Hacettepe Üniversitesi Eğitim Fakültesi Dergisi (H. U. Journal of Education) 31(2): 346-368 [2016]

ISSN: 1300-5340
DOI:10.16986/HUJE.2016015182

Açık Uçlu Sorularla Yapılan Ölçmelerde Klasik Test Kuramı ve Çok Yüzeyli Rasch Modeline Göre Hesaplanan Yetenek Kestirimlerinin Karşılaştırılması

A Comparison of the Ability Estimations of Classical Test Theory and the Many Facet Rasch Model in Measurements with Open-ended Questions

Mustafa İLHAN*

ÖZ: Bu araştırmada, açık uçlu sorularla yapılan ölçmelerde klasik test kuramı (KTK) ve çok yüzeyli Rasch modeline (ÇYRM) göre hesaplanan yetenek kestirimlerinin karşılaştırılması amaçlanmıştır. Araştırma sekizinci sınıfa devam eden 100 öğrenci ile çalışmada puanlayıcı olarak görev alan dört matematik öğretmeninin katılımıyla gerçekleşmiştir. Çalışmanın verileri, açık uçlu altı matematik sorusundan oluşan bir başarı testi ve bu soruların puanlanmasında kullanılan bütüncül bir rubrik yardımıyla toplanmıştır. Açık uçlu matematik sorularına verilen yanıtların puanlanmasıyla elde edilen veriler, hem KTK'ya hem de ÇYRM'ye göre analiz edilmiştir. Dört puanlayıcının verdiği puanların ortalaması alınarak, KTK'ya ilişkin yetenek kestirimleri hesaplanmıştır. Ardından puanlayıcı, birey ve madde şeklinde üç yüzeyli bir desen ile çok yüzeyli Rasch analizi uygulanmıştır. Rasch analizinde ulaşılan ve logit cetvelinde rapor edilen yetenek kestirimlerinin puanlamada kullanılan rubriğin birimlerine dönüştürülmesiyle, iki kurama göre hesaplanan yetenek kestirimleri karşılaştırmaya hazır hale gelmiştir. Araştırma sonucunda; KTK ve ÇYRM'ye göre hesaplanan yetenek kestirimleri arasındaki göreli uyumun son derece yüksek olduğu belirlenmiştir. İki kurama göre hesaplanan yetenek kestirimlerine ait ortalamalar arasında anlamlı fark bulunduğu ve dolayısıyla mutlak bir uyumun söz konusu olmadığı saptanmıştır. ÇYRM'de rapor edilen yetenek kestirimlerinin ölçüt geçerliğinin KTK'dan elde edilen yetenek kestirimlerine kıyasla daha yüksek olduğu tespit edilmiştir.

Anahtar sözcükler: Açık uçlu sorular, klasik test kuramı, çok yüzeyli Rasch modeli, yetenek kestirimi

ABSTRACT: The purpose of this study is to compare the ability estimations of classical test theory (CTT) and the Many Facet Rasch model (MFRM) in measurements conducted with open-ended questions. The study was conducted with 100 eighth graders and four mathematics teachers who rated the students' work. The study's data were obtained using an achievement test with 6 open-ended mathematics questions and a holistic rubric for scoring these questions. The data obtained by rating the open-ended mathematics questions were analyzed using both CTT and MFRM. The ability estimations for CTT were calculated using the score averages of the four raters. Afterwards, Many Facet Rasch analysis was conducted using a three-facet pattern including raters, students and items. The ability estimations calculated by these two theories were prepared to be compared by converting ability estimations obtained from the Rasch analysis and reported in the logit scale into the units of rubric used for scoring. This study found that the relative agreement between the ability estimations calculated using CTT and MFRM was extremely high. It was determined that there was a significant difference between the means of the ability estimations of the two theories, and thus no absolute agreement. Research findings revealed that the concurrent validity of the ability estimations of MFRM is higher than that of CTT's ability estimations.

Keywords: Open-ended questions, classical test theory, many facet Rasch model, ability estimation

1. GİRİŞ

Eğitimde yaygın biçimde kullanılan ölçme araçlarından biri açık uçlu maddelerdir (Güler, 2014). Açık uçlu maddelerde öğrenci cevabı kendisi yapılandırmakta, cevabının gerekçelerini

*Arş. Gör. Dr., Dicle Üniversitesi, Ziya Gökalp Eğitim Fakültesi, Diyarbakır-Türkiye.
e-posta: mustafailhan21@gmail.com

açıklama fırsatı bulmakta ve düşüncelerini daha özgür bir biçimde ifade edebilmektedir (Gronlund, 1998). Açık uçlu maddeler bu yönüyle; problem çözme, problemleri organize etme, yeni ve orijinal fikirler üretme, fikirleri değerlendirme, bilgileri değişik durumlarda işe koşma, neden-sonuç ilişkileri kurma, genellemeler yapma, hipotez üretme, alternatifler arasında karşılaştırmalar yaparak bir yargıya varma gibi üst düzey becerilerin ölçülmesi için en uygun soru türüdür (Tan ve Erdoğan, 2004). Açık uçlu soruların çoktan seçmeli testlerin ötesinde birçok avantajı bulunmaktadır. Açık uçlu soruların ölçme değerlendirme süreçleri açısından sağladığı avantajların çoktan seçmeli testler ile karşılaştırılarak açıklanması, çoktan seçmeli maddelerin gerek sınıf içi değerlendirmelerde gerekse de geniş ölçekli testlerde sıklıkla tercih edilen bir madde türü olmasından kaynaklanmaktadır. Açık uçlu soruların ilk avantajı, çoktan seçmeli testler ile ölçülemeyen bazı üst düzey becerileri etkili bir biçimde ölçebilmesidir (Bahar, Nartgün, Durmuş ve Bıçak, 2010). Şans başarısını ortadan kaldırarak ölçme hatasını azaltması (Turgut ve Baykul, 2012), kısmi puanlamaya uygun oluşu (Bağcan Büyükturan ve Çıkrıkçı Demirtaşlı, 2013) ve çoktan seçmeli maddelere kıyasla daha kolay hazırlanabilmesi açık uçlu maddelerin diğer üstün yönleridir (Başol, 2013). Son olarak, çoktan seçmeli maddelerde öğrenciler eleme yöntemiyle doğru cevaba ulaşabilirken; açık uçlu sorularda böyle bir durum söz konusu değildir (Braun, Bennett, Frye ve Soloway, 1990). Örneğin, $2(X + 4) = 38 – X$ şeklindeki bir cebir problemi çoktan seçmeli soru formatında öğrencilere sunulduğu takdirde, öğrenciler seçeneklerde yer alan X değerlerini sırasıyla yerine koyarak eşitliği sağlayan X değerinin hangisi olduğunu belirleyebilmekte ve doğru cevaba bu şekilde ulaşabilmektedir. Buna bağlı olarak, maddenin çoktan seçmeli soru formatında sunulması, öğrencinin soru ile ölçülmesi hedeflenen özelliği ne derece kazandığını ortaya koymada yetersiz kalabilmektedir (Bridgeman, 1992).

Yukarıda sıralanan avantajlarının yanı sıra açık uçlu maddelerin bir takım dezavantajları bulunmaktadır. Bu dezavantajlarından ilki, uygulanmasının ve puanlanmasının zaman alıcı olmasıdır (Doğan, 2013). Açık uçlu soruların ikinci dezavantajı, kapsam geçerliğini sağlamanın zor oluşudur. Süre sıkıntısından dolayı, açık uçlu maddelerden oluşan sınavlarda sorulabilecek soru sayısı çoktan seçmeli testlere göre daha azdır. Az sayıda soru ile yetinme zorunluluğu açık uçlu maddelerden oluşan testlerde kapsam geçerliğini sağlamayı güçleştirmektedir (Özçelik, 2011). Açık uçlu maddelerin en önemli dezavantajı ise çoktan seçmeli maddeler gibi objektif bir biçimde puanlanamamasıdır (Romagnano, 2001). Öğrencilerin herhangi açık uçlu bir maddeden aldıkları puan, puanlamayı yapan kişiye göre farklılık gösterebilmektedir. Dolayısıyla, açık uçlu maddelerde puanlayıcılar ölçme sonuçlarında hataya yol açan bir değişkenlik kaynağı olmaktadır (Tekin, 2009). Bu durum, açık uçlu maddelerde puanlayıcı kaynaklı faktörlerin ölçme sonuçlarını nasıl etkilediğinin ortaya konulmasını yani puanlayıcı güvenirliğinin belirlenmesini gerekli kılmaktadır (Atılgan, 2004; Kan, 2013). Puanlayıcı güvenirliğinin belirlenmesinde kullanılabilecek çok sayıda farklı yöntem bulunmaktadır. Bu yöntemler; klasik test kuramına (KTK) dayalı yöntemler, genellenebilirlik kuramına dayalı yöntemler ve madde tepki kuramına (MTK) dayalı yöntemler olmak üzere üç başlıkta incelenmektedir (MacMillan, 2000). Ancak bu araştırmada, genellenebilirlik kuramına dayalı yöntemlerden söz edilmemiş; KTK ve MTK'ya dayalı yöntemler üzerinde durulmuştur.

1.1. KTK'ya Dayalı Yöntemler

KTK, bir ölçüme ilişkin gözlenen puanın, gerçek puan ve hata puanı bileşenlerinden meydana geldiği varsayımı üzerine inşa edilmiştir. Bu varsayım, $X = T + E$ eşitliği ile gösterilmektedir (Brennan, 2000). Eşitlikteki X değeri gözlenen puanı yani bireyin testten aldığı puanı temsil etmektedir. T gerçek puana karşılık gelmektedir. Gerçek puan, bir testin bireye sonsuz sayıda uygulanması ve bu uygulamalar arasına öğrenme ya da başka faktörlerin etkisinin karışmaması halinde, bireyin testten alacağı puanların aritmetik ortalamasıdır (Baykul, 2010). Ancak bir testin bireye sonsuz sayıda uygulanması mümkün olmadığından, gerçek puan

hipotetik bir kavramdır (Kline, 2005). E ise hata bileşenini göstermektedir. Puanlama işlemine; maddelerin açık olmaması, madde sayısının yetersiz olması gibi ölçme aracı ile ilgili faktörlerden; yorgunluk, kaygı, dikkat dağınıklığı gibi bireysel faktörlerden ya da açık olmayan yönergeler, zamanın yeterli olmayışı gibi ölçme aracının uygulanması ile ilgili faktörlerden dolayı karışan hatalar, hata varyansını meydana getirmektedir. Objektif olarak puanlanamayan testlerde hata varyansına neden olan etkenlerden biri de puanlayıcılar ile ilgili faktörlerdir. KTK'da, puanlayıcı güvenirliğinin tespitinde; uyuşma yüzdesi, kappa istatistiği, Pearson korelasyon katsayısı ve ortalamaların karşılaştırılması gibi farklı teknikler kullanılabilmektedir (Goodwin, 2001).

Uyuşma yüzdesi puanlayıcılar arasındaki uyumun yüzdesi olup, puanlayıcıların görüş birliği içerisinde oldukları puanlama sayısının toplam puanlama sayısına bölünmesiyle elde edilmektedir (Graham, Milanowski ve Miller, 2012). KTK'da puanlayıcı güvenirliğinin belirlenmesinde kullanılan ikinci bir yöntem kappa istatistiğidir. Kappa istatistiği, uyuşma yüzdesine benzemesine rağmen; puanlayıcılar arasındaki uyumun bir kısmının şanstan kaynaklandığını kabul etmesi ve şanstan kaynaklanan uyumu kontrol altına almak için düzeltme işlemi uygulaması yönüyle uyuşma yüzdesinden ayrılmaktadır (David, 2008). Puanlayıcı güvenirliğini incelemek için kullanılan KTK'ya dayalı tekniklerden biri de Pearson korelasyon katsayısıdır. Bu katsayı, iki puanlayıcının ölçülmek istenen özellik açısından bireyler arasında yaptıkları sıralamanın tutarlılığına ilişkin bir ölçü vermektedir (LeBreton ve Senter, 2008). KTK'da puanlayıcı güvenirliğini saptamak için sıklıkla kullanılan bir diğer teknik de puanlayıcı ortalamalarının karşılaştırılmasıdır. Ortalamalar karşılaştırılırken; iki puanlayıcı olması durumunda ilişkili örneklem *t*-testi, ikiden fazla sayıda puanlayıcı olması durumunda ise tekrarlı ölçümler için varyans analizi kullanılmaktadır (Goodwin, 2001).

1.2. MTK'ya Dayalı Yöntemler

MTK, bireylerin belirli bir alanda doğrudan gözlenemeyen yetenekleri ile ilgili alandaki yeteneklerini ölçmek için hazırlanan test maddelerine verdikleri cevaplar arasındaki ilişkiyi matematiksel olarak ortaya koymaya çalışan bir kuramdır (DeMars, 2010; Urbina, 2004). MTK'ya göre; bir maddenin parametreleri o maddeyi yanıtlayan cevaplayıcı grubundan bağımsız olarak elde edilebildiği gibi, bireylerin yetenek düzeyleri de uygulanan testteki madde örnekleminden bağımsız olarak kestirilebilmektedir (Hambleton, Swaminathan ve Rogers, 1991). MTK, üç ayrı modelden oluşmaktadır. Bu modeller; madde güçlük parametresini içeren bir parametreli model; madde güçlük ve ayırt edicilik parametrelerini içeren iki parametreli model ve madde güçlüğü, madde ayırt ediciliği ile şans parametresini içeren üç parametreli model şeklinde sıralanmaktadır (Van der Linden ve Hambleton, 1997). Bununla birlikte, Rasch modeli olarak da adlandırılan bir parametreli model MTK'nın en temel modelidir (Harvey ve Hammer, 1999). Rasch modeli, ilk olarak doğru/yanlış şeklinde puanlanabilen iki kategorili (1-0) ölçme araçları için geliştirilmiştir (Haiyang, 2010). Ancak sonraları, kısmi puanlama modeli (Andrich, 1978) ve sıralama ölçekli model (Masters, 1982) gibi temel Rasch modelinin farklı uzantıları ileri sürülmüştür. Temel Rasch modelinin uzantılarından biri de; Amerikalı istatistikçi Linacre (1989) tarafından geliştirilen Çok Yüzeyli Rasch Modelidir (ÇYRM).

ÇYRM; bireylerin yetenek düzeyleri ve maddelerin güçlük düzeylerinin yanı sıra puanlayıcılar, puanlama ölçütleri, puanlama anahtarı vb. gibi test puanlarını etkileme olasılığına sahip diğer değişkenlik kaynaklarının da dikkate alınmasını sağlayan bir model olarak tanımlanmaktadır (Lynch ve McNamara, 1998). ÇYRM'de, test puanlarını etkileme potansiyeli bulunan değişkenlik kaynaklarının her biri yüzey olarak adlandırılmaktadır (Sudweeks, Reeve ve Bradshaw, 2005). Örneğin; bir matematik sınavında öğrencilere açık uçlu 10 maddeden oluşan bir test verilmiş ve öğrencilerin sınav kâğıtları üç farklı puanlayıcı tarafından bağımsız olarak

puanlamış olsun. Burada; öğrenciler, maddeler ve puanlayıcılar ölçme sonuçlarını etkileyebilecek değişkenlik kaynaklarıdır. Dolayısıyla; öğrenci yüzeyi, madde yüzeyi ve puanlayıcı yüzeyi şeklinde üç yüzeyli bir model söz konusudur (İlhan, 2015). ÇYRM'de, puanlayıcılar ile ilgili faktörlerin bireyin test puanlarında değişkenliğe yol açabilecek bir yüzey olarak işlem görmesi, bu modeli öznel olarak puanlanan açık uçlu sorular için uygun bir seçenek haline getirmektedir (Mulqueen, Baker ve Dismukes, 2000).

1.3. Açık Uçlu Sorularla Yapılan Ölçmeler için KTK'da ve ÇYRM'de Yetenek Kestirimi

Açık uçlu sorularla yapılan ölçmelerde; puanlayıcı güvenirliğinin belirlenmesinde olduğu gibi yetenek kestirimlerinin hesaplanmasında da KTK ve ÇYRM arasında farklılıklar bulunmaktadır. KTK'ya göre; sınava giren bir katılımcının ham puanı, testin her bir maddesinden aldığı puanların toplamına eşittir (Baker, 2001). Ölçme işleminde çoktan seçmeli sorulardan oluşan bir test kullanılması durumunda, bu puanlama otomatik olarak yapılabilmekte ya da puanlamanın tek bir puanlayıcı tarafından yapılması yeterli olmaktadır. Öte yandan, açık uçlu sorulardan oluşan testlerde öğrencilerin yetenek düzeylerine ilişkin güvenilir kestirimler elde edebilmek için değerlendirme sürecinde birden fazla puanlayıcı kullanılması ve farklı puanlayıcılar tarafından verilen puanların aritmetik ortalamasının alınması gerekmektedir (Ebel, 1951). Buna göre, açık uçlu sorularla yapılan ölçmelerde KTK'ya ilişkin yetenek kestirimleri hesaplanırken, öğrencinin testin her bir maddesinden aldığı puanlar toplanmalı, bu işlem puanlama sürecinde görev alan tüm puanlayıcılar için tekrarlanmalı ve farklı puanlayıcılar tarafından verilen puanların aritmetik ortalaması hesaplanmalıdır. Ancak KTK, bu işlem sonucunda bireylerin hangi güvenirlikte birbirinden ayırt edildiğine ilişkin bir bilgi sunmamaktadır. Daha açık bir anlatımla, KTK'da açık uçlu sorularla yapılan ölçme sonuçlarının güvenirliğinden söz edilirken; test ve puanlayıcılar için hesaplanan güvenirlik katsayılarına yer verilmektedir. Buna karşın; bireylerin yetenek düzeyleri hakkında yapılan kestirimlerin güvenirliğine dair bir bilgi sunulmamaktadır. ÇYRM'de ise ölçme işleminde kullanılan maddeler ve puanlayıcılar tarafından yapılan puanlamaların güvenirliğine ek olarak, bireylerin yetenek düzeyleri hakkında yapılan kestirimlerin güvenirliği de rapor edilmektedir (Linacre, 2014).

ÇYRM'de yetenek kestirimlerini elde etmek amacıyla analize dâhil edilen tüm yüzeyler logit olarak adlandırılan ortak bir metrik üzerine yerleştirilmektedir. KTK'da doğrudan ham puanlar üzerinden işlem yapılırken; ÇYRM'de her bir yüzeye ilişkin ölçümler eşit aralıklı logit ölçeğine dönüştürülmektedir. Sonrasında diğer yüzeylerin bileşenlerine ait ölçümler de dikkate alınarak, öğrencilerin yetenek düzeyleri hakkındaki kestirimlere ulaşılmaktadır (Linacre, 2014). Örneğin; yetenek kestirimlerine ilişkin hesaplamalar yapılmadan önce, puanlayıcıların katılık/cömertlikleri arasındaki farklılıklar istatistiksel olarak kontrol altına alınmaya çalışılmakta ve öğrencilerin yetenek düzeyleri, puanlayıcıların puanlama şiddetindeki farklılıklar ile bu farklılıkları düzeltmek için uygulanan istatistiksel işlemler göz önünde bulundurularak hesaplanmaktadır (Abu Kassim, 2007). Aynı şekilde; yetenek düzeylerine ilişkin kestirimler rapor edilirken maddelerin güçlük düzeyleri de hesaba katılmaktadır (Linacre, 2014). Buna göre; birey, puanlayıcı ve madde şeklinde üç değişkenlik kaynağının bulunduğu bir ÇYRM'de, yetenek kestirimlerinin analizde işlem gören üç yüzey arasında tanımlanan bir fonksiyon yardımıyla elde edildiği söylenebilir. Böylelikle bireylerin yetenek kestirimleri, bütün puanlayıcıların maddelerin tümüne verdiklerin puanların tamamı üzerinden hesaplanmış olmaktadır (Lunz ve Wright, 1997). KTK ve ÇYRM'nin açık uçlu sorularla yapılan ölçmelerde yetenek kestirimlerini hesaplama süreçleri göz önünde bulundurulduğunda, iki kurama göre elde edilen yetenek kestirimleri arasında ne tür benzerliklerin veya farklılıkların bulunduğu merak konusu olmaktadır.

1.4. Araştırmanın Amacı ve Önemi

Bu araştırmada, açık uçlu sorularla yapılan ölçmelerde KTK ve ÇYRM'ye göre hesaplanan yetenek kestirimlerinin karşılaştırılması amaçlanmaktadır. Bu amaç doğrultusunda çalışmada aşağıdaki alt problemlere yanıt aranmıştır.

1. KTK ve ÇYRM'ye göre hesaplanan yetenek kestirimleri arasındaki göreli uyum nasıldır?
2. KTK ve ÇYRM'ye göre hesaplanan yetenek kestirimleri arasındaki mutlak uyum nasıldır?
3. KTK ve ÇYRM'ya göre hesaplanan yetenek kestirimlerinin ölçüt geçerlikleri nasıldır?

Araştırma, amacı itibariyle konu ile ilgili önceki çalışmalardan farklılık göstermektedir. Bundan dolayı, çalışmanın *orijinal* olduğu düşünülmektedir. Literatüre bakıldığında, KTK ile ÇYRM'nin karşılaştırılmasına yönelik çalışmaların bulunduğu görülmektedir. Güler ve Gelbal'ın (2010) KTK ve ÇYRM'nin karşılaştırılmasına dönük çalışmasında, açık uçlu sorularda KTK ve ÇYRM'ye göre hesaplanan madde ve puanlayıcı güvenirlikleri karşılaştırılmıştır. Ancak çalışmada iki kurama göre hesaplanan yetenek kestirimlerinin karşılaştırılmasına yer verilmemiştir. Haiyang (2010) tarafından yapılan çalışmada, açık uçlu maddeler içeren bir İngilizce testi üzerinden KTK ve ÇYRM karşılaştırılmıştır. Yapılan karşılaştırmalar Güler ve Gelbal'ın (2010) çalışmasında olduğu gibi puanlayıcı ve madde güvenirlikleri ile sınırlı tutulmuştur. MacMillan (2000) ile Kadir (2013) tarafından yapılan çalışmalarda, puanlayıcı güvenirliğinin belirlenmesinde KTK, ÇYRM ile genellenebilirlik kuramından elde edilen sonuçlar karşılaştırılmıştır. Huang, Guo, Loadman ve Low (2014) tarafından yürütülen araştırma ise KTK ve ÇYRM'de hesaplanan madde güçlüklerinin, madde ayırt ediciliklerinin ve güvenirlik değerlerinin karşılaştırılması ile sınırlandırılmıştır. Görüldüğü gibi, alanyazında puanlayıcı güvenirliği, madde güvenirliği, madde güçlükleri ile madde ayırt edicilikleri açısından KTK ve ÇYRM'nin karşılaştırıldığı çalışmalar bulunmaktadır. Ancak literatürde, iki kurama göre hesaplanan yetenek kestirimlerinin karşılaştırılmasına yönelik bir çalışmaya rastlanmıştır. Bu bakımdan çalışmanın konu ile ilgili literatüre katkı sağlayacağına inanılmaktadır.

Araştırmada iki farklı ölçme kuramın karşılaştırılması amaçlandığından, çalışmanın *bilimsel bir işlevinin de olacağı* öngörülmektedir. Çünkü bilimin temel işlevlerinden biri; farklı kuramların karşılaştırılıp, işleyen ve işlemeyen yönlerinin tespit edilmesi, karşılaştırılan kuramlar arasındaki benzerlik ve farklılıkların ortaya konulmasıdır (Doğan, 2002). Bu bağlamda, aynı amaçla ortaya atılmış ölçme kuramlarının karşılaştırılması, güçlü ve zayıf yanlarının saptanması, pratikte daha kullanışlı ve doğru olanın belirlenmesi ölçme-değerlendirme bilim dalının ilerlemesi adına bir gereklilik olarak görülmektedir (Atılgan, 2004). Farklı ölçme kuramlarının karşılaştırılmasına yönelik araştırmalar bilimsel işlevleriyle birlikte, *uygulamaya yönelik katkıları* da beraberinde getirmektedir. Örneğin; KTK ve ÇYRM'ye göre hesaplanan yetenek kestirimlerinin göreli ve mutlak uyumlarının karşılaştırılmasıyla, iki kuramın bağıl ve mutlak değerlendirme açısından ne gibi benzer ya da farklı sonuçlar ortaya koyduğu tespit edilebilecektir. Ayrıca, araştırmada KTK ve ÇYRM'ye göre hesaplanan yetenek kestirimlerinden hangisinin ölçüt geçerliğinin daha yüksek olduğu belirlenecektir. Bu sayede araştırma sonuçları, açık uçlu sorularla ölçme yapılırken hangi durumlarda hangi ölçme kuramının kullanılmasının daha uygun olacağına dair bilgiler sunacaktır. Dolayısıyla, çalışmanın uygulamaya yönelik önemli doğurgularının olması beklenmektedir. Özellikle; Öğrenci Seçme ve Yerleştirme Merkezi (ÖSYM) ile Milli Eğitim Bakanlığı (MEB) ilerleyen yıllarda merkezi sınavlarda açık uçlu sorulara yer vermeyi planladığından (ÖSYM, 2015; MEB, 2013), geniş ölçekli test uygulamalarının analizinde de araştırma sonuçlarından yararlanılabileceği tahmin edilmektedir.

2. YÖNTEM

2.1. Araştırma Modeli

Bu çalışma temel bir araştırma niteliğindedir. Temel araştırmalarda; bir kuramın geliştirilmesi veya mevcut kuramların test edilmesi yoluyla (Kaptan, 1998), bilimin teorik yönden ilerletilmesi amaçlanır (Üstdal, Vuillaume, Gülbahar ve Gülbahar, 2004). Diğer bir ifadeyle; temel araştırmalar kuramsal bilgi alanına yenilerinin katılması amacına ağırlık vermekte, uygulamaya yönelme endişesi taşımamaktadır (Kaptan, 1998). Bununla birlikte, temel araştırmaların ileride yapılacak araştırmalara katkı sunacak sonuçlar ortaya koyma potansiyeli yüksektir (Johnson ve Christensen, 2014).

2.2. Araştırma Grubu

Araştırma, ortaokul sekizinci sınıfa devam 100 öğrenci ve çalışmada puanlayıcı olarak görev alan dört matematik öğretmeninin katılımıyla gerçekleşmiştir. Öğrencilerin 46'sı kız ve 54'ü erkektir. Puanlayıcıların cinsiyet, yaş, öğretmenlik mesleğindeki hizmet süresi ve eğitim düzeyi gibi demografik özelliklerine ilişkin bilgiler Tablo 1'de sunulmuştur. Puanlayıcıların tümü ilköğretim matematik öğretmenliği alanından mezun olduğundan, Tablo 1'de puanlayıcıların lisans mezuniyeti ile bir açıklamaya yer verilmemiştir.

Tablo 1. Puanlayıcıların demografik özelliklerine ilişkin bilgiler

Puanlayıcı	Cinsiyet	Yaş	Hizmet Süresi	Eğitim Düzeyi
P1	Erkek	30	6 yıl	Matematik eğitimi alanında yüksek lisans derecesi almıştır.
P2	Erkek	28	6 yıl	Matematik eğitimi alanında yüksek lisans eğitimine devam etmektedir.
P3	Erkek	27	5 yıl	Matematik eğitimi alanında yüksek lisansını sürdürmektedir.
P4	Kadın	27	4 yıl	Eğitim programları ve öğretim alanında yüksek lisans derecesi almıştır.

Araştırma grubundaki öğrenci sayısı belirlenirken DeMars'ın (2010); puanlayıcı sayısına karar verirken ise Turgut ve Baykul'un (2012) önerileri dikkate alınmıştır. DeMars (2010), Rasch analizlerinde 100 ile 200 öğrenciden elde edilen verilerin parametre kestirimleri için yeterli görüldüğünü belirtmiştir. Turgut ve Baykul (2012) ise açık uçlu sorularda en az iki en fazla beş kişinin puanlayıcı olarak görevlendirilmesi gerektiğini dile getirmiştir. Turgut ve Baykul'a (2012) göre, puanlayıcı sayısının daha fazla arttırılması puanlamaların güvenirliğinde önemli bir artış sağlamayacaktır.

2.3. Veri Toplama Araçları

Araştırmada, KTK ve ÇYRM'ye göre hesaplanan yetenek kestirimleri öğrencilerin açık uçlu matematik sorularındaki performansları esas alınarak elde edilmiştir. Dolayısıyla, çalışmanın verileri açık uçlu sorulardan oluşan matematik başarı testi ve bu testteki soruların puanlanmasında kullanılan bütüncül bir rubrik yardımıyla toplanmıştır.

2.3.1. Matematik başarı testi

Araştırmada, öğrencilerin herhangi bir dersteki başarılarının belirlenmesi ile ilgilenilmemektedir. Dolayısıyla, başarı testinin hangi derse yönelik olarak geliştirildiği araştırma sonuçlarını etkileyebilecek bir değişken değildir. Bununla birlikte, açık uçlu maddelere verilen öğrenci yanıtlarını değerlendirecek puanlayıcıların ulaşılabilirliği dikkate alınarak başarı testinin matematik dersine yönelik olarak geliştirilmesine karar verilmiştir. Bu kapsamda, araştırmacı tarafından açık uçlu altı sorudan oluşan bir başarı testi geliştirilmiştir. Geliştirilen

test öğrencilerin akademik başarılarını saptamak için kullanılmayacağından, testin kapsam geçerliği açısından incelenmesine gerek görülmemiştir. Testin yapı geçerliği, ÇYRM'nin varsayımlarından biri olan tek boyutluluk başlığı altında incelenmiştir. Çok yüzeyli Rasch analizi çıktılarında madde yüzeyine ilişkin sunulan ölçüm raporları ise test ile elde edilen ölçümlerin güvenirliğini ortaya koymuştur.

2.3.2. Açık uçlu matematik sorularının puanlanmasında kullanılan rubrik

Öğrencilerin açık uçlu matematik sorularına verdikleri yanıtların puanlanmasında, araştırmacı tarafından geliştirilen bütüncül bir rubrik kullanılmıştır. Rubrikte kullanılacak düzey sayısına karar verilmeden önce konu ile ilgili literatür incelenmiştir. İlgili literatüre bakıldığında, ideal bir rubrikte düzey sayısının kaç olması gerektiğine dair araştırmacılar tarafından çeşitli öneriler getirildiği görülmektedir. Popham (1997) rubrikte üç ile beş arasında bir derecelendirmenin kullanılmasını önerirken; Callison (2000) maksimum dört düzey kullanılmasını tavsiye etmektedir. Stevens ve Levi (2005) en az üç düzeye yer verilmesinin uygun olacağını dile getirmiştir. Kan (2007), puanlamanın zaman ve emek açısından ekonomik olması adına üç ile beş arasındaki düzey sayısının yeterli olacağını ifade etmiştir. Kutlu, Doğan ve Karakaya (2010) ise güvenilir bir puanlama için dört ile yedi arası bir derecelendirmeyi yeterli görmektedir. Sıralanan bu öneriler dört düzeyli bir yapıda kesişmektedir. Bu noktadan hareketle, araştırmada kullanılan rubriğin geliştirilmesinde; *Yetersiz* (1), *Geliştirilmesi Gerek* (2), *İyi* (3) ve *Çok İyi* (4) şeklinde dörtlü bir derecelendirme esas alınmıştır. Rubriğin kategorileri düzenlenirken, problemin ne ölçüde anlaşıldığı; kullanılan çözüm yolunun, çözüm için yapılan işlemlerin ve çözümden elde edilen sonucun doğruluğu ile çözüme nasıl ulaşıldığına ilişkin yapılan açıklamaların yeterliliği dikkate alınmıştır.

Tablo 2. Açık uçlu matematik sorularının puanlanmasında kullanılan rubrik

Puanlama Ölçütleri	
3 *Çok İyi*	-Problem tam olarak anlaşılmıştır. -Uygun çözüm yolu kullanılmıştır. Çözüme yönelik olarak yapılan işlemlerde herhangi bir hata bulunmamaktadır. Doğru sonuca ulaşılmıştır. Problemi çözmek için yapılan işlemler açık, ayrıntılı ve örnek yanıt niteliğindedir.
2 *İyi*	-Problem büyük ölçüde anlaşılmıştır. -Uygun çözüm yolu kullanılmasına rağmen küçük işlem hatalarından ya da anlaşılmayan nedenlerden dolayı doğru sonuca ulaşılmamıştır. -Doğru sonuca ulaşılmıştır. Ancak çözüme nasıl ulaşıldığına dair yeterli açıklama bulunmamaktadır.
1 *Geliştirilmesi Gerek*	-Problem kısmen anlaşılmıştır. -Uygun çözüm yolu ile başlangıç yapılmış, fakat devamı getirilememiştir. -Kullanılan çözüm yolu doğru olmakla birlikte, yapılan işlemlerde önemli hatalar bulunmaktadır. Dolayısıyla doğru sonucuna ulaşılamamıştır.
0 *Yetersiz*	-Problem anlaşılmamıştır. -Problemi cevaplamak için kullanılan stratejiler tamamen yanlıştır ve çözüme yönelik herhangi bir yarar sağlamamaktadır. -Herhangi bir işlem veya açıklama yapılmamıştır. -"Bilmiyorum", "Çok zor bir soru" gibi ifadeler kullanılmış ya da problemde sunulan veriler tekrar edilmiştir.

Puanlama ölçeğindeki dörtlü derecelendirme ve puanlama sırasında dikkat edilecek söz konusu özellikler göz önünde bulundurularak, rubrik için Tablo 2'de sunulan taslak form oluşturulmuştur. Hazırlanan taslak form için biri ölçme değerlendirme ve biri matematik eğitimi alanından olmak üzere iki uzmandan görüş alınmıştır. Ayrıca rubrikte kullanılan dilin anlaşılırlığı ile rubriğin yazım ve noktalama kurallarına uygunluğu bir Türk dili uzmanı tarafından değerlendirilmiştir.

Uzmanlardan alınan görüşler; rubrikte benimsenen düzey sayısının yeterli olduğunu, rubriğin ölçülmesi hedeflenen yapısal çıktıları yansıttığını ve ölçülmek istenen özellik dışında herhangi bir değerlendirme ölçütü içermediğini ortaya koymuştur. Yine uzmanlar, puanlama kategorileri arasındaki farkların açık ve rubrikte kullanılan dilin anlaşılır olduğu yönünde görüş bildirmiştir. Dolayısıyla uzman görüşleri, rubriğin herhangi bir değişikliğe ihtiyaç duyulmadan kullanılabileceği şeklinde yorumlanmıştır.

2.4. İşlem

Araştırma verileri 2014-2015 Öğretim Yılı Bahar Dönemi'nde toplanmıştır. Uygulama öncesinde öğrenciler araştırmanın amacı hakkında bilgilendirilmiştir. Öğrencilere toplanan verilerin yalnızca araştırmanın amacı için kullanılacağı, başka herhangi bir kişi ya da kurumla paylaşılmayacağı ifade edilmiştir. Öğrencilere test sonuçlarının not verme amacıyla kullanılmayacağı belirtilmiştir. Bununla birlikte, geçerli ve güvenilir sonuçlar elde edilebilmesi için test maddelerini gerçek bir sınavdaymış hassasiyetiyle cevaplamalarının önemi vurgulanmıştır. Araştırmanın amacı ile ilgili açıklamalar yapıldıktan sonra çalışmaya katılımın zorunluğu olmadığı öğrencilere hatırlatılmıştır. Bu sayede araştırma grubunun yalnızca gönüllü öğrencilerden oluşması sağlanmıştır. Öğrencilere testteki soruları çözmeleri için 30 dakika süre verilmiştir.

Öğrencilerden verilerin toplanmasının ardından, test kâğıtları numaralandırılmış ve öğrencilerin demografik özelliklerinin yer aldığı bir veri dosyası oluşturulmuştur. Öğrencilerin demografik özellikleri veri dosyasına aktarıldıktan sonra sınav kâğıtlarından silinmiştir. Bu şekildeki bir uygulama ile puanlayıcıların değerlendirme sırasında öğrencinin cinsiyeti ve geçmiş matematik başarısı gibi demografik özelliklerinden etkilenmesinin önüne geçilmesi hedeflenmiştir. Sınav kâğıtlarının her biri dört farklı puanlayıcı tarafından değerlendirileceğinden, fotokopi ile kâğıtların dört adet kopyası oluşturulmuştur. Böylelikle, test kâğıtları puanlama işlemi için hazır hale gelmiştir. Puanlama işlemi öncesinde, puanlayıcılara değerlendirmede kullanılacak rubriğin tanıtıldığı ve puanlama sırasında dikkat edilmesi gereken hususların açıklandığı kısa bir eğitim verilmiştir. Puanlayıcılara, değerlendirme sırasında öğrenciler üzerinden ilerlemek yerine; test maddeleri üzerinden ilerlemeleri gerektiği belirtilmiştir. Örneğin, tüm öğrencilerin birinci soruya verdikleri cevapları puanlandıktan sonra kâğıtları karıştırarak sırasını değiştirmeleri ve ikinci soruya ilişkin puanlamalara geçmeleri istenmiştir. Ek olarak, bir oturuşta bir madde ile ilgili bütün yanıtları puanlamaları; farklı maddeleri ise değişik zamanlarda değerlendirmeleri gerektiği söylenmiştir. Söz gelimi; öğrencilerin birinci maddeye verdikleri cevapları değerlendirirken, tüm kâğıtları tek oturuşta puanlamaları, bütün öğrencilerin birinci maddedeki performanslarını puanladıktan sonra, ikinci maddeye ilişkin puanlamalara hemen başlamayıp bir süre ara vermeleri önerilmiştir. Puanlamalarda dikkat edilmesi gereken bu hususlar için Hogan ve Murphy'nin (2007) "Yapılandırılmış Yanıtlı Maddelerin Hazırlanması ve Puanlanmasına ilişkin Öneriler: Uzmanlar Ne Söylüyor" başlıklı çalışması referans alınmıştır. Puanlama sonrasında, dört puanlayıcı tarafından verilen puanların ortalaması alınarak KTK'ya ilişkin yetenek kestirimleri hesaplanmıştır. Ardından aynı veriler ÇYRM'ye göre analiz edilerek öğrencilerin matematik performansları için Rasch analizinde hesaplanan yetenek kestirimleri elde edilmiştir. Sonrasında KTK ile ÇYRM'ye göre hesaplanan yetenek kestirimleri karşılaştırılmıştır.

2.5. Veri Analizi

Araştırmada, öğrencilerin açık uçlu matematik sorularına verdikleri yanıtlar, dört puanlayıcı tarafından puanlanmıştır. Puanlama sonucunda elde edilen veriler hem KTK'ya hem

de ÇYRM'ye göre analiz edilmiştir. KTK'da; puanlayıcı güvenirliğinin belirlenmesinde puanlayıcılar arası korelasyon katsayısı ve tekrarlı ölçümler için varyans analizinden yararlanılmış; öğrencilerin matematik performanslarına ait yetenek kestirimlerini elde etmek amacıyla dört puanlayıcı tarafından verilen puanlamaların ortalaması alınmıştır.

KTK'ya yönelik analizler tamamlandıktan sonra, ÇYRM'ye ilişkin analizlere başlanmıştır. Analizde; puanlayıcı, öğrenci ve madde şeklinde üç yüzey işlem görmüştür. Analiz gerçekleştirilmeden önce ÇYRM'nin tek boyutluluk, yerel bağımsızlık ve model-veri uyumu varsayımları test edilmiştir. Tek boyutluluk varsayımı için dört puanlayıcı tarafından verilen puanların ortalaması alınmış ve hesaplanan ortalamalar üzerinden Açımlayıcı Faktör Analizi (AFA) uygulanmıştır. AFA yapılmadan önce verilerin faktör analizine uygun olup olmadığı araştırılmıştır. Bu doğrultuda KMO ve Bartlett testlerine yer verilmiştir. Kaiser (1974), 0.50'nin üzerindeki KMO değerlerini kabul edilebilir olarak nitelendirmektedir (Field, 2009). Büyüköztürk'e (2010) göre ise verilerin faktör analizine uygun olabilmesi için KMO değerinin 0.60'tan yüksek olması gerekmektedir. Bu araştırmada; KMO örneklem uygunluk katsayısı 0.60 ve Bartlett testi değeri 56.863 (p<.001, sd=15) bulunmuştur. Buna göre, verilerin faktör analizine uygun olduğu söylenebilir. Bu tespitin ardından temel bileşenler faktörleştirme tekniği ve döngüsüz metot kullanılarak AFA gerçekleştirilmiştir. Döngüsüz metot kullanılması faktör analizi sonucunda tek boyutlu bir yapı elde edileceği öngörüsünden kaynaklanmaktadır. Faktör analizi sonucunda, toplam varyansın %31.18'ini açıklayan ve faktör yükleri 0.51 ile 0.64 arasında değişken tek boyutlu bir yapıya ulaşılmıştır. Buna göre, ÇYRM'nin ilk varsayımı olan tek boyutluluk şartının sağlandığı söylenebilir. Ayrıca, AFA'da açıklanan varyans oranı için %30'un üzerindeki değerlerin yeterli görüldüğü (Bayram, 2009) ve maddelerin faktör yükleri için .32 değerinin alt sınır olarak kabul edildiği (Çokluk, Şekercioğlu ve Büyüköztürk, 2012; Tabachnick ve Fidell, 2007) göz önüne alındığında, matematik başarı testinin yapı geçerliğinin sağlandığı ifade edilebilir.

ÇYRM'nin bir diğer varsayımı yerel bağımsızlıktır. Yerel bağımsızlık, tek boyutluluk ile paralel çalışan bir varsayım olup tek boyutluluk varsayımının sağlandığı durumlarda yerel bağımsızlık varsayımının da karşılandığı kabul edilmektedir (Hambleton, Swaminathan ve Rogers, 1991). Dolayısıyla, araştırmada yerel bağımsızlık varsayımı ayrıca test edilmemiş, tek boyutluluk varsayımının karşılanması yerel bağımsızlık varsayımının sağlandığına yönelik bir gösterge olarak değerlendirilmiştir. ÇYRM'nin üçüncü varsayımı, model ile veri uyumudur. Model ile veri uyumunun karşılanıp karşılanmadığına, analiz sonucunda rapor edilen standartlaştırılmış artık değerleri incelenerek karar verilmektedir. Model ile verinin uyumlu olabilmesi için ±2 aralığının dışında kalan standartlaştırılmış artıkların sayısı toplam veri sayısının yaklaşık %5'ini, ±3 aralığının dışında yer alan standartlaştırılmış artıkların sayısı ise toplam veri sayısının yaklaşık %1'ini aşmamalıdır (Linacre, 2014). Çok yüzeyli Rasch analizine 100 öğrenci, dört puanlayıcı ve altı madde dâhil olduğundan, analiz sonucunda ulaşılacak toplam veri sayısı 2400'dür [100×4×6=2400]. Analiz çıktılarına bakıldığında, ±3 aralığının dışında kalan standartlaştırılmış artık oranının %1.08'e (26 tane) ve ±2 aralığının dışında kalan standartlaştırılmış artık oranının ise %4.75'e (114 tane) karşılık geldiği saptanmıştır. Buna göre, model ile veri uyumu varsayımının da karşılandığı ifade edilebilir.

KTK ve ÇYRM'ye göre öğrencilerin matematik performanslarına ait yetenek kestirimleri hesaplandıktan sonra, iki kurama göre elde edilen yetenek kestirimleri karşılaştırılmıştır. Yetenek kestirimleri arasındaki göreli uyumu belirlemek için Pearson Momentler Çarpımı korelasyon katsayısından, mutlak uyumu saptamak amacıyla ise ilişkili örneklem t-testinden yararlanılmıştır. Son olarak; hesaplanan yetenek kestirimlerinin öğrencilerin matematik karne notları ve Temel Eğitimden Ortaöğretime Geçiş (TEOG) sistemi kapsamında uygulanan merkezi matematik sınavındaki doğru sayıları ile arasındaki korelasyonlara bakılarak, iki kurama göre

hesaplanan yetenek kestirimlerinin ölçüt geçerliği karşılaştırılmıştır. Araştırmada; KTK'ya ilişkin analizlerin gerçekleştirilmesinde, iki farklı kurama göre elde edilen yetenek kestirimlerinin göreli ve mutlak uyumu ile ölçüt geçerliğinin karşılaştırılmasında SPSS paket programı kullanılmıştır. Çok yüzeyli Rasch analizi için ise FACET paket programına başvurulmuştur.

3. BULGULAR

Bu bölümde araştırmada ulaşılan bulgulara yer vermiştir. İlk olarak, KTK'ya göre öğrencilerin matematik performanslarına ilişkin yetenek kestirimleri hesaplanmıştır. Bu amaçla, dört puanlayıcı tarafından yapılan puanlamaların ortalaması alınmış ve elde edilen bulgular Tablo 3'te sunulmuştur.

Tablo 3. Öğrencilerin matematik performansı için KTK'ya göre hesaplanan yetenek kestirimleri

Öğrenci No	Matematik Performansı	Öğrenci No	Matematik Performansı	Öğrenci No	Matematik Performansı	Öğrenci No	Matematik Performansı	Öğrenci No	Matematik Performansı
1	1.42	21	0.83	41	0.83	61	0.83	81	0.79
2	1.50	22	1.08	42	1.13	62	1.00	82	1.25
3	0.96	23	0.50	43	0.42	63	1.08	83	0.83
4	1.54	24	1.58	44	1.67	64	0.88	84	1.33
5	0.50	25	0.33	45	0.63	65	1.21	85	1.25
6	1.46	26	0.21	46	1.50	66	1.54	86	1.25
7	0.92	27	1.83	47	1.29	67	0.67	87	0.75
8	1.42	28	1.46	48	1.08	68	1.04	88	1.25
9	1.67	29	1.29	49	2.17	69	1.04	89	1.13
10	1.25	30	1.38	50	0.63	70	1.29	90	1.00
11	1.46	31	1.21	51	0.79	71	0.58	91	0.38
12	2.21	32	0.88	52	0.79	72	1.75	92	1.83
13	1.33	33	0.75	53	0.92	73	1.21	93	1.42
14	1.04	34	0.75	54	1.08	74	1.04	94	1.33
15	0.13	35	1.17	55	1.04	75	0.96	95	1.83
16	0.58	36	1.08	56	1.21	76	1.17	96	1.79
17	0.71	37	1.42	57	1.21	77	0.75	97	0.83
18	0.83	38	1.29	58	0.96	78	1.83	98	0.58
19	0.88	39	1.13	59	0.46	79	1.04	99	1.08
20	1.42	40	0.88	60	1.04	80	0.83	100	1.00
Ortalama = 1.10						Standart Sapma = 0.40			

Tablo 3'te görüldüğü üzere, öğrencilerin matematik performansları için KTK'ya göre hesaplanan yetenek kestirimleri 0.13 ile 2.21 arasında değişmektedir. Öğrencilerin matematik performanslarına ilişkin ortalama ve standart sapma değerleri ise sırasıyla 1.10 ve 0.40 olarak bulunmuştur. Yetenek kestirimlerinin hesaplanmasının ardından KTK'ya göre puanlayıcılar arası güvenirlik incelenmiştir. Bunun için öncelikle puanlayıcılar arası korelasyon katsayılarına bakılmıştır. Hesaplanan korelasyon katsayıları, her bir puanlayıcının yaptığı puanlamalara ait betimsel istatistikler ile birlikte Tablo 4'te gösterilmiştir.

Tablo 4. Puanlayıcılar arası korelasyon katsayıları ile puanlayıcılara ait betimsel istatistikler

	P1	P2	P3	P4	Ortalama	Standart Sapma	Çarpıklık	Basıklık
P1	1				1.35	0.44	-0.04	-0.32
P2	0.84**	1			1.18	0.48	-0.02	-0.10
P3	0.74**	0.79**	1		0.85	0.43	0.38	-0.24
P4	0.81**	0.81**	0.69**	1	1.02	0.41	0.34	0.57

** $p<0.01$

Tablo 4'e göre; puanlayıcılar arası korelasyon katsayıları 0.69 ile 0.84 arasında değişmekte olup, korelasyon katsayılarının tümü istatistiksel açıdan anlamlıdır. Korelasyon katsayılarının 0.70'in üzerinde veya 0.70'e oldukça yakın değerlere sahip olması (Büyüköztürk, 2010), puanlayıcılar arası güvenirliğin yüksek olduğunu düşündürmektedir. Ancak korelasyon katsayısı ortalamadan bağımsız olarak hesaplanan bir istatistik olduğundan ve puanlamalar arasındaki mutlak uyumu dikkate almadığından, puanlayıcı güvenirliği konusunda bir karara varılmadan önce, puanlayıcı ortalamalarının karşılaştırılması gerekmektedir. Bu kapsamda, tekrarlı ölçümler için varyans analizi uygulanarak puanlayıcı ortalamaları karşılaştırılmış ve elde edilen bulgular Tablo 5'te sunulmuştur.

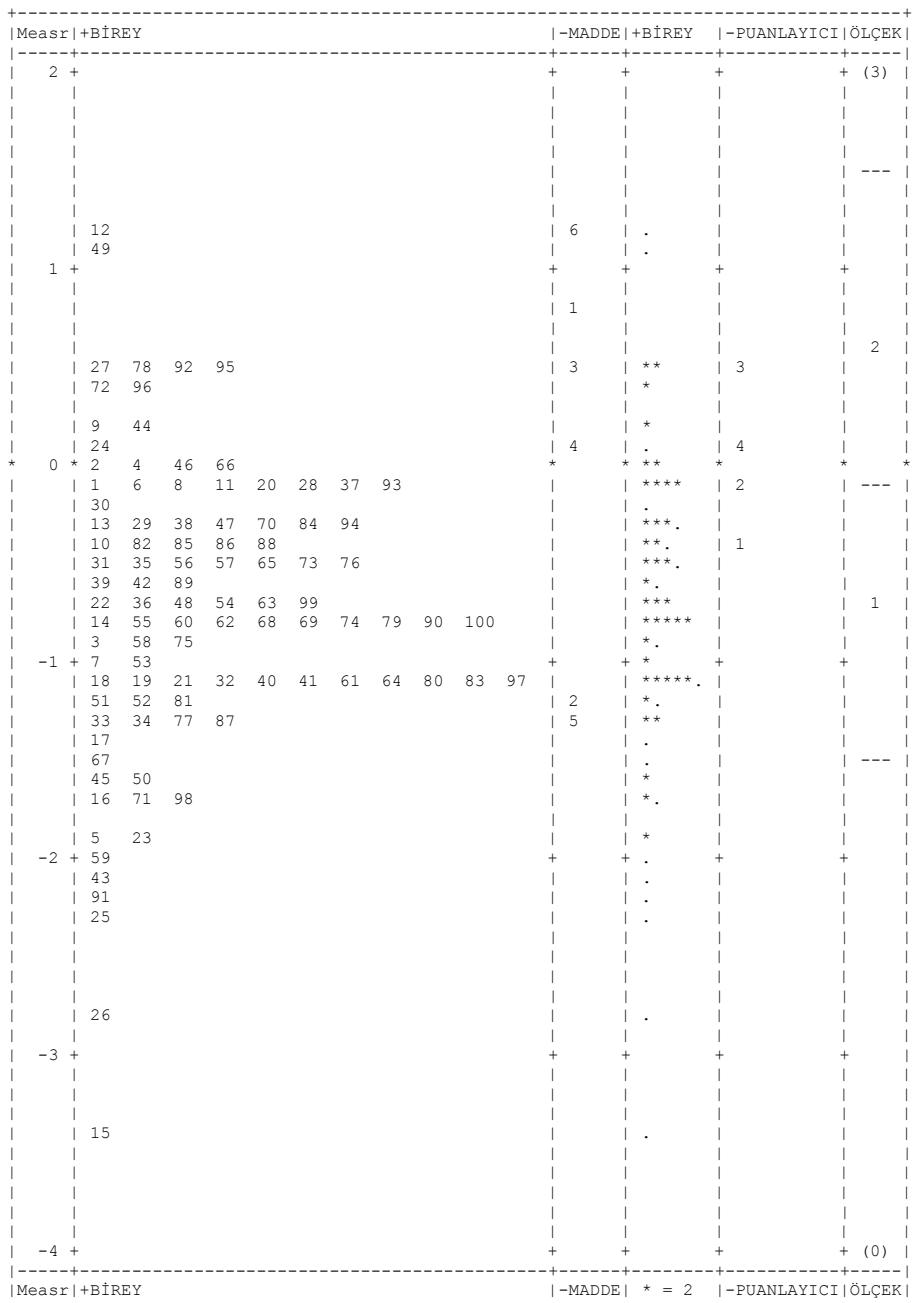
Tablo 5. Puanlayıcı ortalamalarının karşılaştırılmasına yönelik tekrarlı ölçümler için varyans analizi sonuçları

Puanlayıcı	Ortalama	Standart Sapma	Wilks's Lamda	F	Hipotez sd	Hata sd	Eta Kare	Farkın Kaynağı
P1	1.35	0.44	0.23	106.65	3	97	0.77	Tüm puanlayıcılar arasında anlamlı fark vardır
P2	1.18	0.48						
P3	0.85	0.43						
P4	1.02	0.41						

Tablo 5'e bakıldığında, dört puanlayıcının yaptığı puanlamalar arasında anlamlı fark bulunduğu görülmektedir [$F_{(3,97)}$= 106.65, $p<0.01$]. Puanlayıcı ortalamaları dikkate alındığında, puanlama sırasında bir numaralı puanlayıcının diğer puanlayıcılara göre daha cömert; üç numaralı puanlayıcının ise diğer puanlayıcılara göre daha katı davrandığı söylenebilir. Puanlayıcılar arasında güçlü korelasyonlar bulunmasına rağmen; tekrarlı ölçümler için varyans analizi sonuçlarının istatistiksel açıdan anlamlı olması, puanlayıcılar arasındaki mutlak uyumunun göreli uyuma göre daha düşük olduğunu yansıtmaktadır.

Öğrencilerin matematik performanslarına ilişkin yetenek kestirimlerinin ve puanlayıcılar arası güvenirliğin KTK'ya göre incelenmesinin ardından, çok yüzeyli Rasch analizi uygulanmıştır. Rasch analizinde ilk olarak; değişken haritası, birey, madde ve puanlayıcı yüzeylerine ait ölçüm raporları ile rubriğe ilişkin kategori istatistikleri sunulmuştur. Şekil 1'de, çok yüzeyli Rasch analizi sonucunda rapor edilen değişken haritası gösterilmiştir. Şekil 1'in ilk sütunu; öğrencilerin yetenek düzeyi, maddelerin güçlük düzeyi ve puanlayıcıların katılık/cömertliklerine ilişkin ölçüm birimini temsil etmektedir. Anlaşılacağı üzere, ÇYRM'de analize dâhil olan değişkenlik kaynaklarının tamamı logit olarak adlandırılan ortak bir ölçek üzerine yerleştirilmektedir. Şekil 1'in ikinci sütununda öğrenciler matematik performansları açısından sıralanmıştır. Bu sütunda aşağıdan yukarı doğru ilerledikçe öğrencilerin performansları artmaktadır. Buna göre, 12 numaralı katılımcının matematik performansı en yüksek (1.16 logit); 15 numaralı katılımcının ise matematik performansı en düşük (-3.35 logit) öğrenci olduğu söylenebilir. Öğrencilerin matematik performansları için değişken haritasının negatif ve pozitif ucu boyunca uzanan ölçümlerin elde edilmesi matematik performansları farklı olan öğrencilerinin birbirinden başarılı bir şekilde ayırt edildiğini yansıtmaktadır. Şekil 1'in üçüncü sütununda, maddelere ilişkin ölçümler yer almaktadır. Maddelerin güçlük düzeyleri açısından

sıralandığı bu sütunda, aşağıdan yukarı doğru gidildikçe madde güçlüğü artmaktadır. Buna göre, en zor sorunun altı numaralı madde (1.15 logit); en kolay sorunun ise iki numaralı madde (-1.35 logit) olduğu söylenebilir. Maddelerin tek bir noktada kümelenmeyip değişken haritasının farklı noktalarında yer alması; öğrencilerin farklı maddelerdeki performanslarının birbirinden etkili bir şekilde ayırt edilebildiği anlamına gelmektedir. Şekil 1'in beşinci sütunu ise puanlayıcılara ait ölçümleri içermektedir. Bu ölçümler, sütunun pozitif ucunda yer alan ve yüksek logit puanına sahip olan puanlayıcıların puanlamada daha katı; sütunun negatif ucunda yer alan ve düşük logit puanına sahip olan puanlayıcıların puanlamada daha cömert davrandığı şeklinde yorumlanmaktadır. Dolayısıyla, en katı puanlamaların üç numaralı puanlayıcı (.46 logit); en cömert puanlayıcıların ise bir numaralı puanlayıcı (-.45 logit) tarafından yapıldığı söylenebilir.



Şekil 1. Öğrencilerin matematik performanslarının ölçülmesine ilişkin değişken haritası

Değişken haritası; bireylerin yetenek düzeyleri, maddelerin güçlük düzeyleri ve puanlayıcıların katılık/cömertlikleri hakkında önemli ipuçları verse de; birey, madde ve puanlayıcı yüzeylerine ait daha detaylı bilgi edinebilmek için her bir yüzeye ait ölçüm raporlarının incelenmesi gerekmektedir. Bu doğrultuda her bir yüzeye ilişkin ölçüm raporları aşağıda gösterilmiştir. Öncelikle, birey yüzeyine ilişkin ölçüm raporlarına bakılmış ve elde edilen bulgular Tablo 6'da sunulmuştur.

Tablo 6. Matematik performansının ÇYRM'ye göre analiz edilmesiyle birey yüzeyi için elde edilen ölçüm raporları

	Logit Ölçüsü	Standart Hata	Uygunluk İçi	Uygunluk Dışı
Ortalama	-0.72	0.28	1.00	1.02
Standart Sapma (Evren)	0.76	0.04	0.49	0.62
Standart Sapma (Örneklem)	0.76	0.04	0.49	0.62
Model, Evren: RMSE = 0.28	Standart Sapma = 0.70		Ayırma Oranı = 2.48	Güvenirlik = 0.86
Model, Örneklem: RMSE = 0.28	Standart Sapma = 0.70		Ayırma Oranı = 2.49	Güvenirlik = 0.86
Model, Tamamı Aynı Ki Kare= 589.4	sd=99	p=0.00		
Model, Rastgele Normal Ki Kare= 83.1	sd=98	p=0.86		

Tablo 6'ya göre, öğrencilerin matematik performanslarının ortalaması -0.72 ve standart sapması 0.76 logit şeklindedir. Yine Tablo 6'ya göre, uygunluk içi ve uygunluk dışı istatistiklerine ait ortalamalar sırasıyla 1.00 ve 1.02 olarak belirlenmiştir. Uygunluk istatistiklerinin 1'e eşit olması halinde, model ile veri arasındaki uyumun mükemmel olduğu bilinmektedir (Brentari ve Golia, 2008). Dolayısıyla, birey yüzeyi için elde edilen uygunluk istatistikleri model ile veri arasındaki uyumun oldukça yüksek olduğunu yansıtmaktadır. Tablo 6'daki ayırma oranı ve güvenirlik indeksi değerlerine bakıldığında, ayırma oranının 2.49 ve güvenirlik indeksinin 0.86 olduğu görülmektedir. Hesaplanan güvenirlik indeksinin yüksek olması, matematik performansları farklı olan öğrencilerin birbirinden başarılı bir biçimde ayırt edilebildiğine işaret etmektedir. Güvenirlik indeksinin yanı sıra Ki Kare testi sonuçları da matematik performansları farklı olan öğrencilerin etkili bir şekilde ayırt edilebildiğini göstermektedir. Ki Kare testi sonuçlarına göre, matematik performansları açısından öğrenciler arasında istatistiksel olarak anlamlı fark bulunmaktadır [χ^2=589.4, *sd*=99, *p*<0.01].

Birey yüzeyine ilişkin ölçümlerin ardından madde yüzeyine ait ölçümlere bakılmış ve elde edilen bulgular Tablo 7'de sunulmuştur. Tablo 7'ye göre, maddelerin güçlük düzeyleri 1.15 logit ile -1.35 logit arasında değişmekte olup güçlük düzeyleri açısından maddeler arasında 2.50 logitlik bir değişim gözlenmektedir. Maddelerin güçlük düzeylerine ilişkin ortalama 0.00 ve standart sapma 1.05 olarak tespit edilmiştir. Maddeler için hesaplanan uygunluk içi istatistiklerinin 0.82 ile 1.17 arasında değiştiği; uygunluk dışı istatistiklerinin 0.88 ile 1.17 arasında sıralandığı saptanmıştır. Uygunluk içi ve uygunluk dışı istatistiklerine ilişkin ortalamalar ise sırasıyla 0.99 ve 1.02 olarak belirlenmiştir. Linacre'ye (2014) göre, analizde model ile veri uyumunu olumsuz yönde etkileyen madde bulunup bulunmadığına karar vermek için uygunluk istatistiklerine bakılması gerekir. Wright ve Linacre (1994), 0.6 ile 1.4 arasında kalan uygunluk değerlerini kabul edilebilir olarak belirtmiştir. Bu ölçüt dikkate alındığında, analizde model ile veri uyumunu bozan bir madde bulunmadığı söylenebilir. Uygunluk istatistiklerine ilişkin ortalamaların 1.00'a oldukça yakın olması, model ile veri arasındaki uyumun oldukça yüksek olduğu anlamına gelmektedir.

Tablo 7. Matematik performansının ÇYRM'ye göre analiz edilmesiyle madde yüzeyi için elde edilen ölçüm raporları

Madde	Logit Ölçüsü	Standart Hata	Uygunluk İçi	Uygunluk Dışı
M6	1.15	0.08	0.90	0.88
M1	0.79	0.07	1.17	1.17
M3	0.54	0.07	1.04	0.97
M4	0.08	0.06	0.82	1.05
M2	-1.21	0.06	1.14	1.17
M5	-1.35	0.06	0.88	0.89
Ortalama	0.00	0.07	0.99	1.02
Standart Sapma (Evren)	0.96	0.01	0.13	0.12
Standart Sapma (Örneklem)	1.05	0.01	0.15	0.13
Model, Evren: RMSE = 0.07	Standart Sapma = 0.96	Ayırma Oranı = 14.00	Güvenirlik = 0.99	
Model, Örneklem: RMSE = 0.07	Standart Sapma = 1.05	Ayırma Oranı = 15.34	Güvenirlik = 1.00	
Model, Tamamı Aynı Ki Kare = 1178.9 sd=5 p=0.00				
Model, Rastgele Normal Ki Kare = 5.0 sd=4 p=0.29				

Tablo 7'ye göre, madde yüzeyine ilişkin ayırma oranı 15.34 ve güvenirlik indeksi 1.00'dır. Madde yüzeyi için hesaplanan ayırma oranı ve güvenirlik indeksinin yüksek olması, testteki soruların güçlük düzeyleri açısından farklılık gösterdiğine işaret etmektedir. Ayırma oranı ile güvenirlik indeksinin işaret ettiği bu farkın anlamlı olup olmadığını saptamak amacıyla Ki Kare değeri incelenmiştir. Ki-Kare değerinin anlamlı çıkması [χ^2=1178.9, *sd*=5, *p*<0.01], testteki maddelerin güçlük düzeyleri arasında manidar bir fark bulunduğunu göstermektedir.

Puanlayıcı yüzeyine ait ölçüm raporları ise Tablo 8'de sunulmuştur. Tablo 8 incelendiğinde, puanlayıcılara ilişkin logit ölçülerinin 0.46 ile -0.45 arasında değiştiği görülmektedir. Dolayısıyla, puanlayıcıların katılık ve cömertliklerine ilişkin aralık 0.91 logittir [0.46-(-0.45)]. Tablo 8'e göre, uygunluk içi ve uygunluk dışı istatistiklerine ait ortalamalar 1.00 ve 1.02 değerlerine karşılık gelmektedir. Uygunluk istatistikleri puanlayıcıların tümünde, 0.6 ile 1.4 kabul edilebilir aralığı (Myford ve Wolfe, 2003) içerisinde yer almaktadır. Bu noktadan hareketle, model ile veri uyumunu olumsuz yönde etkileyen bir puanlayıcı bulunmadığı söylenebilir.

Tablo 8. Matematik performansının ÇYRM'ye göre analiz edilmesiyle puanlayıcı yüzeyi için elde edilen ölçüm raporları

Puanlayıcı	Logit Ölçüsü	Standart Hata	Uygunluk İçi	Uygunluk Dışı
P3	0.46	0.06	1.05	1.06
P4	0.14	0.06	0.85	0.90
P2	-0.15	0.05	1.08	1.05
P1	-0.45	0.05	0.97	1.08
Ortalama	0.00	0.06	0.99	1.02
Standart Sapma (Evren)	0.34	0.00	0.09	0.07
Standart Sapma (Örneklem)	0.39	0.00	0.10	0.08
Model, Evren: RMSE=0.06	Standart Sapma=0.33	Ayırma Oranı = 6.01	Güvenirlik = 0.97	
Model, Örneklem: RMSE=0.06	Standart Sapma=0.38	Ayırma Oranı = 6.96	Güvenirlik = 0.98	
Model, Tamamı Aynı Ki Kare = 146.8 sd = 3 p = 0.00				
Model, Rastgele Normal Ki Kare = 2.9 sd = 2 p = 0.23				
Puanlayıcılar arası mutlak uyum: %57.10				

Puanlayıcı yüzeyine ilişkin ayırma oranı ve güvenirlik indeksine bakıldığında, ayırma oranı 6.96 ve güvenirlik indeksi 0.98 olarak bulunmuştur. Puanlayıcı yüzeyi için hesaplanan

ayırma oranı ile güvenirlik indeksi, puanlayıcılar arasındaki güvenilir benzerliği değil; güvenilir farkı göstermektedir (Haiyang, 2010). Bundan dolayı, hesaplanan katsayılar puanlayıcıların katılık ve cömertlikleri açısından farklılık gösterdiğini yansıtmaktadır. Tablo 8'deki Ki Kare değerlerinin anlamlı olması [χ^2=146.8, *sd*=3, *p*<0.01], puanlayıcılar arasında gözlenen söz konusu farkın istatistiksel açıdan anlamlı olduğunu ortaya koymaktadır. Son olarak, Tablo 8'e bakıldığında, puanlayıcılar arası mutlak uyumun %57.10 olduğu görülmektedir. Çok yüzeyli Rasch analizi çıktılarında her bir yüzeye ilişkin ölçüm raporlarından sonra kategori istatistikleri rapor edilmektedir.

Öğrencilerin açık uçlu matematik sorularına verdikleri yanıtların dörtlü derecelendirmeye sahip bütüncül bir rubrik kullanılarak puanlanmasıyla elde edilen kategori istatistikleri Tablo 9'da sunulmuştur.

Tablo 9. Matematik performansının ÇYRM'ye göre analiz edilmesiyle elde edilen kategori istatistikleri

Puanlama Ölçeği Kategorileri	Frekans	Yüzde	Yığılmalı Yüzde	Ortalama Ölçüm	Beklenen Ölçüm	Uygunluk Dışı İstatistikleri
0	954	40	40	-1.62	-1.64	1.1
1	557	23	63	-0.87	-0.82	0.80
2	590	25	88	0.08	0.05	1.2
3	299	12	100	0.83	0.83	0.90

Rubriğin etkin bir biçimde çalıştığının söylenebilmesi için puanlama ölçeğinin her bir kategorisinde en az 10 gözlem bulunması gerekmektedir. Tablo 9'daki frekans değerleri bu şartı sağlar niteliktedir. Puanlama ölçeğinin iyi çalıştığına işaret eden bir diğer gösterge, puanlama ölçeğinin kategorileri arttıkça ortalama ölçümlerin de artmasıdır (Linacre, 2014). Tablo 9'daki ortalama ölçümlerin rubrik kategorilerine paralel olarak artması, puanlama ölçeğinin etkin bir şekilde çalıştığını ortaya koymaktadır. Tablo 9'daki uygunluk dışı istatistiklerinin 1'e oldukça yakın olması rubriğin etkin bir biçimde çalıştığını yansıtan göstergelerden bir diğeridir.

Yukarıda da görüldüğü üzere, çok yüzeyli Rasch analizi çıktılarında tüm yüzeylere ilişkin ölçümler logit cetvelinde rapor edilmektedir. Bu durum, KTK'dan elde edilen yetenek kestirimleri ile çok yüzeyli Rasch analizi çıktılarında sunulan yetenek kestirimleri arasındaki göreli uyumun belirlenmesine engel teşkil etmemektedir. Yine ÇYRM'de yetenek kestirimlerinin logit cetvelinde sunulması, iki kurama göre hesaplanan yetenek kestirimlerinin ölçüt geçerliğinin karşılaştırılmasına mani değildir. Çünkü göreli uyum belirlenirken ve ölçüt geçerliği açısından karşılaştırma yapılırken, performansları açısından öğrenciler arasında yapılan sıralamalar temele alınmakta ve yetenek kestirimlerinin rapor edildiği birim bu sıralamalar üzerinde bir farka yol açmamaktadır. Diğer taraftan mutlak uyum incelenirken, öğrenciler arasındaki sıralamanın değil; öğrencilerin matematik performanslarına ilişkin ortalamaların karşılaştırılması söz konusudur. Bu nedenle, çok yüzeyli Rasch analizi çıktılarında hesaplanan yetenek kestirimleri ile KTK'da ulaşılan yetenek kestirimlerinin aynı birimde ifade edilmesi iki kurama göre hesaplanan yetenek kestirimleri arasındaki mutlak uyumun belirlenebilmesi için bir gereklilik olmaktadır. Bu gereklilik doğrultusunda, çok yüzeyli Rasch analizi sonucunda logit cetvelinde sunulan yetenek kestirimleri, puanlamada kullanılan ölçek birimlerine dönüştürülmüştür. Sözü edilen dönüştürme işlemi için Linacre'nin (2014) önerdiği formülden yararlanılmıştır. Linacre'ye (2014) göre; logit cetvelinde rapor edilen yetenek kestirimleri öğrenciler, veliler ve eğitim yöneticileri için anlaşılır olmamaktır. Bu sebeple, logit cetvelinde rapor edilen yetenek kestirimlerinin puanlama ölçeğinin birimlerine dönüştürülmesi gerekir. Linacre'nin (2014) bu dönüştürme işlemi için önerdiği formül doğrultusunda yapılan işlemler Tablo 10'da özetlenmiştir.

Tablo 10. Logit cetvelinde rapor edilen yetenek kestirimlerini puanlama ölçeğinin birimlerine dönüştürmek için uygulanan işlemler

Logit cetvelinde rapor edilen yetenek kestirimlerini puanlama ölçeğinin birimlerine dönüştürürken ilk olarak öğrencilerin yetenek düzeylerine ilişkin aralık belirlenmelidir.	Matematik performansı en yüksek olan öğrencinin yetenek düzeyi 1.16 logit ve en düşük olan öğrencinin yetenek düzeyi -3.35 logit bulunduğundan, öğrencilerin yetenek düzeylerine ilişkin aralık 4.51 logittir.
Puanlamada kullanılan ölçeğin ranj değeri, yetenek düzeyi için hesaplanan aralığa bölünmelidir.	Puanlama ölçeğinin ranjı 3-0=3'tür. $3 \div 4.51 = 0.665$
İkinci adımdaki bölme işlemi sonucunda elde edilen değer ile en düşük yetenek düzeyindeki öğrenciye ait yetenek kestirimi çarpılmalıdır.	$(-3.35) \times (0.665) = -2.228$
Üçüncü adımda ulaşılan değer ile toplandığında, puanlama ölçeğinin en düşük noktasına eşit olacak sabit belirlenmelidir.	-2.228 + (sabit) = 0 olması gerektiğinden, sabit değeri 2.228 olarak bulunur.
Her bir öğrenci için hesaplanan yetenek kestirimi ikinci adımda elde edilen değer ile çarpılıp, çıkan sonuca dördüncü adımdaki sabit eklenmelidir.	$\left(\begin{array}{c}\text{Öğrencinin logit}\\ \text{cetvelindeki yetenek}\\ \text{düzeyi}\end{array}\right) \times (0.665) + 2.228$

Tablo 10'da özetlenen işlemler uygulanarak, her bir öğrenci için logit cetvelinde rapor edilen yetenek kestirimleri puanlama ölçeğinin birimlerine dönüştürülmüştür. Bu dönüştürme işlemi sonucunda hesaplanan yetenek kestirimleri, logit cetvelindeki yetenek kestirimleri ile birlikte Tablo 11'de yer almaktadır.

Tablo 11. Öğrencilerin matematik performansı için ÇYRM'ye göre hesaplanan yetenek kestirimleri

Öğrenci No	Matematik Performansı		Öğrenci No	Matematik Performansı		Öğrenci No	Matematik Performansı		Öğrenci No	Matematik Performansı		Öğrenci No	Matematik Performansı	
	LC	PÖB		LC	PÖB		LC	PÖB		LC	PÖB		LC	PÖB
1	-0.15	2.13	21	-1.14	1.47	41	-1.14	1.47	61	-1.14	1.47	81	-1.22	1.42
2	-0.02	2.21	22	-0.69	1.77	42	-0.62	1.82	62	-0.83	1.68	82	-0.41	1.96
3	-0.91	1.62	23	-1.87	0.98	43	-2.09	0.84	63	-0.69	1.77	83	-1.14	1.47
4	0.05	2.26	24	0.11	2.30	44	0.24	2.39	64	-1.06	1.52	84	-0.28	2.04
5	-1.87	0.98	25	-2.35	0.67	45	-1.57	1.18	65	-0.48	1.91	85	-0.41	1.96
6	-0.08	2.17	26	-2.84	0.34	46	-0.02	2.21	66	0.05	2.26	86	-0.41	1.96
7	-0.99	1.57	27	0.51	2.57	47	-0.34	2.00	67	-1.48	1.24	87	-1.31	1.36
8	-0.15	2.13	28	-0.08	2.17	48	-0.69	1.77	68	-0.76	1.72	88	-0.41	1.96
9	0.24	2.39	29	-0.34	2.00	49	1.08	2.95	69	-0.76	1.72	89	-0.62	1.82
10	-0.41	1.96	30	-0.21	2.09	50	-1.57	1.18	70	-0.34	2.00	90	-0.83	1.68
11	-0.08	2.17	31	-0.48	1.91	51	-1.22	1.42	71	-1.67	1.12	91	-2.21	0.76
12	1.16	3.00	32	-1.06	1.52	52	-1.22	1.42	72	0.38	2.48	92	0.51	2.57
13	-0.28	2.04	33	-1.31	1.36	53	-0.99	1.57	73	-0.48	1.91	93	-0.15	2.13
14	-0.76	1.72	34	-1.31	1.36	54	-0.69	1.77	74	-0.76	1.72	94	-0.28	2.04
15	-3.35	0.00	35	-0.55	1.86	55	-0.76	1.72	75	-0.91	1.62	95	0.51	2.57
16	-1.67	1.12	36	-0.69	1.77	56	-0.48	1.91	76	-0.55	1.86	96	0.44	2,52
17	-1.39	1.30	37	-0.15	2.13	57	-0.48	1.91	77	-1.31	1.36	97	-1.14	1.47
18	-1.14	1.47	38	-0.34	2.00	58	-0.91	1.62	78	0.51	2.57	98	-1.67	1.12
19	-1.06	1.52	39	-0.62	1.82	59	-1.98	0.91	79	-0.76	1.72	99	-0.69	1.77
20	-0.15	2.13	40	-1.06	1.52	60	-0.76	1.72	80	-1.14	1.47	100	-0.83	1.68

LC = Logit Cetveli, PÖB = Puanlama Ölçeği Birimi

Tablo 11 incelendiğinde, öğrencilerin matematik performanslarına ilişkin yetenek kestirimlerinin logit cetvelinde 1.16 ile -3.35 arasında değiştiği; puanlama ölçeği biriminde ise 0.00 ile 3.00 arasında uzandığı görülmektedir. Yetenek kestirimleri logit cetvelinden puanlamada kullanılan rubriğin birimlerine dönüştürüldükten sonra; KTK ile ÇYRM'ye göre hesaplanan yetenek kestirimleri karşılaştırmaya hazır hale gelmiştir. İki kurama göre hesaplanan yetenek kestirimleri arasındaki göreli uyumu saptamak için gerçekleştirilen korelasyon analizi sonuçları ile mutlak uyumu belirlemek amacıyla uygulanan bağımlı gruplar *t* testi sonuçları Tablo 12'de sunulmuştur.

Tablo 12'ye göre, KTK ile ÇYRM'ye göre hesaplanan yetenek kestirimleri arasındaki göreli uyum son derece yüksektir [r=0.99, p<0.01]. Diğer bir deyişle; öğrenciler arasında matematik performansları açısından bir sıralama yapılması durumunda, yetenek kestirimlerinin KTK'ya veya ÇYRM'ye göre hesaplanmış olması yapılan sıralamada bir farka neden olmamaktadır. Öte yandan, Tablo 12'deki bağımlı gruplar *t* testi sonuçlarının anlamlı olması [$t_{(99)}$, p<0.01], iki kurama göre hesaplanan yetenek kestirimleri arasında mutlak bir uyumdan söz etmenin mümkün olmadığına işaret etmektedir.

Tablo 12. KTK ve ÇYRM'ye göre hesaplanan yetenek kestirimleri arasındaki uyumun belirlenmesine yönelik korelasyon analizi ve bağımlı gruplar *t*-testi sonuçları

Kuram	Ortalama	Standart Sapma	N	*r*	*t*
KTK	1.10	0.40	100	0.99**	52.92**
ÇYRM	1.75	0.50			

**p<0.01

KTK ve ÇYRM'ye göre elde edilen yetenek kestirimlerinin ölçüt geçerliği incelenirken, öğrencilerin matematik dersi karne notları ve TEOG kapsamında uygulanan matematik dersi ortak sınavındaki doğru sayıları referans alınmıştır. Hesaplanan yetenek kestirimleri ve söz konusu iki değişken arasındaki ilişki korelasyon analizi ile incelenmiş, analiz sonuçları Tablo 13'te gösterilmiştir.

Tablo 13. Yetenek kestirimlerinin ölçüt geçerliği çalışmasına ilişkin korelasyon analizi sonuçları

	1	**2**	**3**	**4**
1. KTK'ya göre hesaplanan yetenek kestirimi	1			
2. ÇYRM'ye göre hesaplanan yetenek kestirimi	0.99**	1		
3. Matematik Dersi Karne Notu	0.36**	0.40**	1	
4. Merkezi Matematik Sınavındaki Doğru Sayısı	0.48**	0.53**	0.72**	1

**p<0.01

Tablo 13'e göre, hem KTK'ya hem de ÇYRM'ye göre hesaplanan yetenek kestirimlerinin öğrencilerin matematik karne notları ve merkezi ortak sınavda matematik dersi doğru sayıları ile arasındaki ilişkiler istatistiksel açıdan anlamlıdır. Bununla birlikte; ÇYRM'den elde edilen yetenek kestirimlerinin ölçüt olarak alınan iki değişken ile arasındaki korelasyonlar daha yüksektir. Buna göre, ÇYRM'de hesaplanan yetenek kestirimlerinin ölçüt geçerliğinin KTK'dan elde edilen yetenek kestirimlerine kıyasla daha yüksek olduğu söylenebilir.

4. TARTIŞMA VE SONUÇ

Bu araştırmada; açık uçlu sorularla yapılan ölçmelerde KTK ve ÇYRM'ye göre hesaplanan yetenek kestirimleri, göreli ve mutlak uyum ile ölçüt geçerliği açısından karşılaştırılmıştır. Araştırmadan elde edilen bulgular; KTK ve ÇYRM'ye göre hesaplanan

yetenek kestirimleri arasında mükemmele yakın bir göreli uyum bulunduğunu göstermiştir. Bu sonuç; KTK'da ve ÇYRM'de hesaplanan yetenek kestirimleri açısından öğrenciler arasında bir sıralama yapıldığı takdirde, iki kurama göre yapılan sıralamaların birbiriyle örtüşeceği anlamına gelmektedir. Bir başka deyişle; açık uçlu sorularla yapılan ölçme sonuçları bağıl bir değerlendirmeye tabi tutulacaksa, yetenek kestirimlerinin KTK veya ÇYRM'ye göre hesaplanmış olması değerlendirme sonuçlarını etkilemeyecektir. KTK ve ÇYRM'ye göre hesaplanan yetenek ölçülerinin bireyleri benzer şekilde sıraya koyduğuna ilişkin araştırma bulgusu, farklı ölçme kuramlarından elde edilen yetenek kestirimleri arasında yüksek korelasyon bulunduğunu gösteren çalışmalarla desteklenmektedir. Söz gelimi, Zaman, Kashmiri, Mubarak ve Ali (2008) tarafından yapılan çalışmada, KTK ile iki parametreli MTK'dan elde edilen yetenek kestirimleri arasında yüksek korelasyon bulunduğu belirlenmiştir. Benzer şekilde, Çelen (2008) ile Çelen ve Aybek'in (2013) yaptığı çalışmalarda, KTK ve MTK'ya göre hesaplanan öğrenci başarıları arasında yüksek bir korelasyon bulunmuştur. Yine Hwang (2002) tarafından yapılan çalışmada, KTK'ya göre hesaplanan yetenek kestirimleri ile MTK'nın farklı modellerine göre elde edilen yetenek kestirimleri arasında yüksek korelasyonlar saptanmıştır. Özer Özkan (2012) tarafından yapılan araştırmada da, KTK ile tek ve çok boyutlu MTK'ya göre hesaplanan başarı puanları arasında yüksek korelasyonlar tespit edilmiştir. Ancak sıralanan bu çalışmaların, araştırma bulgularını dolaylı olarak destekleyebileceği gözden kaçırılmamalıdır. Çünkü sıralanan çalışmalar çoktan seçmeli testler üzerinde yürütülürken, bu çalışma açık uçlu maddeler üzerinde yürütülmüştür. Ayrıca; hem bahsedilen araştırmalarda kullanılan bir, iki ve üç parametreli modeller hem de bu araştırmada kullanılan ÇYRM, MTK çatısı altında yer alsa da; bu modeller arasında önemli farklılıklar bulunmaktadır.

Araştırmada, KTK ve ÇYRM'ye hesaplanan yetenek kestirimlerine ilişkin ortalamaların farklılık gösterdiği ve dolayısıyla iki kurama göre elde edilen yetenek kestirimleri arasında mutlak bir uyum bulunmadığı sonucuna ulaşılmıştır. Bu sonuca göre; KTK'ya ve ÇYRM'ye göre elde edilen yetenek kestirimlerine dayalı olarak mutlak değerlendirme yapılması durumunda değerlendirme sonuçlarının farklılık göstereceği söylenebilir. ÇYRM'de rapor edilen yetenek kestirimlerinin ortalamasının KTK'da hesaplanan yetenek kestirimlerinin ortalamasına göre daha yüksek olduğu dikkate alındığında; KTK'ya göre başarısız olduğuna karar verilen bir öğrenci ÇYRM'ye göre başarılı bulunabilir. Böylesi bir farklılık, özellikle puanları kesme noktasında olan öğrenciler için telafisi mümkün olmayan ya da çok zor olan sonuçlara yol açabilir. Örneğin, bir son sınıf öğrencisinin açık uçlu bir testten aldığı puan KTK'ya göre belirlendiğinde, öğrencinin başarısız olduğuna karar verilebilir ve bu karar öğrencinin dönem uzatması anlamına gelebilir. Ancak aynı öğrenci için ÇYRM'ye göre yetenek kestirimi hesaplandığında, öğrencinin geçme ölçütünün üzerinde bir puan aldığı ve mezun olması gerektiği sonucuna varılabilir. KTK ve ÇYTRM'ye göre hesaplanan yetenek kestirimleri arasındaki göreli ve mutlak uyum sonuçları bir arada ele alındığında, açık uçlu sorularla yapılan ölçmelerde yetenek kestirimlerinin hangi kurama göre elde edildiğinin bağıl değerlendirmeden çok mutlak değerlendirme sonuçlarını etkileyeceği şeklinde bir çıkarım karşımıza çıkmaktadır. Ancak baraj uygulaması içeren bir bağıl değerlendirme söz konusu ise KTK ve ÇYRM'de hesaplanan yetenek kestirimlerine göre yapılan değerlendirmelerin farklılık gösterebileceği unutulmamalıdır.

Araştırmanın diğer bir sonucu KTK ve ÇYRM'de hesaplanan yetenek kestirimlerinin ölçüt geçerliği ile ilgilidir. Ölçüt geçerliği çalışmasında, KTK ve ÇYRM'ye göre hesaplanan yetenek kestirimlerinin öğrencilerin matematik başarıları ile ilişkisine bakılmıştır. Elde edilen sonuçlar; ÇYRM'ye göre hesaplanan yetenek kestirimlerinin ölçüt geçerliğinin KTK'dan elde edilen yetenek kestirimlerine göre daha yüksek olduğunu ortaya koymuştur. Bu sonuç, ÇYRM'nin puanlayıcı farklılıklarının tespiti ile sınırlı kalmayıp, uyguladığı istatistiksel düzeltmelerle tespit edilen farklılıkları bir dereceye kadar kontrol altına almasıyla (Abu Kassim,

2007; Linacre, Engelhard, Tatum ve Myford, 1994) açıklanabilir. Ölçüt geçerliği çalışması sonuçları, açık uçlu sorularla yapılan ölçmelerde yetenek kestirimlerinin ÇYRM'ye göre hesaplanmasının daha uygun olacağına işaret etmektedir.

5. ÖNERİLER

Alanyazında, KTK ile MTK'nın bir, iki ve üç parametreli modellerinden elde edilen yetenek kestirimleri arasındaki ilişkilerin incelendiği çalışmalar bulunmaktadır. Ancak literatürde, KTK ve ÇYRM'ye göre hesaplanan yetenek kestirimlerinin karşılaştırıldığı bir araştırmaya rastlanmamıştır. Bu araştırmanın başlangıç noktasının oluşturan alanyazındaki söz konusu boşluk, çalışmanın literatürde önemli bir yer edineceğini düşündürmektedir. Bununla birlikte, araştırmanın bir takım sınırlılıklarının bulunduğu göz ardı edilmemelidir. Çalışmaya ilişkin sınırlılıklar ve bu sınırlılıklar doğrultusunda getirilebilecek ileri araştırma önerileri şu şekilde sıralanabilir. Öncelikle, bu araştırmanın çalışma grubu 100 öğrenci ile dört puanlayıcıdan oluşmuştur. Çalışma grubunda, 100 ile 200 katılımcının bulunması, Rasch analizleri için yeterli görülmektedir. Bununla birlikte; MTK'ya dayalı modellerin katılımcı sayısının fazla olduğu gruplarda daha doğru kestirimler ürettiği bilinmektedir (DeMars, 2010). Buna göre; çalışma grubundaki öğrenci sayısı yeterli olsa da; daha büyük örneklemler üzerinde benzer çalışmaların yapılması araştırmada ulaşılan bulguların genellenebilirliğine katkı sunması açısından önem arz etmektedir. İkinci olarak; araştırma kapsamında kullanılan başarı testi, açık uçlu altı madde içermektedir. DeMars'a (2010) göre; MTK'ya dayalı modellerde, örneklem büyüklüğünün yanı sıra testteki madde sayısı da yapılan kestirimleri etkileyebilmektedir. Dolayısıyla, farklı sayıda madde içeren testlerin kullanıldığı ileri araştırmaların yapılması konu ile ilgili literatüre katkı sağlayacaktır. Son olarak, bu çalışmada açık uçlu sorulara verilen öğrenci yanıtları genel ve bütüncül özellik taşıyan bir rubrik yardımıyla puanlanmıştır. Rubriğin genel ya da göreve özel olması, bütüncül veya analitik bir yapı göstermesi puanlayıcılar tarafından yapılan puanlamaları etkileyen önemli bir faktördür (Knoch, 2009; Nitko, 2004). Buna bağlı olarak, çalışmada kullanılan rubrik KTK ve ÇYRM'ye göre hesaplanan yetenek kestirimlerini etkilemiş olabilir. Bu noktadan hareketle, ileri araştırmalarda analitik bir yapı gösteren göreve özel rubriklerin kullanılması önerilebilir.

6. KAYNAKLAR

Abu Kassim, N.L. (2007, June). *Exploring rater judging behaviour using the many-facet Rasch model*. Paper Presented in the Second Biennial International Conference on Teaching and Learning of English in Asia: Exploring New Frontiers (TELiA2), Holiday Villa Beach & Spa Resort, Langkawi. Faculty of Communication and Modern Languages, Universiti Utara Malaysia. [Available online at: http://repo.uum.edu.my/3212/1/Noor1.pdf], Retrieved on October 03, 2015.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, 43(9), 561-573. http://dx.doi.org/10.1007/BF02293814

Atılgan, H. (2004). *Genellenebilirlik kuramı ve çok değişkenlik kaynaklı Rasch modelinin karşılaştırılmasına ilişkin bir araştırma*. Yayınlanmamış Doktora Tezi, Hacettepe Üniversitesi, Sosyal Bilimler Enstitüsü, Ankara.

Bağcan Büyükturan, E., & Çıkrıkçı Demirtaşlı, N. (2013). Çoktan seçmeli testler ile yapılandırılmış gridlerin psikometrik özellikleri bakımından karşılaştırılması. *Ankara Üniversitesi Eğitim Bilimleri Fakültesi Dergisi*, 46(1), 395-415. [Çevrim-içi: http://dergiler.ankara.edu.tr/dergiler/40/1799/19011.pdf], Erişim tarihi: 25 Eylül 2015.

Bahar, M., Nartgün, Z., Durmuş, S., & Bıçak, B. (2010). *Geleneksel-tamamlayıcı ölçme ve değerlendirme teknikleri*. Ankara: Pegem Akademi Yayıncılık.

Baker, F.B. (2001). *The basics of item response theory*. ERIC Clearinghouse on Assessment and Evaluation, University of Maryland, College Park, MD.

Başol, G. (2013). *Eğitimde ölçme ve değerlendirme*. Ankara: Pegem Akademi Yayıncılık.

Baykul, Y. (2010). *Eğitimde ve psikolojide ölçme: Klasik test teorisi ve uygulaması*. Ankara: Pegem Akademi Yayıncılık.

Bayram, N. (2009). *Data analysis through SPSS in social sciences*. Bursa: Ezgi Publishing.

Braun, H.I., Bennett, R.E., Frye, D., & Soloway, E. (1990). Scoring constructed responses using expert systems. *Journal of Educational Measurement,* 27(2), 93-108. http://dx.doi.org/10.1111/j.1745-3984.1990.tb00736.x

Bridgeman, B. (1992). A comparison of quantative questions in open-ended and multiple-choice formats. *Journal of Educaitonal Measurement*, 29(3), 253-271. http://dx.doi.org/10.1002/j.2333-8504.1991.tb01402.x

Büyüköztürk, Ş. (2010). *Sosyal bilimler için veri analizi el kitabı*. Ankara: Pegem Akademi Yayınları.

Callison, D. (2000). Rubrics. *School Library Media Activities Monthly*, 17(2), 34-46.

Çelen, Ü. (2008). Klasik test kuramı ve madde tepki kuramı yöntemleriyle geliştirilen iki testin geçerlilik ve güvenilirliğinin karşılaştırılması. *İlköğretim Online,* 7(3), 758-768. [Çevrim-içi: http://dergipark.ulakbim.gov.tr/ilkonline/article/view/5000038231/5000037088], Erişim tarihi: 29 Eylül 2015.

Çelen, Ü., & Aybek, E.C. (2013). Öğrenci başarısının öğretmen yapımı bir testle klasik test kuramı ve madde tepki kuramı yöntemleriyle elde edilen puanlara göre karşılaştırılması. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, 4(2), 64-75. [Çevrim-içi: http://dergipark.ulakbim.gov.tr/epod/article/view/1040000004], Erişim tarihi: 25 Eylül 2015.

Çokluk, Ö., Şekercioğlu, G., & Büyüköztürk, Ş. (2012). *Sosyal bilimler için çok değişkenli istatistik SPSS ve LISREL uygulamaları*. Ankara: Pegem Akademi Yayınları.

David, A.B. (2008). Comparison of classification accuracy using Cohen's weighted kappa. *Expert Systems with Applications*, 34, 825-832. http://dx.doi.org/10.1016/j.eswa.2006.10.022

DeMars, C. (2010). *Item response theory*. Oxford, UK: Oxford University Press.

Doğan, N. (2002). *Klasik test teorisi ve örtük özellikler kuramının örneklemeler bağlamında karşılaştırılması*. Yayınlanmamış Doktora Tezi, Hacettepe Üniversitesi, Sosyal Bilimler Enstitüsü, Ankara.

Doğan, N. (2013). Yazılı yoklamalar. H. Atılgan (Ed.), *Eğitimde ölçme ve değerlendirme* içinde (145-168). Ankara: Anı Yayıncılık.

Ebel, R.L. (1951). Estimation of the reliability of ratings. *Psychometrika*, 16(4), 407-424. http://dx.doi.org/10.1007/BF02288803

Field, A. (2009). *Discovering statics using SPSS*. London: SAGE Publications Ltd.

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

Graham, M., Milanowski, A., & Miller, J. (2012). *Measuring and promoting inter-rater agreement of teacher and principal performance ratings*. Washington, DC: Center for Educator Compensation Reform. [Available online at: http://cecr.ed.gov/pdfs/Inter_Rater.pdf], Retrieved on June 10, 2013.

Gronlund, N.E. (1998). *Assessment of student achievement*. Boston: Allyn and Bacon.

Güler, N. (2014). Analysis of open-ended statistics questions with many facet Rasch model. *Eurasian Journal of Educational Research*, 55, 73-90. http://dx.doi.org/10.14689/ejer.2014.55.5

Güler, N., & Gelbal, S. (2010). Study based on classic test theory and many facet Rasch model. *Eurasian Journal of Educational Research*, 38, 108-125. [Available online at: *http://www.aniyayincilik.com.tr/main/pdfler/38/7_guler_nese.pdf*], Retrieved on September 10, 2015.

Haiyang, S. (2010). An application of classical test theory and many facet Rasch measurement in analyzing the reliability of an English test for non-English major graduates. *Chinese Journal of Applied Linguistics*, 33(2), 87-102. [Available online at: http://www.celea.org.cn/teic/90/10060807.pdf], Retrieved on August 16, 2013.

Hambleton, R.K., Swaminathan, H., & Rogers, H.J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: SAGE Publications, Inc.

Harvey, R., & Hammer, A. (1999). Item response theory. *The Counseling Psychologist*, 27(3), 353-383. http://dx.doi.org/10.1177/0011000099273004

Hogan, T.P., & Murphy, G. (2007) Recommendations for preparing and scoring constructed-response items: What the experts say. *Applied Measurement in Education*, 20(4), 427-441, http://dx.doi.org/10.1080/08957340701580736

Huang, T.W., Guo, G.J., Loadman, W., & Low, F.M. (2014). Rating score data analysis by classical test theory and many-facet Rasch model. *Psychology Research*, 4(3), 222-231. [Available online at: http://www.davidpublishing.com/show.html?15856], Retrieved on October 01, 2015.

Johnson, B., & Christensen, L. (2014). *Educational research: Quantitative, qualitative, and mixed approaches*. Thousand Oaks, CA: Sage Publications.

İlhan, M. (2015). *Standart ve SOLO taksonomisine dayalı rubrikler ile puanlanan açık uçlu matematik sorularında puanlayıcı etkilerinin çok yüzeyli Rasch modeli ile incelenmesi*. Yayınlanmamış Doktora Tezi, Gaziantep Üniversitesi, Eğitim Bilimleri Enstitüsü, Gaziantep.

Kadir, K.A. (2013). Examining factors affecting language performance: A comparison of three measurement approaches. *Pertanika Journal of Social Sciences & Humanities*, 21 (3), 1149-1162. [Available online at: http://www.pertanika.upm.edu.my/Pertanika%20PAPERS/JSSH%20Vol.%2021%20(3)%20Sep.%202013/19%20Page%201149-1162.pdf], Retrieved on September 25, 2015.

Kan, A. (2007). Performans değerlendirme sürecine katkıları açısından yeni program anlayışı içerisinde kullanılabilecek bir değerlendirme yaklaşımı: Rubrik puanlama yönergeleri. *Kuram ve Uygulamada Eğitim Bilimleri*, 7(1), 129-152. [Çevrim-içi: http://www.edam.com.tr/kuyeb/pdf/tr/99530abf499c979f8fc1b4312f7b4e4fnfull.pdf], Erişim tarihi: 13 Haziran 2013.

Kan, A. (2013). Ölçme araçlarında bulunması gereken nitelikler. H. Atılgan (Ed.), *Eğitimde ölçme ve değerlendirme* içinde (23-80). Ankara: Anı Yayıncılık.

Kaptan, S. (1998). *Bilimsel araştırma ve istatistik teknikleri*. Ankara: Tekışık Web Ofset Tesisleri.

Knoch, U., Read, J., & von Randow, J. (2007). Re-training writing raters online: How does it compare with face-to-face training? *Assessing Writing*, 12(1), 26-43. http://dx.doi.org/10.1016/j.asw.2007.04.001

Kline, T. (2005). *Psychological testing: A practical approach to design and evaluation*. Thousand Oaks: Sage Publications.

Kutlu, Ö., Doğan, C.D., & Karakaya, İ. (2010). *Öğrenci başarısının belirlenmesi: Performansa ve portfolyoya dayalı durum belirleme*. Ankara: Pegem Akademi Yayıncılık.

LeBreton & Senter, (2008). Answers to 20 questions about interrater reliability and interrater agreement. *Organizational Research Methods*, 11(4), 815-852. http://dx.doi.org/10.1177/1094428106296642

Linacre, J.M. (1989). *Many-facet Rasch measurement*. Chicago: MESA Press.

Linacre, J.M. (2014). *A user's guide to FACETS Rasch-model computer programs*. [Available online at: http://www.winsteps.com/a/facets-manual.pdf], Retrieved on July 13, 2015.

Linacre, J.M., Engelhard, G.Jr., Tatum, D.S., & Myford, C.M. (1994). Measurement with judges: Many-faceted conjoint measurement. *International Journal of Educational Research*, 21(6), 569-577. http://dx.doi.org/10.1016/0883-0355(94)90011-6

Van der Linden, W.J., & Hambleton, R.K. (1997). Item response theory: Brief history, common models, and extensions. In W.J. Van der Linden and R.K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 1-28). New York: Springer Verlag.

Lunz, M.E., &Wright, B.D. (1997). Latent trait models for performance examinations. In J. Rost & R. Langeheine (Eds.), *Applications of latent trait and latent class models in the social sciences* (pp. 80-88). Münster, Germany: Waxmann.

Lynch, B.K., & McNamara, T.F. (1998). Using G-theory and many-facet Rasch measurement in the development of performance assessments of the ESL speaking skills of immigrants. *Language Testing*, 15(2), 158-180. http://dx.doi.org/10.1177/026553229801500202

MacMillan, P.D. (2000). Classical, generalizability and multifaceted Rasch detection of interrater variability in large sparse data sets. *The Journal of Experimental Education*, 68(2), 167-190. http://dx.doi.org/10.1080/00220970009598501

Masters, G.N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47(2), 149-174. http://dx.doi.org/10.1007/BF02296272

Milli Eğitim Bakanlığı. (2013). *Temel eğitimden ortaöğretime geçişle ilgili sıkça sorulan sorular*. [Çevrim-içi: http://www.meb.gov.tr/duyurular/duyurular2013/bigb/tegitimdenoogretimegecis/MEB_SSS_20_09_2013.pdf], Erişim tarihi: 06 Ekim 2013.

Mulqueen C., Baker D., & Dismukes, R.K. (2000, April) *Using multifacet Rasch analysis to examine the effectiveness of rater training*. Presented at the 15th Annual Conference for the Society for Industrial and Organizational Psychology (SIOP). New Orleans. [Available online at: http://www.air.org/files/multifacet_Rasch.pdf], Retrieved on September 19, 2013.

Nitko, A.J. (2004). *Educational assessment of students*. Upper Saddle River, NJ: Pearson.

Öğrenci Seçme ve Yerleştirme Merkezi. (2015). *Yazılı sınav (Açık uçlu sorularla sınav).* [Çevrim-içi: http://www.osym.gov.tr/belge/1-23308/yazili-sinav-acik-uclu-sorularla-sinav-04022015.html], Erişim tarihi: 01 Ekim 2015.

Özçelik, D.A. (2011). *Ölçme ve değerlendirme*. Ankara: Pegem Akademi Yayıncılık.

Özer Özkan, Y. (2012). *Öğrenci başarılarının belirlenmesi sınavından (ÖBBS) klasik test kuramı, tek boyutlu ve çok boyutlu madde tepki kuramı modelleri ile kestirilen başarı puanlarının karşılaştırılması*. Yayınlanmamış Doktora Tezi, Ankara Üniversitesi, Eğitim Bilimleri Enstitüsü, Ankara.

Popham, W.J. (1997). What's wrong-and what's right-with rubrics. *Educational Leadership*, 55(2), 72-75. [Available online at: http://pareonline.net/getvn.asp?v=9&n=2], Retrieved on October 02, 2015.

Romagnano, L. (2001). The myth of objectivity in mathematics assessment. *Mathematics Teacher*, 94(1), 31-37. [Available online at: http://www.peterliljedahl.com/wp-content/uploads/Myth-of-Objectivity.pdf], Retrieved on September 11, 2015.

Stevens, D.D., & Levi, A.J. (2005). *Introduction to rubrics: An assessment tool to save grading time, convey effective feedback, and promote student learning*. Sterling, VA; Stylus.

Sudweeks, R.R., Reeve, S., & Bradshaw, W.S. (2005). A comparison of generalizability theory and many-facet Rasch measurement in an analysis of college sophomore writing. *Assessing Writing*, 9(3), 239-261. http://dx.doi.org/10.1016/j.asw.2004.11.001

Tabachnick, B.G., & Fidell, L.S. (2007). *Using multivariate statistics*. Boston, Pearson Education, Inc.

Tan, Ş., & Erdoğan, A. (2004). *Öğretimi planlama ve değerlendirme*. Ankara: PegemA Yayıncılık.

Tekin, H. (2009). *Eğitimde ölçme ve değerlendirme*. Ankara: Yargı Yayınevi.

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

Üstdal, M., Vuillaume, R., Gülbahar, K., & Gülbahar, Y. (2004). *Bilimsel araştırma kılavuzu*. Ankara: Pelikan Yayıncılık.

Zaman, A., Kashmiri, A., Mubarak, M., & Ali, A. (2008, Kasım). *Students ranking, based on their abilities on objective type test: Comparison of CTT and IRT*. EDU-COM International Conference. [Available online at: http://ro.ecu.edu.au/ceducom/52/], Retrieved on September 30, 2015.

Extended Abstract

The purpose of this study is to compare the ability estimations of classical test theory (CTT) and the many facet Rasch model (MFRM) in measurements conducted with open-ended questions. In accordance with this aim, answers were sought for the following sub-problems:

- What is the relative agreement between ability estimations calculated using CTT and MFRM?
- What is the absolute agreement between ability estimations calculated using CTT and MFRM?
- What is the concurrent validity of ability estimations calculated using CTT and MFRM?

The study differs from previous studies because of its aim. Therefore, this study is *original.* There are studies in the literature that examine the relationship between the ability estimations obtained from the models of Classical Test Theory (CTT) and Item Response Theory (IRT) with one, two and three parameters. However, no studies compare the ability estimations calculated by CTT and MFRM. Thus, this study will contribute to the relevant literature. Since the comparison of two different measurement theories is the aim of this study, the study will have also a scientific function. One of the major functions of science is to compare different theories, to determine their operative and non-operative aspects and to reveal similarities and differences between these theories In this context, comparing measurement theories, which have been proposed for the same purpose, detecting their strengths and weaknesses and determining which one is more practical and accurate in practice is a requirement for scientific advancement. In addition to their scientific function, studies that compare measurement theories contribute

to practice. For example, the similarities and differences between these two theories' criterion-referenced and norm-referenced assessments will be determined by comparing the relative and absolute agreement of their ability estimations. Moreover, the ability estimations with the highest concurrent validity will also be determined in this study. Thus, the study results will offer information about which assessment theories are more appropriate for conducting measurements with open-ended questions, and in which situations this is the case. This study will have applicable results since the Student Selection and Placement Center and Ministry of National Education are planning to administer open-ended questions in central exams in the upcoming years, and this study's results can be used for the analysis of large-scale testing.

This study has the characteristics of basic research. The theoretical improvement of science is accomplished by the development of a theory or the testing of available theories in basic research. In other words, basic research focuses on new information in the field of theoretical information and does not focus on its application. Moreover, basic research has a high potential for generating results which will contribute to further studies. The study was conducted with 100 eighth graders and four mathematics teachers who rated the students' work. The study's data were obtained using an achievement test with 6 open-ended mathematics questions and a holistic rubric for scoring these questions. The data obtained by rating the open-ended mathematics questions were analyzed using both CTT and MFRM. The ability estimations for CTT were calculated using the score averages of the four raters. Afterwards, Many Facet Rasch analysis was conducted using a three-facet pattern including raters, students and items. The ability estimations calculated by these two theories were prepared to be compared by converting ability estimations obtained from the Rasch analysis and reported in the logit scale into the units of rubric used for scoring. Pearson's product-moment correlation coefficient and the dependent samples *t*-test were used to determine the relative agreement and the absolute agreement between the ability estimations, respectively. Finally, the concurrent validity of the ability estimations calculated by these two theories was compared by considering the correlations between them, the students' mathematics grades and their scores on the national high school entrance mathematics examination.

This study found that the relative agreement between the ability estimations calculated using CTT and MFRM was extremely high. This means that if that the students are ranked according to these ability estimations, these ranks will correspond to each other. In other words, if the measurement results carried out with the open-ended questions are subjected to criterion-referenced assessment, calculating ability estimations with either CTT or MFRM will not affect the results. It was determined that there was a significant difference between the means of the ability estimations of the two theories, and thus no absolute agreement. This result indicates that norm-referenced assessment of CTT and MFRM's ability estimations will differ. When the relative and absolute agreements of CTT and MFRM's ability estimations are dealt with together, it should not be forgotten that the assessments of CTT and MFRM's ability estimations can differ in criterion-referenced assessments, including threshold applications. It was determined that the concurrent validity of the ability estimations of MFRM is higher than that of CTT's ability estimations. This can be explained by the fact that MFRM is not only limited by the raters' differences, but also control the differences determined with the statistical corrections to a certain degree. The results of the concurrent validity study show that MFRM's ability estimations in measurements conducted with open-ended questions are better.

Kaynakça Bilgisi

İlhan, M. (2016). Açık uçlu sorularla yapılan ölçmelerde klasik test kuramı ve çok yüzeyli rasch modeline göre hesaplanan yetenek kestirimlerinin karşılaştırılması. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi [Hacettepe University Journal of Education]*, 31(2), 346-368.

Citation Information

İlhan, M. (2016). A comparison of the ability estimations of classical test theory and the many facet rasch model in measurements with open-ended questions [in Turkish]. *Hacettepe University Journal of Education [Hacettepe Üniversitesi Eğitim Fakültesi Dergisi],* 31(2), 346-368.