

Bilimsel Araştırma Ödevlerinin Çok Yüzeyle Rasch Ölçme Modeli ile Değerlendirilmesi

Ramazan BAŞTÜRK*

Pamukkale Üniversitesi

Özet

Bu araştırmanın amacı, öğrencilerin bilimsel araştırma ile ilgili teorik kazanımlarını göstermek için performansa dayalı olarak oluşturdukları bilimsel araştırma ödevlerini çok yüzeyle Rasch ölçme yöntemi ile değerlendirmektir. Rasch ölçme modeli bir soruya doğru cevap verme olasılığının, o kişinin yeteneğine bağlı olduğunu ifade eder. Çok yüzeyle Rasch ölçme modeli bu modele, puanlayıcı katılığı/cömertliği, soru güclüğü, gibi faktörleri de ekleyerek tek parametrelili modeli çok parametrelili bir model haline getirmiştir. Bu çalışmada öğrencilerin Bilimsel araştırma becerileri, puanlayıcıların katılık/cömertliklerine ve davranış sergileme gücüne göre değerlendirilmiştir. Araştırmada Pamukkale Üniversitesi, Eğitim Fakültesi, Eğitim Bilimleri Bölümü, Rehberlik ve Psikolojik Danışmanlık anabilim dalında “Bilimsel Araştırma Yöntemleri” dersine devam eden 20 öğrencinin katılımcı olarak kullanılmıştır. Verilerin analizi sonucunda, hazırlanan bilimsel araştırma ödevleri içerisinde “uygunluk içi” ve “uygunluk dışı” istatistiklere sahip ödevlerin olmadığı, ödevlerin istatistiksel olarak anlamlı yaklaşık 3 ödev kalite tabakasına ayrılabilmesi gözlemlenmiştir. Puanlayıcıların yapmış oldukları puanlamalarda ise istatistiksel olarak katılık/cömertlik bakımından birbirlerinden farklılıklar göstermediği, yalnızca bir puanlayıcı için “uygunluk dışı” kareler toplamının beklenen değerden farklı olduğu belirlenmiştir. Araştırmada, ödevlerinin kalitelerini belirlemede kullanılan “Bilimsel Araştırma Ödevlerini Değerlendirme Formu” içerisinde yer alan soruların da amaca hizmet eder nitelikte oldukları belirlenmiştir. Sonuç olarak yüksek öğrenimde özellikle performansa dayalı öğrenmelerin değerlendirilmesinde Çok-Yüzeyle Rasch ölçme modelinin etkin olarak kullanılabilmesi sonucuna varılmıştır.

Anahtar kelimeler: Ödev, yüksek öğrenim, çok yüzeyle Rasch Ölçme modeli, Madde Tepki Kuramı

Abstract

The purpose of this study is to investigate the usefulness of the many-facet Rasch model (MFRM) in evaluating the quality of performance related to preparing Research assignment in higher education. The Rasch Model utilizes item response theory stating that the probability of a correct response to a test item/task depends largely on a single parameter, the ability of the person. MFRM extends this one-parameter model to other facets, for example, rater severity, rating scale format, task difficulty levels. This paper specifically investigated research preparation ability in terms of items/task difficulty and rater severity/laxity. Fourth year counseling students prepared research assignments during the autumn semester of the 2009-2010 school years in the “Research Methods in Education” course. Six judges evaluated each students assignments using “Research Assignment Evaluation Rubric”. The results of this study demonstrated that the MFRM is a powerful tool for handling polytomous data in performance and peer assessment in higher education.

Key words: Research assignment, Higher education, Many-facet Rasch measurement, Item response theory

Yüksek öğrenimde öğrencilere kazandırılmaya çalışılan bilişsel, duyuşsal ve devinişsel alanların üst düzey davranış örüntülerini değerlendirmek için geleneksel olarak kullanılan ölçme araçlarının (yazılı, sözlü, çoktan seçmeli, eşleştirmeli, vb.) yetersiz kaldığı değişik araştırmalar tarafından belirlenmiştir (Berberoğlu, 1988; White, 1999; Wiliam, 2003, Baştürk, 2008). Eğitimde geleneksel olarak kullanılan ölçme araçları, öğrencilerin bilgilerini ölçmek için doğal olmayan (hayali)

* Doç. Dr. Pamukkale Üniversitesi, Eğitim Fakültesi, rbasturk@pau.edu.tr

bir ortam oluşturmakta ve öğrencilerin bu hayali ortamdaki oluşumlara verdikleri tepkiler dikkate alınarak öğrencilerin başarıları değerlendirilmektedir (Shepard, 2000). Buna paralel olarak geliştirilen performansa dayalı ölçme ve değerlendirmenin ana felsefesi ise, geleneksel olarak kullanılan ölçme araçlarının tersine, öğrenci için doğal bir ölçme ortamı oluşturmakta ve bu ortamda öğrencinin iş ya da eylemi gerçek olarak yapıp yapmadığını hem süreç hem de ürün bazında müzakere edilerek değerlendirmektedir. Bu durum, gerçekten de daha derin bir öğrenmenin öğrencilerde oluşup-oluşmadığını belirlemekte ve bazı bilgi ya da kavramları ezberleyerek önceden hazırlanan düzeneklere tepki vermek yerine öğrencileri ölçme işleminin öznesi hale getirerek değerlendirme işleminin gerçekleştirmektedir (Fosnot, 1996; Korkmaz ve Kaptan, 2005).

Ölçme sonuçlarını analiz etmede kullanılan kuramlar Klasik Test Teorisi (KTT) (Classical Test Theory) ve Örtük Özellikler Teorisi (ÖÖT) (Latent Trait Models) adı altında iki başlık altında toplamak mümkündür (Berberoğlu, 1988; Van der Linden ve Hambleton, 1997; Englehard, 1990). KTT'ye alternatif olarak geliştirilen ÖÖT'nin altında da "Madde Tepki Kuramı" (MTT) (Item response Theory) ve "Rasch" olmak üzere günümüzde halen üzerinde çalışılan ve gelişime açık iki farklı model yer almaktadır. Rasch tarafından (1980) geliştirilen modele (Bireylerin Yetenek Düzeyleri – Soruların Güçlük Düzeyleri), Linacre (1993), puanlayıcıların Katılığı/Cömertliği yüzeyini ekleyerek, Çok-Yüzeyle (Many-Facet) Rasch Ölçme Modelini geliştirmiştir.

Bu model geleneksel ölçme yöntemlerine göre birtakım üstün özelliklere sahip bulunmaktadır (Rasch, 1980; Linacre, 1993). Bunları kısaca özetlemek gerekirse: Çok-Yüzeyle Rasch modeli ham puanlar yerine kişilere ait (ölçme hatalarından arındırılmış) ölçme değerlerini kullanmaktadır. Çok-Yüzeyle Rasch modeli aynı zamanda her bir yüzey için (örneğin bu araştırmada bilimsel araştırma ödevlerinin kalitesi, değerlendirme sorularının özellikleri, puanlayıcı davranışı) yüzeyler arası olması gereken doğrusal bir bağlantının kurallarını belirlemekte ve bu bağlantıları oluşturmaktadır (Hambleton ve Swaminathan, 1985). Kısacası, Çok-Yüzeyle Rasch modeli, yansız ve etkili bir ölçme elde etmek için yüzeyleri ortak bir düzlemde birleştirerek standart hale getirmekte ve bireylerin bilimsel araştırma becerilerini, soruların güçlüğü ve puanlayıcı katılık ya da cömertlik davranışlarını aynı anda karşılaştırma olanağı sunmaktadır.

Bu araştırmanın amacı, bilişsel, duyuşsal ve devinişsel alan davranışlarının hepsinin üst basamaklarını da kapsayacak şekilde öğrencilerin kazanmış oldukları davranışları sergileyebildikleri ve öğrencilerin performanslarına dayalı olarak oluşturdukları bilimsel araştırma ödevlerini Çok yüzeyle Rasch ölçme yöntemi ile değerlendirmektir.

Yöntem

Çalışma Grubu

Bu araştırmanın çalışma grubunu, Pamukkale Üniversitesi, Eğitim Fakültesi, Eğitim Bilimleri Bölümü, Rehberlik ve Psikolojik Danışmanlık Anabilim dalında, 2009 – 2010 eğitim yılında "Bilimsel Araştırma Yöntemleri" dersini alan ve ikinci öğretime kayıtlı 20 öğrenci oluşturmaktadır.

Araştırma Verileri

Öğrencilerin bilimsel araştırma ödevlerini puanlamak için araştırmacı tarafından geliştirilen "Bilimsel Araştırma Ödevlerini Değerlendirme Formu" kullanılmıştır. Araştırmada 6 öğrenci puanlayıcı olarak tesadüfi yöntemle seçilerek analize dahil edilmiştir. Analizde 1) bireysel araştırmaların başarı ölçüleri ve uygunlu istatistikleri 2) puanlayıcıların katılık/cömertlik ve uygunluk istatistikleri, 3) Bilimsel Araştırma Ödevlerini değerlendirme formundaki soruların güçlük düzeyleri ve uygunluk istatistikleri belirlenmeye çalışılmıştır. Verilerin analizler için FACETS (Linacre, 1989-1998) bilgisayar programı kullanılmıştır

Bulgular

Model Veri Uyumu

Analizde kullanılan verilerin modelle uyumlu olabilmesi için standartlaştırılmış değerlerin (z-puanı) yaklaşım %5'inden daha azının mutlak değerce ikiden büyük veya ikiye eşit olması veya %1'inden daha azının mutlak değerce üçten büyük ya da eşit olması şartı aranır (Linacre, 2003). Analizde kullanılan toplam 1800 verinin standartlaştırılmış değerinin +/- 3'den büyük ya da eşit olanlarının sayısı 32 (% 0.02), +/- 2'den büyük olanların sayısı ise 56 (% 0.03) olarak elde edilmiş ve model veri uyumu sağlanmıştır.

Şekil 1'de, "Puanlayıcılar", "Araştırmalar" ve araştırmaları değerlendirmede kullanılan "Kriterler" ayrı sütunlar içerisinde gösterilmiştir. Puanlayıcıların yer aldığı sütun incelendiğinde, P1 kodlu puanlayıcının "en katı" ve P4 kodlu puanlayıcının ise "en cömert" puanlayıcı davranışına sahip oldukları, genel olarak puanlayıcıların birbirinden farklı puanlayıcı özelliklere sahip oldukları söylenebilir.

"Araştırmalar" sütunu incelendiğinde, puanlayıcılar tarafından en başarılı bulunan araştırma ödevinin A19 kodlu araştırma olduğu, bu araştırmayı A1 ve A4 kodlu araştırma ödevlerinin izlediği görülmektedir. Yine puanlayıcılara göre en başarısız olarak değerlendirilen araştırma ödevi ise A2 ve A3 kod numaraları ile verilen araştırmalar olarak yer almaktadır.

Bilimsel Araştırma ödevlerinin Çok-yüzeyle Rasch Ölçme Modeli ile Değerlendirilmesi 01-30-2010 16:09:06
Table 6.0 All Facet Vertical "Rulers".

Vertical = (1A,2A,3A) Yardstick (columns,lines,low,high)= 0,10,-1,2

Measr	Puanlayıcı	Araştırmalar	Kriterler	S.1
2				(5)
		A19		4
1		A1 A4	Bulguların tartışılması	
		A18 A8 A13 A17 A7	Yeni önerilerin geliştirilmesi	
		A15 A14	Bulguların yorumu	
	P4	A20	Örneklem / Çalışma grubunun seçimi özgünlük	---
	P2	A11 A6 A9	Problem ifadesi	
	P5 P6	A12 A16 A5	Teorik çerçeve	
0		A2 A3	Tablo ve şekillerin kullanımı	3
			Bulguların ifade edilmesi	
			Dil ve anlaşılabilirlik	
			Amacın belirtilmesi	---
	P3		Başlık ve içerik uyumu	
	P1		Analizlerin uygunluğu	2
			Yöntem	
-1			Veri toplama araçları	(0)
Measr	Puanlayıcı	Araştırmalar	Kriterler	S.1

Şekil 1. Veri Kalibrasyon Haritası

Bilimsel araştırma ödevlerini değerlendirmede kullanılan "Kriterlerin" yer aldığı sütun incelendiğinde ise, öğrenciler tarafından ödevlerinde yapılması en güç olarak karşılaşılan kriterin elde edilen "bulguların tartışılması" olduğu, zorluk bakımından bu kriteri "Yeni önerilerin geliştirilmesi"nin izlediği görülmektedir. Diğer yandan Öğrenciler tarafından en kolay bir şekilde gerçekleştirilen kriter ise "veri toplama aracının belirlenmesi" olarak görülmekte, bu kriteri araştırma "yönteminin" seçilmesi izlemektedir.

Bilimsel Araştırma Ödevleri

Öğrencilerin hazırlamış oldukları bilimsel araştırma ödevleri ile ilgili bilgiler Şekil 2’de ayrıntılı bir şekilde verilmiştir. Öğrencilerin bilimsel araştırma ödevleri kalitelerine en başarılı ödevden en başarısız ödevde göre sıralanmıştır. Buna göre araştırma kalitesi yönünden en üst düzeyde yer alan “A19” kodlu ödev iken, en düşük kaliteye sahip olarak görülen ödev, “A3” olarak kodlanan ödev olarak söylenebilir. Şekil 2’deki araştırma ödevlerinin kaliteleri ile ilgili bilgiler detaylı olarak incelendiğinde, ödevlerin kaliteleri ile ilgili logit değerlerine ait standart hata (RMSE, Root Mean Square Standard Error) değerinin 0,12 olduğu belirlenmiştir. RMSE, aşırı uçlarda yer alan değerler hariç bütün veriler için ölçme hatasını gösterir. Bu rakamın 0.12 olması, standart hatanın oldukça düşük olduğunu gösterir. Bu hata oranı dikkate alınarak düzeltilmiş standart sapma değeri de kritik değer olan 1.0’in altında bulunmuştur (0.33).

Bilimsel Araştırma Ödevlerinin Çok-yüzeyle Rasch Ölçme Modeli ile Değerlendirilmesi 01-30-2010 16:09:06
Table 7.2.1 Araştırmalar Measurement Report (arranged by 2MN).

obsvd Score	obsvd Count	obsvd Average	Fair Avrage	Measure	Model S.E.	Infit MnSq ZStd	Outfit MnSq ZStd	Nu Araştırmalar		
352	90	3.9	3.96	1.27	.14	1.3	1	19 A19		
344	90	3.8	3.89	1.12	.13	1.4	1	1 A1		
337	90	3.7	3.82	.99	.13	0.8	-1	4 A4		
334	90	3.7	3.79	.94	.13	1.1	0	8 A8		
331	90	3.7	3.76	.89	.13	1.0	0	18 A18		
328	90	3.6	3.73	.85	.13	1.5	2	7 A7		
328	90	3.6	3.73	.85	.13	0.9	0	13 A13		
324	90	3.6	3.70	.78	.12	1.0	0	17 A17		
321	90	3.6	3.67	.74	.12	1.3	1	15 A15		
311	90	3.5	3.58	.60	.12	1.0	0	14 A14		
306	90	3.4	3.53	.53	.11	1.3	1	20 A20		
293	90	3.3	3.41	.37	.11	0.7	-1	9 A9		
292	90	3.2	3.40	.36	.11	0.8	-1	6 A6		
292	90	3.2	3.40	.36	.11	0.9	0	11 A11		
291	90	3.2	3.39	.34	.11	0.9	0	5 A5		
288	90	3.2	3.36	.31	.11	1.3	1	16 A16		
286	90	3.2	3.34	.29	.11	0.7	-2	12 A12		
282	90	3.1	3.31	.24	.11	1.1	0	10 A10		
263	90	2.9	3.12	-.04	.10	1.0	0	2 A2		
257	90	2.9	3.05	-.02	.10	1.0	0	3 A3		
obsvd Score	obsvd Count	obsvd Average	Fair Avrage	Measure	Model S.E.	Infit MnSq ZStd	Outfit MnSq ZStd	Nu Araştırmalar		
308.0	90.0	3.4	3.55	.59	.12	1.1	0.2	1.0	-0.0	Mean (Count: 20)
26.2	0.0	0.3	0.25	.35	.01	0.2	1.2	0.2	1.0	S.D.

RMSE (Model) .12 Adj S.D. .33 Separation 2.83 Reliability .89
Fixed (all same) chi-square: 179.4 d.f.: 19 significance: .00
Random (normal) chi-square: 18.9 d.f.: 18 significance: .40

Şekil 2. Araştırma Ödevlerinin Kalitelerine Göre Karşılaştırılması

Rasch analizinde güvenilirlik (Reliability) kat sayısı, tıpkı KR-20 veya Cronbach Alpha gibi aynı anlamda hesaplanarak yorumlanır (Gözlenen varyansın gerçek varyansa oranı). Rasch analizinden elde edilen güvenilirlik kat sayısı, hangi güvenilirlikle okul öncesi kurumlarının kalite sıralamasının elde edildiğini gösterir. Güvenirlik kat sayısı olarak elde edilen 0.89 değeri, öğrencilerin ödevlerinin kalitelerine göre sıralamasının oldukça yüksek bir güvenilirlikle elde edildiğini göstermektedir.

Ayırma İndeksi (Separation Index) 2.83 ve güvenilirlik kat sayısı 0.89 ile sabit etkiye ait (Fixed effect) “öğrencilerin ödevlerinin kaliteleri arasında anlamlı farklılık vardır” hipotezi Kay-Kare ile test edildiğinde ($\chi^2 = 179.4$, $sd = 19$, $p = 0.00$) reddedilmiştir. Bir başka ifade ile, öğrencilerin ödevlerinin kalitelerinin birbirinden istatistiksel olarak farklılık gösterdiği söylenebilir.

Rasch analizi aynı zamanda yüzeyler ile ilgili “uygunluk içi” ve “uygunluk dışı” istatistiksel değerleri de vermektedir. Her iki değer içinde belirtilen kalite kontrol limiti 0.6 ile 1.4 aralığıdır (Wright, & Linacre, 1994). Uygunluk içi indeksi (Infit) karar verme noktasında beklenmeyen cevaplar için uygunluk dışı (Outfit) indeksi de uzaktaki beklenmeyen cevaplara (outlier) hassaslık gösteren değerlerdir. Şekil 2 incelendiğinde, her iki indeks içinde belirlenen sınırı aşan cevaplara rastlanmamıştır. Bu durumda öğrencilerin hazırlamış oldukları her bir bilimsel araştırma ödevi için

belirtilen uygunluk içi ve uygunluk dışı değerler beklenen kalite kontrol değerleri içinde yer aldığı ve uygun olarak kabul edileceği söylenebilir.

Puanlayıcı Analizi

Puanlayıcıların bilimsel araştırma ödevi hazırlayan öğrencilerin ödev hasırlama performansları ile ilgili bilgiler Şekil 3’de ayrıntılı bir şekilde verilmiştir. Puanlayıcılar en katı olandan en cömert olana doğru sıralandığında P1 kodlu puanlayıcının “en katı” puanlayıcı, P4 kodlu puanlayıcının ise “en cömert” puanlayıcı davranışına sahip olduğu söylenebilir.

Puanlayıcılar için katılık/cömertliklerine ait standart hata (RMSE), aşırı uçlarda yer alan değerler hariç bütün veriler için ölçme hatası için hesaplanan değerdir. Elde edilen rakamın 0.07 olması, standart hatanın oldukça düşük olduğunu gösterir. Bu hata oranı dikkate alınarak düzeltilmiş standart sapma değeri de kritik değer olan 1.0’in altında bulunmuştur (0.41). Puanlayıcıların puanlama davranışları ile ilgili güvenilirlik kat sayısı 0.98 olarak hesaplanmıştır. Bu sonuçta puanlayıcıların, puanlayıcı davranışlarının oldukça yüksek bir güvenilirlikle gerçekleştirildiğini göstermektedir.

Bilimsel Araştırma Ödevlerinin Çok-yüzeyle Rasch Ölçme Modeli ile Değerlendirilmesi 01-30-2010 16:09:06
Table 7.1.1 Puanlayıcı Measurement Report (arranged by 1MN).

Obsvd Score	Obsvd Count	obsvd Average	Fair Average	Model Measure	S.E.	Infit MnSq	Zstd	Outfit MnSq	Zstd	N	Puanlayıcı
1137	300	3.8	3.85	.46	.07	1.0	0	1.0	0	4	P4
1099	300	3.7	3.74	.26	.07	1.6	4	1.5	4	2	P2
1086	300	3.6	3.70	.20	.07	0.8	-2	0.8	-2	5	P5
1086	300	3.6	3.70	.20	.07	0.8	-2	0.7	-3	6	P6
923	300	3.1	3.24	-.42	.06	1.0	0	0.9	-1	3	P3
829	300	2.8	2.95	-.70	.05	1.1	1	1.2	1	1	P1
1026.7	300.0	3.4	3.53	.00	.06	1.0	0.2	1.0	-0.2	Mean (Count: 6)	
111.3	0.0	0.4	0.32	.41	.01	0.3	2.5	0.3	2.6	S.D.	

RMSE (Model) .07 Adj S.D. .41 Separation 6.26 Reliability .98
Fixed (all same) chi-square: 284.8 d.f.: 5 significance: .00
Random (normal) chi-square: 5.0 d.f.: 4 significance: .29

Şekil 3. Puanlayıcıların Puanlama Performanslarına Göre Karşılaştırılması

Ayırma İndeksi 6.26 ve güvenilirlik kat sayısı 0.94 ile sabit etkiye ait “Puanlayıcıların katılık/cömertlik dereceleri yönünden aralarında anlamlı farklılık vardır.” hipotezi Kay-Kare ile test edildiğinde ($\chi^2 = 284,8$ sd = 5, p = 0.00) yokluk hipotezi reddedilmiştir. Bir başka ifade ile puanlayıcıların katılık/cömertlik dereceleri bakımından kendi aralarında istatistiksel olarak anlamlı farklılıklar gösterdiği belirlenmiştir.

Yüzeyler ile ilgili “uygunluk içi” ve “uygunluk dışı” istatistiksel değerler Şekil 3’de incelendiğinde, sadece P2 kodlu puanlayıcının uygunluk içi (infit) ve uygunluk dışı (outfit) değerinin belirlenen sınırın dışında kaldığı (1.4 - 0.6 aralığı) belirlenmiştir. Bu durumda, her 6 puanlayıcı için belirtilen uygunluk içi ve uygunluk dışı değerler, beklenen kalite kontrol değerleri içinde yer aldığı ve uygun olarak kabul edileceği, fakat P2 kodlu puanlayıcının uygunluk içi ve uygunluk dışı kareler ortalamasının beklenen değerlerden daha yüksek olduğu, bir başka ifade ile, P2 kodlu puanlayıcının, öğrencilerin bilimsel araştırma ödevlerini değerlendirmede tutarlı puanlayıcı davranışına sahip olmadığı söylenebilir.

Bilimsel Araştırma Ödevlerini Değerlendirmede Kullanılan Kriterlerin Analizi

Öğrencilerin bilimsel araştırma ödevi hazırlama becerilerini ölçmek amacı ile kullanılan kriterlerin amaca uygunluğu ile ilgili bilgiler Şekil 4’de ayrıntılı bir şekilde verilmiştir.

Bilimsel Araştırma Ödevlerinin Çok-yüzeyle Rasch Ölçme Modeli ile Değerlendirilmesi 01-30-2010 16:09:06
Table 7.3.1 Kriterler Measurement Report (arranged by 3MN).

Obsvd Score	Obsvd Count	Obsvd Average	Fair Avrage	Measure	Model S.E.	Infit Mnsq	Zstd	Outfit Mnsq	Zstd	Nu Kriterler
288	120	2.4	2.51	1.02	.08	1.0	0	0.9	0	13 Bulguların tartışılması
317	120	2.6	2.78	.84	.08	1.8	5	1.9	5	14 Yeni önerilerin geliştirilmesi
357	120	3.0	3.11	.56	.09	0.9	0	0.9	0	12 Bulguların yorumu
374	120	3.1	3.24	.43	.09	0.9	0	0.9	0	6 Özgünlük
381	120	3.2	3.29	.37	.09	1.2	1	1.3	1	7 Örneklem / Çalışma grubunun seçimi
388	120	3.2	3.34	.31	.09	1.4	2	1.5	2	3 Problem ifadesi
415	120	3.5	3.53	.06	.10	0.8	-1	0.9	0	5 Teorik çerçeve
421	120	3.5	3.58	-.01	.10	1.1	0	1.2	1	15 Tablo ve şekillerin kullanımı
430	120	3.6	3.64	-.11	.11	0.6	-2	0.6	-2	11 Bulguların ifade edilmesi
437	120	3.6	3.69	-.19	.11	0.4	-4	0.4	-4	2 Dil ve anlaşılabilirlik
447	120	3.7	3.77	-.32	.11	1.0	0	0.9	0	4 Amacın belirtilmesi
458	120	3.8	3.85	-.46	.12	0.5	-3	0.6	-3	1 Başlık ve içerik uyumu
469	120	3.9	3.94	-.62	.12	0.7	-2	0.7	-2	10 Analizlerin uygunluğu
487	120	4.1	4.08	-.91	.13	1.6	3	1.4	2	9 Yöntem
491	120	4.1	4.11	-.98	.13	1.0	0	1.0	0	8 veri toplama araçları
410.7	120.0	3.4	3.50	.00	.10	1.0	-0.2	1.0	-0.2	Mean (Count: 15)
57.7	0.0	0.5	0.44	.58	.02	0.4	2.5	0.4	2.5	S.D.

RMSE (Model) .11 Adj S.D. .57 Separation 5.41 Reliability .97
Fixed (all same) chi-square: 480.5 d.f.: 14 significance: .00
Random (normal) chi-square: 14.0 d.f.: 13 significance: .38

Şekil 4. Bilimsel Araştırma Ödevlerini Değerlendirme Formunun Madde İstatistikleri

Şekil 4 incelendiğinde genel olarak öğrencilerin bilimsel araştırma yapma becerileri yönünden en zayıf kaldıkları noktaların elde edilen “bulguların tartışılması”, araştırma sonunda elde edilen sonuçlara göre “yeni önerilerin geliştirilmesi” ve “bulguların yorumu” konularının olduğu söylenebilir. Aynı şekilde araştırma ödevi hazırlayan öğrencilerin en kolay buldukları ve gerçekleştirmede zorlanmadıkları kriterler ise “veri toplama araçlarını” belirlemek, araştırma, “yöntemini” seçmek ve verileri uygun bir şekilde “analiz etmek” olarak karşımıza çıkmaktadır.

Öğrenciler tarafından hazırlanan bilimsel araştırma ödevlerini değerlendirmede kullanılan soruların analizi ile ilgili standart hatanın (RMSE) 0.11 olması, kalite belirleme ile ilgili standart hatanın oldukça düşük olduğunu gösterir. Bu hata oranı dikkate alınarak düzeltilmiş standart sapma değeri de kritik değer olan 1.0’in altında bulunmuştur (0.57). Öğrencilerin bilimsel araştırma ödevlerini değerlendirmede kullanılan kriterler ile ilgili güvenilirlik kat sayısı 0.97 olarak hesaplanmıştır. Bu sonuçta göstermektedir ki kullanılan kriterler, öğrencilerin bilimsel araştırma hazırlama ile ilgili becerilerini belirlemede oldukça güvenilir sonuçlar vermektedir.

Ayrırma İndeksi 5.41 ve güvenilirlik kat sayısı 0.97 ile sabit etkiye ait “öğrencilerin bilimsel araştırma ödevlerinin kalitelerini belirlemede kullanılan kriterlerin güçlükleri yönünden anlamlı farklılık vardır.” hipotezi Kay-Kare ile test edildiğinde ($\chi^2 = 480,5$, $sd = 14$, $p = 0.00$) yokluk hipotezi reddedilmiştir. Bir başka ifade ile kriterlerin, öğrencilerin araştırmalarını kalitelerini belirlemede, ödevlere ait farklı özellikleri ölçtükleri ve kriterlerin yerine getirilmeleri ile ilgili güçlükleri arasında da istatistiksel olarak anlamlı farklılıkların olduğu belirlenmiştir.

Yüzeyler ile ilgili “uygunluk içi” ve “uygunluk dışı” istatistiksel değerler Şekil 4’de incelendiğinde, her iki indeks içinde belirlenen sınır değerlerini aşan verilere rastlanmıştır. Önce Uygunluk içi (infit) değerleri incelendiğinde, 14, 2, 1 ve 9 kodlu soruların beklenen sınırların dışında değerler aldığı belirlenmiştir. Bu sorular için öğrencilerin ödev hazırlama performanslarını belirlerken tutarsızlıklar gösterdiği söylenebilir. Şekil 4’de yer alan Uygunluk Dışı (outfit) değerleri incelendiğinde ise, 14, 3 ve 2 numarası ile kodlanan soruların beklenen değerlerden farklı olduğu görülmektedir. Bu sorular dışında kalan soruların tamamının kabul edilebilir kullanım özelliklerine sahip oldukları ve hepsinin uygunluk içi veya uygunluk dışı kareler ortalamasının beklenen değerlerde yer aldıkları belirlenmiştir.

Tartışma ve Yorum

Bu araştırmada kullanılan çok - yüzeyle Rasch ölçme modeli ile öğrencilerin hazırlanmış oldukları bilimsel araştırma ödevleri ile ilgili analiz sonucunda, hazırlanan bilimsel araştırma ödevleri içerisinde uygunluk içi ve uygunluk dışı istatistiklerine sahip ödevlerin olmadığı, ödevlerin istatistiksel olarak anlamlı yaklaşık 3 ödev kalite tabakasına ayrılabilceği gözlenmiştir. Puanlayıcıların yapmış oldukları puanlamalarda ise istatistiksel olarak katılık/cömertlik bakımından birbirlerinden farklılıklar göstermediği belirlenmiştir. Yalnızca bir puanlayıcı için uygunluk dışı kareler toplamının beklenen değerden farklı olduğu belirlenmiştir. Araştırma ödevlerinin kalitelerini belirlemede kullanılan "Bilimsel Araştırma Ödevlerini Değerlendirme Formu" içerisinde yer alan soruların da amaca hizmet eder nitelikte oldukları bulunmuştur. Rasch ölçme modeli Cronbach alpha güvenilirlik katsayısına eşdeğer bir güvenilirlik sonucu vermektedir. Bir başka ifade ile Rasch ölçme modeli, grupların performanslarını kalitelerine göre, soruların zorluk ve kolaylıklarına göre ve puanlayıcıları katılık ya da cömertlik seviyelerine göre ayırırken bu işlemleri ne kadar bir güvenilirlik ile sağladığının istatistiksel değerini vermektedir. Geleneksel güvenilirlik sonuçlarını yorumlamada olduğu gibi Rasch güvenilirlik katsayısı da + 1.00'e yaklaştıkça güvenilirlik artmaktadır. Bu araştırmada, ödevlerin kalitelerini belirlemede 0.89, soruların kalitelerini belirlemede 0.97 ve puanlayıcıların katılık/cömertlik seviyelerini belirlemede ise 0.98'lik güvenilirlik katsayıları elde edilmiştir. Sonuç olarak yüksek öğrenimde özellikle performansa dayalı değerlendirmelerde Çok-Yüzeyle Rasch ölçme modelinin etkin olarak kullanılabilceği sonucuna varılmıştır.

Kaynaklar

- Baştürk, R. Sözlü Sunuma Dayalı Davranışların Çok-Yüzeyle Rasch Ölçme Modeli İle Analizi, I. Ulusal Eğitimde ve Psikolojide Ölçme ve Değerlendirme Kongresi, Cilt I, 671-683, Ankara, 2008.
- Berberoğlu, G. (1988). Seçme amacıyla kullanılan testlerde Rasch modelinin katkıları. Yayınlanmamış Doktora Tezi. Hacettepe Üniversitesi, Ankara.
- Englehard, G., Jr. (1990). Thorndike, Thurstone and Rasch: A comparison of their approaches to item-invariant measurement. Paper Presented at the American Educational Research Association conference, Boston.
- Fosnot, C. (1996) *Constructivism: Theory, perspectives, and practice*. New York: Teachers College Press.
- Hambleton, R.K. & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston: Kluwer. Hambleton, R. K., Swaminathan
- Korkmaz, H. ve Kaptan, F. (2005). Fen Eğitiminde Öğrencilerin Gelişimini Değerlendirmek İçin Elektronik Portfolyo Kullanımı Üzerine Bir İnceleme, *The Turkish Online Journal of Educational Technology*, 4(1), 101-107.
- Linacre, M. (1993). Generalizability theory and many-facet Rasch measurement. Paper presented at the annual meeting of the American educational research association, Atlanta, GA.
- Linacre, M. (1998). *A User guide to FACETS Rasch Measurement Computer Program and FACFORM Data formatting Computer Program*. Chicago, IL. : MESA Press
- Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests*. Chicago, IL.: MESA Press.
- Shepard, L. A. (2000). The role of assessment in a learning culture, *Educational Researcher*, 29(7), 4-14.
- White, J. (1999). Thinking about Assessment, *Journal of Philosophy of Education*, 33(2), 201.
- Wiliam, D. (2003). National curriculum assessment: How to make it better. *Research Papers in Education*, 18(2), 129-136.
- Van der Linden, W. J. & Hambleton, R. K. (1997). "Item response theory: Brief history, common models and extensions". In van der Linden, W. J. & Hambleton, R. K. (Eds.), *Handbook of Modern Item Response Theory*. New York: Springer.