar başlığı altında sınanan yerine öğrenci ifadesi kullanılmıştır.

Kısaltmalar ve Semboller

NEAP	National Assessment Of Educational Progress
SAT	Scholastic Aptitude Test
AP	Advanced Placement
GRE	Graduate Record Examinations
TİMMS	Trends İn International Mathematics And Science Study
PISA	Programme For International Student Assessment
ÇDKRÖM	Çok Değişkenlik Kaynaklı Rasch Ölçme Modeli
HPM	Hiyerarşik Puanlayıcı Modeli
GRM	Graded Response Model
AŞM	Aşamalı Tepki Modeli
RRM	Rating Response Model
DTM	Dereceli Tepki Modeli
PCM	Partial Credit Model
KKM	Kısmi Kredi Modeli
GPCM	Generalized Partial Credit Model
GKKM	Genelleştirilmiş Kısmi Kredi Modeli
OECD	The Organisation For Economic Co-Operation And Development
BUGS	Bayesian İnference Using Gibbs Sampling
BBK	Bayes Bilgi Kriteri
SBK	Sapma Bilgi Kriteri
MCMC	Markov Chain Carlo Estimation
RMSE	Hataların Ortalama Karekökü
P_{nijk}	K Kategorisinde Puanlanma Olasılığı
P_{nijk}-1	K-1 Kategorisinde Puanlanma Olasılığı
B_n	N Sınanına Ait Yetenek Düzeyi
D_i	İ Maddesinin Güçlük Düzeyi
C_j	J Puanlayıcısının Katılık Düzeyi
F_k	Kategoriler Arası Geçişin Güçlük Düzeyi
θ	Sınanan Yeterliliği
SS	Standart Sapma
G	Ayırma Oranı
R	Ayırma İndeksi Güvenirliği
İ	HPM İçin Sınanan
J	HPM İçin Madde
R	HPM İçin Puanlayıcı
X_{ijr}	Atanan Puan
ξ_{ij}	İdeal Puan
ϕ	Puanlayıcı Katılığı
ψ	Puanlayıcı Değişkenliği
B	Madde Güçlüğünü
γ_{jk}	Kategorilere Ait Adım Güçlüğü
K-1	Kategori Eşik (Parametre) Sayısı
D	Dağılımlar Arasındaki Uzaklık
K	Gözlenen Puanlar
τ	Tau
M	Öğrenci Yeterlilik Ortalaması
σ^2	Öğrenci Yeterlilik Varyansı

BÖLÜM II

YÖNTEM

Bu bölümde, araştırma modeli, çalışma grubu, veriler ve toplanması, verilerin analizi ile ilgili açıklamalara yer verilmiştir.

Araştırma Modeli

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

Araştırma, elde edilecek bulgular doğrultusunda örnekleme ilişkin bilgilerin evrene genellenmesi amacını taşımamaktadır. Bu bağlamda, araştırma grubu: 2012-2013 eğitim-öğretim yılı II. yarıyılında (bahar dönemi) Çizelge 2'de belirtilen okullarda öğrenim gören, 15 yaş grubu öğrencilerinden oluşmaktadır. Ulaşım kolaylığı gözetilerek, Ankara ili Çankaya ilçesinde yer alan okullar tercih edilmiştir. Her iki model için literatürde parametre kestirimlerinin yapılabilmesi amacıyla ihtiyaç duyulan örneklem büyüklüğüne dair bir bilgiye ulaşılamamıştır. Araştırmada yer alan öğrenci sayısı, her iki model için tamamen çaprazlanmış desende (Fully-Crossed Design) gerçekleştirilmiş çalışmalardaki (Patz ve diğerleri, 2000; Patz ve diğerleri, 2002; Turner, 2003) öğrenci, puanlayıcı ve madde sayılarıyla birlikte öğrenci, puanlayıcı ve madde sayılarının çarpımı sonucu elde edilen toplam veri sayısı dikkate alınarak belirlenmiştir. Bu bağlamda, öğrencilerden gönüllü olanların katılımıyla dersler esnasında gerçekleştirilen uygulamalar ile 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 üç, 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 esasına dayalı olarak çalışmaya dahil edilmiş beş ortaöğretim matematik öğretmeninden oluşmaktadır. Çalışmaya gönüllü olarak katılmak isteyen öğretmenlerin, ortaöğretim matematik öğretmenliği lisans derecesine sahip olması gerekliliği gözetilmiştir.

Çizelge 2. Okul Adları ve Öğrenci Sayıları

Okullar	Öğrenci Sayısı		Toplam
	Analize Dahil Olan	Analiz Dışında Tutulan	
100. Yıl Kız Teknik ve Meslek Lisesi	29	3	32
50.Yıl Anadolu Sağlık Meslek Lisesi	31	3	34
Kurtuluş Anadolu Lisesi	45	3	48
Ayrancı Anadolu Lisesi	43	3	46
Ankara Çankaya Aziz Altın Pınar Lisesi	25	3	28
Cumhuriyet Anadolu Öğretmen Lisesi	44	3	47
Ali-Hasan Coşkun Kız Teknik Ve Meslek Lisesi	23	3	26
Bahçelievler Anadolu Lisesi	51	3	54
Anıttepe Anadolu Lisesi	32	3	35
Çankaya Cumhuriyet Ticaret Meslek Lisesi	27	3	30
Toplam	350	30	380

Veriler ve Toplanması

Araştırmanın amaç ve alt amaçlarının gerçekleştirilebilmesi için açık uçlu maddelere ihtiyaç duyulmuştur. Araştırma kapsamında maddelerin yazılması yerine hâlihazırda uzman bir grup tarafından geliştirilmiş ve uygulanmış maddelerin kullanımı tercih edilmiştir. Bu bağlamda, Ekonomik İşbirliği ve Kalkınma Teşkilatı

(The Organisation for Economic Co-operation and Development-OECD) tarafından üç dönem hâlinde uygulaması yapılan uluslararası bir çalışma olan, PISA (EARGED, 2004) kapsamında kullanılmış ve açıklanmış maddeler incelenmiştir. Var olan açık uçlu maddelerin niceliği ve araştırma kapsamında kullanılabilirliği dikkate alınmış ve PISA ikinci dönem uygulamasında yer almış, 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, aşağıda sıralanan gerekçeler dikkate alınarak bütünsel dereceli puanlama anahtarının kullanılması tercih edilmiştir. Bu gerekçeler:

- Ürünün ya da performansın bir bütün olarak ve daha hızlı puanlanmasına olanak sağlaması (Mertler, 2001; Nitko, 2001; Reynolds, Livingston ve Willson 2006; Volk, 2002);
- Maddelerle ölçülmesi amaçlanan performansın, alt bileşenlerinin ölçülmesine ihtiyaç duyulmadığı durumlarda uygulanabilir olması (Finson, 1998 aktaran Rezaei ve Lovorn, 2010; Jonsson ve Svingby, 2007; Mertler, 2001; Quinlan, 2011);
- Araştırma grubunda sayıca fazla bireyin bulunmasıdır (Jonsson ve Svingby, 2007; Lund ve Veal, 2013; Quinlan, 2011).

Bütünsel dereceli puanlama anahtarının hazırlanması sürecinde litaratürde konuyla ilgili kaynaklarda (Airasian, 2001; Mertler, 2001; Nitko, 2001; Popham, 1997; Stevens ve Levi, 2005) yer alan adımlar dikkate alınmıştır. 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 araştırma kapsamı dışında tutulmuştur. Öncelikle, her bir maddeye ilişkin işlem basamakları, olası performansların gözlenebilir özellikleri ve önemli yönleri açık bir biçimde belirlenmiştir. Ardından, her maddenin her bir kategorisi için öğrenci yanıtlarına ve yardım alınan kişilerin görüşlerine dayanılarak gözlenebilir biçimde ifade edilebilen, tanımlamalar geliştirilmiştir. Literatürde geçen dereceli puanlama anahtarı hazırlama sürecinde yapılmaması gereken hatalar (Benjamin, 2013; Cunningham, 1998; Popham, 1997; Reynolds ve diğerleri, 2006) ışığında her bir kategori tekrar düzeltilmiştir. 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. Bu aşamalar dikkate alınarak hazırlanan dereceli puanlama anahtarları Ek B'de yer almaktadır.

Maddeleri içeren soru kitapçıkları, öğrencilerin okumakta zorlanmayacakları ve yanıtlarını rahatlıkla yazabilecekleri biçimde hazırlanmıştır. Soru kitapçığı, Ek A'da verilmiştir. 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 yanıtlanmıştır. 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, sekiz açık uçlu maddeye verdikleri yanıtlara ait beş puanlayıcının puan ataması sonucunda toplam 14000 veri elde edilmiştir.

Verilerin Analizi

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

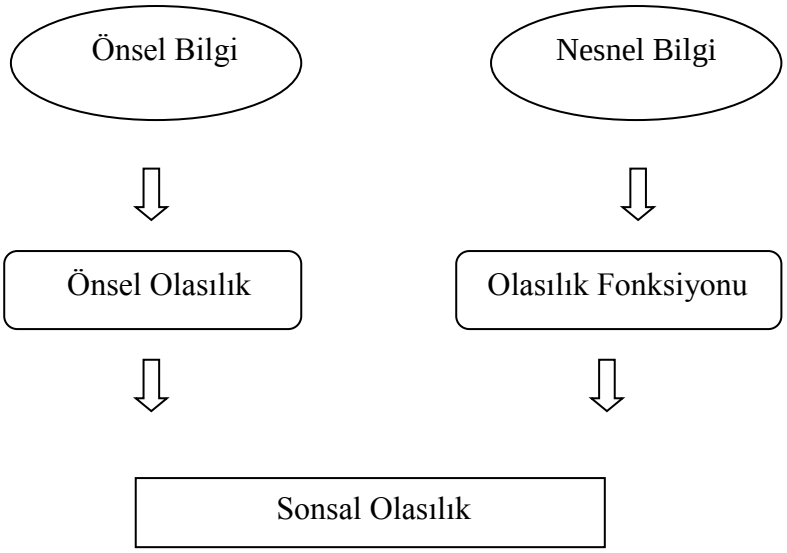
ÇDKRÖM kapsamında öğrenciler, puanlayıcılar ve maddeler değişkenlik kaynağı olarak ele alınmıştır. ÇDKRÖM, Rasch ailesine mensup bir modeldir. Bu modelin model veri uyumu Rasch yaklaşımlarında olduğu gibi ele alınır. Rasch analizlerinin öncelikli amacı model veri uyumunu maksimum noktasına çıkarmak değil, genellenebilir doğrusal ölçümler oluşturabilmektir. Bunun göstergesi olarak da standart hata (güvenirlik) ve geçerlilik [uyum(fit)] kullanılmaktadır (Linacre, 1994, 1998). ÇDKRÖM doğası gereği her bir değişkenlik kaynağı için standart hata ve geçerlilik değerlerini vermektedir. Ayrıca Linacre (1994) tarafından kullanımı tavsiye edilmiş, model veri uyumu göstergesi olarak ele alınan, standartlaştırılmış artık değerlere de ulaşılmaktadır. Modele ait elde edilen bu değerlerin tamamına ilişkin açıklamalar, bulgularla bütünlük taşıması bakımından bulgular başlığı altında rapor edilmiştir. ÇDKRÖM analizleri, farklı programlarla gerçekleştirilmekle birlikte, kullanım ve ulaşım kolaylığı nedeniyle araştırma kapsamında FACETS programından faydalanılmıştır.

ÇDKRÖM analizi gerçekleştirilirken, birisi hariç (genellikle öğrenci yeterliliği) tüm değişkenlik kaynakları ortak bir orijine (genellikle 0 olan) merkezlenmektedir. Eğer birden daha fazla değişkenlik kaynağı merkezlenmezse (noncentered), referans çerçevesi yeterince sınırlanmadığı için sonuçlar belirsiz

olabilir (Linacre, 1998). Dağılım merkezinin manipüle edilmesi, gerçekte araştırma sonuçlarını değiştirmez. Sadece, verilen değişkenlik kaynakları için kestirilen, tüm parametre dağılımının ek olarak yer değiştirmesine yol açar (Engelhard ve Myford, 2003). Bu çalışmada da öğrenci yeterliliği merkezlenmemiş, puanlayıcı katılığı ve madde güçlüğü merkezlenmiştir.

HPM uygulamaları, farklı kestirim yöntemleri kullanılarak gerçekleştirilebilir. Literatür incelendiğinde; Hombo ve Donoghue (2001) çalışmalarında, marjinal en çok olabilirlik (marginal maximum likelihood) kestirimini, Patz ve diğerleri (2002) Bayes (Bayesian) kestirimini kullanırken DeCarlo ve diğerleri (2011) HPM-STD model için kısmen Bayes yaklaşımına dahil olan, sonsal mode (posterior mode) kestirimini kullanmıştır. Bu çalışmada Bayes kestirimi kullanılmıştır. Çünkü Bayes model kurulumu ve model uyumu dikkate alındığında, örtük değişkenlere ait parametrelerin kestirilmesine yapısı gereği oldukça uygundur (Patz ve Junker, 1999a, 1999b; Patz ve diğerleri, 2002). Bunun yanı sıra Bayes kestirimi sınır değer problemiyle (boundary value problems) başa çıkma noktasında oldukça kullanışlıdır (DeCarlo ve diğerleri, 2011; Gelman, Carlin, Stern ve Rubin, 1995; Wikle, Berliner ve Milliff, 2003). Kolaylıkla ulaşılabilen bir çok yazılım aracılığıyla da kestirimler gerçekleştirilebilir. Ayrıca, literatürde yer alan HPM uygulamalarında büyük çoğunlukla Bayes kestirimi kullanılmıştır (Casabianca ve Junker, 2013; Casabianca ve diğerleri, 2014; Mariano, 2002; Patz ve Junker, 1999a, 1999b; Patz ve diğerleri, 2002).

Bayes, istatistiksel karar verme (Efron, 1986) ve çıkarsama yöntemlerinden biridir (Congdon, 2001). Bayes yaklaşımı, bilinmeyen parametrelerin sonsal dağılımlarının kestirilmesi amacıyla iki bilgi kaynağını olabildiğince verimli bir şekilde birleştirmeyi amaçlamaktadır. Bu kaynaklardan ilki, veride yer alan nesnel bilgileri temsil ederken; ikincisi, gerçekliği kabul edilen fikir, teori ya da önsel bir bilgi olarak ifade edilebilir (Efron, 1986). Parametrelerin kestirilmesi sürecinde, öznelliği temsil eden önsel bilgiler, uygun olan dağılımlar aracılığıyla modele dâhil edilmektedir. Bayes yaklaşımı, önsel bilgilerden sonsal bilgilere doğru bir geçişi içermektedir (O'Hagan, 1986). Başka bir ifadeyle, kestirilmesi amaçlanan parametrelere ait koşullu dağılım olan sonsal dağılım; çalışma verilerinin örneklem bilgisi ile önsel bilginin birleşimi şeklinde yorumlanabilir (Gelman ve diğerleri, 2004; Gelman ve diğerleri, 2013).



Şekil 4. Bayes Bilgi Şeması

Bayes kestiriminin kullanılabilmesi amacıyla, modelde bulunan parametreler için önsel dağılımların belirlenmesi gerekmektedir. Önsel dağılımların seçiminde belirgin bir kural bulunmamaktadır. Araştırmacı, çalışmasının koşulları doğrultusunda, bilgi içeren (informative) veya bilgi içermeyen (non-informative) önsel dağılımlardan yararlanabilir (Gelman ve diğerleri, 2004; Zuur ve diğerleri, 2009). Bu çalışmada, önsel dağılımlar için benzer koşullara sahip olunması bakımından, Patz ve diğerleri (2002)ne ait olan çalışmada yer alan önsel dağılımlardan faydalanılmıştır. Puanlayıcı parametreleri için güçlü ön bilgilere sahip olunmadığından; puanlayıcı katılık parametresi (ϕ_r) ortalaması 0, varyansı 10 olan bilgilendirici olmayan (uninformative) normal dağılım N(0,10); puanlayıcı değişkenlik parametresi (ψ_r) log-normal yoğunluk (log-Normal density) $\log(\psi_r) \sim N(0,10)$ olarak belirlenmiştir. Madde güçlük parametresi (β_j) önsellerinin geliştirilmesi için konum belirsizliğinin (location indeterminacy) dikkate alınması gereklidir. Tanımlanmış modelin elde edilebilmesi amacıyla, öğrenci yeterlilik ortalaması (μ) veya βj ya kısıt getirilmelidir. Bu nedenle, μ için N(0,10) önsel dağılımı belirlenirken βj için de bağımsız özdeşçe dağılmış (independent identically distributed-i.i.d.) benzer normal dağılım kullanılmış ve sıfıra giden toplam (sum to zero) $\beta_J = -\sum_{j=1}^{J-1} \beta_j$ kısıtı getirilmiştir. Bu, kestirimin daha kararlı hâle gelmesini sağlayacaktır (Gilks, Richardson ve Spiegelhalter, 1995 akt. Patz ve diğerleri, 2002). K-1 madde adım parametrelerinin (γ_{jk}) sonuncusu hariç, bağımsız özdeşçe dağılmış

N(0,10) önselleri kullanılmıştır. Madde adım parametreleri için de konum belirsizliği bulunmaktadır. Tüm maddeler için sonuncu madde adım parametreleri ise diğer madde parametrelerinin doğrusal fonksiyonudur. Bu nedenle, sıfıra giden toplam (sum to zero) $\gamma_{j(K-1)} = - \sum_{K=1}^{K-2} \gamma_{jk}$ kısıtı getirilmiştir. Son olarak, öğrenci yeterliliklerinin varyansı (σ^2) için $1/\sigma^2 \sim$ Gamma (α,η) önsel dağılımı kullanılmıştır. σ^2 ilişkin küçük bir miktar önsel bilginin yansıtılması amacıyla $\alpha=\eta=1$ tercih edilmiştir. HPM parametreleri için kullanılan bilgilendirici olmayan önsel dağılımlar Çizelge 3'te özetlenmiştir.

Çizelge 3. 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	$1/\sigma^2 \sim$ Gamma (α,η), $\alpha=\eta=1$

Bayes yaklaşımı altında, bilinmeyen parametrelere ilişkin önsel bilgilerin temsil edilmesi, belli olasılık dağılımlarıyla sınırlandırıldığı durumlarda bile yüksek dereceden integraller alınmasını gerektirir. Bunların çözümü oldukça güçtür. Uygun olmayan önsel bilgiler kullanıldığında ise sonsal dağılıma ulaşılamayacağı için, analitik işlemlerin gerçekleştirilmesi imkânsız hâle gelecektir. Bu problemin ortadan kaldırılması amacıyla, bilgisayar teknolojilerinin ilerlemesiyle birlikte koşullu dağılımlardan örneklem çekerek parametre tahminlerinin elde edilmesini sağlayan, Markov Chain Monte Carlo (MCMC) kestirim yöntemleri geliştirilmiştir (Gemerman, 1997). MCMC yöntemleri, Markov Zincirleri'nin kullanımıyla Monte Carlo integrasyonunun gerçekleştirilmesi temeline dayanır (Walsh, 2002). Bu yöntemle sonlu sayıda gözlem değeri kullanılarak, sonsuz sayıda veri elde etmek mümkündür (Yardımcı ve Erar, 2005). Böylece, analize önsel bilgilerin dâhil edilmesi mümkün hâle gelirken, Bayes yaklaşımının kullanımına da olanak sağlanır (Browne ve Rasbash, 2004).

HPM kestirimi, Bayes kapsamında MCMC yöntemi kullanılarak gerçekleştirilmiştir. Öncelikle, Patz ve Junker (1999a, 1999b) yayınladıkları

çalışmalarında, ideal puanlar değişkenlerini veren kısmi kredi modeli MCMC kestirimini doğrudan elde edilebilir hale getirmişlerdir. Daha sonra Johnson, Cohen ve Junker (1999) kısmi kredi model parametrelerinin, MCMC kestirimini Bayesian Inference Using Gibbs Sampling (BUGS) (Spiegelhalter, Thomas, Best ve Gilks, 1996)da gerçekleştirmiştir. BUGS, 1989 yılında başlamış, MCMC yöntemlerini kullanan, karmaşık istatistiksel modellerin Bayes analizleri için geliştirilmiş, esnek bir yazılım programıdır (BUGS, 2014). Son olarak, Patz ve diğerleri (2002) tarafından tamamlanmış bağımlı dağılımdan (complete conditional distrubution) ideal puanların ve puanlayıcı parametrelerinin elde edilmesini sağlayan adımlar eklenerek, HPM'de yer alan kısmı kredi modeli için MCMC yöntemleri genişletilmiştir.

HPM analizinin gerçekleştirilebilmesi için ihtiyaç duyulan kodlar, Patz ve diğerleri (2002) tarafından yayınlanan çalışmada yer alan kodların araştırma kapsamında, tekrar düzenlenmesiyle elde edilmiştir. Puanlayıcı değişkenlik parametresi, model kodları içinde birin tau (τ) parametresine bölümünün karekökü ($\tau=1/\psi^2$) şeklinde tanımlanmıştır. Analiz sonucunda, model içinde tanımlanma biçimi nedeniyle, puanlayıcı değişkenlik parametresi değil tau elde edilmiştir. Çizelge oluşturulurken yukarıda verilmiş olan eşitlik doğrultusunda puanlayıcı değişkenlik parametreleri hesaplanmış ve bu değerlere çizelgede yer verilmiştir.

HPM analizleri, öncelikle R (R Development Core Team, 2013) programı RUBE (Seltman, 2010) paketi aracığıyla BUGS ailesinin bir üyesi olan, WinBUGS (Lunn, Thomas, Best ve Spiegelhalter, 2000) arayüzü ile gerçekleştirilmiştir. Fakat gerek iterasyon sayısının takip edilememesi gerekse daha fazla zaman alması nedeniyle yine BUGS ailesinden OpenBUGS programında tamamlanmıştır. Analiz sonuçlarının, geçerli ve güvenilir olup olmadığının görülebilmesi için en az iki Marcov zinciri kullanılmalıdır. Zincir sayısının artmasıyla analiz süresinin artacağı da dikkate alınmalıdır. Analizde, MCMC yöntemleriyle başlangıç değerlerine dayanılarak üretilen örneklemlerin dağılımı, denge dağılımını yakınsaması bakımından Marcov zincirleri yeterince uzun çalıştırılmalıdır. Başka bir deyişle, iterasyon sayısı arttırılmalıdır. Ayrıca zincirlerin başlangıç değerlerini unutabilmesi açısından burn-in periodu olarak adlandırılan analizin, ilk belli sayıdaki iterasyonu örneklemden çıkarılmalıdır. Ardışık iki iterasyon ($\theta^{(t)}$, $\theta^{(t+1)}$) ya da aralarında bir iterasyon bulunan iki iterasyon arasında ($\theta^{(t)}$, $\theta^{(t+2)}$) yöntemin işleyişi nedeniyle

korelasyon görülebilir. Zincirlerin oto korelasyon gösterip göstermediğine ilişkin grafik çizdirilebilir. Eğer zincirler oto korelasyon gösteriyorsa, her 5. ya da 10. iterasyon seçilebilir. Gerçekleştirilen sonuncu analizde bu değerler kullanılabilir. Bu işleme seyreltme oranı (thinning rate) adı verilmektedir (Zuur, Saveliev ve Ieno, 2012). Tüm bu bilgiler dikkate alınarak analizde, 5000 burn-in periodu, 30,000 iterasyon sayısı, üç Marcov zinciri ve 10 seyreltme kullanılmıştır. 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).

BGR, 1998'de Brooks ve Gelman tarafından önerilmiş olan, zincir içi ve zincirler arası varyansı karşılaştıran, ANOVA tipi tanılayıcı bir istatistiktir. Genellikle görsel olarak incelenen, grafikte yer alan değerin 1 veya yaklaşık olarak 1 bulunması yakınsamanın göstergesiyken, 1.1 ve çevresi kabul edilebilir değerler olarak ele alınmaktadır (Gelman ve Hill, 2007; Kéry, 2010).

Bir diğer gösterge olan; zaman serileri diyagramından, seçkisiz gürültülerin (random noise) belirlenebilmesi için faydalanılmaktadır. Bu diyagramlar görsel olarak incelenmelidir (Spiegelhalter ve diğerleri, 2003; Lunn, Spiegelhalter, Thomas ve Best, 2009). Örneklenen değerlerin (sampled values), ortak bir ortalama değeri etrafında seçkisiz olarak hareketli (bouncing randomly) olması, istenen bir durumdur (Gelman, Carlin, Stern ve Rubin, 2004; Gelman ve diğerleri, 2013). Analizde yer alan tüm Markov zincirlerinin binişik ya da bir noktada birleşmiş olması, yakınsamanın göstergesi olarak kabul edilebilir (Lee ve Wagenmakers, 2014). Eğer örneklenen değerler, tek bir yönde ve sabit bir biçimde ilerliyorsa bu, zincirin yakınsamadığını ifade etmektedir (Gelman ve diğerleri, 2004).

HPM için model uyumunun ölçülmesi, diğer karmaşık (kompleks) hiyerarşik yapıdaki modellerin uyumunun ölçülmesiyle benzerdir. Model veri uyumu göstergeleri olan kriterler, genellikle model sapma (model deviance) değerine dayanır ve modellerin yuvalanmış (nested) olup olmama durumlarına göre farklılık gösterirler (Casabianca ve Junker, 2013). Bilgi teorisine dayanan bu kriterler,

literatürde model seçimi için de kullanılmaktadır (Ucal, 2006). Birden fazla modelin karşılaştırılmasında olabilirlik oran testlerine dayanan klasik yaklaşımlar, sahip oldukları yorumlama güçlükleri nedeniyle son yıllarda daha az tercih edilir hâle gelmiş (Raftery, 1995); buna karşın evrensel tek bir kriter olamamakla birlikte, Akaike Bilgi Kriteri (Akaike, 1973), Schwartz Bilgi Kriteri olarak da adlandırılan Bayes Bilgi Kriteri (BBK) (Schwarz, 1978), Bayes Faktörü (Kass ve Raftery, 1995; akt. Casabianca, 2012), Sapma Bilgi Kriteri (SBK) (Spiegelhalter, Best, Carlin ve Van der Linde, 1998) kullanımı artmıştır (Zuur ve diğerleri, 2012; Berg, Meyer, Yu, 2004; Ward, 2008; Zuur ve diğerleri, 2009). ÇDKRÖM ve HPM sonuçlarının doğrudan karşılaştırılması, gerek modelde yer alan parametre sayılarının ve yorumlanma biçimlerinin gerekse her iki model için kullanılan kestirim yöntemlerinin farklı olması nedeniyle gerçekçi olmayacaktır. HPM uygulamalarının yer aldığı çalışmalar (Mariano ve Junker, 2007; Patz ve diğerleri, 2002) incelendiğinde, model veri uyumunun belirlenebilmesi ve model karşılaştırması için BBK değerinden faydalanıldığı görülmektedir. BBK, Gideon E. Schwarz tarafından 1978 yılında geliştirilmiş olup artık kareler toplamı ve tahmin edilen parametre sayısının artan bir fonksiyonu olarak, doğru modeli seçme noktasında yüksek olasılığa sahip bir prosedür ortaya koymaktadır (Schwarz, 1978). Karmaşık hiyerarşik modeller için BBK kullanıldığında, serbestlik derecesinin belirlenmesi noktasında problem yaşanmaktadır (Berg ve diğerleri, 2004; Casabianca, 2012; Spiegelhalter, Best, Carlin ve Van Der Linde, 2002). Bu nedenle özellikle karmaşık hiyerarşik modeller için SBK kullanımı önerilmektedir (Berg ve diğerleri, 2004). SBK, modelde uygun olmayan önseller kullanıldığında ya da oldukça fazla bilinmeyen parametrenin yer aldığı durumlarda, hesaplanması zor olan Bayes Faktör ve BBK'ya alternatif olarak ortaya konulmuş bir kriterdir (Ardia, 2008). Dempster (1997) tarafından sunulmuş, sapma (deviance)nın genişletilmiş bir hâli olan SBK, uyum iyiliği (goodness of fit) ile karmaşıklığın (complexity) toplamı prensibine dayanır (Berg ve diğerleri, 2004; Spiegelhalter ve diğerleri, 2002; Zhu ve Carlin, 2000). Uyum, sapma aracılığıyla aşağıda yer alan formülle hesaplanabilir;

(11)

$$\overline{D} = E_{\theta|y}[D]$$

Karmaşıklık ise modelde etkili olan parametre sayısıyla; başka bir deyişle sonsal ortalama sapmasından, parametrelerin sonsal ortalamalarından hesaplanan

sapmanın çıkarılmasıyla elde edilir. Formülü aşağıdaki gibidir (Spiegelhalter ve diğerleri, 2002):

(12)

$$pD = E_{\theta|y}[D] - D(E_{\theta|y}[\theta])$$
$$= \overline{D} E_{\theta|y}[\overline{D}] - D(\theta)$$

Böylece SBK değerine aşağıdaki formül aracılığıyla erişilebilir;

(13)

$$SBK = \overline{D} + pD$$

$\overline{\mathbf{D}}$: Uyum
pD : Karmaşıklık

SBK, kolayca hesaplanabilmesi ve birçok model için kullanılabilmesi açısından oldukça pratiktir (Berg ve diğerleri, 2004; Casabianca, 2012; Li ve Yu, 2012; Li, Zeng ve Yu, 2014). Ayrıca farklı parametre sayılarına sahip modellerin birlikte değerlendirilebilmesi açısından da kullanışlıdır (Liddle, 2007; Li ve Yu, 2012). Daha küçük SBK değeri, daha güçlü model veri uyumunun ve daha iyi parametre tahminin göstergesidir (Berg ve diğerleri, 2004; Liddle, 2007; Spiegelhalter ve diğerleri, 2002; Zhu ve Carlin, 2000). Bu çalışmada da HPM'nin model veri uyumunun ölçülmesi ve HPM ile ÇDKRÖM sonuçlarının karşılaştırılması için SBK'dan faydalanılmıştır. Bu nedenle, araştırmanın son alt amacının gerçekleştirilebilmesi için HPM ve ÇDKRÖM SBK değerleri OpenBUGS üzerinden kestirilmiştir.

ÇDKRÖM ve HPM için kestirilen parametrelerin, rapor edilebilir niteliğe sahip olup olmadığına dair bilgi veren ve her iki modelin birlikte değerlendirilmesinde kullanılan göstergeler, genel bir çerçevede çizelge 4'te özetlenmiştir. Çizelge 4'te yer alan göstergelerin her biri, farklı kaynaklardan elde edilen bilgilere dayanmaktadır. Bu kaynaklara ilişkin bilgiler, çizelgede net bir değer aralığına sahip olmadığı için yer verilemeyen diğer göstergeler ve göstergelere ait daha detaylı açıklamalar metin içinde verilmiştir.

Çizelge 4. ÇDKRÖM ve HPM ile İki Modelin Birlikte Değerlendirilmesine İlişkin Kullanılan Göstergeler ve Aralık Değerleri

ÇDKRÖM	
Gösterge	Değer/Değer Aralığı
Düzeltilmiş Standart Sapma	≤ 1.0
Uyum İçi ve Uyum Dışı İstatistikleri	0.8 – 1.2
HPM	
Gösterge	Değer/Değer Aralığı - İstenen Durum
BGR Grafikleri	≈ 1
Zaman Serileri Diyagram	Zincirlerin seçkisiz olarak hareketli ve binişik olması
MC hatası	$<$ 0.05
Her İki Modelin Birlikte Değerlendirilmesi	
Gösterge	İstenen Durum
SBK	Daha küçük bir değer alması
$\overline{D}$	Daha küçük bir değer alması
pD	Daha büyük bir değer alması

BÖLÜM III

BULGULAR VE YORUMLAR

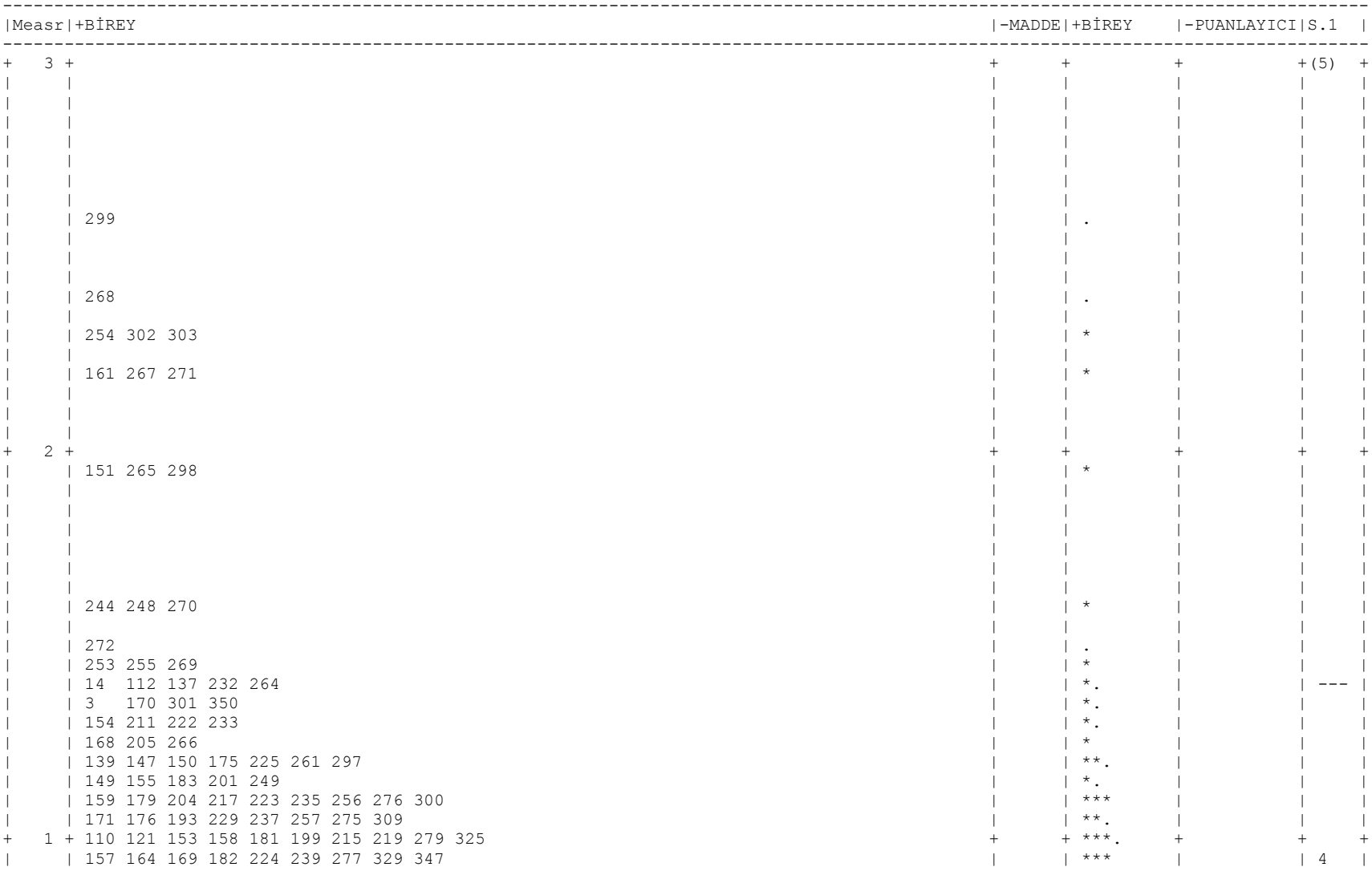
Bu bölümde, araştırmanın verilerinin, araştırmanın amaç ve alt amaçlarına dayalı olarak çözümlenmesiyle ulaşılan bulgular ve yorumlar yer almaktadır.

Çok Değişkenlik Kaynaklı Rasch Ölçme Modeli Bulguları

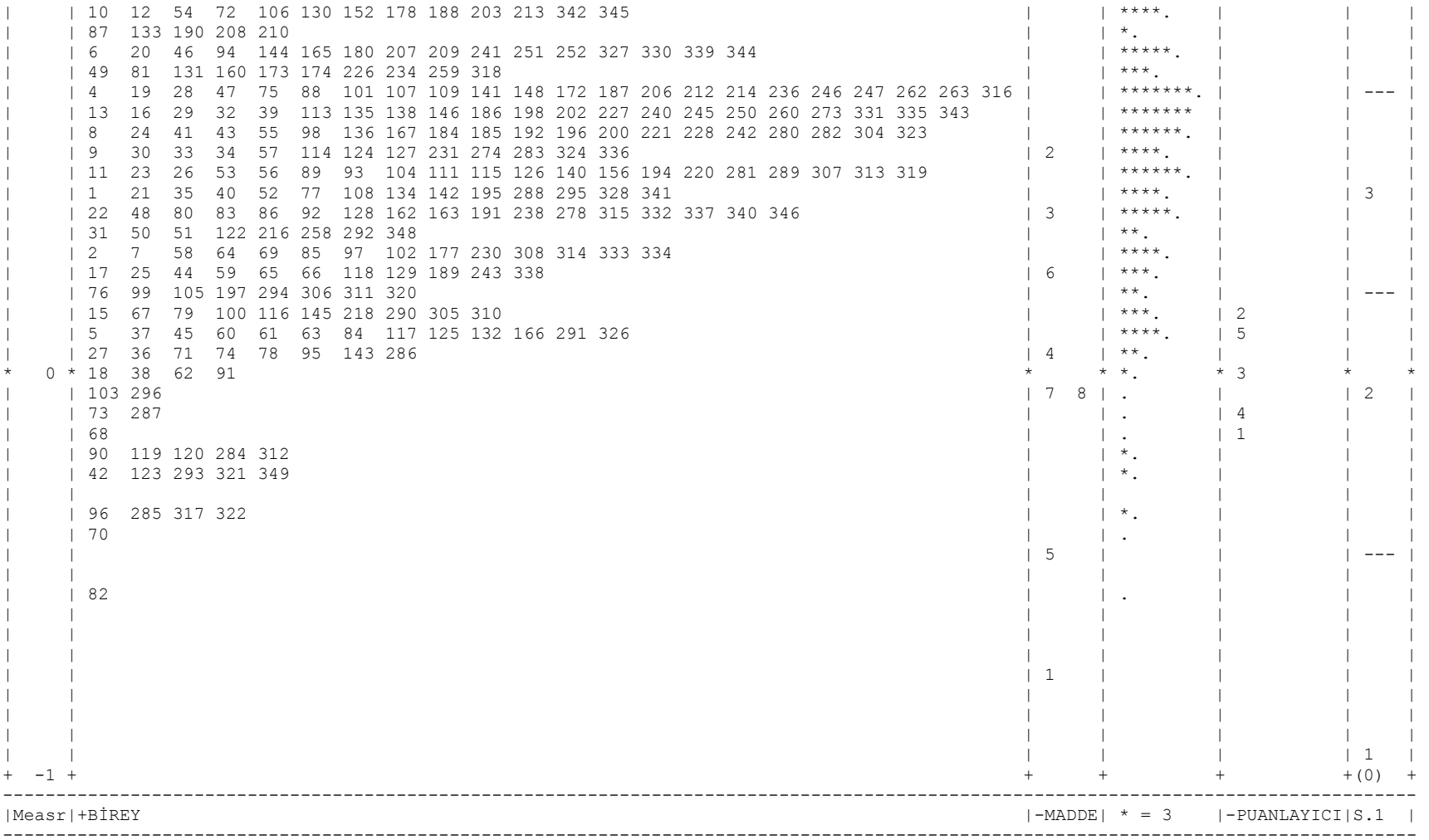
Araştırmanın verileri için birinci alt problem doğrultusunda, çok değişkenlik kaynaklı Rasch modeli ile elde edilen model veri uyumu istatistikleri, öğrenci yetenek ve uygunluk istatistikleri, puanlayıcıların katılık/cömertlik ve uygunluk istatistikleri ile madde güçlükleri ve uygunluk istatistikleri belirlenmiş ve açıklanmıştır.

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. ÇDKRÖM analizi sonucu elde edilen ve verilerin tamamına ilişkin tüm değişkenlik kaynaklarını içeren veri kalibrasyon haritası, Şekil 5'te verilmiştir.


|Measr|+BİREY
|-MADDE|+BİREY
|-PUANLAYICI|S.1
3
2
1
299
268
254 302 303
161 267 271
151 265 298
244 248 270
272
253 255 269
14 112 137 232 264
3 170 301 350
154 211 222 233
168 205 266
139 147 150 175 225 261 297
149 155 183 201 249
159 179 204 217 223 235 256 276 300
171 176 193 229 237 257 275 309
110 121 153 158 181 199 215 219 279 325
157 164 169 182 224 239 277 329 347
(5)

4



Şekil 5. Çok Değişkenlik Kaynaklı Rasch Modeli Veri Kalibrasyon Haritası

Şekil 5'te verilmiş olan veri kalibrasyon haritası; öğrenci yeterliliğinin, puanlayıcı katılık/cömertliğinin ve madde güçlüğünün tek bir aralıklı (logit) ölçek üzerindeki kalibrasyonunu göstermektedir (Iramaneerat ve diğerleri, 2008). Nakamura (2000)e göre; veri kalibrasyon haritası aynı ölçek üzerinde öğrenci, puanlayıcı ve maddeler arasındaki ilişkiye dair hızlı bir biçimde öz ve uygun bilgi edinmemizi sağlamaktadır. Bireylerin yer aldığı sütun incelendiğinde; her bir bireyin analizde adlandırıldıkları numaralarıyla yetenek puanlarına göre logit ölçek üzerinde sıralandıkları görülmektedir. Bu sıralamaya göre, en yüksek logit değere sahip olan (2.59 logit birim ile) 299 numaralı birey, yetenek düzeyi en yüksek olan bireydir. Benzer biçimde en düşük logit değere sahip olan (-0.57 logit birim ile) 82 numaralı birey, yetenek düzeyi en düşük olan bireydir. Maddelerin güçlük düzeylerine göre sıralanışlarının verilmiş olduğu sütun dikkate alındığında; diğer maddelere göre ölçeğin en üstünde yer alan (0.56 logit birim ile) iki numaralı madde, en zor maddedir. Ölçeğin en altında yer alan (-0.74 logit birim ile) bir numaralı madde ise en kolay maddedir. Son olarak puanlayıcı davranışlarının yer aldığı sütun incelendiğinde; logit puanı (0.14) diğer puanlayıcılara göre yüksek olan iki numaralı puanlayıcı, en katı puanlayıcıyken, logit puanı (-0.16) diğer puanlayıcılara göre düşük olan bir numaralı puanlayıcı, en cömert puanlayıcıdır. Bu değerler dikkate alındığında, bireylerin çoğunluğu için analizde yer alan maddelerin kolay olduğu ifade edilebilir.

Öğrenci Yetenek ve Uygunluk İstatistikleri

ÇDKRÖM uygulaması sonucunda elde edilen öğrenci yetenek ve uygunluk istatistikleri incelendiğinde, analizde yer alan 350 bireyin logit değerleri 2.59 ile -0.57 arasında değiştiği veri kalibrasyon haritasında görülmektedir. Öğrencilere ait logit değerlerin ortalaması 0.64, standart sapması 0.50'dir. Bireylerin logit değerlerine ait standart hata (Root Mean Square Error) 0.14'tür. Bu değer aşırı uçlarda yer alan değerler hariç bütün veriler için ölçme hatasını gösterir (Baştürk, 2010). Elde edilen bu değer doğrultusunda standart hatanın oldukça düşük olduğu ifade edilebilir. Paralel olarak, düzeltilmiş standart sapma değeri de 0.48'dir. Bu değer, kritik değer 1.0'ın altındadır. Öğrencilerin değişkenlik kaynaklarına göre logit ölçek üzerinde birbirlerinden ne ölçüde ayrıldıklarını gösteren ayırma indeksi (Separation Index-G) (Nakamura, 2002; Turner, 2003) değeri, 3.48 bulunmuştur. Bu

değer doğrultusunda öğrencilerin 3 farklı yetenek seviyesine ayrıldığı belirtilebilir (Engelhard ve Myford, 2003). Ayırma indeksi güvenirliği ise 0.92'dir. Bu değerin KR-20 ve Cronbach Alfa değerine benzer biçimde hesaplandığı ve yorumlandığı söylenebilir (Iramaneerat ve diğerleri, 2009). Oldukça yüksek olan bu değer; iç tutarlılık anlamında testin güvenirliğinin yüksek olduğunun (Bond ve Fox, 2001) ve aynı zamanda testte merkezi eğilim hatası (central tendency error) olmadığının (Myford ve Wolfe, 2004) göstergesi olarak kabul edilebilir. Değişkenliğin manidar olup olmadığının belirlenebilmesi için ayırma indeksi ve güvenirliği test edildiğinde, sabit etki hipotezi ki-kare testiyle (X^2=3377.7, sd=349, p=0.00) reddedilmiştir. Bu sonuç doğrultusunda elde edilen değişkenliğin, istatistiksel olarak manidar olduğu ifade edilebilir.

Model kesinliğinin göstergesi olan uygunluk içi ($\overline{X}$ =1, s.s.=0.4) ve uygunluk dışı ($\overline{X}$ =1, s.s.=0.45) istatistikleri incelendiğinde, öğrencilerden elde edilen verilerin modele uyum sağladığı sonucuna ulaşılabilir (Linacre, 2012). Bu iki değer öğrencilerin tamamı (350) için incelendiğinde %16,8'i (59) uygunluk içi kareler ortalamaları ve %18,5'i (65) uygunluk dışı kareler ortalamaları 0.8-1.2 sınır değerlerinin (Linacre, 1994) dışındadır. Aykırı değerlere karşı daha hassas olan uygunluk dışı kareler ortalamaları (Linacre, 2007) dikkate alındığında öğrencilerin %10,2'si (36) aşırı uyum gösterirken, %8,3'ü (29) uyum göstermemektedir. Elde edilen tüm bu değerler göz önünde bulundurulduğunda analizde kullanılan verilerin ÇDKRÖM çerçevesinde uygun olarak kabul edileceği (Linacre ve diğerleri, 1990) ve ölçmenin değişmezliğine ilişkin kanıt sunduğu (Myford ve Wolfe, 2004) ifade edilebilir.

Puanlayıcı Katılık/Cömertlik ve Uygunluk İstatistikleri

Puanlayıcılara ait ÇDKRÖM analizi sonuçları Çizelge 5'te verilmiştir. Çizelge 5'te yer alan bulgular doğrultusunda yorumlamalar yapılmıştır.

Çizelge 5. ÇDKRÖM Analizi Puanlayıcı Ölçme 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			

Çizelge 5'te 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 üstünde bir değer bulunmuştur. Puanlayıcılar dikkate alındığında istenmeyen varyansın tanımlanmasındaki alternatif bir yol olan (Sudweeks ve diğerleri, 2004) bu değer, puanlayıcılar arası puan atama noktasında farklılıkların olduğunun göstergesidir (Nakamura, 2002). Başka bir ifadeyle puanlayıcıların katılık ve cömertlik düzeylerine göre farklılaştığını ve puanlayıcıların atadıkları puanlarda cömertlik/katılık hatasının yer aldığını (Engelhard ve Myford, 2003; Iramaneerat ve diğerleri, 2009) yansıtmaktadır. İstenen değer ise puanlayıcılar arası uyum sağlandığında ortaya çıkan, 0.00 ve 0.00'a yakın değerlerdir (Myford ve Wolfe, 2004; Nakamura 2000). Literatürde, her hangi bir üst sınır değerine rastlanmamakla birlikte, 0.00 değerinin oldukça üzerinde puanlayıcı ayırma indeksine sahip farklı çalışmalar yer almaktadır (Engelhard ve Myford, 2003; Nakamura, 2000, 2002). Puanlayıcı ayırma indeksi güvenirliği 0.98'dir. Bu, puanlayıcılar arası istenmeyen varyansın göstergesi olarak yorumlanabilir (Engelhard

2002). Myford ve Wolfe (2004)a göre bu indeks, puanlayıcı katılık düzeyindeki puanlayıcılar arası istenmeyen değişkenliği (variation) yansıtmaktadır. Engelhard (2002), puanlayıcılardaki değişkenliğe ilişkin bilgi sağladığını ve tanımlayıcı bir indeks olduğunu belirtmiştir. McNamara (1996) ise puanlayıcılar arası uyumun ya da güvenirliğin derecesini belirtmediğini, puanlayıcıların katılık düzeyleri arasındaki farklılığın derecesini yansıttığını ortaya koymuştur. Her iki değer göz önünde bulundurulduğunda, puanlayıcıların birbirleri yerine geçmeleri durumunun sakıncalı olabileceği (Linacre ve diğerleri, 1990) 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. Bu bağlamda 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 (Sudweeks ve diğerleri, 2004) 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). Bu bağlamda puanlayıcıların, puan atama noktasında birbirleriyle ve kendi içlerinde tutarlılık gösterdikleri ifade edilebilir.

Madde Güçlük ve Uygunluk İstatistikleri

Araştırma kapsamında kullanılan maddelere ilişkin ÇDKRÖM analizi sonuçları, Çizelge 6'da verilmiştir. Çizelge 6'da yer alan bulgular doğrultusunda yorumlamalar yapılmıştır.

Çizelge 6. ÇDKRÖM Analizi Madde Ölçme Sonuçları

			Madde Güçlüğü		Uygunluk İçi		Uygunluk Dışı	
Madde No	Madde Ort.	Madde Toplam r	Logit Ölçüsü	S.H.	Kareler Ort.	Z Std.	Kareler Ort.	Z Std.
2	2.3	2.24	.56	.02	0.8	-7.0	0.7	-7.0
3	2.6	2.53	.41	.02	0.9	-4.0	0.9	-2.0
6	2.8	2.84	.26	.02	0.9	-5.0	0.8	-5.0
4	3.2	3.31	.04	.02	0.8	-6.0	0.8	-6.0
7	3.4	3.48	-.04	.02	1.5	9.0	1.6	9.0
8	3.4	3.51	-.06	.02	1.1	3.0	1.1	1.0
5	4.0	4.16	-.44	.02	1.1	2.0	1.0	0.0
1	4.3	4.49	-.74	.02	0.9	-1.0	0.8	-2.0
Ortalama	3.3	3.32	.00	.02	1.0	-1.5	1.0	-2.0
SS	0.6	0.72	.40	.00	0.3	5.7	0.3	5.2
RMSE (Model) = .02		SS = .40		Ayırma İndeksi = 20.71			Güvenirlik = 1.00	
Tamamı Aynı Ki-Kare = 2961.1				Sd = 7		p = .00		

Çizelge 6'da, sekiz maddeye ait ÇDKRÖM analizi sonucunda elde edilen logit değerler, uyum içi ile uyum dışı istatistikleri ve diğer istatistikler yer almaktadır. Çizelge 6'da görüldüğü üzere, maddelerin logit değerleri 0.56 ile -0.74 arasında değişmektedir ($\overline{X}$ =0.00, s.s.=0.4). Bu değerler doğrultusunda en zor madde ikinci madde olurken, en basit madde ise birinci maddedir. Öğrencilere ait logit değerlerin ortalaması (0.64) dikkate alındığında, bireylerin çoğunluğunun analizde yer alan maddeleri doğru yanıtladığı söylenebilir. Maddelere ait ayırma indeksi 20.71 olarak bulunmuştur. Maddelere ait ayırma indeksinin, başlangıç değeri olarak önerilen 2.00'nin oldukça üzerinde bir değer alması, maddelerin işlevsel olarak öğrenciler üzerinde dağıldığı ve maddelerin farklı yetenek seviyelerine sahip öğrencilere ilişkin ölçmeler yapabileceğini ifade etmektedir (Nakamura, 2002). Ayırma güvenirliği ise 1.00'dir. Elde edilen bu değer merkezi eğilim (central

tendecy) sorununun olmadığının (Myford ve Wolfe, 2004) ve madde güçlük indeksinin öğrenciler için kesin bir biçimde kestirildiğinin (Engelhard ve Myford, 2003) göstergesi olarak yorumlanabilir. Sabit etki (fixed effect) hipotezi, ki-kare testiyle (X^2=2961.1, sd=7, p=0.00) reddedilmiştir. Bu bağlamda madde güçlük dereceleri arasındaki tüm farklılıkların, istatistiksel olarak manidar olduğu ifade edilebilir.

Maddelere ait uygunluk içi istatistikleri, 0.8 ile 1.5 aralığında değerler alırken ($\overline{X}$ =1.0, s.s.=0.3), uygunluk dışı istatistikleri, 0.7 ile 1.6 aralığında değerler almıştır ($\overline{X}$ =1.0, s.s.=0.3). Uygunluk içi ve uygunluk dışı istatistikleri birlikte incelenmekle birlikte uç değerlere daha duyarlı olması ve maddeler için elde edilen uygunluk içi istatistiklerini kapsaması bakımından uygunluk dışı istatistikleri, dikkate alınarak bulgular yorumlanmıştır. Uygunluk dışı istatistikleri kareler ortalamaları, iki madde (2. ve 4.) hariç diğer tüm maddeler için nitelik kontrol (quality-control) değerleri (0.8-1.2) aralığında yer almaktadır. Nitelik kontrol değerleri dışında bir değer (0.7) alan, ikinci maddenin diğer maddelerle karşılaştırılınca bağımsızlığa sahip olmadığı başka bir ifadeyle aşırı uyum sağladığı (overfitting) belirtilebilir (Engelhard ve Myford, 2003). Ayrıca bu değer, "modellenenden %30 daha az değişkenlik göstermektedir" biçiminde de yorumlanabilir (Linacre, 1989). Fakat bu değerin, nitelik kontrol değerine oldukça yakın olduğu göz önünde bulundurulmalı ve farklı kaynaklara (Iramaneerat ve diğerleri, 2008; Nakamura, 2000; Nakamura, 2002; Wright ve Linacre 1994) göre bu değerin nitelik kontrol değerleri aralığında yer aldığı da dikkate alınmalıdır. Diğer nitelik kontrol değerleri dışında bir değer (1.6) alan, dördüncü maddenin beklenenin oldukça üzerinde (modellenenden %60 daha fazla) değişkenlik gösterdiği (Linacre, 1994) ve diğer maddelerle uyum göstermediği (misfitting) (Iramaneerat ve diğerleri, 2008) ifade edilebilir.

Hiyerarşik Puanlayıcı Modeli Bulguları

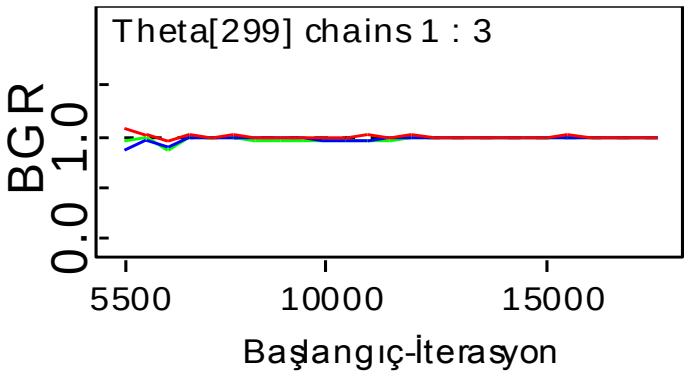
Araştırmanın verileri için ikinci alt problem doğrultusunda hiyerarşik puanlayıcı modeli ile elde edilen öğrenci yetenek istatistikleri, puanlayıcı katılık ve değişkenlik istatistikleri, madde güçlük istatistikleri ve bunlara ait BGR grafikleri ile zaman serileri diyagramları verilmiş ve açıklanmıştır.

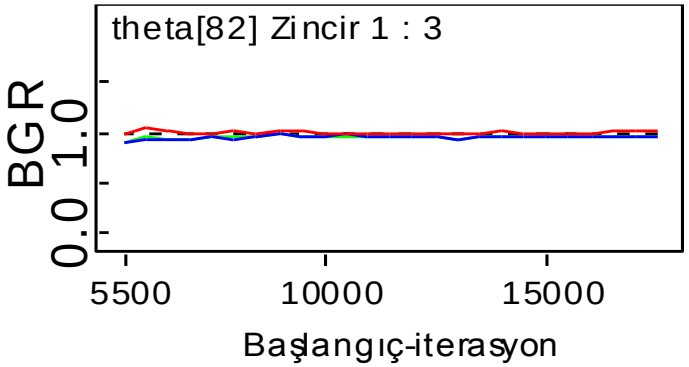
Öğrenci Yetenek İstatistikleri

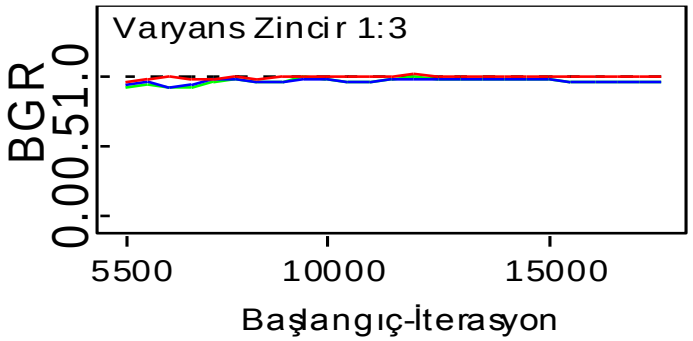
Analizde yer alan parametreler rapor edilmeden önce ilgili parametreler için analizin gerçekleştirildiği üç zincirin yakınsama durumu kontrol edilmelidir. Sonsal dağılımdan geçerli örneklem elde edildiğine dair kanıtlara ulaşıldıktan sonra gözlemlere dayanan parametrelere ilişkin çıkarımlarda bulunulabilir (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) tanısalından (diagnostic) faydalanılmıştır (Kéry, 2010).

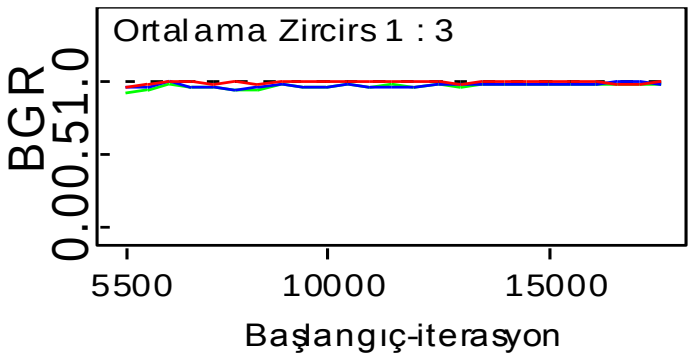
Araştırmaya katılan tüm öğrencilerin yetenek düzeylerine ilişkin elde edilen grafikler, diyagramlar ve değerler kaplayacağı alandan kaynaklanan okuma zorluğu nedeniyle rapor edilmemiş, sadece sonsal dağılıma göre; en yüksek ve en düşük yetenek düzeyine sahip öğrencilere ilişkin grafikler, diyagramlar ve değerler rapor edilmiştir. Aşağıda en yüksek ve en düşük yetenek düzeyine sahip öğrenciler ile öğrenci yetenek dağılımının ortalama ve varyansına ilişkin BGR grafiklerine yer verilmiştir.

Grafik 1. En Yüksek ve En Düşük Yetenek Düzeyine Sahip Öğrenciler ile Öğrenci Yetenek Dağılımının Ortalama ve Varyans Parametrelerine İlişkin BGR Grafikleri





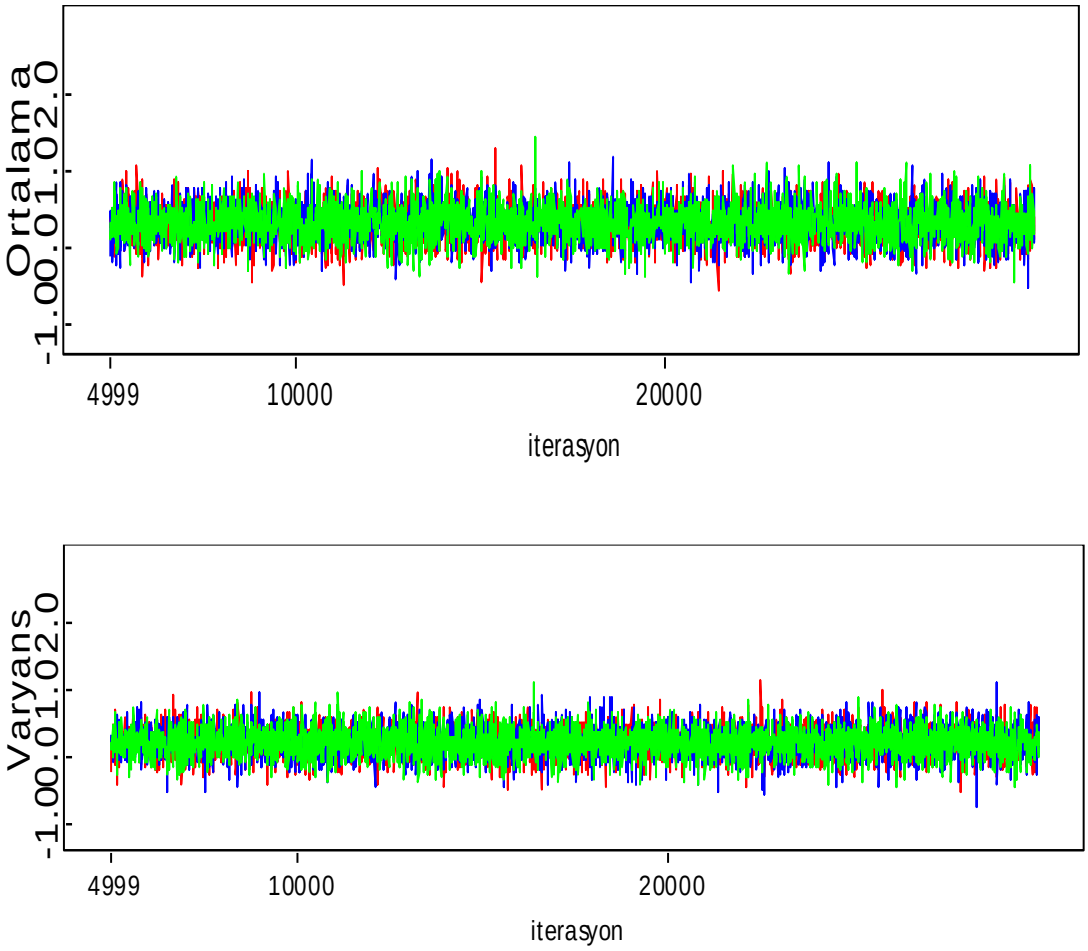


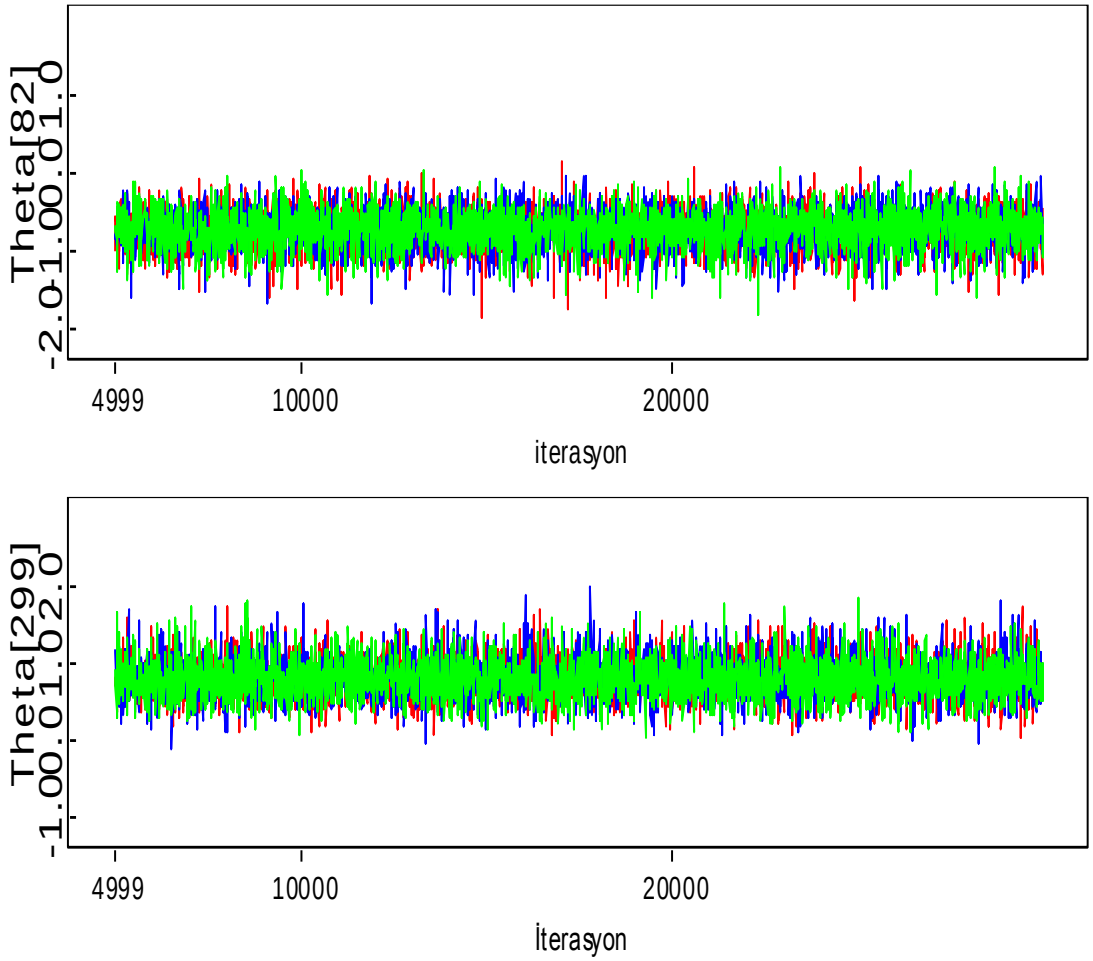


Yukarıda verilen grafiklerle birlikte burada rapor edilmeyen her bir öğrencinin yetenek düzeyine ilişkin BGR grafikleri incelenmiş ve her birinin, bir ya da bir değerine oldukça yakın değerler aldığı görülmüştür. Bu bulgular doğrultusunda her bir öğrencinin yetenek düzeyi ile öğrenci yetenek dağılımının ortalama ve varyans parametreleri için üç zincirin yakınsadığı ve sonuçların rapor edilebilir niteliğine sahip olduğu ifade edilebilir.

Markov zincirlerinin yakınsayıp yakınsamadığının bir diğer göstergesi ise zaman serileri diyagramıdır. Bu diyagramlar, görsel olarak incelenmelidir (Spiegelhalter ve diğerleri, 2003; Lunn ve diğerleri, 2009). Elde edilen zaman serileri diyagramları aşağıda verilmiştir.

Diyagram 1. En Yüksek ve En Düşük Yetenek Düzeyine Sahip Öğrenciler ile Öğrenci Yetenek Dağılımının Ortalama ve Varyans Parametrelerine İlişkin Zaman Serileri Diyagramları





Elde edilen tüm zaman serileri diyagramları incelendiğinde; modelde yer alan birey parametrelerinin sonsal dağılımı yakınsadığı, örneklenen değerlerin ortak bir ortalama değeri etrafında seçkisiz olarak hareketli olması ve tüm Markov zincirlerinin binişik olmasına dayanılarak belirtilebilir. Bu bağlamda, parametrelerin rapor edilebilir niteliğe sahip olduğu bulgusu desteklenmektedir (Zuur ve diğerleri, 2012; Lee ve Wagenmakers, 2014).

Bayes nokta kestirimi olarak, genellikle sonsal ortalama ya da sonsal medyan ile birlikte sonsal standart sapma rapor edilmektedir. Sonsal dağılımın yüzde 2.5 ile 97.5 arasındaki ranjı ise Bayes güven aralığıdır. Bu aralık, "credible interval" olarak adlandırılmaktadır (Congdon, 2001). Bayes kestirimleri için farklı tür güven aralığı yöntemleri [Bayesian credible interval, frequentist confidence interval ve highest posterior density interval (HPDI)] kullanılmakla birlikte, daha kolay ve OpenBUGS, WinBUGS gibi programlar aracılığıyla da hesaplanabilmesi açısından (Kéry, 2010) Bayes güven aralığı değerleri bulgulanmıştır. Bir diğer incelenmesi tavsiye edilen

istatistik ise Monte Carlo (MC) hatasıdır. MC hatası, Markov zinciriyle yapılan kestirimdeki değişkenliğin niceliğini belirterek, yakınsamanın ne kadar iyi olduğuna ilişkin bilgi vermektedir (Lunn ve diğerleri, 2009; Lunn, Jackson, Best, Thomas ve Spiegelhalter, 2012). MC hatası; sonsal dağılımın gerçek varyansına, sonsal örneklem büyüklüğüne (MCMC iterasyon sayısı), MCMC örneklemindeki otokorelasyona dayanmaktadır. Bu değer mümkün olduğunca küçük (0.05'ten küçük) olmalıdır. Model parametreleri için MC hatasının, standart sapmanın %5'inden küçük olması beklenmektedir (Ntzoufras, 2009; Spiegelhalter ve diğerleri, 2003).

Çizelge 7'de, en yüksek ve en düşük yetenek düzeyine sahip öğrenciler ile yetenek dağılımının, ortalama ve varyans değerlerine ait ortalama, standart sapma, MC hatası, medyan ve %95 güven aralığı değerleri verilmiştir. Literatürde yer alan çalışmalar incelendiğinde genellikle medyan ve Bayes güven aralığı değerlerinin yorumlandığı görülmüştür (Casabianca ve Junker, 2013; Patz ve Junker, 1999a, 1999b; Patz ve diğerleri, 2002). Bu nedenle araştırma kapsamında, bulguların yorumlanmasında medyan değeri dikkate alınmıştır.

Çizelge 7. HPM Öğrenci Yetenek Değerleri Ortalaması, Varyansı ve Öğrenci Yetenekleri MCMC Kestirimi Sonsal Değerleri

Parametreler	Ortalama	Standart Sapma	MC Hatası	Medyan	Güven Aralığı
					Alt sınır - Üst Sınır
Ortalama (μ)	0,313	0.2325	0.00250	0,309	-0,139 – 0.787
Varyans (σ^2)	0.213	0.1243	0.00235	0.212	-0.224 – 0.657
θ (82)	-0.728	0.2518	0.00299	-0.719	-1.254 – -0.259
θ (299)	0.796	0.2644	0.00310	0.784	0.315 – 1.366

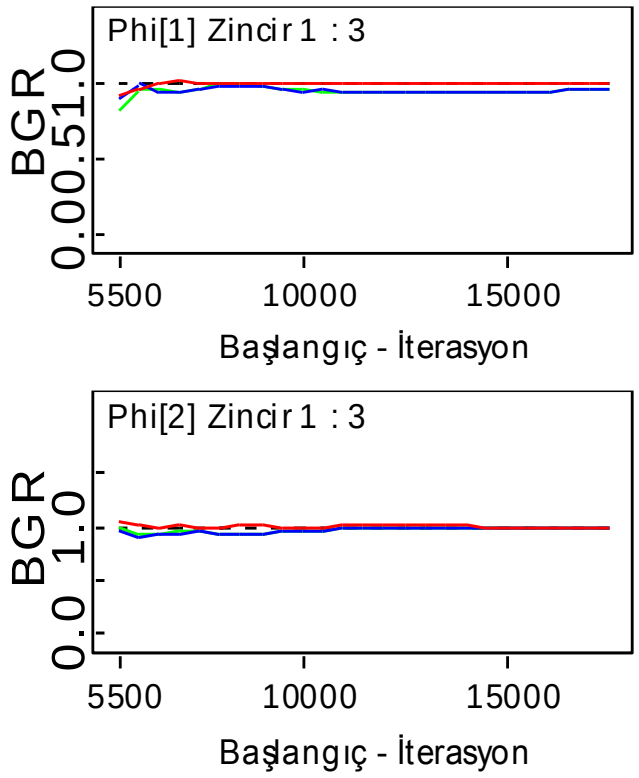
Çizelge 7'de yer alan parametrelere ait MC hataları incelendiğinde, tamamının .05'ten küçük olduğu görülmektedir. Ayrıca bu değerlerin hepsi ilgili parametrenin standart hatasının %5'inden de küçüktür. Elde edilen bu sonuç; yukarıda verilen sonsal dağılımın yakınsamasına ilişkin bulguları, destekler niteliktedir. Parametrelerin medyan değerleri doğrultusunda; yetenek düzeyi en yüksek olan öğrencinin, 0.784 değeri ile 299 numaralı öğrenci olduğu görülmektedir. En düşük yetenek düzeyine sahip olan öğrenci ise -0.719 değeri ile

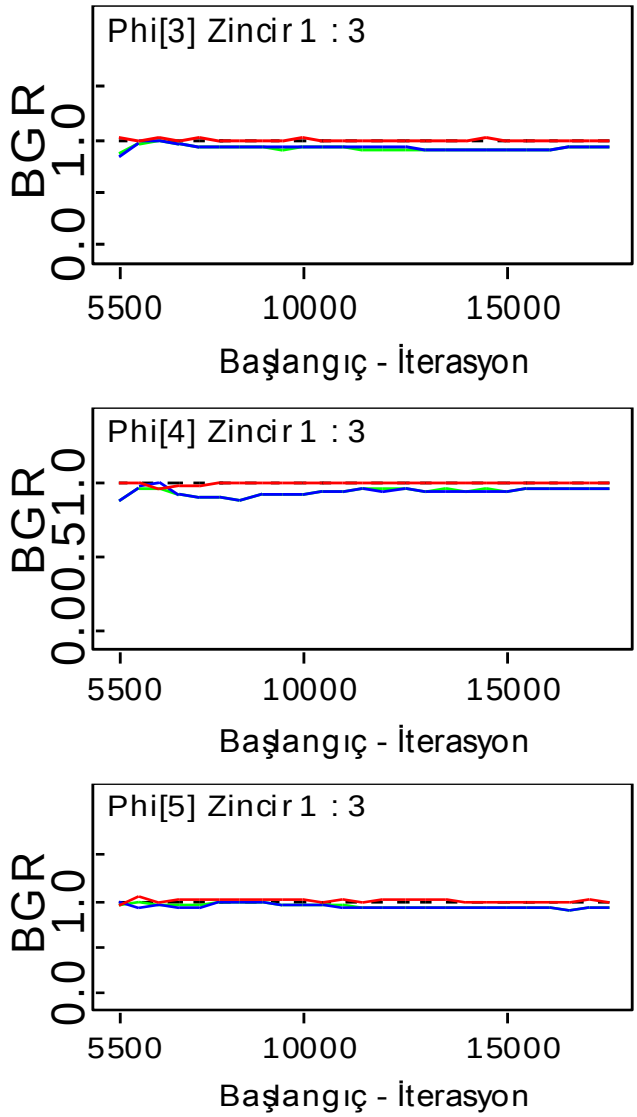
82 numaralı öğrencidir. Öğrenci yetenek dağılımının ortalaması, 0,309 iken varyansı 0.212 olarak elde edilmiştir.

Puanlayıcı Katılık/Cömertlik ve Değişkenlik İstatistikleri

Literatürde kullanıldığı modele göre farklı isimlerle anılan, Patz ve diğerleri (2002) tarafından katılık (severity) ya da yanlılık (bias) olarak da adlandırılan Phi (ϕ_r) parametresi; bu alt bölümde kullanım yaygınlığı nedeniyle, puanlayıcı katılık parametresi olarak ifade edilmiştir. Puanlayıcı parametreleri verilmeden önce BGR grafikleri ve zaman serileri diyagramlarına aşağıda yer verilmiştir.

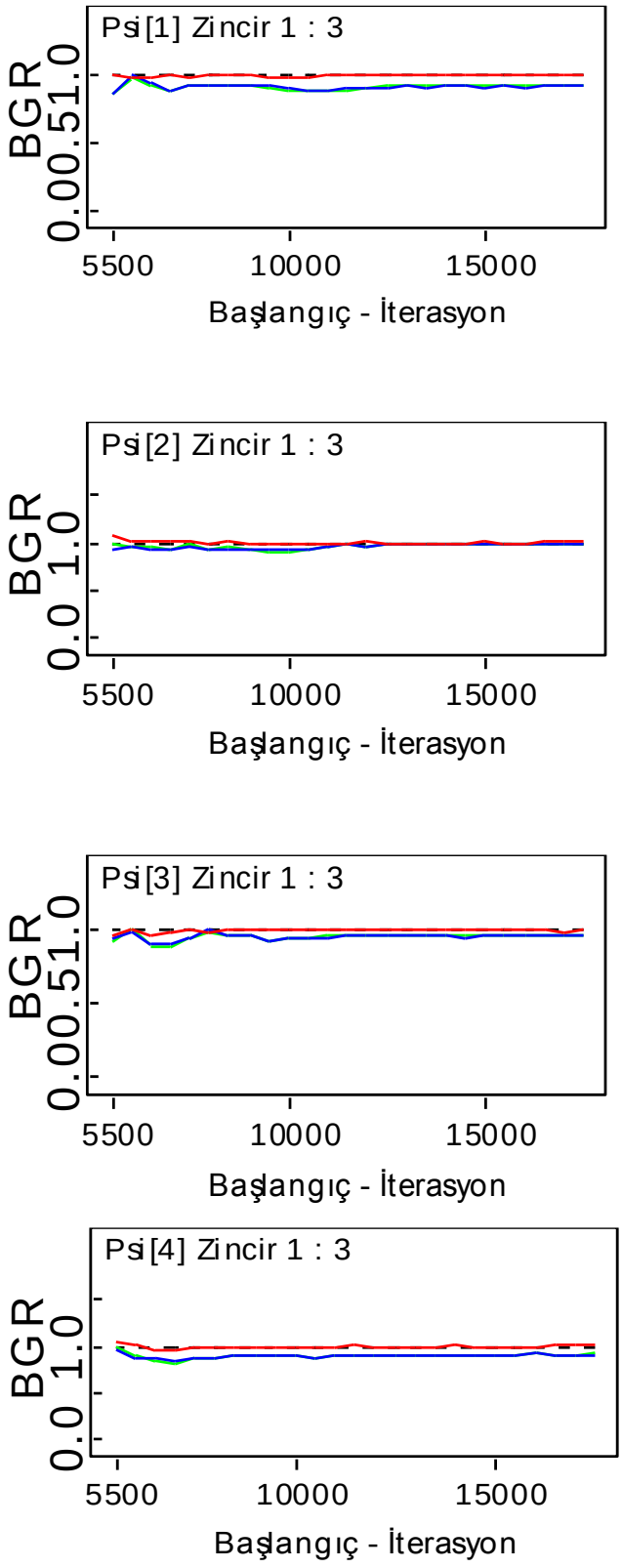
Grafik 2. Puanlayıcı Katılık Parametrelerine İlişkin BGR Grafikleri

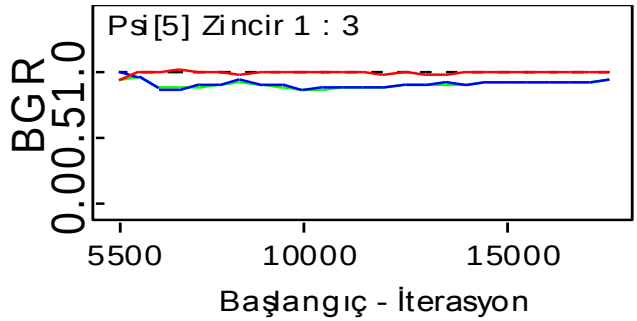




Puanlayıcı katılık parametreleri, BGR grafikleri incelendiğinde, özellikle ikinci puanlayıcının neredeyse bir, diğer puanlayıcıların da bire çok yakın değerler aldığı görülmektedir. Sadece dördüncü puanlayıcıya ait katılık parametresi, BGR grafiğinin bir değerinden çok az da olsa uzaklaştığı gözlenmiş 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.

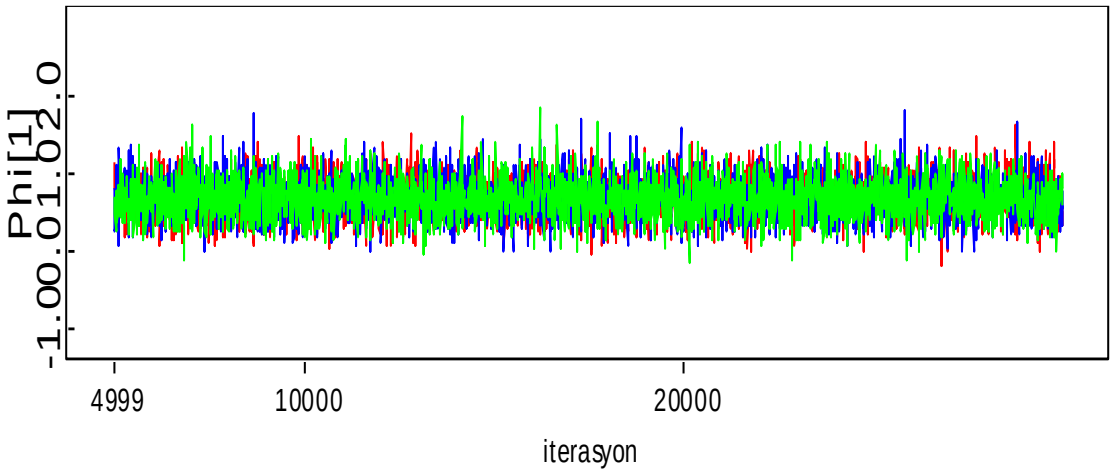
Grafik 3. Puanlayıcı Değişkenlik Parametrelerine İlişkin BGR Grafikleri

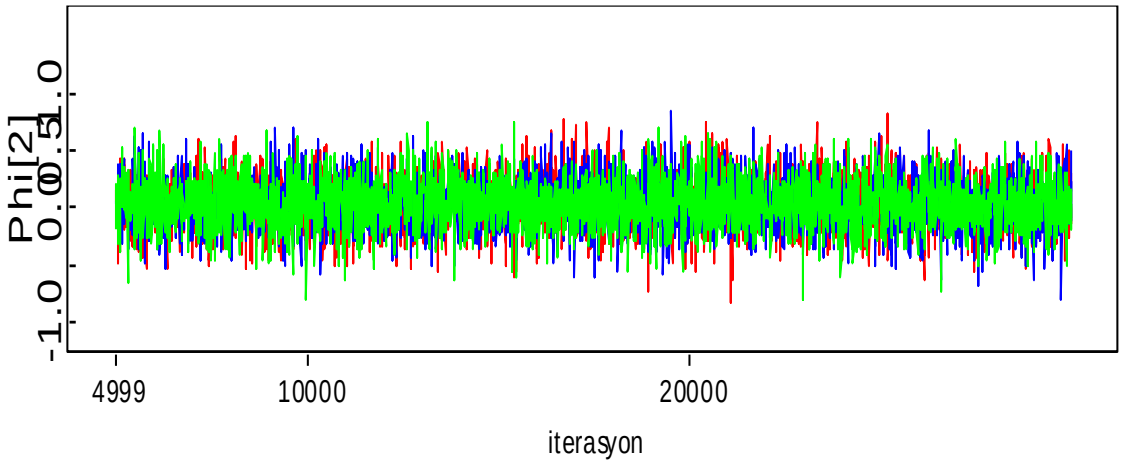


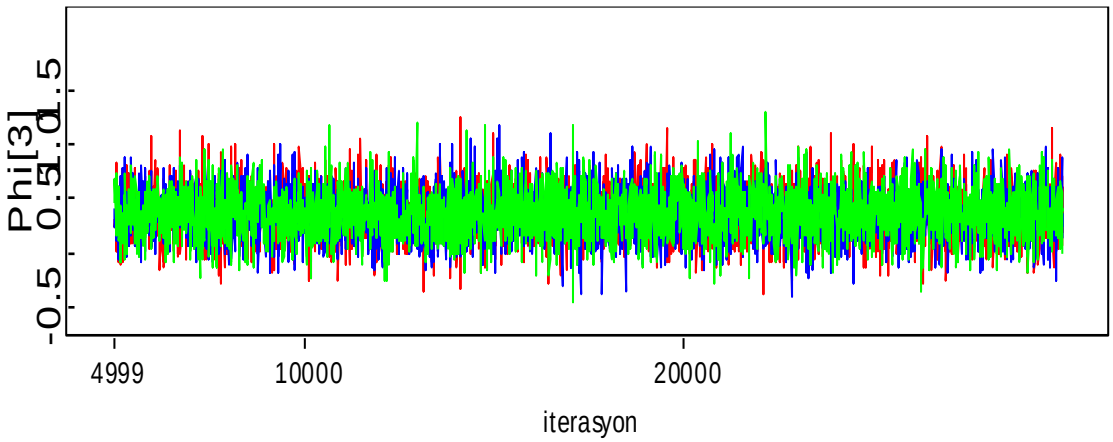


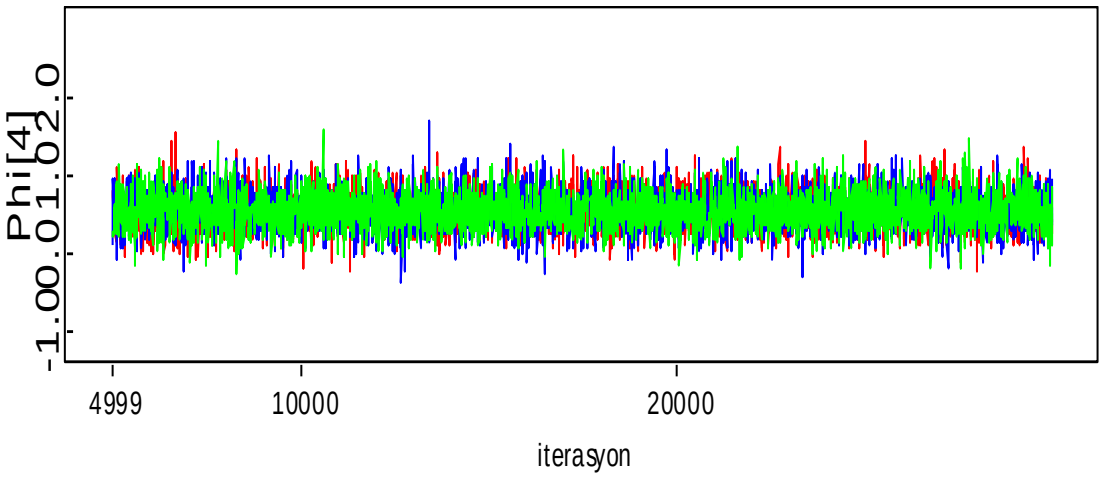
Puanlayıcı katılık parametrelerine benzer olarak, puanlayıcı değişkenlik parametreleri, BGR grafikleri için ikinci puanlayıcının neredeyse bir, diğer puanlayıcıların da bire oldukça yakın değerler aldığı görülmektedir. Fakat puanlayıcı bir ve puanlayıcı beşin diğer puanlayıcılara göre bir değerinden 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.

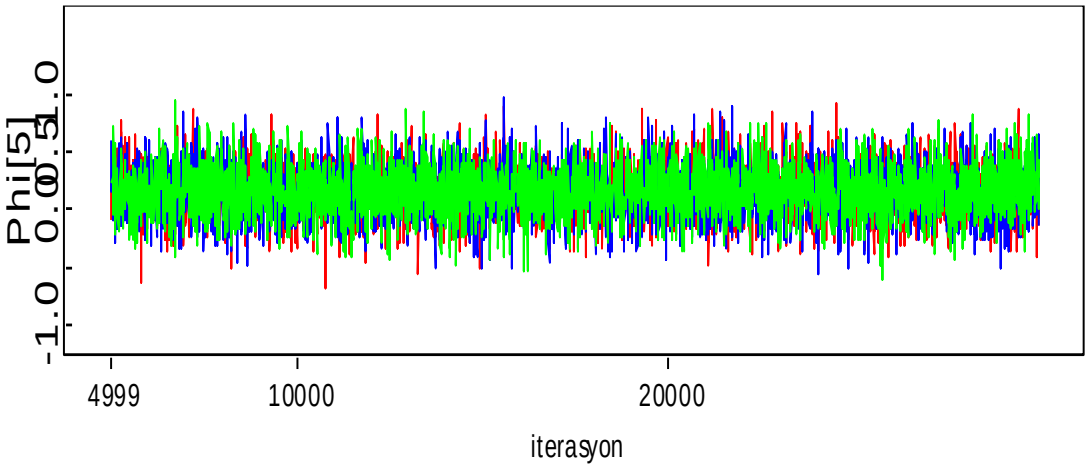
Diyagram 2. Puanlayıcı Katılık Parametrelerine İlişkin Zaman Serileri Diyagramları



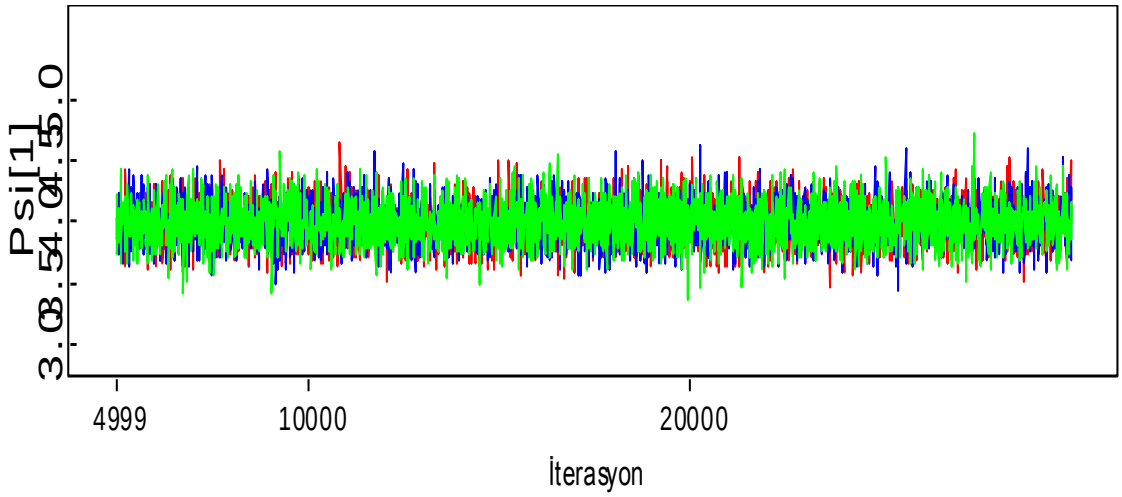

Phi[2]
-1.0
0.0
0.5
1.0
4999
10000
20000
iterasyon

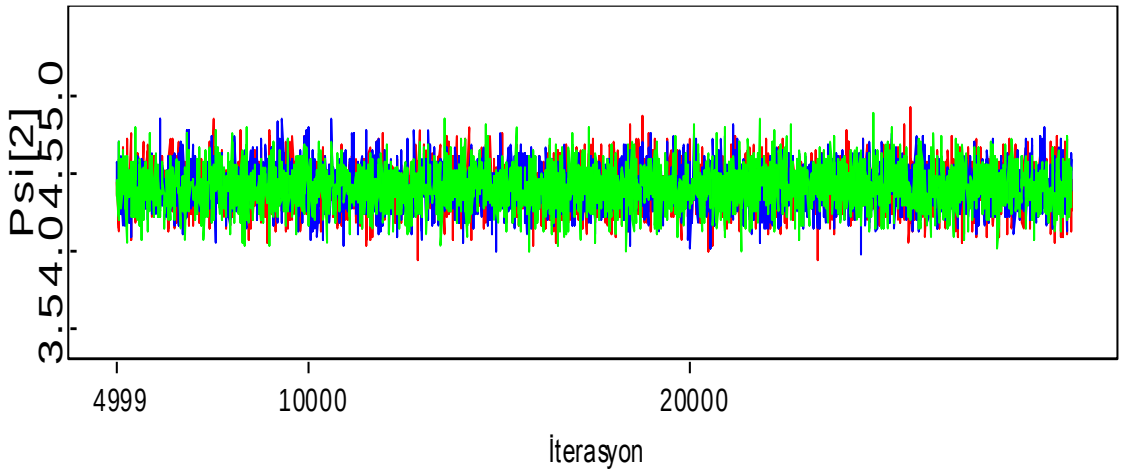

Phi[3]
-0.5
0.5
1.0
1.5
4999
10000
20000
iterasyon

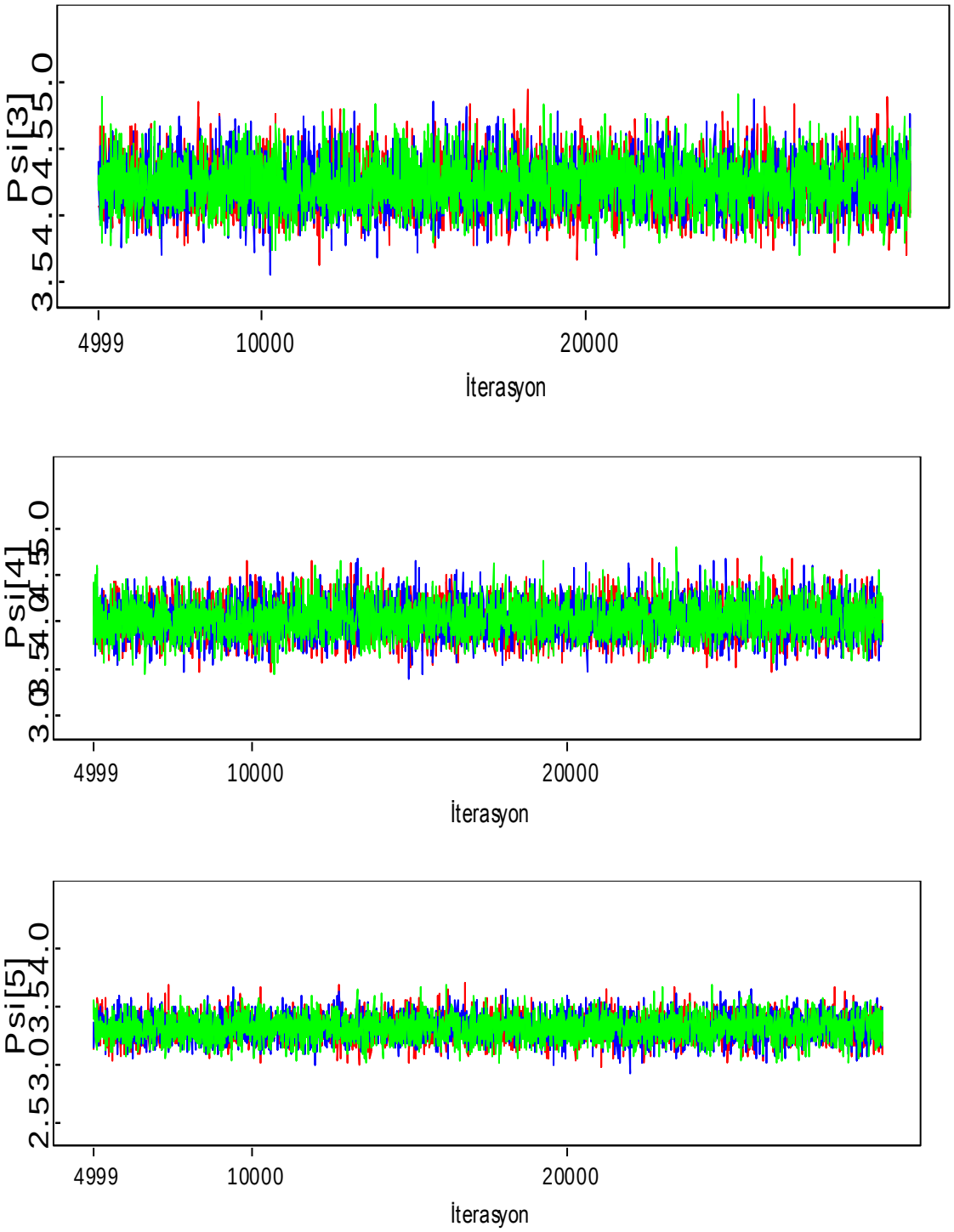

Phi[4]
-1.0
0.0
1.0
2.0
4999
10000
20000
iterasyon



Diyagram 3. Puanlayıcı Değişkenlik Parametrelerine İlişkin Zaman Serileri Diyagramları







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 Çizelge 8'de, puanlayıcıların katılık ve değişmezlik değerlerine ait ortalama, standart sapma, MC hatası, medyan ve güven aralığı değerleri verilmiştir.

Çizelge 8. HPM Puanlayıcı Katılığı ve Değişkenliği 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	0.6621 – 0.7124
Psi (ψ_1)	0.5008	0.16780	0.001789	0.5208	0.4801 –0.5218
Phi (ϕ_2)	0.0346	0.14505	0.001580	0.0347	0.0049 – 0.0628
Psi (ψ_2)	0.4735	0.13510	0.001659	0.4766	0.4624 – 0.4911
Phi (ϕ_3)	0.3558	0.12500	0.001433	0.3559	0.3313 – 0.3799
Psi (ψ_3)	0.4865	0.18051	0.002178	0.4869	0.4659 – 0.5071
Phi (ϕ_4)	0.5476	0.11760	0.002310	0.5476	0.5245 – 0.5707
Psi (ψ_4)	0.4988	0.18590	0.002142	0.4991	0.4771 – 0.5216
Phi (ϕ_5)	0.1400	0.14470	0.001530	0.1401	0.1111 – 0.1686
Psi (ψ_5)	0.5493	0.10311	0.001226	0.5493	0.5334 – 0.5671

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_r<ξ), başka bir ifadeyle ideal kategori göz önünde bulundurulduğunda daha katı bir puanlayıcı davranışı sergilediği belirtilebilir. Bu değer, 0.5'ten büyük olduğunda (ϕ_r>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, sıfıra 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 ϕ_r ve ψ_r 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.

Çizelge 8'de yer alan değerler incelendiğinde; ikinci, üçüncü ve beşinci puanlayıcıların, puanlayıcı katılık parametrelerinin, mutlak değerce 0.5'ten küçük ($|\phi_r|<0.5$) olduğu görülmektedir. Bu bağlamda, ilgili puanlayıcıların diğer kategorilerden ziyade ideal puan kategorisinde ya da bu kategoriye yakın bir kategoride puanlama yaptığı belirtilebilir. Ayrıca bu üç puanlayıcının, maddelere verilen öğrenci yanıtlarına puan atama noktasında birbirleriyle uyum içinde oldukları da ifade edilebilir. Bu puanlayıcılar arasında ikinci 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 (birinci ve dördüncü) 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 ifade edilebilir. Fakat dördüncü puanlayıcının birinci 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, birinci puanlayıcı olduğu ifade edilebilir. Puanlayıcılardan hiç birinin negatif yanlılık göstermediği ya da öğrencilere ideal puan değerinden daha düşük değerler verme eğiliminde olmadığı da saptanmıştır.

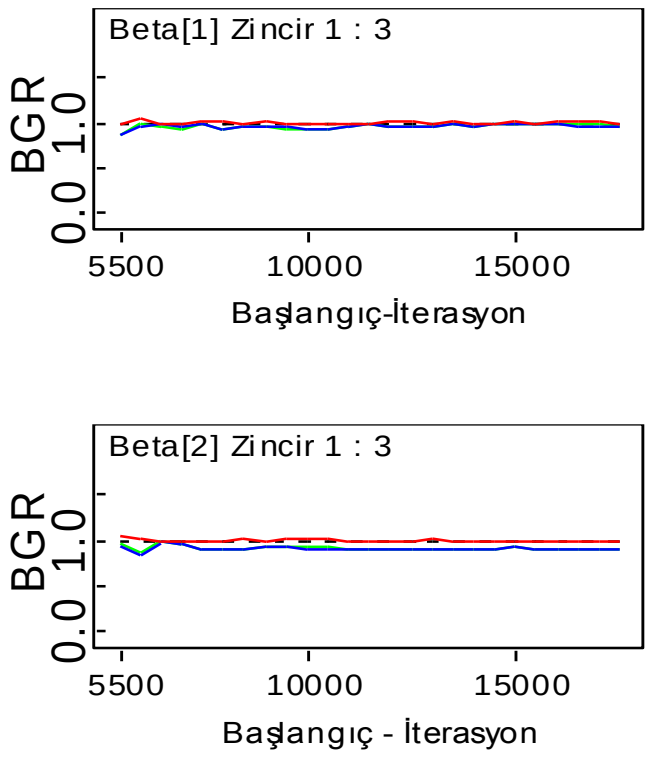
Puanlayıcı katılık parametrelerinin aksine, puanlayıcı değişkenlik parametrelerinin birbirine oldukça yakın değerler aldığı Çizelge 8'de görülmektedir. Puanlayıcı değişkenlik parametresi değeri, 0'a en yakın olan puanlayıcı 0.4766 değeri ile ikinci puanlayıcıdır. Başka bir ifadeyle, ikinci puanlayıcının en güvenilir

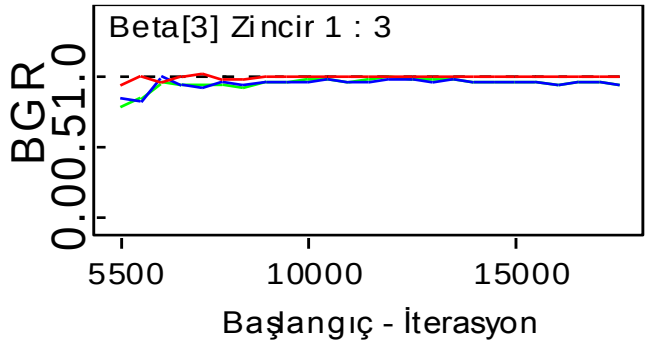
puanlayıcı olduğu belirtilebilir. Ayrıca ikinci puanlayıcının, Çizelge 8'de yer alan her iki parametre değeri incelendiğinde diğer puanlayıcılara nispeten hem daha güvenilir hem de daha az yanlı puanlar atadığı da ifade edilebilir. Puanlayıcı katılık parametresi dikkate alındığında ideal kategoriye oldukça yakın puan atayan beşinci puanlayıcı; değişkenlik parametresi dikkate alındığında en yüksek değere sahip ve dolayısıyla en az güvenilir olan puanlayıcıdır. Başka bir ifadeyle, beşinci puanlayıcının, aynı niteliğe sahip öğrenci yanıtlarına puan atama noktasında daha az tutarlı olduğu belirtilebilir. Elde edilen bu bulgunun, sadece katılık parametresi dikkate alınarak puanlayıcılara dair yargıya varmanın, zaman zaman yanıltıcı olabileceğinin göstergesi olması bakımından önemli olduğu düşünülmektedir.

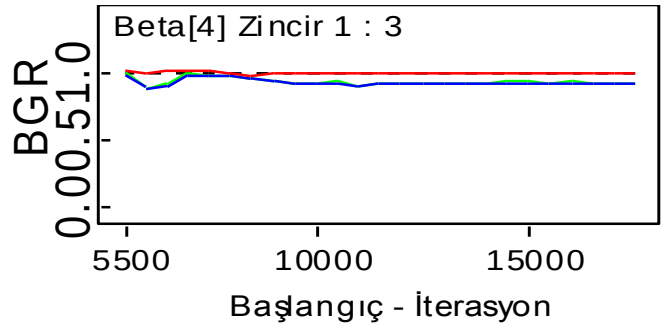
Madde Güçlük İstatistikleri

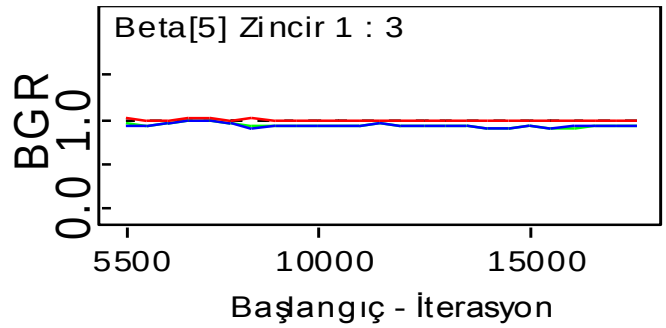
Madde güçlük istatistikleri verilmeden önce BGR grafikleri ve zaman serileri diyagramlarına aşağıda yer verilmiştir.

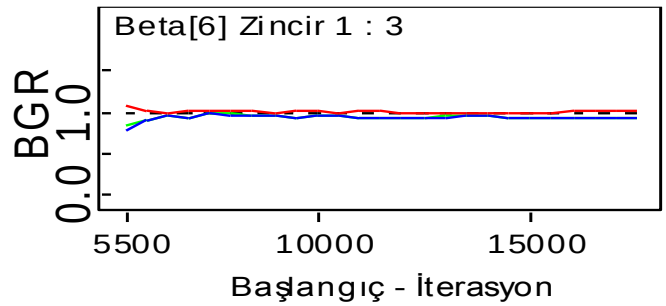
Grafik 4. Madde Güçlük Parametrelerine İlişkin BGR Grafikleri

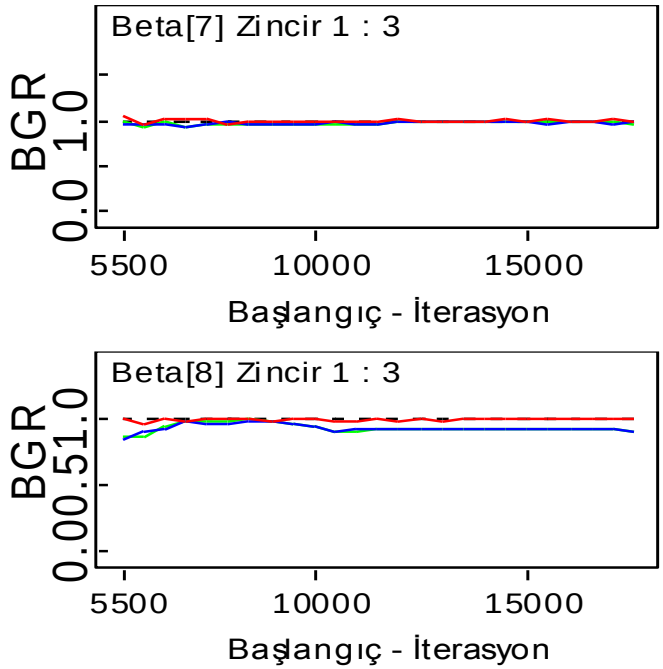



Beta[3] Zincir 1 : 3
BGR
0.0 0.5 1.0
5500
10000
15000
Başlangıç - İterasyon


Beta[4] Zincir 1 : 3
BGR
0.0 0.5 1.0
5500
10000
15000
Başlangıç - İterasyon

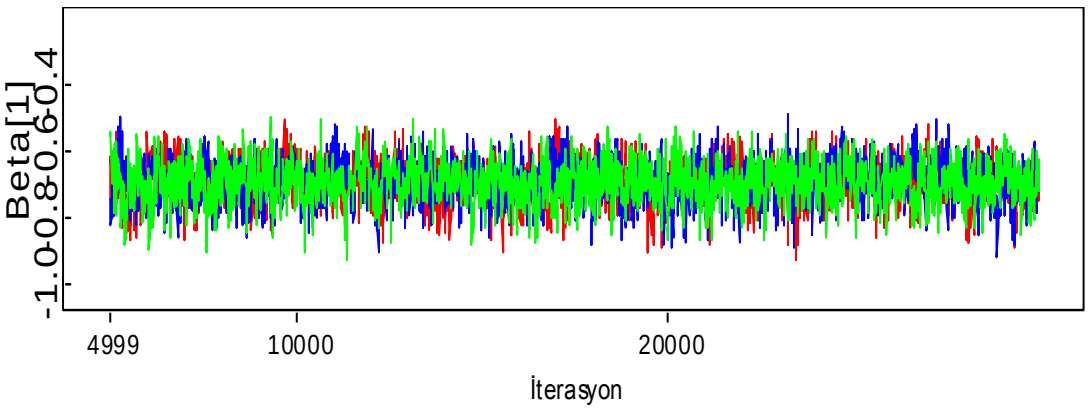

Beta[5] Zincir 1 : 3
BGR
0.0 1.0
5500
10000
15000
Başlangıç - İterasyon

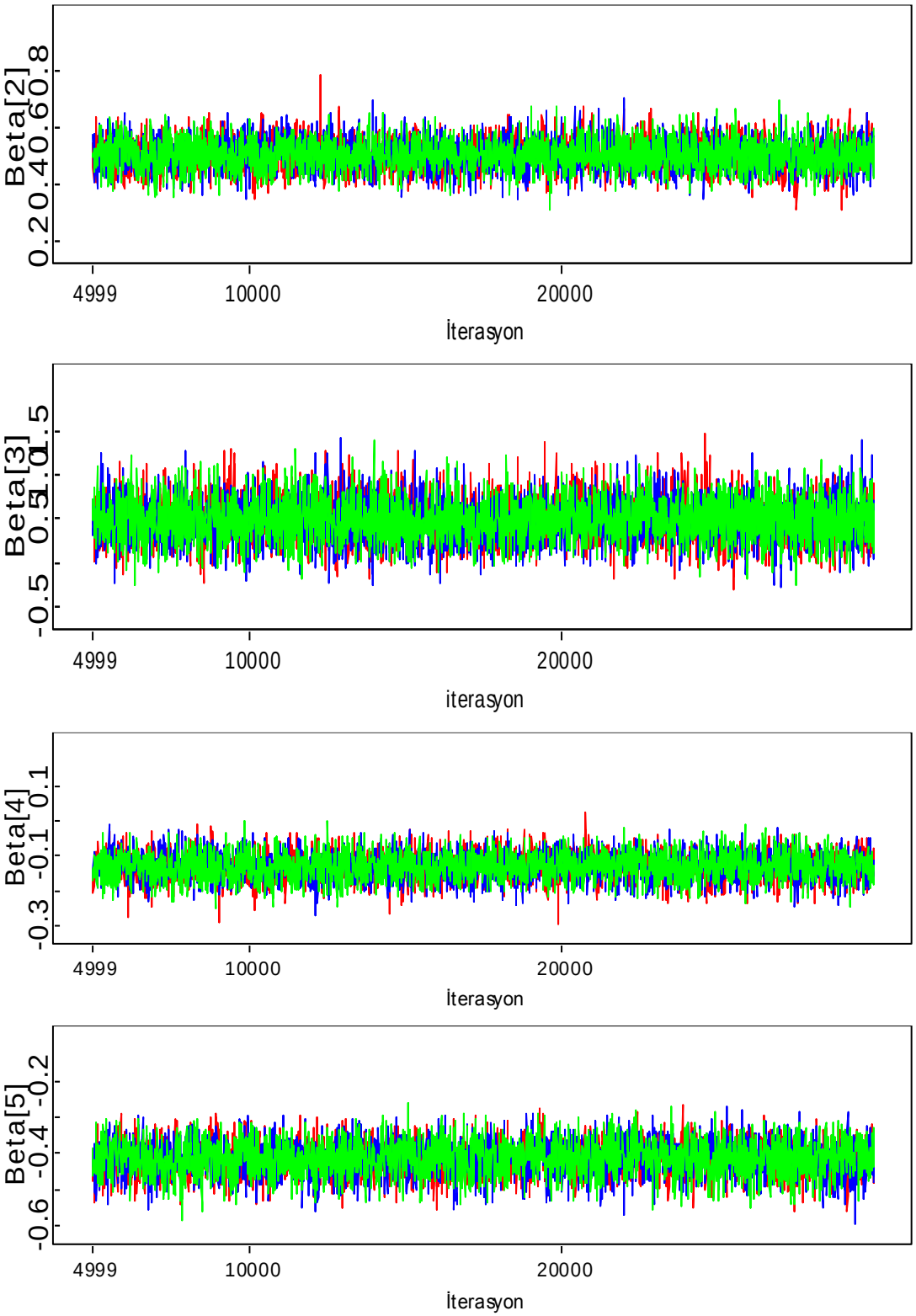

Beta[6] Zincir 1 : 3
BGR
0.0 1.0
5500
10000
15000
Başlangıç - İterasyon

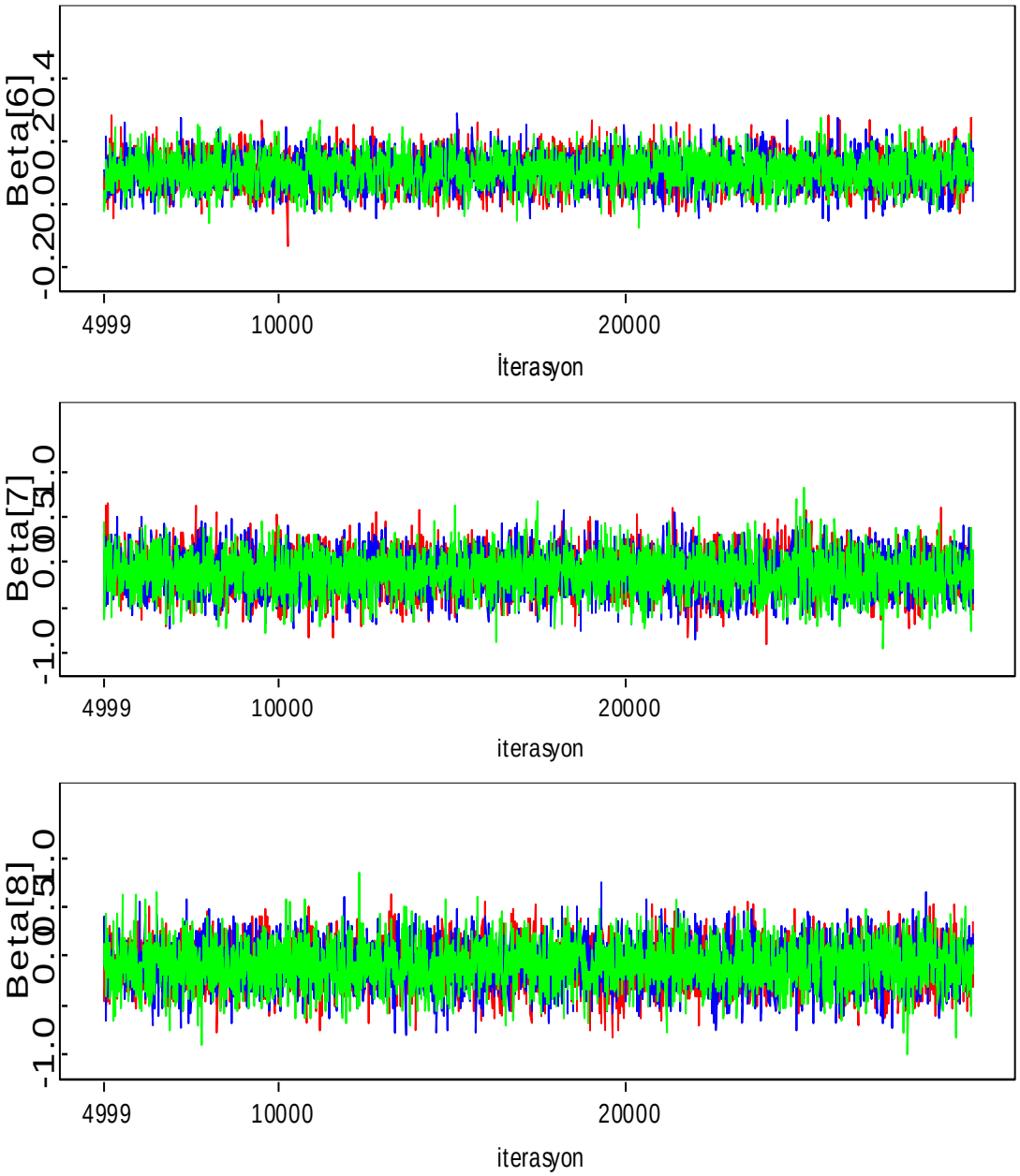


Madde güçlük parametrelerine ait BGR grafikleri incelendiğinde, dördüncü ve sekizinci madde hariç tüm maddelerin bire yakın, dördüncü ve sekizinci maddelerinse az da olsa birden uzaklaştığı görülmektedir. Bu nedenle iki maddeye ait BGR değerleri incelenmiş ve son iterasyon aralığı için bu değerler sırasıyla, 1.002 ve 1.001 olarak elde edilmiştir. Tüm BGR değerlerinin, istenilen aralıkta (1.0-1.1) yer alması doğrultusunda madde güçlük parametrelerinin raporlanabilir olduğu söylenebilir.Bu bulguya ilişkin daha fazla kanıt elde etmek amacıyla, madde güçlük parametrelerine ait zaman serileri diyagramları incelenmek üzere aşağıda verilmiştir.

Diyagram 4. Madde Güçlük Parametrelerine İlişkin Zaman Serileri Diyagramları




Beta[2]
0.2
0.4
0.6
0.8
4999
10000
20000
İterasyon
Beta[3]
-0.5
0.5
1.0
1.5
4999
10000
20000
iterasyon
Beta[4]
-0.3
-0.1
0.1
4999
10000
20000
İterasyon
Beta[5]
-0.6
-0.4
-0.2
4999
10000
20000
İterasyon



Zaman serileri diyagramlarının, tüm maddeler için istenen şekle sahip olduğu belirtilebilir. Tüm maddeler için bu sonuçlar dikkate alındığında 3 zincirin yakınsadığı ve ilgili parametrelerin BGR grafiklerinden elde edilen bulgularla paralel olarak rapor edilebilir niteliğe sahip olduğu ifade edilebilir.

Madde parametrelerinin yakınsak bir dağılımdan geldiği belirlendikten sonra Çizelge 9'da, madde güçlük değerlerine ait ortalama, standart sapma, MC hatası, medyan ve güven aralığı değerleri verilmiştir.

Çizelge 9. HPM Madde Güçlüğü MCMC Kestirimi Sonsal Değerleri

Madde No	Ortalama	Standart Sapma	MC Hatası	Medyan	Güven Aralığı Alt sınır - Üst Sınır
1 (β_1)	-0.6946	0.0631	0.00153	-0.6930	-0.825 – -0.576
2 (β_2)	0.5481	0.1443	0.00164	0.5359	0.438 – 0.996
3 (β_3)	0.5015	0.0522	0.00195	0.5005	0.401 – 0.604
4 (β_4)	-0.0609	0.0376	0.00153	-0.0611	-0.135 – -0.012
5 (β_5)	-0.4140	0.0456	0.00180	-0.4126	-0.507 – -0.327
6 (β_6)	0.1020	0.0505	0.00185	0.1012	0.006 – 0.202
7 (β_7)	-0.1027	0.0523	0.00109	-0.1017	-0.208 – -0.005
8 (β_8)	-0.0894	0.0436	0.00134	-0.0887	-0.176 – -0.005

Öncelikle Çizelge 9'da yer alan MC hataları incelenmiş; tüm maddelerin MC hatalarının 0.05'ten küçük olduğu görülmüştür. Benzer biçimde, bu değerlerin hepsi ilgili parametrenin, standart hatasının %5'inden de küçüktür. Elde edilen bu bulgu, yukarıda yer alan BGR grafikleri ve zaman serileri diyagramlarıyla paraleldir. Bu bağlamda sonsal dağılımın yakınsamasına ilişkin yeterli kanıtların elde edildiği ifade edilebilir. Madde güçlük düzeyleri, 0.5359 ile -0.693 değerleri arasında değişmektedir. En zor madde, 0.5359 değeriyle ikinci madde olurken, -0.693 değeriyle birinci madde ise en kolay maddedir. Özellikle Çizelge 10'da verilen, ham veriler (raw data) incelendiğinde maddeler arası güçlük farklılıkları görülmektedir. Madde iki ve madde üçün öğrenci yetenek ortalamasının (0.309) biraz üzerinde yer aldığı görülmektedir. Fakat genel olarak, öğrencilerin uygulanan maddelerin büyük bir kısmına doğru yanıt verdikleri söylenebilir.

Çizelge 10'da, birinci maddeye ait, Çizelge 11'de, ikinci maddeye ait ham veriler verilmiştir.

Çizelge 10. Madde 1 Ham Veriler

Puan	Puanlayıcı Kombinasyonu										
Kombinasyonu	1-2	1-3	1-4	1-5	2-3	2-4	2-5	3-4	3-5	4-5	Toplam
0-0	0	0	0	0	0	0	0	0	0	0	0
0-1	0	0	0	0	0	0	0	0	0	0	0
0-2	0	0	0	0	0	0	0	0	0	0	0
0-3	0	0	0	0	0	0	0	0	0	0	0
0-4	0	0	0	0	0	0	0	0	0	0	0
0-5	0	0	0	0	0	0	0	0	0	0	0
1-0	0	0	0	0	0	0	0	0	0	0	0
1-1	4	4	1	4	8	1	12	1	15	1	51
1-2	0	0	3	0	8	13	4	17	4	0	49
1-3	0	0	0	0	0	2	0	1	0	0	3
1-4	0	0	0	0	0	0	0	0	0	0	0
1-5	0	0	0	0	0	0	0	0	0	0	0
2-0	0	0	0	0	0	0	0	0	0	0	0
2-1	7	12	0	14	11	0	10	0	7	20	81
2-2	13	8	18	6	23	18	27	13	24	11	161
2-3	0	0	2	0	4	20	1	19	1	0	47
2-4	0	0	0	0	0	0	0	0	0	0	0
2-5	0	0	0	0	0	0	0	0	0	0	0
3-0	0	0	0	0	0	0	0	0	0	0	0
3-1	5	3	0	4	0	0	0	0	0	1	13
3-2	24	23	10	25	1	0	2	1	5	22	113
3-3	1	4	22	1	3	2	3	5	2	1	44
3-4	1	2	0	2	2	4	1	4	3	3	22
3-5	1	0	0	0	0	0	0	0	0	0	1
4-0	0	0	0	0	0	0	0	0	0	0	0
4-1	0	0	0	0	0	0	0	0	0	0	0
4-2	1	1	0	2	0	0	0	0	0	0	4
4-3	3	6	3	2	3	2	0	2	1	3	25
4-4	24	16	19	25	34	22	64	14	32	22	272
4-5	10	15	16	9	55	68	28	35	18	5	259
5-0	0	0	0	0	0	0	0	0	0	0	0
5-1	0	0	0	0	0	0	0	0	0	0	0
5-2	0	0	0	0	0	0	0	0	0	0	0
5-3	2	0	01	0	1	0	0	0	0	0	4
5-4	67	33	11	60	15	4	22	12	52	62	338
5-5	187	223	245	195	183	193	176	226	186	199	2013

Çizelge 11. Madde 2 Ham Veriler

Puan	Puanlayıcı Kombinasyonu										
Kombinasyonu	1-2	1-3	1-4	1-5	2-3	2-4	2-5	3-4	3-5	4-5	Toplam
0-0	3	3	3	3	3	3	3	3	3	3	30
0-1	0	0	0	0	0	0	0	0	0	0	0
0-2	0	0	0	0	0	0	0	0	0	0	0
0-3	0	0	0	0	0	0	0	0	0	0	0
0-4	0	0	0	0	0	0	0	0	0	0	0
0-5	0	0	0	0	0	0	0	0	0	0	0
1-0	0	0	0	0	0	0	0	0	0	0	0
1-1	25	21	4	23	124	7	45	7	31	5	292
1-2	1	5	22	3	52	167	131	117	95	2	595
1-3	0	0	0	0	0	0	0	0	0	0	0
1-4	0	0	0	0	0	0	0	0	0	0	0
1-5	0	0	0	0	0	0	0	0	0	0	0
2-0	0	0	0	0	0	0	0	0	0	0	0
2-1	150	105	3	28	2	0	6	0	20	46	360
2-2	35	78	176	155	84	75	78	124	109	192	1106
2-3	0	2	6	2	11	21	13	16	11	7	89
2-4	0	0	0	0	0	1	0	0	0	0	1
2-5	0	0	0	0	0	0	0	0	0	0	0
3-0	0	0	0	0	0	0	0	0	0	0	0
3-1	1	0	0	0	0	0	0	0	0	0	1
3-2	59	55	45	50	4	3	1	4	6	16	243
3-3	12	16	26	22	10	16	18	16	15	22	173
3-4	0	1	1	0	6	1	1	1	0	1	12
3-5	0	0	0	0	0	0	0	0	0	0	0
4-0	0	0	0	0	0	0	0	0	0	0	0
4-1	0	0	0	0	0	0	0	0	0	0	0
4-2	2	2	2	2	0	0	0	0	0	0	8
4-3	8	3	7	7	0	0	0	5	5	2	37
4-4	6	7	3	5	4	5	7	2	3	3	45
4-5	3	7	7	5	14	13	11	5	4	2	71
5-0	0	0	0	0	0	0	0	0	0	0	0
5-1	0	0	0	0	0	0	0	0	0	0	0
5-2	0	0	0	0	0	0	0	0	0	0	0
5-3	0	0	0	0	0	0	0	0	0	0	0
5-4	12	4	3	5	2	0	2	4	7	6	45
5-5	33	41	42	40	34	36	34	44	41	43	388

Çok Değişkenlik Kaynaklı Rasch Ölçme Modeli ve Hiyerarşik Puanlayıcı ModeliAnaliz Sonuçlarının Birlikte Değerlendirilmesi

Her iki modelden elde edilen ve ortak olan öğrenci, puanlayıcı ve madde parametreleri arasında gözlenen korelasyon değerleri ile birlikte her iki modele ait Sapma Bilgi Kriteri değerleri üçüncü alt problem doğrultusunda verilmiş ve açıklanmıştır.

Çok Değişkenlik Kaynaklı Rasch Ölçme Modeli ve Hiyerarşik Puanlayıcı Modelinden Elde Edilen Öğrenci, Puanlayıcı ve Madde parametreleri Korelasyon Değerleri

Her iki modelde de ortak olan parametreler; öğrenci yetenek parametresi, puanlayıcı katılık parametresi ve madde güçlük parametresidir. Sadece HPM'de bulunan puanlayıcı değişkenlik parametresine, bu alt amaçta yer verilmemiştir. Çizelge 12'de, ortak olan parametrelere ilişkin iki modele ait sonuçlar arasındaki Spearman sıra farkları korelasyon katsayısı değerleri verilmiştir.

Çizelge 12. Öğrenci, Puanlayıcı ve Madde Parametreleri Korelasyon Değerleri

ÇDKRÖM - HPM	N	r	p
Öğrenci yetenek parametresi	350	0.969	0.00
Puanlayıcı katılık parametresi	5	-1.000	0.00
Madde güçlük parametresi	8	0.976	0.00

Çizelge 12 incelendiğinde; ÇDKRÖM ve HPM'ye ait öğrenci, puanlayıcı ve madde parametreleri arasındaki korelasyon değerlerinin çok yüksek olduğu görülmektedir. En yüksek korelasyon değerinin, puanlayıcı katılık parametresine ait olduğu ve iki modelin bu değer doğrultusunda negatif yönlü ve manidar bir ilişki gösterdiği ifade edilebilir ($r= -1.00$, $p<0.01$). "Bulguların Analizi" başlığı altında da açıklandığı üzere modellerin puanlayıcı katılık parametresini yorumlaması noktasında işaret değişikliği mevcuttur. Bu nedenle, negatif yönlü gözlenen bu ilişkinin esasında pozitif yönlü olduğu ifade edilebilir. Nitekim modeller doğrultusunda puanlayıcıların sahip oldukları puanlayıcı katılık parametrelerine ait yorumlamaların, oldukça benzer olması da bu durumun kanıtı olarak değerlendirilebilir. İki modelin madde güçlük parametresi ($r= 0.98$, $p<0.01$) ve öğrenci yetenek parametresi ($r= 0.97$, $p<0.01$) korelasyon değerleri dikkate alındığında, pozitif yönlü ve manidar ilişki gösterdiği belirtilebilir. Maddeler iki model için madde güçlük parametrelerine göre zordan kolaya doğru sıralandığında; ÇDKRÖM'de -0.04 değeri ile beşinci sırada olan yedinci madde, HPM'de -0.10 değeri ile altıncı sıradadır. HPM'de -0.09 değeri ile beşinci sırada olan sekizinci madde ise ÇDKRÖM'de -0.06 değeri altıncı sıradadır. İki model için maddelerin parametre güçlük sıralamaları birbirinden farklı olsa da aldıkları değerlerin birbirine yakın olduğu ifade edilebilir.

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. SBK, benzer amaçlara yönelik geliştirilmiş, daha geniş kullanım alanına sahip, AIC'in bayes benzeşiği olarak tanımlanabilir ve modellerin tahmininin beklenen kayıpları arasındaki farklılığın kestirimini sunar (Spiegelhalter ve diğerleri, 2002). Çizelge 13'de SBK ve SBK'nın dayandığı değerlere yer verilmiştir.

Çizelge 13. ÇDKRÖM ve HPM'ye ait Sapma Bilgi Kriteri Değerleri

	pD	$\overline{D}$	SBK
ÇDKRÖM	36	3657	3693
HPM	44	3641	3685

Çizelge 13'de, her iki model için de pD, $\overline{D}$ ve SBK değerleri yer almaktadır. Model karşılaştırmalarında Literatürde yer alan çalışmaların bir kısmında (Zhu ve Carlin, 2000) 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 (Barry, Brooks, Catchpole ve Morgan, 2003; Berg ve diğerleri, 2004; Celeux, Forbes, Robert ve Titterington, 2006; Claeskens ve Hjort, 2008; Liddle, 2007; Mason, Richardson ve Best, 2012; Spiegelhalter ve diğerleri, 1998; Spiegelhalter ve diğerleri, 2002; Spiegelhalter, 2006; Zhu ve Carlin, 2000). 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). Daha spesifik olarak, yuvalanmış model karşılaştırmaları için gözlenen iki birimden az farklılık manidar kabul edilmezken, özellikle gözlenen farklılıklar 10 birimden daha fazla ise büyük değere sahip olan

model reddedilir. AIC'e ait olan bu kural, SBK için de geçerlidir (Spiegelhalter, Best, Carlin ve Van der Linde, 2002). Yuvalanmamış modeller için benzer manidar bir kural mevcut olmamakla birlikte AIC ve BIC de olduğu gibi, elde edilen daha küçük bir SBK değeri daha iyi modelin göstergesidir (Berg ve diğerleri, 2004; Liddle, 2007; Spiegelhalter ve diğerleri, 2002; Wilberg ve Bence, 2008; Zhu ve Carlin, 2000). 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.

BÖLÜM IV

SONUÇ VE ÖNERİLER

Bu bölümde, araştırma bulgularına dayalı sonuçlar ve sonuçlar doğrultusunda sunulan önerilere yer verilmiştir.

Sonuçlar

Bu araştırmada, aynı öğrenciler tarafından açık uçlu maddelere verilen yanıtların, birden fazla puanlayıcı tarafından puanlanması durumunda, elde edilen çok değişkenlik kaynaklı Rasch ölçme modeli ve hiyerarşik puanlayıcı modeli sonuçlarının belirlenmesi ve bu sonuçların birlikte değerlendirilmesi amaçlanmıştır.

Araştırmanın amacı doğrultusunda ulaşılan sonuçlar, alt amaçların sırasına uygun olarak, maddeler halinde aşağıda özetlenmiştir.

1. a) Araştırmanın verileri için ÇDKRÖM model veri uyumu sağlanmıştır.

   b) Araştırma kapsamında gerçekleştirilen uygulamaya katılan tüm öğrencilerin, %18,5'i uygunluk istatistikleri sınır değerlerinin dışında yer almıştır. Ayırma indeksi ve güvenirliği doğrultusunda öğrencilerin, istatistiksel olarak manidar beş farklı yetenek seviyesine ayrıldığı sonucuna ulaşılmıştır.

   c) 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.

   d) Uygulamada kullanılan sekiz maddenin, çok değişkenlik kaynaklı Rasch ölçme modeli analizi doğrultusunda; öğrenciler tarafından çoğunlukla doğru yanıtlandığı, farklı güçlük derecelerine sahip oldukları ve bu madde güçlük indekslerinin istatistiksel olarak manidar olduğu belirlenmiştir. Toplam sekiz

maddeden biri (dördüncü madde) uygunluk istatistikleri sınır değerlerinin dışında yer almıştır. Bu maddenin, diğer maddelerle aynı doğrultuda bulunmadığı, olası farklı bir boyuta veya puanlayıcılar arası belirli bir düzeyde uyuşmazlığa işaret ettiği sonucu elde edilmiştir.

2. a) Araştırmaya katılan öğrencilerin yetenek düzeylerinin belirlenebilmesi amacıyla HPM analizleri, markov zincirleri monte carlo yöntemi aracılığıyla üç zincir üzerinden gerçekleştirilmiştir. Brooks-Gelman-Rubin (BGR) grafikleri, zaman serileri diyagramı ve MC hataları bağlamında tüm öğrenciler için üç zincirin yakınsadığı; başka bir ifadeyle, elde edilen sonuçların geçerli ve güvenilir olduğu sonucuna ulaşılmıştır.

   b) 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 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.

   c) Araştırma kapsamında kullanılan sekiz madde için BGR grafikleri, zaman serileri diyagramları ve MC hataları doğrultusunda üç zincirin de sonsal dağılımın yakınsadığı belirlenmiştir. Maddelerin birbirinden farklı güçlük düzeylerine sahip oldukları, öğrenciler tarafından genel olarak doğru yanıtlandıkları sonucuna varılmıştır.

3. a) Her iki modele ilişkin yapılan analizler sonucunda ulaşılan değerlerden benzer amaca yönelik olanlar öğrenci, puanlayıcı ve madde değişkenlik kaynaklarına göre öncelikle korelasyon katsayıları aracılığıyla incelenmiştir. Bu bağlamda, iki modele ait üç değişkenlik kaynağı sonuçlarının genel olarak örtüştüğü; en

yüksek ilişkinin puanlayıcı katılık parametrelerinde görüldüğü ve bu değişkenlik kaynağı için puanlayıcı katılık/cömertlik sıralamalarının aynı olduğu tespit edilmiştir. Madde değişkenlik kaynağı için korelasyon katsayısının, yüksek düzeyde ilişkiye işaret ettiği ve değerleri birbirine yakın olmakla birlikte iki madde hariç sıralamanın benzer olduğu; öğrenci değişkenlik kaynağına göre de yüksek düzeyde ilişkinin görüldüğü sonuçları elde edilmiştir.

b) 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.

Öneriler

ÇDKRÖM ve HPM sonuçlarının ayrı ayrı ve birlikte değerlendirilmesiyle ulaşılan önerilere aşağıda yer verilmiştir.

1. Puanlayıcı davranışlarının incelenmesinin önemli görüldüğü durumlarda, HPM'nin kullanımı daha detaylı bilgi sunması bakımından yararlı olabilir.

2. Her iki modelden, sıralama düzeyinde benzer sonuçlar elde edilmiş olmakla birlikte, veriye daha iyi uyum sağlaması ve elde edilen sonuçların veriyi daha iyi yansıtması bakımlarından HPM'nin kullanımı önerilebilir.

3. 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.

4. Daha az sayıda puanlayıcının kullanılması durumunda, her iki modele ilişkin kestirimlerin gerçekleştirilebilmesi ve istenilen düzeyde güvenirliğe erişilebilmesi amacıyla kullanılan madde sayısı arttırılabilir.

5. İş gücü, zaman, kullanışlılık, ekonomik koşullar ve uygun güvenirlik düzeyi dikkate alındığında, beş puanlayıcı kullanmak yerine üç puanlayıcının kullanıldığı bir araştırma gerçekleştirilebilir.

6. Bayes kestirimi aracılığıyla HPM ile verilerin analiz edilmesi, ÇDKRÖM'ne göre oldukça uzun sürede tamamlanmaktadır. HPM kodları, farklı kestirim yöntemlerinin kullanımına izin verecek biçimde düzenlenip, analizler gerçekleştirilebilir. Bu analiz ile elde edilen sonuçlarla, aynı örneklem üzerinde hali hazırda kullanılan kestirim yöntemi ile ulaşılan sonuçlar karşılaştırılıp, farklılık gösterip göstermediği incelenebilir.

7. Araştırma, tamamen çaprazlanmış desenden faydalanılarak gerçekleştirilmiştir. Özellikle geniş ölçekli testler kapsamında kullanılan desenlerden faydalanılarak benzer çalışmaların gerçekleştirilmesi önerilebilir.

8. Araştırmada, bütünsel dereceli puanlama anahtarı kullanılarak veriler elde edilmiştir. Puanlayıcıya, puanlama sürecine ilişkin daha detaylı bilgi vermesi açısından analitik dereceli puanlama anahtarı kullanılarak, benzer çalışmaların gerçekleştirilmesi faydalı olabilir.

9. HPM kodları madde ayırıcılık ve şans parametrelerini içerecek biçimde yeniden düzenlenip, bu kodlar aracılığıyla çalışmaların gerçekleştirilmesi yararlı olabilir.

10. Farklı araştırma koşulları altında, HPM'ne ait parametre kestirimlerinin gerçekleştirilebilmesi amacıyla ihtiyaç duyulan örneklem büyüklüklerine ilişkin, çalışmaların yapılması faydalı olabilir.

11. Aynı örneklem kullanılarak, ÇDKRÖM ve HPM ile elde edilen kestirimler üzerinde, yanlılık analizleri gerçekleştirilip, sonuçların farklılaşıp farklılaşmadığı araştırılabilir.

KAYNAKÇA

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

Akın, Ö. ve 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.

Anıl, D. ve Büyükkıdık, S. (2012). Genellenebilirlik kuramında dört facetli karışık desen kullanımı için örnek bir uygulama. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, 3 (6), 291-296.

Ardia, D. (2008). *Financial risk management with bayesian estimation of GARCH models: Theory and applications*. Berlin: Springer.

Atılgan, H. (2005a). Genellenebilirlik kuramı ve puanlayıcılar arası güvenirlik için örnek bir uygulama. *Eğitim Bilimleri ve Uygulama*, 4 (7), 95-108.

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.

Baştürk, R. (2010). Bilimsel araştırma ödevlerinin çok yüzeyli Rasch ölçme modeli ile değerlendirilmesi. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, 1 (1), 51-57.

Barry, S.C., Brooks, S.P., Catchpole, E.A. and Morgan, B.J.T. (2003). The analysis of ring-recovery data using random effects. *Biometrics*, 59 (1), 54-65.

Benjamin, A. (2013). *English teacher's guide to performance tasks and rubrics: middle school.* New York: Routledge.

Berg, A., Meyer, R. and Yu, J. (2004). Deviance information criterion for comparing stochastic volatility models. *Journal of Business & Economic Statistics*, 22(1), 107-120.

Bond, T.G. and Fox, C.M. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences*. NJ: Lawrence Erlbaum Associates.

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. (1998). Raw-score conditional standard errors of measurement in generalizability theory. *Applied Psychological Measurement*, 22 (4), 307-331.

Brennan, R.L. (2000a). (Mis) Conception about generalizability theory. *Educational Measurement: Issues and Practice*, 19 (1), 5-10.

Brennan, R.L. (2000b). Performance assessments from the perspective of generalizability theory. *Applied Psychological Measurement*, 24 (4), 339-353.

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

Browne, W. and Rasbash, J. (2004). Multilevel modelling. Hardy, M. and Bryman, A. (Ed.). *Handbook of data analysis*. London: Sage Publications.

BUGS. (2014). Bayesian inference Using Gibbs Sampling. MRC Biostatistics Unit. 03.01.2014 tarihinde http://www.mrc-bsu.cam.ac.uk/software/bugs/ adresinden ulaşılmıştır.

Cardinet, J., Tourneur, W., and Allal, L. (1976). The symmetry of generalizability theory: Applications to educational measurement. *Journal of Educational Measurement*, 13 (2), 119-135.

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

Casabianca, J.M. and 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. and 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.

Casabianca, J.M., Junker, B.W. and Patz, R. (2014). The hierarchical rater model. Invited chapter for W. J. van der Linden and R. K. Hambleton (Ed.). *Handbook of modern item response theory*. Boca Raton, FL: Chapman and Hall/CRC.

Celeux, G., Forbes, F., Robert, C.P. and Titterington, D.M. (2006). Deviance information criteria for missing data models. *Bayesian Analysis*, 1 (4), 651-673.

Chan, G.C. and Reeve, B.B. (2005). Item response theory and ıts applications to patient-reported outcomes measurement. *Evaluation & The Health Professions*, 28 (3), 264-282.

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

Claeskens, G. and Hjort, N.L. (2008). *Model selection and model averaging*. Cambridge: Cambridge University Press.

Congdon, P. (2001) *Bayesian statistical modelling*. NJ: Wiley.

Cooper, P.L. (1984). The assessment of writing ability: A review of research. *ETS Research Report Series*, (1), i-46. Princeton, NJ: Educational Testing Service.

Crocker, L. and Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Holt, Reinhart and Wilson.

Cronbach, L.J., Gieser, G.C., Nanda, H. and Rajarantnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: Wiley.

Cunningham, G.K. (1998). *Assessment in the classroom: Constructing and interpreting texts*. Psychology Press.

Çıkrıkçı, N. (2010). Üst düzey düşünme becerilerinin ölçülmesinde gündelik yaşam unsuru. *Cito Eğitim: Kuram ve Uygulama*. Ocak-Mart, 9-26.

Davis, L.L. (2004). Strategies for controlling item exposure in computerized adaptive testing with the generalized partial credit model. *Applied Psychological Measurement*, 28 (3), 165-185.

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., and 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.

Donoghue, J.R. and Hombo, C.M. (2000). *A comparison of different model assumptions about rater effects*. Annual Meeting ofthe National Council on Measurement in Education'da sunulan bildiri. New Orleans, LA.

Downing, S.M. and Haladyna, T.M. (2004). Validity threats: overcoming interference with proposed interpretations of assessment data. *Medical Education*, 38 (3), 327-333.

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.

Efron, B. (1986). Why isn't everyone a bayesian? *American Statistician*, 40, 1-11.

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.

Ercikan, K., Schwarz, R.D., Julian, M.W., Burket, G.R., Weber, M.M. and Link, V., (1998). Calibration and scoring of tests with multiple-choice and constructed-response ıtem types. *Journal of Educational Measurement*, 35 (2), 137-154.

Engelhard, G. (2002). Monitoring raters in performance assessments. In G. Tindal and T. Haladyna (Ed.), *Large-scale assessment programs for all examinees: Validity, technical adequacy, and implementation* (261–287). Mahwah, NJ: Lawrence Erlbaum.

Engelhard, G. and 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.

Ferrando, P.J. (2009). A graded response model for measuring person reliability. *British Journal of Mathematical and Statistical Psychology*, 62 (3), 641-662.

Fischer, G.H. and Molenaar, I.W. (1995). *Rasch models: Foundations, recent developments, and applications*. NY: Springer Science & Business Media.

Forero, C.G. and Maydeu-Olivares, A. (2009). Estimation of IRT graded response models: limited versus full information methods. *Psychological Methods*, 14 (3), 275.

Gamerman, D. (1997). *Markov chain monte carlo*. USA: Chapman and Hall.

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

Gelman, A., Carlin, J.B., Stern, H.S. ve Rubin, D.B. (2004). *Bayesian data analysis*. Boca Raton, FL: Chapman & Hall/CRC.

Gelman, A. and Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models.* Cambridge: Cambridge University Press.

Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A. and Rubin, D.B. (2013). *Bayesian data analysis*. CRC Press.

Güler, N. (2009). Genellenebilirlik kuramı ve SPSS ile GENOVA programlarıyla hesaplanan G ve K çalışmalarına ilişkin sonuçların karşılaştırılması. *Education and Sciences*, 34 (154), 93-103.

Güler, N. ve Gelbal, S. (2010). Studying reliability of open ended mathematics ıtems according to the classical test theory and generalizability theory. *Educational Sciences: Theory and Practice*, 10 (2), 1011-1019.

Güler, N. (2011). Rasgele veriler üzerinde genellenebilirlik kuramı ve klasik test kuramına göre güvenirliğin karşılaştırılması. *Education and Sciences*, 36 (162), 225-234.

Haladyna, T.M. (1997). *Writing test ıtems to evaluate higher order thinking*. USA: A Pearson Education Company.

Hays, R.D., Morales, L.S. ve Reise, S.P. (2000). Item response theory and health outcomes measurement in the 21st century. *Medical Care*, 38 (9), II-28.

Hombo, C. and Donoghue, J.R. (2001). *Applying the hierarchical raters model to NAEP*. Annual Meeting of the National Council on Measurement in Education'da sunulan bildiri. Seattle, Washington.

Iramaneerat, C., Yudkowsky, R., Myford, C.M. ve Downing, S.M. (2008). Quality control of an OSCE using generalizability theory and many-faceted Rasch measurement. *Advances İn Health Sciences Education*, 13 (4), 479-493.

Iramaneerat, C., Myford, C.M., Yudkowsky, R., and 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.

Johnson, M.S., Cohen, W. and Junker, B.W. (1999). *Measuring appropriability in research and development with item response models*. InPrepared for the

Committee on the Foundations of Assessment, National Research Council. 01.05.2013 tarihinde http://www. stat. cmu. edu/˜ brian/nrc/cfa adresinden ulaşılmıştır.

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

Junker, B.W. ve Patz, R.J. (1998). *The hierarchical rater model for rated test items.*Annual North American Meeting of the Psychometric Society'de sunulan bildiri. Champaign-Urbana, IL.

Kastner, M. ve 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.

Lee, Y.W. and Kantor, R. (2003). *Investigating differential rater functioning for academic writing samples: an MFRM approach*. Annual Meeting of National Council on Measurement in Education'da sunulan bildiri. Chicago, IL.

Lee, M.D. and Wagenmakers, E.J. (2014). *Bayesian cognitive modeling: A practical course*. Cambridge University Press, UK.

Li, Y. and J. Yu (2012). Bayesian hypothesis testing in latent variable models. *Journal of Econometrics*, 166, 237–246.

Li, Y., Zeng, T., and Yu, J. (2014). Robust deviance information criterion for latent variable models. *CAFE Research Paper*, 13 (19), 1-44.

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

Ligtvoet, R. (2012). An isotonic partial credit model for ordering subjects on the basis of their sum scores. *Psychometrika*, 77 (3), 479-494.

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

Linacre, J.M., Wright B.D. and 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. (1990). *A facet model for judmental scoring*. MESA Memo 61.

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

Linacre, J.M. (1998). *A user's guide to FACETS: Rasch measurement computer program*. 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.

Linacre, J.M. and Wright, B.D. (2004). *Construction of measures from many-facet data*. Chicago: MESA Press.

Linacre, J.M. (2007). *Realiability and separations. A users guide to Winsteps/Ministep Rasch–model computer programs*. Chicago: Winsteps. Com.

Linacre, J. M. (2012). *Winsteps® (Version 3.71. 0)[Computer Software] user manual*. Beaverton, Oregon: Winsteps. com. Retrieved February 1, 2013.

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

Lunn, D.J., Thomas, A., Best, N. and Spiegelhalter, D. (2000). WinBUGS –a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing*, 10, 325–337.

Lunn, D., Spiegelhalter, D., Thomas, A. and Best, N. (2009). The BUGS project: evolution, critique and future directions. *Statistics in Medicine*, 28(25), 3049-3067.

Lunn, D., Jackson, C., Best, N., Thomas, A. and Spiegelhalter, D. (2012). *The BUGS book: A practical introduction to bayesian analysis*. USA: CRC pres.

Lynch, B.K. and 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. and 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.

Martinez, M.E. (1999). Cognition and The Questions Of Test İtem Format. *Educational Psychologist*, 34 (4), 207-218.

Mason, A., Richardson, S. and Best, N. (2012). Two-pronged strategy for using DIC to compare selection models with non-ignorable missing responses.*Bayesian Analysis*, 7 (1), 109-146.

Masters, G.N. (1982). A Rasch Model for Partial Credit Scoring. *Psychometrika*, *47*, 149- 174.

McNamara, T.F. (1996). *Measuring second language performance*. New York: Addisonn Wesley Longman.

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.

MRC Biostatistics Unit. (2014). Medical research council, The BUGS Project, DIC: Deviance ınformation criterion. Cambridge, UK. 02.01.2014 tarihinde http://www.mrc-bsu.cam.ac.uk/software/bugs/the-bugs-project-dic/ adresinden erişilmiştir.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, 16 (2), 159-176.

Muraki, E. (1993). Information functions of the generalized partial credit model. *Applied Psychological Measurement*, 17 (4), 351-363.

Mushquash, C. and O'Connor, B.P. (2006). SPSS and SAS programs for generalizability theory analyses. *Behavior Research Methods*, 38 (3), 542-547.

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

Myford, C.M. and Wolfe, E.W. (2004). Detecting and measuring rater effects using many-facet Rasch measurement: Part II. *Journal of Applied Measurement*, 5, 189–227.

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

Nakamura, Y. (2002). Teacher assessment and peer assessment in practice. *Educational Studies*, 44, 203-215.

Nitko, A.J. (2001). *Educational assessment of students*. USA: Pearson Education.

Ntzoufras, I. (2009). *Bayesian modeling using WinBUGS*. Hoboken, NJ: John Wiley & Sons.

O'Hagan, A. (1986). *Probability methods and measurement*. Chapman and Hall, London.

Patz, R. J. and 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. and 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. and 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. and 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), 341-384.

Penfield, R.D., Myers, N.D. and Wolfe, E.W. (2008). Methods for assessing item, step, and threshold invariance in polytomous items following the partial credit model. *Educational and Psychological Measurement*, 68 (5), 717-733.

Pollack, J.M., Rock, D.A. and Jenkins, F. (1992). *Advantages and disadvantages of constructed-response item formats in large-scale surveys*. Inannual meeting of the American Educational Research Association'da sunulan bildir.San Francisco, California.

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.

R Development Core Team. (2013). R: A Language and Environment for Statistical Computing (Tech. Rep.). Vienna, Austria: R Foundation for Statistical

Computing. 02.06.2013 tarihinde http://www.r-project.org/ adresinden ulaşılmıştır.

Raftery, A.E. (1995). Bayesian model selection in social research. *Sociological Methodology, 25*, 111-164.

Reynolds, C.R., Livingston, R.B. and Willson, V. (2006). *Measurement and assessment in education*. Boston, MA: Allyn and Bacon.

Rezaei, A.R. and Lovorn, M. (2010). Reliability and validity of rubrics for assessment through writing. *Assessing Writing*, 15 (1), 18-39.

Roid, G.H. and 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.

Samejima, F. (2008). Graded response model based on the logistic positive exponent family of models for dichotomous responses. *Psychometrika*, 73 (4), 561-578.

Seltman, H. (2010). *R package rube* (really useful WinBUGS enhancer). 03.10.2013 tarihinde http://www.stat.cmu.edu/˜hseltman/rube/. (Version 0.2-13) adresinden ulaşılmıştır.

Shavelson, R.J., Webb, N.M. and Rowley, G.L. (1989). Generalizability theory. *American Psychologist*, 44 (6), 922.

Smith, E.V. and Smith, R.M. (2004). *Introduction to Rasch measurement*. Maple Grove, MN: JAM.

Spiegelhalter, D., Thomas, A., Best, N. and Gilks, W. (1996). BUGS 0.5: Bayesian inference using Gibbs sampling manual (version ii). MRC Biostatistics Unit, Cambridge.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. and Van der Linde, A. (1998). Bayesian deviance, the effective number of parameters, and the comparison of arbitrarily complex models. *Research Report*, 98-009.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P. and Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64 (4), 583-639.

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

Spiegelhalter, D.J. (2006). *Two brief topics on modelling with WinBUGS.* InIceBUGS Conference Proceedings'de sunulan bildiri. Hanko, Finland.

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

Sudweeks, R.R., Reeve, S. and 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.

Tekin, H. (1987). *Eğitimde ölçme ve değerlendirme*. Ankara: Meso Yayınevi.

Turgut, M.F. ve Baykul, Y. (2012). *Eğitimde ölçme ve değerlendirme*. Ankara: Pegem Akademi.

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

Ucal, M.S. (2006). Ekonometrik model seçim kriterleri üzerine kısa bir inceleme. *C.Ü. İktisadi ve İdari Bilimler Dergisi*, 7 (2), 41-56.

Verhelst, N. and 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.

Volk, J. (2002). *Assessment strategies.* Regina, SK: Saskatchewan Learning-Region 3.

Walsh, B. (2002). *Markov Chain Monte Carlo and Gibbs Sampling* (ders notları*)*, 07.02.2013tarihindenitro.biosci.arizona.edu/courses/EBB596/handouts/Gibbs.pdf/ adresinden ulaşılmıştır.

Wang, Z.G. (2012). *On the use of covariates in a latent class signal detection model, with applications to constructed response scoring*(Doktora Tezi). Columbia University, New York.

Ward, E.J. (2008). A review and comparison of four commonly used Bayesian and maximum likelihood model selection tools. *Ecological Modelling*, 211 (1), 1-10.

Wikle, C.K., Berliner, L.M. and Milliff, R.F. (2003). Hierarchical Bayesian approach to boundary value problems with stochastic boundary conditions.*Monthly Weather Review*, 131 (6), 1051-1062.

Wilberg, M.J. and Bence, J.R. (2008). Performance of deviance information criterion model selection in statistical catch-at-age analysis. *Fisheries Research*, 93 (1), 212-221.

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

Wright, B.D. and Masters, G.N. (1982). *Rating scale analysis: Rasch measurement.* Chicago: MESA Press.

Wright, B.D. and Linacre, J.M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, 8 (3), 370.

Yardımcı, A. ve Erar, A. (2005). Aykırı deger varlıgında dogrusal regresyonda degisken seçimine gibbs örneklemesi yaklasımı. *Gazi Üniversitesi Fen Bilimleri Dergisi*, 18 (4), 603-611.

Yelboga, A. and Tavsancil, E. (2010). The Examination of Reliability According to Classical Test and Generalizability on a Job Performance Scale. *Educational Sciences: Theory and Practice*, 10 (3), 1847-1854.

Yılmaz Nalbantoğlu, F. ve Gelbal, S. (2011). İletişim becerileri istasyonu örneğinde Genellenebilirlik Kuramı'yla farklı desenlerin karşılaştırılması. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi*, 41, 509-518.

Yılmaz Nalbantoğlu, F. ve Tavsancil, E. (2014). Making intramuscular injection station Comparison of generalizability theory Dengelenmemis balanced eating pattern with Data.*Educatıon and Scıence*, 39 (175), 285-295.

Zhu, L. and Carlin, B.P. (2000). Comparing hierarchical models for spatio-temporally misaligned data using the deviance information criterion. *Statistics in Medicine*, 19 (17-18), 2265-2278.

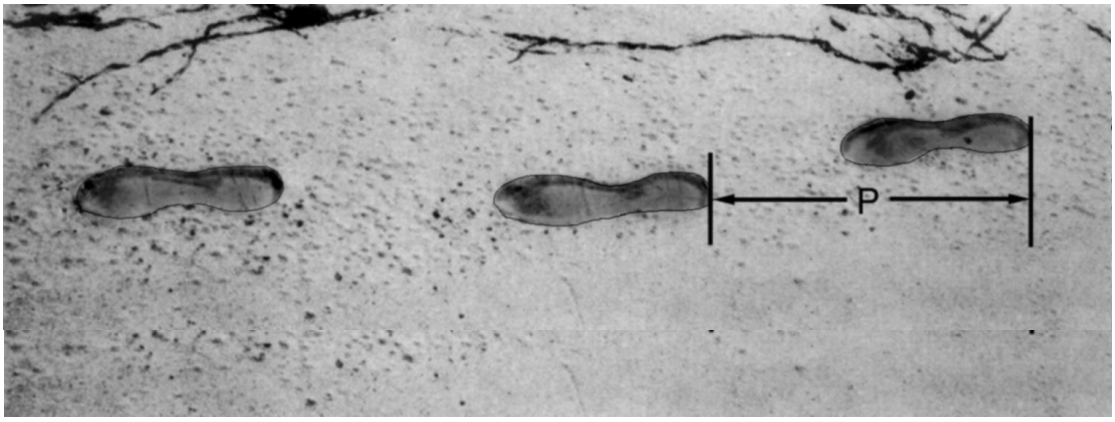
Zuur, A., Ieno, E. N., Walker, N., Saveliev, A. A. and Smith, G. M. (2009).*Mixed effects models and extensions in ecology with R*. Springer Science & Business Media, New York.

Zuur, A., Saveliev, A. and Ieno, E. (2012). *Zero inflated models and generalized linear mixed models with R*. Newburgh, U.K.: Highland Statistics Ltd.

EK A

Araştırma Kapsamında Kullanılan Açık Uçlu Maddeler

Yürüyüş



Resim, yürüyen bir erkeğin ayak izlerini gösteriyor. Adım uzunluğu *P*, ardışık iki ayak izinin topukları arasındaki mesafedir.

n = bir dakikadaki adım sayısı

P = adım uzunluğunu metre olarak belirtirse;

Erkekler için, $\frac{n}{P} = 140$ formülü, *n* ve *P* arasındaki yaklaşık bir ilişkiyi gösterir.

1. ve 2. soruyu yukarıda yer alan bilgiler doğrultusunda yanıtlandırınız.
1) Eğer formül Hakkı'nın yürüyüşüne uygulanırsa ve Hakkı dakikada 70 adım atarsa, Hakkı'nın bir adım uzunluğu ne olur? İşleminizi gösteriniz.
Çözüm:

2) Burak, adım uzunluğunun 0,80 metre olduğunu biliyor. Formül Burak'ın yürüyüşüne uygulanır.

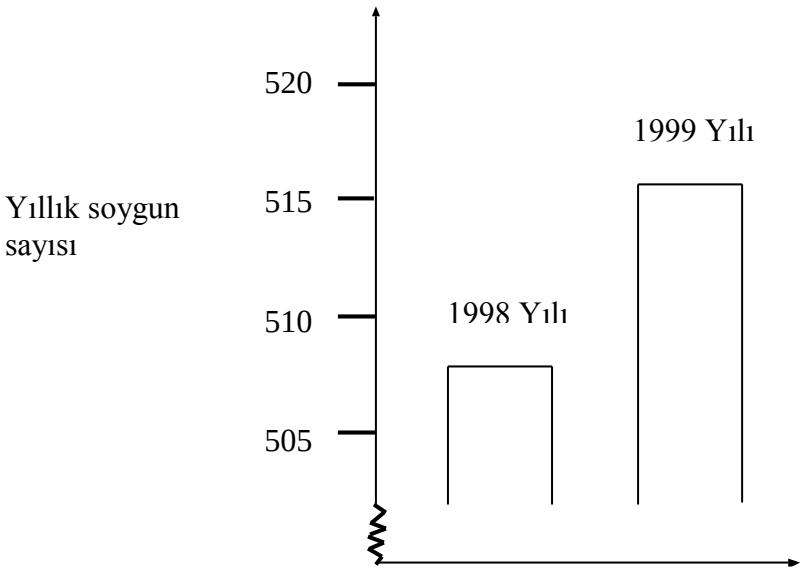
Burak'ın bir dakikadaki yürüme hızını metre olarak ve bir saatteki yürüme hızını kilometre olarak hesaplayınız. İşleminizi gösteriniz.

Çözüm:

Soygunlar

Bir televizyon muhabiri, bu grafiği gösterdi ve şöyle dedi:

“Bu grafik 1998 yılından 1999’a kadar soygunların sayısında çok büyük bir artış olduğunu göstermektedir.”



3) Muhabirin sözlerinin grafiğin kabul edilebilir bir yorumu olduğunu düşünüyor musunuz? Yanıtınızı desteklemek için bir açıklama yapınız.

Çözüm:

Başkana Destek

Zed ülkesinde, yapılacak seçimlerde Başkana verilecek desteğin oranını öğrenmek için bir kamuoyu yoklaması gerçekleştirilmiştir. Dört gazete, ülke genelinde ayrı ayrı kamuoyu yoklaması yapmıştır. Dört gazetenin kamuoyu araştırma sonuçları aşağıda gösterilmiştir.

1.Gazete: % 36,5 (Kamuoyu yoklaması oy kullanma hakkı olanlar arasından rastlantısal olarak seçilen 500 kişi üzerinde 6 Ocak'ta gerçekleştirilmiştir.)

2. Gazete. % 41,0 (Kamuoyu yoklaması oy kullanma hakkı olanlar arasından rastlantısal olarak seçilen 500 kişi üzerinde 20 Ocak'ta gerçekleştirilmiştir.)

3. Gazete: % 39,0 (Kamuoyu yoklaması oy kullanma hakkı olanlar arasından rastlantısal olarak seçilen 1000 kişi üzerinde 20 Ocak'ta gerçekleştirilmiştir.)

4. Gazete:% 44,5 (Kamuoyu yoklaması telefonla arayıp oy veren 1000 okuyucuyla 20 Ocak'ta gerçekleştirilmiştir.)

4) Eğer seçim 25 Ocak'ta yapılırsa, hangi gazetenin sonucu Başkana verilen desteğin oranını en iyi biçimde kestirebilir? Yanıtınızı desteklemek için iki neden gösteriniz.

Çözüm:

Alana Göre Ödeme

Bir binada yaşayan insanlar binadaki daireleri satın almaya karar veriyorlar. Paralarını dairelerinin alanı ile orantılı olarak ödeyeceklerdir.

Örneğin, katlardaki tüm dairelerin alanının beşte biri kadar alanı bulunan bir dairede yaşayan kişi, binanın toplam fiyatının beşte biri kadar bir tutarı ödeyecektir.

Binada üç daire vardır. En büyük olan Daire 1'in toplam alanı $95m^2$ dir. Daire 2 ve 3'ün alanları ise sırasıyla $85m^2$ ve $70m^2$ 'dir. Binanın satış fiyatı 300 000 zeddir.

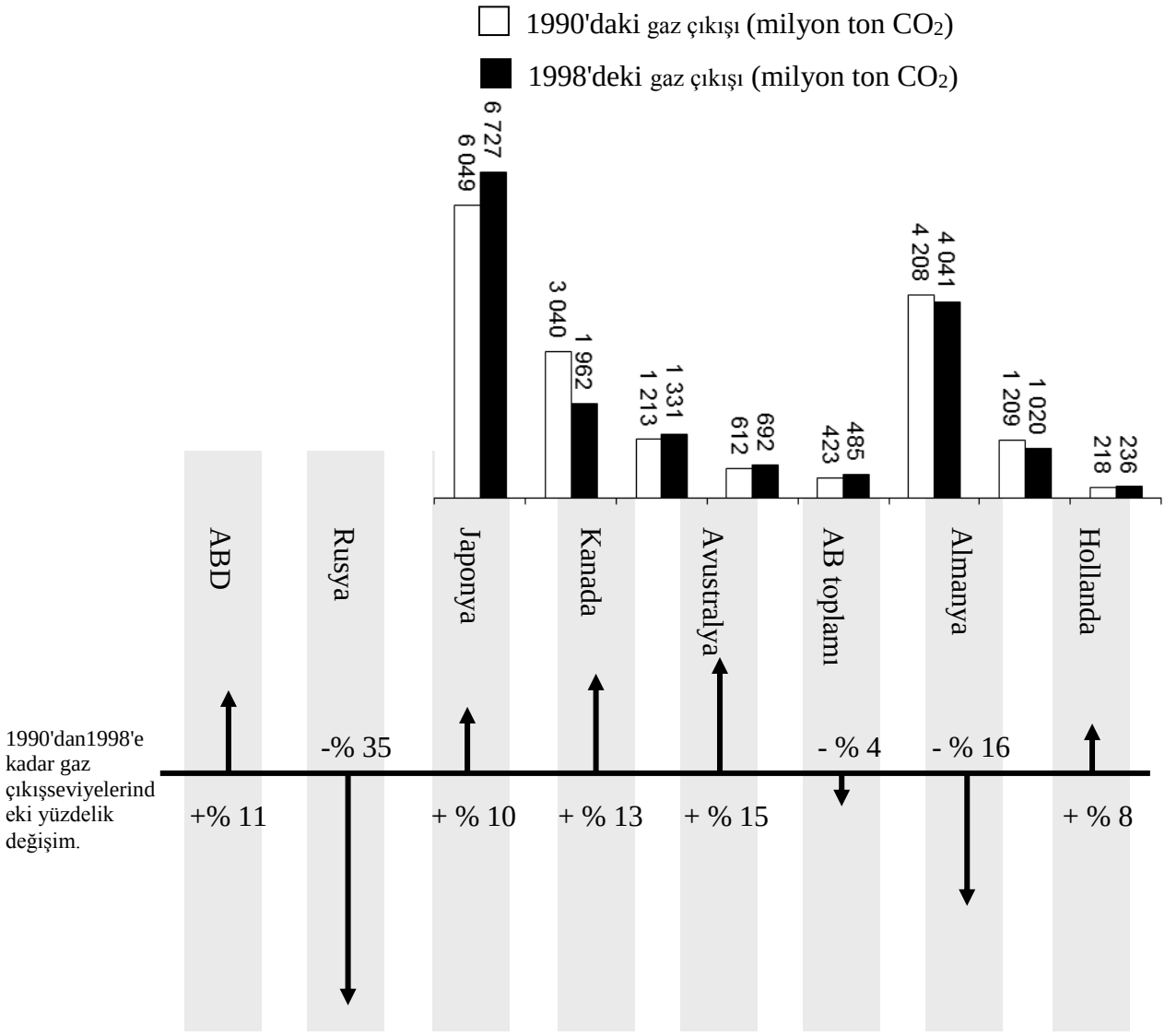
5) Daire 2'nin sahibinin ne kadar ödemesi gerekir? İşleminizi gösteriniz.

Çözüm:

Co_2 Seviyelerini Azaltmak

Bir çok bilim adamı, atmosferimizde artan CO_2 gazı seviyesinin iklim değişikliklerine sebep olmasından korkmaktadırlar.

Aşağıdaki grafik, çeşitli ülkeler (veya bölgeler) için 1990'da çıkan CO_2 seviyelerini (beyaz sütunlar), 1998'deki gaz çıkış seviyelerini (siyah sütunlar) ve 1990 ile 1998 yılları arasında gaz çıkış seviyelerindeki yüzdelik değişimi (yüzdeliklerle verilen oklar) göstermektedir.



6. ve 7. Soruyu aşağıda yer alan bilgiler doğrultusunda yanıtlandırınız.

6) Bu grafikten ABD'de, 1990'dan 1998'e kadar karbon dioksit gazının çıkış seviyesindeki artışın % 11 olduğunu okuyabilirsiniz. % 11'in nasıl elde edildiğini ortaya koymak için hesaplamayı gösteriniz.

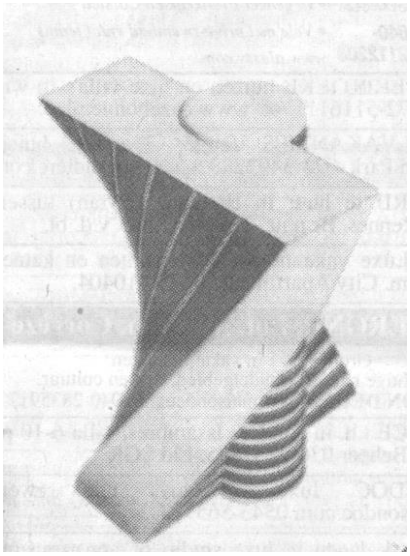
Çözüm:

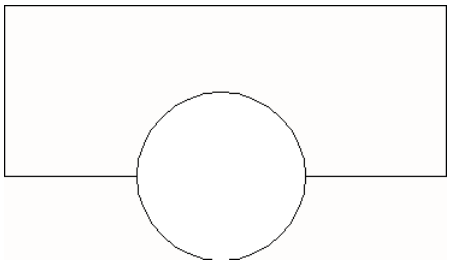
7) Melek ve Nazan, CO_2 gazı çıkışındaki en büyük **artışa** sahip ülkenin (veya bölge) hangisi olduğunu tartıştılar. İkisi de, grafiğe dayanan farklı sonuçlar çıkardılar. Bu soru için olası iki doğru yanıtı veriniz, bu yanıtların her birini nasıl elde ettiğinizi açıklayınız.

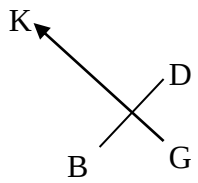
Çözüm:

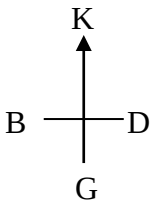
Burgulu Bina

Modern mimaride, binalar genellikle alışılmışın dışındaki şekildedirler. Aşağıdaki resim bir ‘burgulu binanın’ bilgisayar modelini ve zemin katın plânını göstermektedir. Pusula uçları, binanın konumunu göstermektedir. Binanın zemin katında ana giriş ve mağazalar









için ayrılmış yerler vardır. Zemin katın üstünde, apartman dairelerinin bulunduğu 20 kat vardır.

Her katın plânı, zemin katın plânına benzemekte ama bir altındaki kattan biraz farklı konumdadır. Silindir şeklindeki bölümde asansör boşluğu ve her kata bir sahanlık bulunmaktadır.

8) Binanın toplam yüksekliğini metre olarak tahmin ediniz. Yanıtınızı nasıl bulduğunuzu açıklayınız.

Çözüm:

EK B

Araştırma Kapsamında Kullanılan Açık Uçlu Maddelere Ait Bütünsel Dereceli Puanlama Anahtarıİstatistikleri

1., 2., 5. ve 6. maddelere ait dereceli puanlama anahtarı

Kriterler	**Derecelendirme**
Tam, uygun ve tutarlı olarak yanıtlandırılmış. Problemin; matematiksel süreç ve fikrini anladığını yansıtmış.	5
Problemin çözümünde kullanılan yöntem uygun ancak, işlem hatalarından (çarpma, bölme, toplama, çıkarma) dolayı doğru sonuca ulaşılamamış. Probleme ilişkin istenilen açıklamalar kısmen eksik.	4
Probleme uygun başlanılmış fakat sonuca ulaşılamamış, yarım bırakılmış.	3
Problemin anlaşıldığına ilişkin net(açık) göstergeler yok. Büyük matematiksel hatalar (oran orantı kurma, denklem çözümü vb.) yapılmış. Problemin çözümünde kullanılan yöntem uygun değil.	2
Problemin anlaşıldığına ilişkin herhangi bir gösterge yok. Çözüm yok, sadece doğru yanıt var.	1
Yanıtlanmamış (boş bırakılmış).	0

3. maddeye ait dereceli puanlama anahtarı

Kriterler	**Derecelendirme**
Hayır ya da kabul edilemez ifadelerini kullanmış. Grafiğin sadece küçük bir parçasının gösterilmiş olması gerçeği üzerinde odaklanmış ve tüm grafik üzerinden kıyaslamalar yapmış. Doğru sayısal birimleri dayanak olarak kullanmış, tam ve uygun açıklamalar ile desteklemiş.	5
Hayır ya da kabul edilemez ifadelerini kullanmış. Sadece problemde geçen iki farklı yıl arasında kıyaslamalar yapmış. Doğru sayısal birimleri dayanak olarak kullanmış fakat yaptığı açıklamalar tatmin edici düzeyde değil.	4
Hayır ya da kabul edilemez ifadelerini kullanmış. Yanlış sayısal birimleri dayanak olarak kullanmış. Yaptığı açıklamalar tatmin edici düzeyde değil.	3
Hayır ya da kabul edilemez ifadelerini kullanmış. Sayısal herhangi bir birime yer vermemiş. Eksik ya da yanlış açıklamalar yapmış.	2
Sadece; Hayır ya da kabul edilemez ifadelerini kullanmış. Herhangi destekleyici bir açıklama yapılmamış.	1
Yanıtlanmamış (boş bırakılmış). Evet ya da kabul edilebilir ifadelerini kullanmış.	0

4. maddeye ait dereceli puanlama anahtarı

Kriterler	**Derecelendirme**
Doğru yanıtı yazmış. Verilen yanıtı tam ve uygun açıklamalar içeren 2 neden ile desteklemiş.	5
Doğru yanıtı yazmış. Verilen yanıtı destekleyen nedenlerden biri tam ve uygun açıklamalar içerirken, diğer neden tatmin edici düzeyde açıklamalar içermemekte.	4
Doğru yanıtı yazmış. Verilen yanıtı destekleyen nedenlerden biri tam ve uygun açıklamalar içerirken, diğer neden yanlış açıklamalar içermekte	3
Doğru yanıtı yazmış. Verilen yanıtı destekleyen nedenlerden biri tatmin edici düzeyde açıklamalar içermezken, diğer neden yanlış açıklamalar içermekte.	2
Doğru yanıtı yazmış. Verilen yanıtı yanlış açıklamalar içeren 2 neden ile desteklemiş.	1
Yanıtlanmamış (boş bırakılmış). Yanlış yanıt yazılmış.	0

7. maddeye ait dereceli puanlama anahtarı

Kriterler	**Derecelendirme**
Verilen iki yanıtta doğru. Her iki yanıtı da tam ve uygun açıklamalar ile desteklemiş.	5
Verilen iki yanıtta doğru. Yapılan her iki açıklama da tatmin edici düzeyde değil.	4
Verilen iki yanıtta doğru. Yapılan açıklamalardan biri yanlış/uygun olmayan, diğeri ise tatmin edici düzeyde değil. Ya da; Verilen yanıtlardan biri doğru diğeri yanlış. Doğru olan yanıt tam ve uygun açıklamalar ile desteklemiş.	3
Verilen iki yanıtta doğru. Her iki yanıtı da yanlış/uygun olmayan açıklamalar ile desteklemiş.	2
Verilen iki yanıtta doğru. Yanıtlara ilişkin herhangi bir açıklama yapılmamış. Ya da; Verilen iki yanıttan biri doğru. Yanıtlara ilişkin herhangi bir açıklama yapılmamış ve ya yapılan açıklamalar yanlış.	1
Yanıtlanmamış (boş bırakılmış). Yanlış yanıt yazılmış.	0

8. maddeye ait dereceli puanlama anahtarı

Kriterler	**Derecelendirme**
Verilen yanıta binanın 21 kat olduğu öngörülerek ulaşılmış. Her bir kat için öngörülen yükseklik 2,4 m ve üstü ile 4,3 m ve altı aralığında yer alıyor. Yanıt tam ve uygun açıklamalar ile desteklemiş.	5
Verilen yanıta binanın 21 kat olduğu öngörülerek ulaşılmış. İşlem hatalarından (çarpma, bölme, toplama, çıkarma) dolayı doğru sonuca ulaşılamamış. Her bir kat için öngörülen yükseklik 2,4 m ve üstü ile 4,3 m ve altı aralığında yer alıyor. Yapılan açıklamalar tatmin edici düzeyde değil.	4
Verilen yanıta binanın 20 kat olduğu öngörülerek ulaşılmış. Her bir kat için öngörülen yükseklik 2,4 m ve üstü ile 4,3 m ve altı aralığında yer alıyor. Yanıt tam ve uygun açıklamalar ile desteklemiş.	3
Verilen yanıta binanın 21 kat olduğu öngörülerek ulaşılmış. Her bir kat için öngörülen yükseklik 2,4 m ve üstü ile 4,3 m ve altı aralığında yer alıyor. Herhangi bir açıklama yapılmamış ve ya yapılan açıklama yanlış.	2
Verilen yanıta binanın 20 kat olduğu öngörülerek ulaşılmış. Her bir kat için öngörülen yükseklik 2,4 m ve üstü ile 4,3 m ve altı aralığında yer alıyor. Herhangi bir açıklama yapılmamış ve ya yapılan açıklama yanlış.	1
Yanıtlanmamış (boş bırakılmış). Her bir kat için öngörülen yükseklik 2,4 m ve altı ile 4,3 m ve üstü aralığında yer alıyor.	0

EK C

Maddelere Ait Ham Veriler

3. maddeye ait ham veriler

Puan Kombinasyonu	Puanlayıcı Kombinasyonu										
	1-2	**1-3**	**1-4**	**1-5**	**2-3**	**2-4**	**2-5**	**3-4**	**3-5**	**4-5**	**Toplam**
0-0	34	34	34	34	34	34	34	34	34	34	340
0-1	0	0	0	0	0	0	0	0	0	0	0
0-2	0	0	0	0	0	0	0	0	0	0	0
0-3	0	0	0	0	0	0	0	0	0	0	0
0-4	0	0	0	0	0	0	0	0	0	0	0
0-5	0	0	0	0	0	0	0	0	0	0	0
1-0	0	0	0	0	0	0	0	0	0	0	0
1-1	33	33	28	33	50	35	57	34	51	35	389
1-2	0	0	5	0	23	39	17	16	0	0	100
1-3	0	0	0	0	1	0	0	0	0	0	1
1-4	0	0	0	0	0	0	1	0	0	0	1
1-5	0	0	0	0	0	0	0	0	0	0	0
2-0	0	0	0	0	0	0	0	0	0	0	0
2-1	40	18	7	25	0	0	0	1	7	22	120
2-2	7	30	40	23	49	59	66	68	69	78	489
2-3	3	2	1	1	26	13	8	11	3	3	71
2-4	0	0	2	1	1	4	2	0	1	1	12
2-5	0	0	0	0	0	0	0	0	0	0	0
3-0	0	0	0	0	0	0	0	0	0	0	0
3-1	1	0	0	0	1	0	1	0	0	0	3
3-2	58	45	53	63	8	6	10	20	23	15	301
3-3	18	31	17	10	75	24	50	17	40	19	301
3-4	0	1	7	4	26	80	49	66	40	3	276
3-5	0	0	0	0	2	2	2	0	0	0	6
4-0	0	0	0	0	0	0	0	0	0	0	0
4-1	0	0	0	0	0	0	0	0	0	1	1
4-2	11	5	6	7	0	0	0	0	1	0	30
4-3	86	69	19	46	1	0	1	9	16	37	284
4-4	9	30	76	52	26	25	32	38	31	69	388
4-5	0	2	5	1	16	18	10	11	10	7	80
5-0	0	0	0	0	0	0	0	0	0	0	0
5-1	0	0	0	0	0	0	0	0	0	0	0
5-2	0	0	0	0	0	0	0	0	0	0	0
5-3	5	1	0	2	0	0	0	0	0	0	8
5-4	34	27	29	29	5	5	3	9	14	13	168
5-5	11	22	21	19	6	6	8	15	10	13	131

4. maddeye ait ham veriler

Puan	**Puanlayıcı Kombinasyonu**										
Kombinasyonu	**1-2**	**1-3**	**1-4**	**1-5**	**2-3**	**2-4**	**2-5**	**3-4**	**3-5**	**4-5**	**Toplam**
0-0	37	37	37	37	37	37	37	37	37	37	370
0-1	0	0	0	0	0	0	0	0	0	0	0
0-2	0	0	0	0	0	0	0	0	0	0	0
0-3	0	0	0	0	0	0	0	0	0	0	0
0-4	0	0	0	0	0	0	0	0	0	0	0
0-5	0	0	0	0	0	0	0	0	0	0	0
1-0	0	0	0	0	0	0	0	0	0	0	0
1-1	41	41	25	41	79	42	89	35	79	42	514
1-2	0	0	16	0	10	47	0	44	0	0	117
1-3	0	0	0	0	0	0	0	0	0	0	0
1-4	0	0	0	0	0	0	0	0	0	0	0
1-5	0	0	0	0	0	0	0	0	0	0	0
2-0	0	0	0	0	0	0	0	0	0	0	0
2-1	48	38	17	48	0	0	1	7	11	47	217
2-2	0	10	31	0	1	0	0	3	0	1	46
2-3	0	0	0	0	1	2	1	1	0	0	5
2-4	0	0	0	0	0	0	0	0	0	0	0
2-5	0	0	0	0	0	0	0	0	0	0	0
3-0	0	0	0	0	0	0	0	0	0	0	0
3-1	0	0	0	1	0	0	0	0	0	1	2
3-2	2	1	1	1	0	1	1	1	1	1	10
3-3	1	2	2	1	2	1	1	2	2	3	17
3-4	0	0	0	0	1	1	1	0	0	0	3
3-5	0	0	0	0	1	1	1	0	0	0	3
4-0	0	0	0	0	0	0	0	0	0	0	0
4-1	0	0	0	0	0	0	0	0	0	0	0
4-2	0	0	0	0	0	0	1	0	0	0	1
4-3	1	1	1	1	0	1	0	0	0	0	5
4-4	2	1	1	3	6	10	29	3	7	11	73
4-5	5	6	6	4	64	59	40	8	4	9	205
5-0	0	0	0	0	0	0	0	0	0	0	0
5-1	0	0	0	0	0	0	0	0	0	0	0
5-2	0	0	0	1	0	0	0	0	1	0	2
5-3	2	0	2	2	0	1	2	2	2	1	14
5-4	68	10	19	91	4	9	64	17	87	83	452
5-5	143	203	192	119	144	138	82	190	119	114	1444

5. maddeye ait ham veriler

Puan	**Puanlayıcı Kombinasyonu**										
Kombinasyonu	**1-2**	**1-3**	**1-4**	**1-5**	**2-3**	**2-4**	**2-5**	**3-4**	**3-5**	**4-5**	**Toplam**
0-0	1	1	1	1	2	1	1	1	1	1	11
0-1	0	0	0	0	0	0	0	0	0	0	0
0-2	0	0	0	0	0	0	1	0	1	0	2
0-3	0	0	0	0	0	1	0	1	0	0	2
0-4	0	0	0	0	0	0	0	0	0	0	0
0-5	0	0	0	0	0	0	0	0	0	0	0
1-0	0	0	0	0	0	0	0	0	0	0	0
1-1	25	20	5	17	29	5	33	6	20	6	166
1-2	0	6	21	9	24	48	20	24	10	0	162
1-3	0	0	0	0	0	0	0	0	0	0	0
1-4	0	0	0	0	0	0	0	0	0	0	0
1-5	1	0	0	0	0	0	0	0	0	0	1
2-0	1	1	0	0	0	0	0	0	0	0	2
2-1	27	10	1	18	0	0	1	0	15	29	101
2-2	8	23	29	15	17	15	18	34	22	29	210
2-3	0	2	5	3	14	13	10	6	4	6	63
2-4	0	0	1	0	1	4	3	1	0	1	11
2-5	0	0	0	0	0	0	0	0	0	0	0
3-0	0	0	0	0	0	0	0	0	0	0	0
3-1	1	0	0	0	0	0	0	0	0	0	1
3-2	19	10	11	13	0	2	0	7	6	10	78
3-3	4	15	9	9	15	4	13	12	17	10	108
3-4	2	1	6	4	10	18	12	23	19	0	95
3-5	0	0	0	0	0	1	0	0	0	0	1
4-0	0	0	0	0	0	0	0	0	0	0	0
4-1	0	0	0	0	0	0	0	0	0	0	0
4-2	5	2	4	2	0	0	0	0	0	0	13
4-3	20	24	5	13	13	2	4	1	5	11	98
4-4	14	12	29	25	8	17	24	19	21	35	204
4-5	4	5	5	3	27	29	20	13	7	3	116
5-0	0	0	0	0	0	0	0	0	0	0	0
5-1	0	0	0	0	1	1	1	0	0	0	3
5-2	0	0	0	0	0	0	0	0	0	0	0
5-3	1	1	1	2	0	0	0	0	1	0	6
5-4	32	20	13	34	14	10	24	6	23	27	203
5-5	185	197	204	182	175	179	165	196	178	182	1843

6. maddeye ait ham veriler

Puan	**Puanlayıcı Kombinasyonu**										
Kombinasyonu	**1-2**	**1-3**	**1-4**	**1-5**	**2-3**	**2-4**	**2-5**	**3-4**	**3-5**	**4-5**	**Toplam**
0-0	13	13	13	13	13	13	13	13	13	13	130
0-1	0	0	0	0	0	0	0	0	1	1	2
0-2	0	0	0	0	0	0	0	2	1	0	3
0-3	0	0	0	0	0	0	0	0	0	0	0
0-4	0	0	0	0	0	0	0	0	0	0	0
0-5	0	0	0	0	0	0	0	0	0	0	0
1-0	0	0	0	0	1	1	0	1	0	0	3
1-1	24	21	9	22	37	14	47	13	35	13	235
1-2	3	5	18	5	20	43	11	23	2	1	131
1-3	0	1	0	0	0	0	0	0	0	0	1
1-4	0	0	0	0	0	0	0	0	0	0	0
1-5	0	0	0	0	0	0	0	0	0	0	0
2-0	0	1	1	0	1	0	0	0	0	0	3
2-1	34	16	5	29	0	0	6	1	15	39	145
2-2	71	49	64	71	64	89	118	66	68	95	755
2-3	0	38	35	5	74	51	16	22	6	16	263
2-4	0	1	0	0	1	0	0	0	0	0	2
2-5	0	0	0	0	0	0	0	0	0	0	0
3-0	0	1	0	0	0	0	0	0	0	0	1
3-1	0	0	0	2	0	0	0	0	2	0	4
3-2	63	33	65	63	5	18	13	59	71	46	436
3-3	23	49	21	21	24	14	26	37	25	21	261
3-4	0	3	0	0	13	6	3	2	0	0	27
3-5	0	0	0	0	13	17	13	1	1	0	45
4-0	0	0	0	0	0	0	0	0	0	0	0
4-1	0	0	0	0	0	0	0	0	0	0	0
4-2	3	2	3	3	0	0	0	0	0	0	11
4-3	15	9	10	15	0	2	2	8	12	5	78
4-4	3	10	6	2	14	2	7	5	5	1	55
4-5	4	4	6	5	18	28	23	23	19	2	132
5-0	0	0	0	0	0	0	0	0	0	0	0
5-1	0	0	0	0	0	0	0	0	0	0	0
5-2	0	0	0	0	0	0	0	0	0	0	0
5-3	17	2	1	4	1	0	1	0	2	3	31
5-4	29	22	2	12	8	0	4	1	9	13	100
5-5	48	70	91	78	43	52	47	73	63	81	646

7. maddeye ait ham veriler

Puan	**Puanlayıcı Kombinasyonu**										
Kombinasyonu	**1-2**	**1-3**	**1-4**	**1-5**	**2-3**	**2-4**	**2-5**	**3-4**	**3-5**	**4-5**	**Toplam**
0-0	9	9	9	9	10	10	10	10	10	10	96
0-1	0	0	0	0	0	0	0	0	0	0	0
0-2	0	0	0	0	0	0	0	0	0	0	0
0-3	0	0	0	0	0	0	0	0	0	0	0
0-4	0	0	0	0	0	0	0	0	0	0	0
0-5	0	0	0	0	0	0	0	0	0	0	0
1-0	0	0	0	0	0	0	0	0	0	0	0
1-1	6	6	6	6	9	9	12	16	16	23	109
1-2	0	0	0	0	3	3	0	1	1	2	10
1-3	0	0	0	0	0	0	0	0	0	0	0
1-4	0	0	0	0	0	0	0	0	0	0	0
1-5	0	0	0	0	0	0	0	0	0	0	0
2-0	0	0	0	0	0	0	0	0	0	0	0
2-1	6	11	18	22	8	16	18	9	13	7	128
2-2	25	20	13	9	97	39	164	27	87	35	516
2-3	0	0	0	0	80	131	4	64	0	0	279
2-4	0	0	0	0	1	0	0	0	0	0	1
2-5	0	0	0	0	0	0	0	0	0	0	0
3-0	1	1	1	1	0	0	0	0	0	0	4
3-1	0	0	0	1	0	0	0	0	1	0	2
3-2	159	79	28	155	0	0	1	13	76	128	639
3-3	3	83	134	6	6	6	5	73	9	9	334
3-4	0	0	0	0	0	0	0	0	0	0	0
3-5	0	0	0	0	0	0	0	0	0	0	0
4-0	0	0	0	0	0	0	0	0	0	0	0
4-1	0	0	1	1	0	0	0	0	0	0	2
4-2	2	1	1	1	0	0	0	1	1	0	7
4-3	3	3	3	3	0	0	0	0	0	0	12
4-4	6	18	10	13	6	8	10	9	14	11	105
4-5	17	6	13	10	8	6	4	12	7	4	87
5-0	0	0	0	0	0	0	0	0	0	0	0
5-1	0	0	0	1	0	0	1	0	1	1	4
5-2	0	0	0	0	0	0	0	0	0	0	0
5-3	0	0	0	0	0	0	0	0	0	0	0
5-4	8	4	5	41	15	7	44	6	40	43	213
5-5	105	109	108	71	107	115	77	109	74	77	952

8. maddeye ait ham veriler

Puan	**Puanlayıcı Kombinasyonu**										
Kombinasyonu	**1-2**	**1-3**	**1-4**	**1-5**	**2-3**	**2-4**	**2-5**	**3-4**	**3-5**	**4-5**	**Toplam**
0-0	6	6	6	6	7	6	7	6	7	6	63
0-1	0	0	0	0	0	0	0	0	0	0	0
0-2	0	0	0	0	0	0	0	0	0	0	0
0-3	0	0	0	0	0	1	0	1	0	0	2
0-4	0	0	0	0	0	0	0	0	0	0	0
0-5	0	0	0	0	0	0	0	0	0	0	0
1-0	0	0	0	0	0	0	0	0	0	0	0
1-1	30	29	12	30	57	27	63	23	57	26	354
1-2	4	5	19	4	11	38	5	34	3	2	125
1-3	0	0	2	0	0	2	0	2	0	0	6
1-4	0	0	0	0	0	0	0	0	0	0	0
1-5	0	0	1	0	0	1	0	1	0	0	3
2-0	1	1	0	1	0	0	1	0	1	0	5
2-1	37	30	16	34	3	1	2	5	8	38	174
2-2	15	21	26	18	33	17	47	21	40	22	260
2-3	0	1	10	0	15	32	1	23	1	2	85
2-4	0	0	0	0	0	0	0	0	0	0	0
2-5	0	0	1	0	0	1	0	1	0	0	3
3-0	0	0	0	1	0	0	0	0	0	2	3
3-1	1	1	0	3	0	0	2	0	2	2	11
3-2	32	24	17	42	6	7	12	7	21	39	207
3-3	24	31	40	12	22	38	33	29	14	30	273
3-4	1	2	1	0	18	2	0	1	0	0	25
3-5	0	0	0	0	1	0	0	0	0	0	1
4-0	0	0	0	0	0	0	0	0	0	0	0
4-1	0	0	0	0	0	0	0	0	0	0	0
4-2	0	0	0	0	0	0	0	0	0	0	0
4-3	22	4	20	22	0	0	3	17	21	9	118
4-4	3	21	3	4	23	40	50	14	22	43	223
4-5	2	2	4	1	42	25	12	22	10	1	121
5-0	0	0	0	0	0	0	0	0	0	0	0
5-1	0	0	0	0	0	0	0	0	0	1	1
5-2	0	0	0	0	0	0	0	0	0	1	1
5-3	1	1	18	0	0	5	1	6	1	1	34
5-4	61	30	49	66	12	11	20	38	48	27	362
5-5	110	141	122	98	100	101	87	104	89	98	1050

ÖZGEÇMİŞ

Adı ve Soyadı :Müge ULUMAN

Doğum Tarihi :02.01.1987

İletişim Bilgileri :03123633350/5018

E posta :mugeulumann@gmail.com :

Öğrenim Durumu Doktora

Derece	**Bölüm/Program**	**Üniversite**	**Yıl**
Lisans	İlköğretim Bölümü, Sınıf Öğretmenliği Programı	Abant İzzet Baysal Üniversitesi	2004-2007

Derece	**Bölüm/Program**	**Üniversite**	**Yıl**
Yüksek Lisans	Eğitim Bilimleri Bölümü, Ölçme ve Değerlendirme Programı	Abant İzzet Baysal Üniversitesi	2007-2009

İş Deneyimi **:**

Unvan	Görev Yeri	Yıl
Arş. Gör.	Kastamonu Üniversitesi Eğitim Fakültesi	2010-2012
Arş. Gör.	Marmara Üniversitesi Eğitim Fakültesi	2012-2013
Misafir Araştırmacı	Ohio State University Educational Faculty	2014-2015
Arş. Gör. (ÖYP kapsamında 35. maddeyle)	Ankara Üniversitesi Eğitim Bilimleri Fakültesi	2013-