Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
ISSN: 1309 - 6575

Performans Görevinin Akran Gruplar ve Öğretmen Yaklaşımları Doğrultusunda Çok-Yüzeyli Rasch Ölçme Modeli İle Analizi*

A Many-Facet Rasch Measurement Approach to Analyze Peer and Teacher Assessment for Authentic Assessment Task*

Ahmet Volkan YÜZÜAK ** Betül YÜZÜAK *** Fitnat KAPTAN****

Öz
Bu araştırmanın amacı Genetik ve Biyoteknoloji konulu performans görevinin (poster) akran grupları ve öğretmen tarafından çok-yüzeyli Rasch ölçme modeliyle değerlendirmektir. Araştırmada tarama modeli kullanılmıştır. Araştırma 2013-2014 eğitim yılı güz döneminde Bartın ilinde çalışan bir öğretmen ve 7. sınıfta öğrenim gören 50 öğrenci ile yürütülmüştür. Değerlendirme için beşer kişilik 10 öğrenci grubu oluşturulmuştur. Her grup kendi içerisinde birer materyal hazırlamıştır. Araştırmada veri toplama aracı olarak öğretmen tarafından revize edilen form kullanılmıştır. Verilerin analizi için FACETS programı kullanılmıştır. Ölçütler çerçevesinde elde edilen analiz sonuçlarına göre Grup 74'ün en cömert puanlayıcı grup olduğu, öğretmenin ise en katı puanlayıcı olduğu; P4'ün ölçütleri en iyi sağlayan görev olduğu, P6'nın ise ölçütleri daha iyi sağlamayan görev olduğu; Bilgi-doğruluk kodlu ölçütün en kolay karşılandığı, Kaynakça kodlu ölçütün ise en zor karşılandığı ifade edilebilir. Çok-yüzeyli Rasch ölçme modelinin fen eğitiminde performans ve akran grup değerlendirmeleri için etkili bir şekilde kullanılabileceği sonucuna ulaşılmıştır.

Anahtar Kelimeler: Fen eğitimi, performans görevleri, çok-yüzeyli Rasch ölçme modeli

Abstract
Aim of this study was to evaulate performance tasks related to Genetics and Biotechnology subject in view of peer groups and a science teacher through many-facet Rasch model (MFRM). Survey method was used. Students prepared these posters during the autumn semester of the 2013-2014 school years. Rasch model's surfaces are respectively: 10 peer groups and one science teacher, criterias and performance tasks which were handmade posters related to genetics and biotechnology. Each group has prepared one material. A form which was revised by the teacher was used as data collection tool. FACETS program was used for data analysis. Findings reveal that Group 74 is the most lenient group, teacher was the most severe assessor; P4 was found to be the most successful and P6 was found most unsuccessful; Knowledge-accuracy criteria was reached easily and reference criteria was reached hard. In conclusion, it is thought that the results of this study demonstrated MFRM can be used in elementary science education effectively to handle polytomous data in performance and peer assessment.

Key Words: Elementary science education, performance tasks, many-facet Rasch model

GİRİŞ

Eğitim süreci hedef, içerik, öğrenme-öğretme süreci ve değerlendirme aşamalarından oluşmaktadır. Ölçütler çerçevesinde karar verme anlamına gelen değerlendirme eğitim sisteminin sürekliliği, gelişebilmesi ve belirlenen kazanımlar doğrultusunda başarıya ulaşabilmesi için önemlidir (Baykul, 2010). Eğitimin değerlendirme aşamasında geleneksel veya alternatif ölçme ve değerlendirme araçlarından yararlanılmaktadır.

* Bu çalışma "YILDIZ International Conference on Educational Research and Social Studies" kongresinde sözlü bildiri olarak sunulmuştur.

**Araş. Gör., Bartın Üniversitesi, Eğitim Fakültesi, Bartın-TÜRKİYE, e-posta: ahmetvolkanyuzuak@gmail.com
***Öğretmen, Hendekyanı Ortaokulu, Bartın-TÜRKİYE, e-posta: betulyumrutas@gmail.com
****Prof. Dr., Hacettepe Üniversitesi, Eğitim Fakültesi, Ankara-TÜRKİYE, e-posta: fitnat@hacettepe.edu.tr

Geleneksel ölçme-değerlendirme araçları öğrenciyi düşünmeye sevk etme konusunda yeterli olamamaktadır. Bu sebeple, öğrencilerin düşünme becerileri istenilen seviyede gelişememektedir (Lawson, 1995; Çalışkan ve Kaptan, 2012). Bu durum bizi alternatif ölçme ve değerlendirme araçlarına yönlendirmektedir. Alternatif ölçme kavramı, yazılı yoklamalar, kısa cevaplı sınavlar, doğru-yanlış testleri, çoktan seçmeli testlerin yerine kullanılmaktadır. Alternatif ölçme-değerlendirme araçlarını değerlendirmede bazı olumsuz yanları olmakla birlikte akran gruplar sürece dahil edilebilir. Karar verme sürecine katılan öğrenciler, öğrenme sürecine aktif bir şekilde katılacaktır. Bu noktada ölçütlerin kesin bir şekilde belirlenmesi ve öğretmenin tutumu önemlidir (Peters ve Stout, 2011; Turgut ve Baykul, 2012).

Ölçme araçları değerlendirilirken genelde iki kuramdan faydalanılmaktadır: Klasik Test Kuramı ve Örtük Özellikler Kuramı altında yer alan Madde Tepki Kuramı. Madde Tepki Kuramı ve Klasik Test Kuramı arasında benzerlikler olmasına rağmen Madde Tepki Kuramının avantajları olduğu söylenebilir (Hambleton ve Swaminathan, 1985). Klasik Test Kuramında bireyin maddenin zorluk veya kolaylık derecesine bakılmaksızın doğru cevaplandırması tam puan alması, yanlış cevaplandırması ise puan alamaması anlamına gelmektedir ve farklı maddeleri doğru cevaplayarak aynı puanı almış iki öğrencinin bilgi ve beceri seviyeleri eşit olarak değerlendirilmektedir (Kaptan, 1994; Bahar, 2001). Madde Tepki Kuramını temel alan modeller aracılığıyla ise doğrusal ilişki elde edilebilir, kayıp verilerin üstesinden gelinebilir, ölçümün hassasiyeti tahmin edilebilir ve standart değerlere uymayan sonuçlar değerlendirilebilir (Elhan ve Atakurt, 2005; Aziz ve Masodi, 2010).

Madde Tepki Kuramını temel alan, gruptan bağımsız istatistikler verebilen ve bir parametreli lojistik model olarak bilinen Rasch modeli, George Rasch tarafından "Yetenek ve Zeka Testleri İçin Olasılık Modeli (Some Probabilistic Models for Intelligence and Attainment Tests" adlı kitapta açıklanmıştır (Rasch, 1960). Rasch ölçme modeli objektif bir ölçme işlemi için, ham puanları logit değerlere dönüştürmekte ve analizde bu değerleri kullanmaktadır. Çok yüzeyli Rasch Modeli ise bir parametreli Rasch Modeline puanlayıcıların katılığı/cömertliği yüzeyinin eklenmesi sonucu geliştirilmiştir (Linacre, 1993).

Rasch analizi tıp, müzik eğitimi ve beden eğitimi gibi alanlarda kullanılmaktadır (Elhan ve Atakurt, 2005; Akın ve Baştürk, 2012, Arsan, 2012). Çok-yüzeyli Rasch modeli ile öğrencilerin sunumlarının, bilimsel ödevlerinin, eğitim gördükleri bölüme ilişkin görüşlerinin, mikro öğretim uygulamalarının değerlendirildiği araştırmalar bulunmaktadır (Baştürk, 2009; Semerci 2011a; Semerci, 2011b; Semerci, Semerci ve Duman, 2013). Örneğin, Akın ve Baştürk (2012) Güzel Sanatlar ve Spor Lisesi öğrencilerinin keman eğitiminde kazanmış oldukları temel becerileri değerlendirmiştir. Semerci (2011) "Özel Öğretim Yöntemleri" dersini alan öğretmen adaylarının mikro öğretim uygulamalarını değerlendirmiştir.

Akran grupların yapmış olduğu değerlendirmelerde Çok-yüzeyli Rasch ölçme modelinden yararlanılabilir. 2013 Fen Bilimleri Dersi Öğretim Programında ölçme ve değerlendirme anlayışı olarak "tamamlayıcı ölçme araç ve tekniklerinin kullanımı ile birlikte sürece dönük değerlendirme yaklaşımına önem verilerek öğrencinin kendini ve akranını değerlendirme şansı bulduğu öz ve akran değerlendirme yaklaşımları benimsenmiştir" (MEB, 2013:4) ifadesi yer almaktadır. Bu araştırmanın amacı; fen eğitimi dersinde hazırlanan, öğrenme ürünü olarak da ifade edilebilen performans görevlerinin akran gruplar aracılığıyla değerlendirilmesini objektif bir biçimde yapmaktır. Bu amaç çerçevesinde aşağıdaki sorulara cevap aranmıştır:

1. Öğrenci gruplarının performans görevine ilişkin görüşlerinin genel analizi nasıldır?
2. Öğrenci grup puanlayıcılarının katılıkları veya cömertliklerine ilişkin analizi nedir?

3. Performans görevlerine ilişkin ölçütlerin gerçekleşme düzeyleri nasıldır?
4. Öğrenci gruplarının yanlılık analizi sonuçları nelerdir?

YÖNTEM

Araştırmada tarama modeli kullanılmıştır. "Tarama modelleri, geçmişte ya da halen varolan bir durumu varolduğu şekliyle betimlemeyi amaçlayan araştırma yaklaşımlarıdır. Araştırmaya konu olan olay, birey ya da nesne, kendi koşulları içinde ve olduğu gibi tanımlanmaya çalışılır" (Karasar, 2012). Bu amaçla hazırlanan likert tipi form yardımıyla akran gruplarının ve öğretmenin performans görevleri hakkındaki görüşleri belirlenmiştir.

Çalışma Grubu

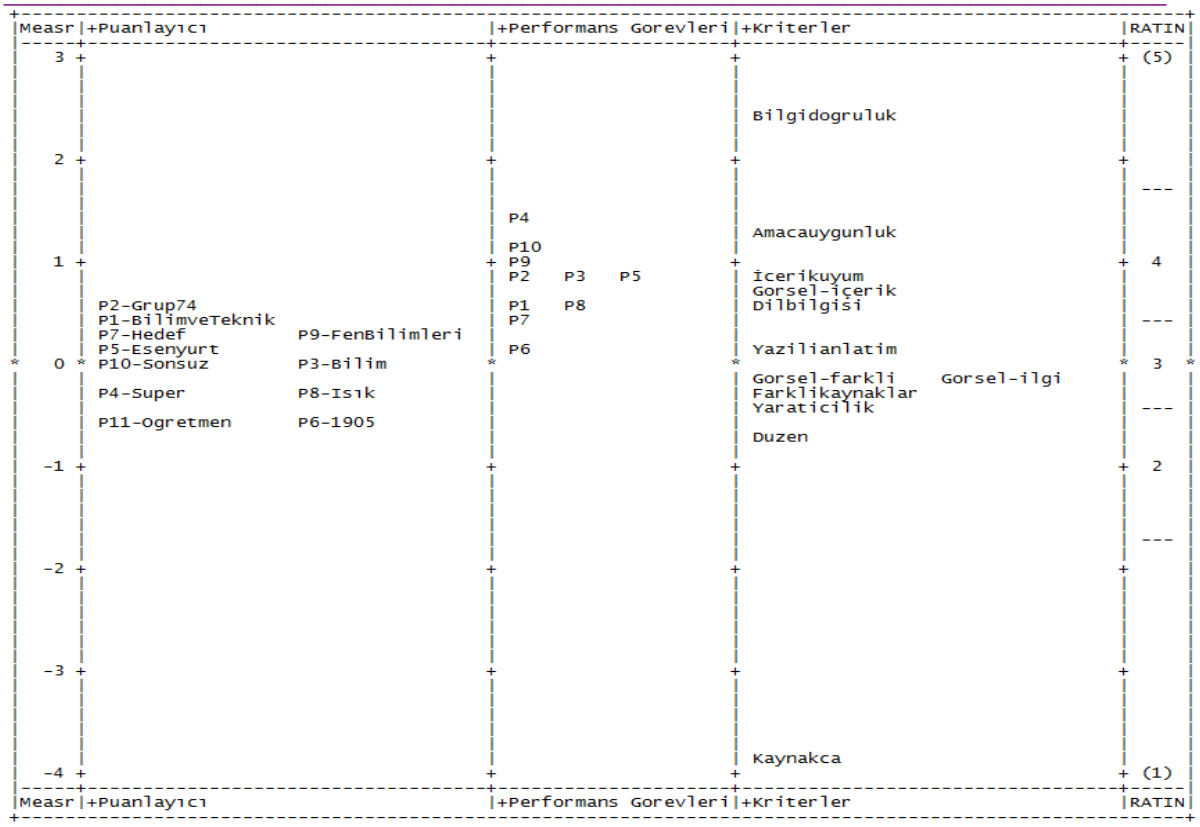
Rasch ölçme modelinde örneklemden elde edilen verilerin sonuçlarının evrene genelleme gibi bir varsayımı bulunmamaktadır (Linacre, 1993). Bu sebeple araştırmada evren ve örneklem tayinine gidilmemiştir. Araştırmanın çalışma grubunu, bir Fen Bilgisi öğretmeni ve 2013-2014 güz yarıyılında Bartın ilinde merkeze bağlı bir köy okulunda okumakta olan 7. sınıf 50 öğrenci (10 grup) oluşturmaktadır. Grup isimleri: Bilim ve Teknik, Grup 74, Bilim, Süper, Esenyurt, Grup 1905, Hedef, Grup Işık, Fen Bilimleri, Sonsuz'dur.

Veri Toplama Teknikleri

Veri toplama aracı için öğretmen tarafından hazırlanan taslak form incelenmiş ve Ölçme ve Değerlendirme alanında iki uzmana (Doktor), Fen Bilgisi eğitimi alanında yüksek lisansını tamamlamış bir uzmana ve iki adet Fen Bilgisi öğretmenine gösterilerek uzman görüşleri alınmıştır. Projenin amaca uygunluğu ölçütü amaca uygunluk, proje adının içerikle uyumu içerik uyum, farklı kaynaklardan bilgi toplama farklı kaynaklar, derlenen bilgilerin analiz/şema ve grafiklerin çizimi duzen, yazılı anlatımın açık-seçik, özlü, akıcı ve sürükleyici olması dilbilgisi, bilgilerin doğruluğu bilgi-dogruluk, yaratıcılık yeteneğini kullanma yaratıcılık, görsel öğelerin içeriğe uygunluk düzeyi görsel-içerik, görsel ögelerin ilgi çekme düzeyi görsel-ilgi, farklı görsel ögelere yer verilme düzeyi görsel-farklı, kaynakça belirtme kaynakça, Türkçe yazım ve noktalama kurallarına uyma dilbilgisi şeklinde kodlanmıştır. Belirtilen ölçütler hazırlanan formda, kesinlikle katılmıyorum: 1, katılmıyorum: 2, kısmen katılıyorum: 3, katılıyorum: 4 ve tamamen katılıyorum: 5 aralığında derecelendirilmiştir. Elde edilen verilerin analizi Linacre (1993) tarafından geliştirilen ve Rasch ölçme modelini temel alan FACETS 3.71.4 (Linacre, 2014) programı ile yapılmıştır.

BULGULAR

Araştırmanın bu kısmında bulgulara yer verilmiştir. Bulgularda 10 öğrenci grubu, bir Fen Bilgisi öğretmeni, ölçüt olarak 12 madde ve ürün olarak 10 performans görevi dikkate alınmıştır. Şekil 1′ de analize ilişkin genel bir kalibrasyon haritası gösterilmiştir.



Şekil 1. Veri kalibrasyon haritası

Şekil 1'de görüldüğü üzere; Grup 74, Bilim ve Teknik, Hedef, Fen Bilimleri, Esenyurt orta düzeyinin üstünde; Bilim, Süper, 1905, Işık ve Öğretmen ise orta düzeyin altında puanlama yapmışlardır. Belirlenen ölçütlere göre daha yeterli olan P4, P10 iken P7 ve P6 daha yetersiz performans görevleridir. Tüm performans görevleri orta düzeyin üstünde bir yığılma göstermiştir. Belirlenen ölçütler doğrultusunda performans görevlerinin yeterliliğinin tespiti için her bir puanlayıcı grubu, performans görevi ve ölçüt için logit değerleri hesaplanmıştır. Tablo 1'de ilgili logit değerleri gösterilmiştir.

Tablo 1. *Puanlayıcılar, Performans Görevleri ve Ölçütler İçin Logit Değerleri*

Puanlayıcılar	Logit Değerleri	Performans Görevleri	Logit Değerleri	Ölçütler	Logit Değerleri
BilimveTeknik	0.39	P1	0.62	Amacauygunluk	-1.22
Grup74	0.6	P2	0.8	İcerikuyum	-0.79
Bilim	- 0.02	P3	0.8	Farklikaynaklar	0.35
Super	- 0.21	P4	1.39	Duzen	0.78
Esenyurt	0.2	P5	0.92	Yazilianlatim	-0.15
1905	- 0.51	P6	0.08	Bilgidogruluk	-2.5
Hedef	0.22	P7	0.46	Yaraticilik	0.5
Isık	- 0.33	P8	0.51	Gorsel-içerik	-0.65
Fen Bilimleri	0.22	P9	1.06	Gorsel-ilgi	0.2
Sonsuz	0.06	P10	1.21	Gorsel-farkli	0.16
Ogretmen	- 0.61			Kaynakca	3.83
				Dilbilgisi	-0.51

Tablo 1'de ifade edilen logit değerlerine göre en cömert grup puanlayıcısı Grup 74 (logit Grup 74=0.6) iken en katı puanlayıcı Öğretmen (logit Ogretmen=-0.61)' dir. P4 (logit P4=1.39) kodlu performans görevi en yeterli iken, P6 (logit P6=0.08) kodlu performans

görevi daha yetersizdir. Ayrıca, en zor gerçekleşen ölçüt Kaynakça (logit Kaynakca=3.83) iken en kolay gerçekleşen ölçüt Amacauygunluk (logit Amacauygunluk=-1.22) olmuştur.

Öğrenci Grup Puanlayıcılarının Katılıkları veya Cömertliklerine Yönelik Bulgular

Tablo 2'de puanlayıcıların katılık ve cömertliklerine ilişkin analiz sonuçları gösterilmiştir.

Tablo 2. *Puanlayıcıların Katılık/cömertlik Karşılaştırması*

Total Score	Total Count	Obsvd Average	Fair(M) Average	Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Correlation PtMea	PtExp	Nu	Puanlayıcı
491	120	4.09	4.32	.60	.13	1.00	.0	.80	-1.0	1.08	.77	.76	2	P2-Grup74
477	120	3.97	4.18	.39	.12	1.13	.8	1.13	.7	.95	.73	.76	1	P1-BilimveTeknik
465	120	3.88	4.05	.22	.11	.92	-.5	.90	-.5	1.10	.78	.75	7	P7-Hedef
465	120	3.88	4.05	.22	.11	1.27	1.7	1.00	.0	1.09	.75	.75	9	P9-FenBilimleri
463	120	3.86	4.03	.20	.11	1.17	1.1	1.15	.8	.98	.74	.75	5	P5-Esenyurt
452	120	3.77	3.92	.06	.11	.93	-.4	1.66	3.2	.82	.73	.75	10	P10-Sonsuz
445	120	3.71	3.84	-.02	.11	1.01	.0	1.10	.6	.91	.73	.74	3	P3-Bilim
428	120	3.57	3.66	-.21	.10	1.13	1.0	1.01	.1	1.05	.73	.73	4	P4-Super
417	120	3.47	3.53	-.33	.10	.76	-1.9	.80	-1.2	1.15	.76	.73	8	P8-Isık
400	120	3.33	3.34	-.51	.10	1.06	.5	.97	-.1	1.04	.73	.72	6	P6-1905
390	120	3.25	3.22	-.61	.10	.92	-.6	.98	.0	.88	.69	.72	11	P11-Ogretmen
444.8	120.0	3.71	3.83	.00	.11	1.03	.2	1.05	.2		.74			Mean (Count: 11)
30.8	.0	.26	.34	.36	.01	.14	1.0	.22	1.1		.02			S.D. (Population)
32.3	.0	.27	.35	.38	.01	.14	1.0	.23	1.2		.02			S.D. (Sample)

Model, Populn: RMSE .11 Adj (True) S.D. .34 Separation 3.10 Strata 4.47 Reliability .91
Model, Sample: RMSE .11 Adj (True) S.D. .36 Separation 3.27 Strata 4.69 Reliability .91
Model, Fixed (all same) chi-square: 118.7 d.f.: 10 significance (probability): .00
Model, Random (normal) chi-square: 9.2 d.f.: 9 significance (probability): .42

Tablo 2'ye göre puanlayıcı ayırma indeksi 3.10 ve güvenirlik katsayısı 0.91 ile sabit etkiye ait 'Puanlayıcıların katılık veya cömertlikleri arasında farklılık vardır' hipotezi ki-kare ile test edildiğinde (χ^2=118.7, sd=10, p=0.00) yokluk hipotezi reddedilmiştir. Bu 10 grubun puanlamalarının katılık/cömertlikleri arasında istatistiksel olarak anlamlı bir farklılığın bulunduğu anlamına gelmektedir. Grup 74 gözlenen puan 491 ile en cömert, Öğretmen gözlenen puan 390 ile en katıdır. En cömert puanlamadan en katı puanlamaya doğru sıralanışı şu şekildedir: Grup 74, Bilim ve Teknik, Hedef, Fen Bilimleri, Esenyurt, Sonsuz, Bilim, Super, Işık, Grup 1905, Öğretmendir.

Performans Görevlerine İlişkin Ölçütlerin Gerçekleşme Düzeylerine Yönelik Bulgular

Tablo 3'te performans görevlerine ilişkin ölçüm raporları gösterilmiştir.

Tablo 3. *Performans Görevlerine İlişkin Ölçüm Raporu*

Total Score	Total Count	Obsvd Average	Fair(M) Average	Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Correlation PtMea	PtExp	Nu	Performans Gorevleri
541	132	4.10	4.32	1.39	.12	1.05	.3	.88	-.6	1.04	.76	.76	4	P4
528	132	4.00	4.21	1.21	.11	.93	-.4	.88	-.6	1.06	.77	.76	10	P10
516	132	3.91	4.10	1.06	.11	1.05	.3	1.02	.1	1.03	.75	.75	9	P9
504	132	3.82	3.98	.92	.11	1.09	.7	1.64	3.2	.84	.71	.75	5	P5
493	132	3.73	3.88	.80	.10	.96	-.3	1.00	.0	.99	.74	.74	2	P2
493	132	3.73	3.88	.80	.10	1.30	2.1	1.46	2.5	.58	.66	.74	3	P3
476	132	3.61	3.71	.62	.10	.95	-.3	.87	-.8	1.19	.77	.74	1	P1
465	132	3.52	3.60	.51	.10	1.11	.9	1.15	.9	.92	.71	.73	8	P8
459	132	3.48	3.53	.46	.10	.75	-2.2	.66	-2.3	1.38	.79	.73	7	P7
418	132	3.17	3.11	.08	.09	1.04	.3	.92	-.4	1.07	.72	.71	6	P6
489.3	132.0	3.71	3.83	.79	.11	1.02	.2	1.05	.2		.74			Mean (Count: 10)
34.5	.0	.26	.34	.37	.01	.14	1.1	.28	1.6		.04			S.D. (Population)
36.4	.0	.28	.36	.39	.01	.14	1.1	.29	1.7		.04			S.D. (Sample)

Model, Populn: RMSE .11 Adj (True) S.D. .35 Separation 3.34 Strata 4.79 Reliability .92
Model, Sample: RMSE .11 Adj (True) S.D. .37 Separation 3.54 Strata 5.05 Reliability .93
Model, Fixed (all same) chi-square: 122.1 d.f.: 9 significance (probability): .00
Model, Random (normal) chi-square: 8.4 d.f.: 8 significance (probability): .40

Tablo 3'e göre Rasch analizinde performans görevlerine ilişkin güvenirlik katsayısı 0.92'dir. Ayırma indeksi 3.34 ve güvenirlik katsayısı 0.92 ile sabit etkiye ait 'Maddeler

açısından performans görevlerinin etkinliği arasında anlamlı bir fark vardır' hipotezi ki-kare ile test edildiğinde (χ^2 =122.1, Sd=9, p=0.00) yokluk hipotezi reddedilmiştir. Bu anlamda, performans görevlerinin arasında istatistiksel açıdan anlamlı bir farklılık vardır. Performans görevlerinin yeterlilik sıralanışı şu şekildedir: P4, P10, P9, P5, P2, P3, P1, P8, P7, P6'dır.

Performans Görevlerine İlişkin Ölçütlerin Gerçekleşme Düzeylerine Yönelik Bulgular

Tablo 4'te performans görevlerinin değerlendirilmesinde kullanılan ölçütlerin istatistikleri gösterilmiştir.

Tablo 4.*Performans Görevlerinin Değerlendirilmesinde Kullanılan Ölçütlerin İstatistikleri*

Total Score	Total Count	Obsvd Average	Fair(M) Average	Model Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Correlation PtMea	Correlation PtExp	Nu	Kriterler
536	110	4.87	4.89	2.50	.27	.81	-.5	.75	-.7	1.05	.30	.18	6	Bilgidogruluk
503	110	4.57	4.61	1.22	.15	1.23	1.2	1.15	.8	.99	.38	.31	1	Amacauygunluk
481	110	4.37	4.43	.79	.13	1.23	1.3	1.14	.8	1.00	.46	.36	2	İcerikuyum
472	110	4.29	4.35	.65	.12	1.16	1.0	1.07	.4	1.01	.50	.38	8	Gorsel-içerik
462	110	4.20	4.26	.51	.12	1.09	.6	1.12	.7	.86	.26	.40	12	Dilbilgisi
433	110	3.94	4.00	.15	.11	.74	-2.1	.78	-1.6	1.17	.37	.43	5	Yazilianlatim
402	110	3.65	3.71	-.16	.10	1.10	.8	1.11	.8	.88	.44	.46	10	Gorsel-farkli
398	110	3.62	3.67	-.20	.10	.90	-.8	.90	-.7	1.24	.65	.46	9	Gorsel-ilgi
382	110	3.47	3.51	-.35	.10	.75	-2.2	.78	-1.8	1.22	.38	.47	3	Farklikaynaklar
365	110	3.32	3.35	-.50	.09	.77	-2.0	.77	-2.0	1.39	.60	.47	7	Yaraticilik
332	110	3.02	3.02	-.78	.09	1.39	3.0	1.44	3.3	.25	.14	.48	4	Duzen
127	110	1.15	1.14	-3.83	.24	1.72	2.2	1.55	1.6	.97	.24	.19	11	Kaynakca
407.8	110.0	3.71	3.74	.00	.13	1.07	.2	1.05	.1		.39		Mean (Count: 12)	
102.4	.0	.93	.95	1.44	.06	.29	1.7	.25	1.5		.14		S.D. (Population)	
106.9	.0	.97	.99	1.50	.06	.30	1.8	.26	1.6		.15		S.D. (Sample)	

Model, Populn: RMSE .15 Adj (True) S.D. 1.43 Separation 9.75 Strata 13.34 Reliability .99
Model, Sample: RMSE .15 Adj (True) S.D. 1.49 Separation 10.19 Strata 13.92 Reliability .99
Model, Fixed (all same) chi-square: 599.2 d.f.: 11 significance (probability): .00
Model, Random (normal) chi-square: 10.7 d.f.: 10 significance (probability): .38

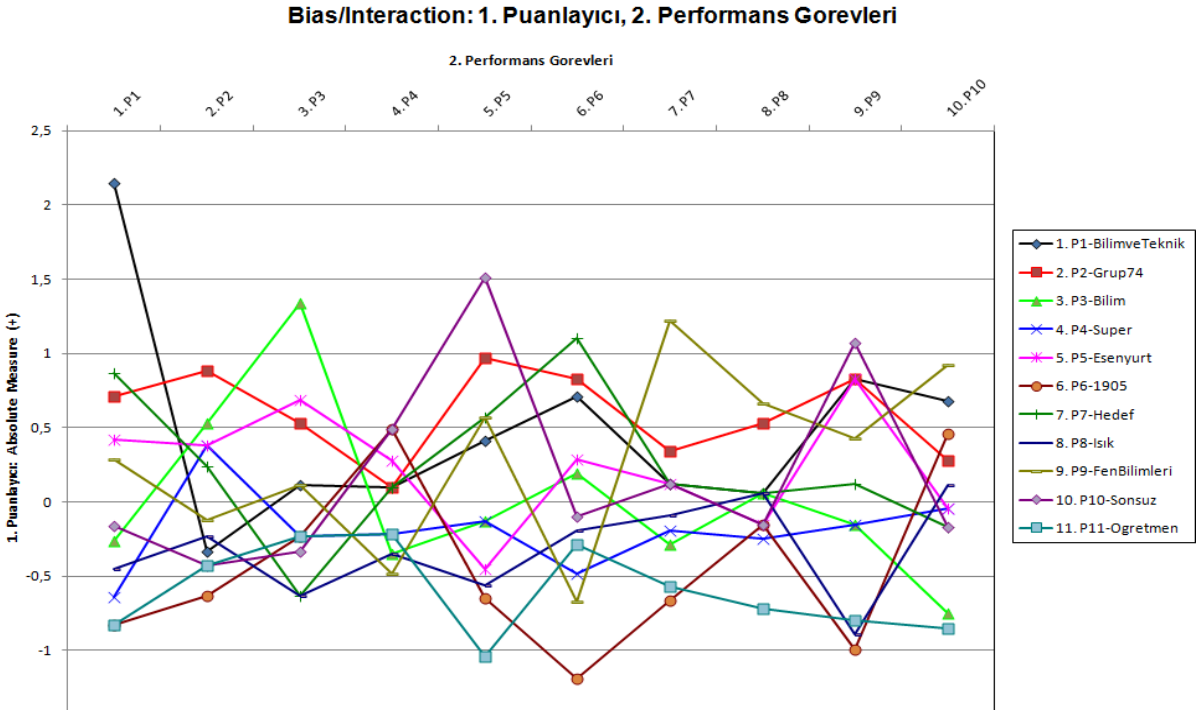
Tablo 4'e göre ayırma indeksi 9.75 ve güvenirlik katsayısı 0.99 ile sabit etkiye ait "Performans görevlerinin değerlendirilmesinde kullanılan maddelerin gerçekleşme düzeyleri arasında anlamlı bir fark vardır' hipotezi ki-kare ile test edildiğinde (χ^2 =599.2, sd=11, p=0.00) yokluk hipotezi reddedilmiştir. Buna göre, performans görevlerinin değerlendirilmesinde kullanılan maddelerin gerçekleşme düzeyleri arasında istatistiksel açıdan anlamlı farklılık bulunmaktadır. Performans görevlerinde en zor gerçekleşen madde "Kaynakça Belirtme-Kaynakça" ve "Derlenen bilgilerin analiz/şema ve grafiklerin çizimi-Düzen" iken en kolay gerçekleşen maddeler "Bilgilerin doğruluğu-Bilgidogruluk" ve "Projenin amaca uygunluğu-Amcauygunluk" olmuştur. Rasch analizi ile yüzeylere ait uygunluk içi ve uygunluk dışı değerleri 0.5-1.5 aralığında olmalıdır (Wright, Linacre, Gustafson ve Martin-Lof, 1994). " Kaynakça" ölçütünün uygunluk içi ve uygunluk dışı değerlerin beklenen kontrol değerleri içinde yer almadığı söylenebilir. Tablo 5'de puanlayıcılar ile değerlendirilmesi yapılan performans görevlerilerin etkileşim analizi, Şekil 2 de ise bu analizin grafiği gösterilmiştir.

Tablo 5.*Puanlayıcılar ile Değerlendirilmesi Yapılan Performans Görevlerinin Etkileşim Analizi*

Observd Score	Expctd Score	Observd Count	Obs-Exp Average	Bias Size	Model S.E.	t	d.f.	Prob.	Infit MnSq	Outfit MnSq	Puanlayıcı Sq	Puanlayıcı Nu	Puanlayıcı	Performans Go measr	Performans Nu	Performans Per measr
31	40.37	12	-.78	-.89	.31	-2.85	11	.0157	1.2	1.0	64	9	P9-FenBilimleri	.22	6	P6 .08
39	46.85	12	-.65	-.85	.31	-2.75	11	.0189	1.5	1.4	29	7	P7-Hedef	.22	3	P3 .80
42	48.08	12	-.51	-.72	.32	-2.26	11	.0455	1.0	.8	102	3	P3-Bilim	-.02	10	P10 1.21
42	48.05	12	-.50	-.72	.32	-2.24	11	.0468	.5	.8	12	1	P1-BilimveTeknik	.39	2	P2 .80
26	32.68	12	-.56	-.69	.34	-2.03	11	.0673	.5	.5	61	6	P6-1905	-.51	6	P6 .08
42	47.57	12	-.46	-.65	.32	-2.03	11	.0674	1.1	.9	49	5	P5-Esenyurt	.20	5	P5 .92
46	50.69	12	-.39	-.70	.35	-2.00	11	.0706	1.6	1.2	42	9	P9-FenBilimleri	.22	4	P4 1.39
53	47.59	12	.45	1.01	.52	1.96	11	.0756	1.8	2.0	98	10	P10-Sonsuz	.06	9	P9 1.06
52	45.78	12	.52	1.00	.48	2.10	11	.0594	.8	.6	39	6	P6-1905	-.51	4	P4 1.39
51	44.24	12	.56	.97	.44	2.19	11	.0511	.6	.4	105	6	P6-1905	-.51	10	P10 1.21
51	43.99	12	.58	1.00	.44	2.25	11	.0458	.5	.4	75	9	P9-FenBilimleri	.22	7	P7 .46
48	40.37	12	.64	.88	.38	2.32	11	.0408	.4	.4	62	7	P7-Hedef	.22	6	P6 .08
54	46.53	12	.62	1.45	.56	2.56	11	.0264	5.1	3.0	54	10	P10-Sonsuz	.06	5	P5 .92
53	44.84	12	.68	1.36	.52	2.63	11	.0234	2.0	2.1	25	3	P3-Bilim	-.02	3	P3 .80
55	46.75	12	.69	1.77	.62	2.87	11	.0152	.7	.4	1	1	P1-BilimveTeknik	.39	1	P1 .62
44.5	44.47	12.0	.00	.05	.36	.00			.9	.9	Mean (Count: 110)					
5.6	4.41	.0	.28	.47	.06	1.16			.6	.5	S.D. (Population)					
5.6	4.43	.0	.29	.47	.06	1.16			.6	.5	S.D. (Sample)					

Fixed (all = 0) chi-square: 146.9 d.f.: 110 significance (probability): .01

Şekil 2 de ise bu analizin grafiği gösterilmiştir.



Şekil 2. Puanlayıcılar ile Performans Görevlerinin Etkileşim Grafiği

Tablo 5 ve Şekil 2 incelendiğinde Bilim ve Teknik, Bilim, Esenyurt, 1905, Hedef, Fen Bilimleri, Sonsuz grupları yanlı davranımlarda bulunmuştur. Çok yüzeyli Rasch analizine göre, Bilim ve Teknik grubunun, P2 kodlu performans görevine 48.05 vermesi gerekirken 42 puan verdiği; P1 kodlu performans görevine 46.75 puan vermesi gerekirken 55 puan verdiği, Bilim grubunun P3 kodlu performans görevine 44.84 puan vermesi gerekirken 53 puan verdiği, P10 kodlu performans görevine 48.08 puan vermesi gerekirken 42 puan verdiği; Esenyurt'un P5 kodlu performans görevine 47.57 puan vermesi gerekirken 42 puan verdiği, 1905'in P4 kodlu performans görevine 45.78 puan vermesi gerekirken 52 puan verdiği; P10 kodlu performans görevine 44.24 puan vermesi gerekirken 51 puan verdiği, Hedef'in P3 kodlu performans görevine 46.85 puan vermesi gerekirken 39 puan verdiği; P6 kodlu performans görevine 40.37 puan vermesi gerekirken 48 puan verdiği, Fen Bilimlerinin P6 kodlu performans görevine 40.37 puan vermesi gerekirken 31 puan verdiği, P7 kodlu performans görevine 43.99 puan vermesi gerekirken 51 puan verdiği; Sonsuz'un P5 kodlu performans görevine 46.53 puan vermesi gerekirken 54 puan verdiği, P9 kodlu performans görevine 47.59 puan vermesi gerekirken 53 puan verdiği görülmektedir.

SONUÇLAR ve TARTIŞMA

Bu araştırmada, Fen ve Teknoloji dersinde hazırlanan 10 performans görevinin akran gruplar ve öğretmen tarafından değerlendirilmiştir. Araştırma, performans görevinin hazırlanma süreci ile birlikte toplam 7 hafta, 28 ders saati sonunda tamamlanmıştır. Verilerin analizi öğrencilerle paylaşılmış ve araştırma aracılığıyla öğrencilerin derse karşı motivasyonlarının arttığı gözlemlenmiştir.

Puanlayıcıların katılık/cömertlik karşılaştırmasına ilişkin RMSE (Root Mean Square Standart Error) değeri 0.11, performans görevlerine İlişkin RMSE değeri 0.11 ve performans

görevlerinin değerlendirilmesinde kullanılan maddelerin istatistiklerine ilişkin RMSE değeri 0.15 bulunmuştur. İlgili logit değerlerine ait standart hata değerlerinin düşük olması standart hatanın oldukça düşük olduğunu göstermektedir. Analiz sonuçları uygunluk içi ve uygunluk dışı değerler açısından incelendiğinde ise sadece kaynakça belirtme ölçütü belirtilen aralığın dışında kalmaktadır. Bu durum çoğu performans görevinde kaynakça yazımının beklenilen seviyede olmadığı şeklinde düşünülebilir. Dolayısıyla öğrencilere kaynak belirtmenin önemi eğitim hayatının ilk yıllarından itibaren vurgulanmalıdır. Öğrencilerin kendi yaptıkları performans görevlerini yeterli derecede düzenli ve yaratıcı bulmamaları da dikkat çekicidir.

Çok-yüzeyli Rasch ölçme modelinin kullanıldığı çalışmaların bir çoğunda jüriler veya puanlayıcılar bazen objektif, bazen de yanlı davranış göstermektedirler (Akın ve Baştürk, 2012; Baştürk, 2010;Baştürk, 2009, Baştürk, 2008, Baştürk ve Işıkoğlu, 2007; Semerci, 2011a, Semerci, 2011b, Semerci 2011c, Semerci, Semerci ve Duman, 2013; Yüzüak, Şahin ve Semerci, 2013). Bu araştırmada ise Bilim ve Teknik P1'e; Bilim P3'e, 1905 P4 ve P10'a, Hedef P6'a, Fen Bilimleri P7'e, Sonsuz P5 ve P9'a pozitif yönde puanlama yapmıştır. Diğer taraftan Bilim ve Teknik P2'e, Bilim P10'a, Esenyurt P5'e, 1905 P6'a, Hedef P3'e, Fen Bilimleri P4 ve P6'a negatif yönde puanlama yapmıştır.

Rasch analizi sonucu elde edilen güvenirlik katsayısı Cronbach alpha ve KR 20 güvenirlik katsayıları gibidir (Linacre, 1997). Analiz sonuçları ve güvenirlik katsayıları birlikte değerlendirildiğinde grupların tutarlı bir şekilde puanlama yaptıkları ve bu çerçevede hazırlanan formun amaca hizmet ettiği söylenebilir. Sonuç olarak, Çok-yüzeyli Rasch ölçme modelinin fen eğitiminde akran grupları değerlendirmek için etkin bir şekilde kullanılabileceği ve objektif sonuçların alınabileceği görülmüştür.

KAYNAKLAR

Abdul Aziz, A. ve Masodi, M.S. (2010). *Workshop on Rasch analysis: a practical guide to winsteps.* Retrieved from http://www.docstoc.com/docs/55062758/Rasch-Workshop-Booklet---Structu.

Akın, Ö. ve Baştürk, R. (2012). Keman eğitiminde temel becerilerin Rasch ölçme modeli ile değerlendirilmesi. *Pamukkale Üniversitesi Eğitim Fakültesi Dergisi, 31*(1), 175-187.

Arsan, N. (2012). Buz pateninde hakem değerlendirmelerinin genellenebilirlik kuramı ve Rasch Modeli ile incelenmesi (Doktora tezi, Hacettepe Üniversitesi, Eğitimde Ölçme ve Değerlendirme Bilim Dalı, Ankara.

Bahar, M. (2001). Çoktan seçmeli testlere eleştirel bir yaklaşım ve alternatif metotlar. *Kuram ve Uygulamada Eğitim Bilimleri, 1*(1), 24-38.

Baştürk, R. (2008). Sözlü sunuma dayalı davranişların çok-yüzeyli Rasch ölçme modeli ile analizi, I. Ulusal Eğitimde ve Psikolojide Ölçme ve Değerlendirme Kongresi'nde sunulan bildiri, Ankara

Baştürk, R. (2009). Applying the many – facet Rasch model to evaluate powerpoint presentation performance ın higher education. *Assesment and Evaluation In Higher Education, 33*(4), 431–444.

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

Baştürk, R. ve Işıkoğlu, N. (2007). Okul öncesi eğitim kurumlarının işlevsel kalitelerinin çok-yüzeyli Rasch modeli ile analizi. *Kuram ve Uygulamada Eğitim Bilimleri, 7*(3), 727 – 752.

Baykul, Y. (2010). *Eğitimde ve Psikolojide Ölçme: Klasik Test Teorisi ve* Uygulanması *(2. Baskı).* Pegem A Akademi. Ankara.

Çalışkan, İ. Ö. ve Kaptan, F. (2012). Fen öğretiminde performans değerlendirmenin bilimsel süreç becerileri, tutum ve kalıcılık açısından yansımaları. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi*, *43*, 117-129.

Elhan A. H., ve Atakurt Y. (2005). Ölçeklerin değerlendirilmesinde niçin Rasch analizi kullanılmalıdır? *Ankara Üniversitesi Tıp Fakültesi Mecmuası*. *58*, 47-50.

Hambleton, R.K., & Swaminathan, H.(1985). I*tem response theory principles and applications.* Kluwer-Nijhoff Publishing, USA.

Kaptan, F. (1994). Rasch modeli madde parametrelerini klullanarak en yüksek olabilirlik yöntemiyle yeteneğin kestirilmesi. *Hacettepe Üniversitesi Eğitim Fakültesi. 10*, 95-97.

Karasar, N. (2012). *Bilimsel araştırma yöntemi*. Nobel Yayıncılık. Ankara.

Lawson, A. (1995). *Science teaching and the development of thinking.* USA: Wadsworth Inc.

Linacre, J. M. (1993, April). *Generalizability theory and many facet Rasch measurement.* Annual Meeting of the American Educational Research Association. Atlanta Georgia.

Linacre, J. M. (1997). Kr-20/Cronbach alpha or rasch person reliability: which tells the "truth"? *Rasch Measurement Transactions,11*(3), 580-1.

Linacre, M. (2014). *A user's guide to FACETS. Rasch model computer programs*. Chicago, IL.

MEB, (2013). *İlköğretim Kurumları (İlkokullar ve Ortaokullar) Fen Bilimleri Dersi (3, 4, 5, 6, 7, ve 8. Sınıflar) Öğretim Programı.* Ankara.

Peters, J. M., & Stout, D. L. (2011). *Science in elementary education methods, concepts, and inquiries*, Pearson, USA.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Chicago: The University of Chicago Press.

Semerci, Ç. (2011a, Elazığ). *Öğrencilerin BÖTE bölümüne ilişkin Rasch ölçme modeline gore değerlendirilmesi* (Fırat Üniversitesi Örneği), 5th International Computer & Instructional Technologies Symposium'da sunulan bildiri, Fırat Üniversitesi, Elazığ.

Semerci, Ç. (2011b). Mikro öğretim uygulamalarının çok-yüzeyli Rasch ölçme modeli ile analizi. *Eğitim ve Bilim/Education and Science*, *36*(161), 14-25.

Semerci, Ç. (2011c). Doktora yeterlikler çerçevesinde öğretim üyesi, akran ve öz değerlendirmelerin Rasch ölçme modeliyle analizi. Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi, 2(2), 164-171.

Semerci, Ç., Semerci, N. ve Duman, B. (2013). Yüksek lisans öğrencilerinin seminer sunu performanslarının çok-yüzeyli Rasch modeli ile analizi. *Sakarya Üniversitesi Eğitim Fakültesi Dergisi, 25*, 7-22.

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

Wright, B.D., Linacre, M., Gustafson, J.E. ve Martin-Lof, P. (1994). Reasonable mean-square fit values. *Rasch measurement Transactions*, *8*(3):370.

Yüzüak A.V., Şahin A.E. ve Semerci, Ç. (2013, Eylül). *Kimya laboratuvar uygulamalarının çok-yüzeyli Rasch ölçme modeliyle değerlendirilmesi*. 22. Ulusal Eğitim Bilimleri Kurultayı'nda sunulan bildiri, Eskişehir Osmangazi Üniversitesi, Eskişehir.

EXTENDED ABSTRACT

Introduction

Educational process includes objectives, content, teaching-learning process and evaulation parts. Evaulation, means giving decision according to criterias, is important for education continuity, development and to realize objectives (Baykul, 2010). For measurement and evaulation, traditional and authentic measurement-assesment tools can be used. However traditional measurement and assesment tools are insuffient to enhance students' thinking skills due to this reason these skills can not be developed at the desired level (Lawson, 1995; Çalışkan and Kaptan, 2012). This situation leds us to authentic measurement-assesment tools. Even though there is negative aspects, peer group assesment can be thought to evaulate authentic tasks. Students who participate in decision making process will be active for learning process. At this point criterias should be determined precisely and teacher's attitude is also important (Peters and Stout, 2011).

To evaulate measurement-assesment tools generally two theorems are used: Classical Test Theory and Item Response Theory. Even though there are similarities between these two theorems, it can be said that Item Response Theory has advantages (Hambleton and Swaminathan, 1985). Through the models based on Item Response Theory lineer relations can be obtained, missing data can be overcome, the sensitivity of measurement can be estimated and the results which do not obey standart value can be evaulated (Elhan and Atakurt, 2005; Masodi, 2010).

Rasch model which is one parameter logistic model depends on Item Response Theory and explained in "Some Probabilistic Models for Intelligence and Attainment Tests"

book by George Rasch (Rasch, 1960). Severity or leniency of the juries was added to Rasch model for Many-Facet Rasch Measurement (Linacre, 1993). The vision of science curriculum which was published in 2013 emphasizes the importance of peer evaulation. Many-facet Rasch model can be used to evaulate peer evaulation. The aim of this study is to evaulate students' performance tasks by using MFRM objectively.

Method

In this research, survey model was used. The general aim of the survey model can be said as describe the situation (Karasar, 2012). The study was conducted in 2013-2014 autumn semester in Bartın. Many-facet Rasch model's surfaces are respectively; 10 peer groups whose names are: Bilim ve Teknik, Grup 74, Bilim, Süper, Esenyurt, Grup 1905, Hedef, Grup Işık, Fen Bilimleri, Sonsuz and one elementary science teacher, criterias and performance tasks. A likert type form was used to evaluate for peer and teacher evaluation. Moreover, criterias were coded such as, the aim of the project suitability was coded as fitnesswithpurpose, name of the project compliance with content name was coded as fitnesswithcontent, collecting information from different sources was coded as differentsources, analysis of information/ drawing charts and diagrams was coded as straight etc. FACETS analysis program was used. Groups gave the points for criteria as "insufficient: 1", "Very few sufficient: 2", "Partially sufficient: 3", "Great amount of sufficient: 4", "Completely sufficient: 5".

Results and Discussion

Results can be examined in five parts which are calibration map, logit values, judge measurement report, criterias measurement report and judge bias analysis. The general information concerning these surfaces are given in Figure 1. It can be said that performance task coded P4 (logit P4=1.39) is successful at higher level, performance task coded P6 (logit P6=0.08) is successful at lowest level. The severest member is teacher and the most lenient group is Grup 74. It was observed that the most difficult criteria is "references" and "straight". In other words, it could be said that, these criterions were met at lower level. The easiest criterion is "information accuracy" among the evaluation criterions.

Severity/leniency comparison of evaluators is given in Table 2. Evaluators separation index is 3.10 and reliability coefficient is 0.91 in Table. Null hypothesis was rejected when "there is distinction between severity/leniency of evaluators" hypothesis was tested with chi-square test (χ^2=118.7, sd=10, p=0.00). There is a significant distinction between severity/leniency points of the eleven evaluators statistically. It could be said that the scorer coded Grup 74 is "the most lenient" and scorer numbered Ogretmen is "the severest" when evaluators are sequenced from the most lenient towards the severest in Table. Evaluators may be sequenced from the severest to the most lenient as Grup 74, Bilim ve Teknik, Hedef, Fen Bilimleri, Esenyurt, Sonsuz, Bilim, Super, Işık, Grup 1905, Öğretmen.

A detailed measurement report including performance tasks of 8th grade students are shown in Table 3. Reliability was calculated as 0.92 and this number shows that students graded in a high reliability. Null hypothesis was rejected when "there is a measurable distinction amongst the performance tasks of students" hypothesis that belongs to fixed effect with separation index 3.34 and reliability coefficient 0.92 was tested with chi-square test (χ^2=122.1, Sd=9, p=0.00).

For criterias, separation index was found 9.75 and reliability coefficient was found 0.99. Null hypothesis was rejected when "there is significant distinctions between difficulties of articles used in evaluation of performance tasks" hypothesis was tested with chi square (χ^2=599.2, sd=11, p=0.00). There is a significant distinctions between articles

used in evaluation of performance tasks statistically. The most difficult article is reference and the simplest articles are informationaccuracy, fitnesswithpurpose, fitnesswithcontent.

It could be said that, some of the jury members were extremely severe or lenient against some post graduate students according to bias analysis. For example, Bilim ve Teknik made a severe scoring by giving 42 point for the performance task P2 even though they were expected to give 48.05. Bilim made an lenient scoring by giving 53 point for the performance task P3 even though they were expected to give approximately 45 point. Bilim made a severe scoring by giving 42 point for the performance task P10 even though they were expected to give nearly 48. 1905 made an lenient scoring by giving 52 point for the performance task P4 even though they were expected to give approximately 46 point. Hedef made a severe scoring by giving 39 point for the performance task P3 even though they were expected to give 46.85.

To conclude, the analysis of performance tasks with Many-Facet Rasch measurement model was conducted in the study. The surfaces used in researches simultaneously (performance tasks, severity/leniency of the juries and criteria used). Grup 74 is the most lenient jury, Öğretmen is the most severest jury. P4 coded was found to be the most successful and P7 coded was found the most unsuccessful. The most difficult article is reference which is out of infit and outfit values. Reliability coeffient of Rasch analysis is similar to Cronbach alpha or KR-20 (Linacre, 1997). Therefore, according to results groups' and teacher's ideas about performance tasks were analyzed consistently. Moreover, the form which was revised by the teacher was served its objective. The results of the analysis were shared with 7^{th} grade students and students' motivation for science education was increased.