Eurasian Journal of Educational Research, 20, pp, 62-73 /2005

Müzik Öğretmenliği Özel Yetenek Seçme Sınavının Çok-Yüzeyli Rasch Modeli ile Analizi (İnönü Üniversitesi Örneği)

Analysis of Special Ability Sellection Examination for Music Education Department Using Many-Facets Rasch Measurement (İnönü University Case)

Hakan ATILGAN*

ÖZ

Bu araştırmada, İnönü Üniversitesi Eğitim Fakültesi Müzik Öğretmenliği Özel Yetenek Seçme Sınavları Çok-Yüzeyli Rasch Modeli ile analiz edilerek; birey yetenek ölçüleri, puanlayıcıların katılık/cömertlikleri, görev güçlük düzeyleri ve uygunluk istatistikleri ile puanlayıcı-birey, görev-birey ve puanlayıcı-görev etkileşimleri yanlılıkları araştırılmıştır. Deşifre, çalma ve söyleme olmak üzere üç boyutta, adayların birbirinden bağımsız puanlama yapan, müzik öğretimi alanından dört öğretim üyesi yapması ile yapılan Özel Yetenek Seçme Sınavından elde edilen veriler kullanılmıştır. Elde edilen bulgulara göre; puanlayıcıların bütün bireyler için puanlamalarının birbirlerine yakın fakat manidar düzeyde farklı olduğu, 10 birey ve bir görevin uygun olmadığı, puanlayıcı-birey, görev-birey ve puanlayıcı-görev etkileşimlerinde yanlılıkların bulunduğu sonucuna varılmıştır.

Anahtar Sözcükler: Çok-Yüzeyli Rasch Modeli, Madde Yanlılığı, Puanlayıcı Yanlılığı, Ortak Etki Yanlılığı, Özel Yetenek Seçme Sınavları

ABSTRACT

In this research, individual ability measure, severity of the raters, task diffuculty levels and fit statistics for main effect's, and bias of the raters-individuals, tasks-individuals, raters tasks interaction effects were investigated by analyzing Special Ability Sellection Examination for İnönü University Faculty of Educatinon Music Education Department using Many-Facets Rasch Measurement (MFMR). The data of the research were from Special Ability Sellection Examination, which was in three dimensions such as decoding, playing and singing, and which was rated by independent raters who were lecturers in education department. According to the findings, it was concluded that ratings for the candidates by raters were so close to each other but signaficantly differs, that10 individuals and one task were not fit and that there were biases in interactions between the raters-individuals, tasks-individuals and raters-tasks.

Key Words: Many-Facet Rasch Measurement, Item Bias, Rater Bias, Interaction Effect Bias, Special Ability Selection Examination

GİRİŞ

Türkiye'de yüksek öğretim programlarına öğrenci seçilmesi ve yerleştirilmesi; Öğrenci Seçme ve Yerleştirme Merkezi (ÖSYM) tarafından merkezi olarak yapılan Öğrenci Seçme Sınavı (ÖSS) sonuçları ve adayların tercihleri doğrultusunda yapıl-

*Ege Üniversitesi, Eğitim Fakültesi Eğ.t Bilim. Blm Eğt. Ölç. ve Değ hakan_atilgan@hotmail.com

maktadır. Ancak özel yetenek gerektiren yüksek öğretim programlarına öğrenci alımı, ilgili üniversitenin kendisi tarafında yapılan özel yetenek seçme sınavları ile yapılmaktadır. (ÖSYM 2003). Bu tür sınavlar "Özel Yetenek Seçme Sınavları (ÖYSS)" olarak adlandırılmaktadır.

Üniversitelerin eğitim fakültelerinin Müzik Öğretmenliği Programına öğrenci seçilmesi ÖYSS ile yapılmaktadır. Bu sınavlarda adaylardan belirlenmiş olan görevleri yerine getirmeleri istenmekte ve jüri üyeleri tarafından adayın görevdeki başarısına göre puan verilmektedir. Bu bağlamda Müzik Öğretmenliği bölümü özel yetenek seçme sınavları; adayların yerine getirmeleri istenilen görevlerin, birbirinden bağımsız puanlayıcılar tarafından puanlanmasının söz konusu olduğu bir ölçme desenine sahiptir. Adayların gösterdikleri performansa göre jüri üyelerinin verdikleri puanlardan hareketle, özel yetenek seçme sınavlarının sonunda verilecek kararlar, bireylerin ilgili programa kabul edilmesi ya da edilmemesi yönünde, oldukça önemli niteliktedir.

Müzik Öğretmenliği programlarına öğrenci seçmek amaçlı kullanılan ÖYSS'larında kullanılan performans testlerinde her bir adayın başarısını etkileyen, bireyin kendisi, yerine getirilen görevler, görevlerin alt basamakları ya da maddeleri olmak üzere geniş bir değişkenlik-yüzey (facet) içeriğine sahiptir. Böylesi ölçme durumlarının analizinde kullanılabilen madde tepki kuramı modeli Çok-Yüzeyli Rasch Modelidir.

Çok-Yüzeyli Rasch Modeli

Madde tepki kuramı içinde yüzeylerin çok olması durumunda kullanılabilen Çok-Yüzeyli Rasch Modeli (ÇYRM); Danimarkalı matematikçi Rasch tarafından geliştirilen ve kendi adıyla da anılan bir parametreli lojistik modelden, Linacre (1989) tarafından türetilmiştir. Rasch modeli iki yüzeyli (facet) bir modeldir ve bir parametreli lojistik model bireylerin yetenek düzeyi ve maddelerin güçlük düzeyini aynı zamanda tanımlamaya çalışılır (Hambleton ve Swaminathan 1985; Linacre 1990). Başka bir ifadeyle Rasch modeli, bireyin yetenek düzeyi ve madenin güçlüğü üzerine kurulan bir lojistik olasılık fonksiyonudur. Bu modelin matematiksel özellikleri koruyarak Andrich (1978) Rasch modelini dereceleme ölçekleri için kullanılacak halde genişletilmiştir. Özellikle performansa dayalı ölçmelerle yetenek kestirimini etkileyen bir faktör de puanlayıcıların katı ya da ılımlı puanlama yapmasıdır. Linacre (1989), kısmi puanlama modeline her bir puanlayıcının dereceleme ölçeği uygulamasını tanımlayan parametreleri de dahil ederek kısmi puanlama modelini genişletmiştir. Bireyin yeteneği ve madde güçlüğüne ek olarak puanlayıcıların da yüzey olarak modele eklenmesiyle Çok-Yüzeyli Rasch Modeli ortaya çıkmıştır (Linacre 1989). Çok-Yüzeyli Rasch Modeli; bireyin örtülü özelliğini kestirmeyi etkileyebilen puanlayıcı katılığı gibi ek değerlendirme değişkenlerinin (yüzeyin) de hesaplamaya katılmasını olanaklı hale getirmiştir.

ÇYRM, bir yüzeyler içindeki öğelerden uygunsuz ya da problemli olanların tanımlanmasına olanak sağlar. Bu problemli ya da uygun olmayan öğe, puanlamasında sistematik olmayan bir tutarsızlık sergileyen puanlayıcı ya da ölçme gözlemi süreci içinde sistematik olmayan güçlükte bir görev veya tutarsız tepkileri olan bir birey olabilir. Bu modelde yüzey elemanlarının değişik kombinasyonları için

yanlılık (bias) analizi yapılarak yanlılıklar tanımlanabilir (Linacre 1989-2003; Lynch ve McNamara 1998).

Amaç

Bu araştırma ile İnönü Üniversitesi Eğitim Fakültesi Müzik Öğretmenliği Özel Yetenek Seçme Sınavları analiz edilerek; (a) birey yetenek ölçüleri ve uygunluk istatistikleri, (b) puanlayıcıların katılık/cömertlikleri ve uygunluk istatistikleri, (c) görev güçlük düzeyleri ve uygunluk istatistikleri, (d) puanlayıcı x birey etkileşimi yanlılıkları, (e) görev x birey etkileşimleri yanlılıkları, ve (f) puanlayıcı x görev etkileşimi yanlılıkları belirlenmeye çalışılmıştır. Araştırma bulgularına dayalı önerilerle sınavın daha uygun hale getirilebileceği düşünülmektedir.

YÖNTEM

Çalışma Grubu

Bu araştırmanın çalışma grubunu, 2003-2004 öğretim yılında İnönü Üniversitesi Eğitim Fakültesi Güzel Sanatlar Eğitimi Bölümü Müzik Öğretmenliği programına başvuran 689 adaydan birinci aşamada başarılı olarak ikinci aşama sınavına giren 249 aday oluşturmaktadır.

Araştırma Verileri

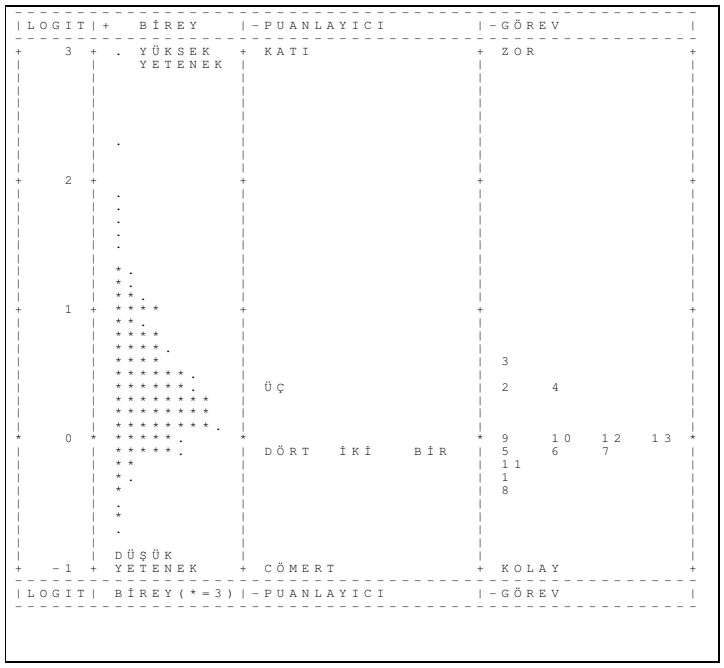
Araştırmada, iki aşamalı olarak yapılan İnönü üniversitesi Eğitim fakültesi güzel sanatlar eğitimi bölümü Müzik öğretmenliği programı özel yetenek seçme sınavının ikinci aşamasında elde edilen veriler kullanılmıştır. İkinci aşama sınavı; deşifre, söyleme ve çalma olmak üzere üç boyuttan oluşturulmuştur. Sınavın puanlama ölçeği; "Deşifre" dört ölçü her ölçü 13 puan,"Söyleme"; sağlıklı ses 30 puan, sesin tını, gürlük ve genişliği 30 puan, doğru ve temiz söyleme 20 puan, konuşmada anlaşılırlık 10 puan ve müzikalite 10 puan, "çalma" ise; doğru ve temiz çalma 30 puan, bütünlük-teknik 20 puan, müziksel yorum 25 puan ve eserin düzeyi 25 puan olarak belirlenmiştir. Adayların belirtilen boyutlarda puanlanmasını, birbirinden bağımsız puanlama yapan, biri profesör, diğer üçü yardımcı doçent olan, müzik öğretimi alanından dört öğretim üyesi yapmıştır. Araştırmada 2003-2004 öğretim yılında İnönü Üniversitesi Eğitim Fakültesi Güzel Sanatlar Eğitimi Bölümü Müzik Öğretmenliği programı Özel Yetenek Seçme Sınavından yukarıda belirtilen puanlama işlemleri ile elde edilen veriler kullanılmıştır.

Verilerin Analizi

ÇYRM analizleri için MINIFAC, FACDOS ve FACETS (Linacre 1991-1993) bilgisayar programları kullanılmıştır. Verilerin düzenlenmesi ve analize hazır hale getirilmesinde FACFORM (Linacre 1991-1993) veri düzenleme programından yararlanılmıştır. Araştırmada, B_n; n bireyinin yeteneği, A_g; g görevi ile baş etme, D_g; g görevinin güçlüğü, C_p; p puanlayıcısının katılığı, F_k; k kategorisi ile ilgili k-1 kategorisinde gözlenen sınır olmak üzere, $\log(P_{nggpk}/ P_{nggp(k-1)})=B_n-A_gD_i-C_p-F_k$ ÇYRM deseni kullanılmıştır.

BULGULAR

Model-Veri Uyumu: Analizde kullanılan verilerin modelle uyumlu olabilmesi için, standartlaştırılmış artıkların (Z puanı) yaklaşık % 5'ininden daha azının mutlak değerce ikiden büyük ya da eşit olması veya % 1'ininden daha azının mutlak değerce üçten büyük ya da eşit olması şartı aranır (Linacre 2003). Analizde kullanılan toplam 12948 verinin standartlaştırılmış artık değerinin +/- 3'den büyük ya da eşit olanlarının sayısı 134 (% 0,10), +/- 2'den büyük olanlarının sayısı ise 544 (% 4,20) olarak bulunmuş ve model-veri uyumu sağlanmıştır.



Şekil 1. ÖYSS Kalibrasyon Haritası

Bireylerin Yetenek Ölçüleri ve Uygunluk İstatistikleri: 249 bireyin yetenek düzeyleri (logit) ortalaması 0,41, standart sapması 0,48 ve logit değerleri aralığı -0,70 ile 5,20'dir. Elde edilen birey ayırma indeksi (G) 4,98 ve güvenirlik 0,96'dır. Bu ayırma indeksi için "sınavla bireylerin farklı yetenek tabakalarına ayrılamadığı" yönünde oluşturulan sabit etki (fixed effect) hipotezi Kay-kare ile test edildiğinde (χ^2=4887,7 ve sd=246, p=0,00) reddedilmiştir. Bireylerin istatistiksel olarak anlamlı olan yetenek ayrımının kaç yetenek tabakasından oluştuğunun belirlenmesi için (4G+1)/3 formülü (Lee ve Kantor 2003) kullanılarak, ÖYSS ile bireylerin istatistiksel olarak anlamlı yaklaşık yedi (6,97) ayrı yetenek tabakasına ayrılabildiği görülmüştür. Standartlaştırılmamış uygunluk içi kareler ortalaması (infit mean square) 0,4 ile 4,8 arasındayken, standartlaştırılmış uygunluk içi Z puanları –3 ile 8 arasında değişmektedir.

Bununla birlikte, 249 bireyden 10'unun (% 4,02) uygunluk içi standart Z puanlarının mutlak değeri 2'den büyüktür. Olağan olmayan uygunluk istatistiklerinin Z puanlarının pozitif olması uygunsuzluğu, negatif olması ise aşırı uygunluğu göstergesi olması nedeniyle, Z puanları pozitif olan 10 bireyin tamamının uygun olmadığı sonucuna varılmıştır.

Puanlayıcıların Katılıkları ve Uygunluk İstatistikleri: Puanlayıcıların katılık/cömertlik logit değerleri ortalaması 0,00 ve standart sapması 0,21, birinci puanlayıcının -0,14, ikinci puanlayıcının -0,09, üçüncü puanlayıcının 0,36 ve dördüncü puanlayıcının -0,13 logittir. Puanlayıcı ayırma ineksi 17,80 ve güvenirlik 1,00 sabit etki (fixed effect) "puanlayıcıların katılık/cömertlikleri arasında anlamlı farklılık vardır" hipotezi Kay-kare ile test edildiğinde (χ^2=1298,9, sd=3, p=0,00) reddedilmiştir. Başka bir ifadeyle, dört puanlayıcının puanlamalarının katılık/cömertlikleri arasında istatistiksel olarak fark bulunmaktadır. Şekil 1'de görüldüğü gibi puanlayıcıların logit değerlerinin katılık/cömertlik ölçeğinde 1 logit uzaklıkta kümelenmiş olması puanlayıcıların katılık/cömertlik farklılıklarının oldukça ılımlı olduğunu göstermektedir (Lee ve Kantor 2003). Bu sonuç, sınava giren bütün bireyler boyunca dört puanlayıcının puanlarının anlamlı düzeyde farklı, ancak göreceli olarak yakın puanlamalar yaptıklarını göstermektedir.

Tablo1.

ÖYSS Puanlayıcı Ölçüm Raporu

Puanlayıcı		Puanlayıcı Katılığı		Uygunluk İçi		Uygunluk Dışı	
No	$\overline{X}$	Logit	S.H.	Kareler $\overline{X}$	Z	Kareler $\overline{X}$	Z
1	10,2	-0,14	0,01	1,3	9	1,3	9
4	10,1	-0,13	0,01	0,8	-6	1,0	0
2	9,5	-0,09	0,01	1,1	2	1,0	-1
3	7,5	0,36	0,01	1,2	5	1,1	2
$\overline{X}$	9,3	0,00	0,01	1,1	2,8	1,1	2,4
SS	1,1	0,21	0,00	0,2	5,7	0,1	4,1
RMSE (Model)= 0,01 SS: 0,21				Ayırma İndeksi=17,80 Güvenirlik=1,00			
Tamamı Aynı Kay-Kare= 1298,9				Sd= 3 p= 0,00			

Tablo 1 incelendiğinde, uygunluk kareler ortalamalarının ortalaması 1,1 ve standart sapması 0,2 olmakla birlikte, bütün puanlayıcıların Z puanlarının mutlak değer olarak 2'den büyük olması, uygunluk istatistiklerinin olağan olmadığını göstermektedir. Ancak negatif işaretli Z puanları ya da kareler ortalaması 1 ve 1'den küçük olan puanlayıcılar, pozitif işaretli Z puanlarından ya da kareler ortalaması 1 ve 1'den büyük olanlardan daha az problemli olarak değerlendirilir (Engelhard ve Myford 2003; Myford ve Wolfe 2000). Bu bağlamda; ikinci, üçüncü ve dördüncü puanlayıcıların uygunluk kareler ortalamaları Myford ve Wolfe (2000) tarafından belirtilen

kalite kontrol limiti 0,7 ile 1,3 aralığındadır. Ancak birinci puanlayıcının uygunluk kareler ortalaması (1,3) ise bu aralık değerlerinin üst sınırdadır. Bu durumda dört puanlayıcının da belirtilen kareler ortalamaları kalite kontrol sınır değerlerinde olduğu ve uygun olarak kabul edilebileceği söylenebilir. Lynch ve McNamara (1998) uygunluğun değerlendirilmesinde, her bir uygunluk kareler ortalamalarının ilgili yüzey (facet) için, uygunluk kareler ortalaması ve standart sapmasıyla yeniden açıklanması gerektiğini vurgularlar ve bu amaçla bir değerin ortalamaya iki standart sapma eklenmesiyle elde edilenden daha büyük olması durumunda uygun olmayan (misfitting) olarak değerlendirilebileceğini belirtirler. Bu durumda Tablo 1'de görüldüğü gibi; uygunluk kareler ortalamasının ortalaması 1,1 ve standart sapması 0,2 olduğundan, (1,1+/-1,96x0,2) kabul edilebilirlik aralığı 0,71 ile 1,49'dur. Bu kabul edilebilirlik aralığının üst sınırından mutlak değerce büyük olan uygunluk kareler ortalaması uygun olmayan olarak değerlendirilmesi gerekir. Bütün puanlayıcıların uygunluk kareler ortalaması 1,49'dan küçük olduğundan, yukarıdaki yargıda olduğu gibi uygun olarak değerlendirilebilir.

Görev Güçlükleri ve Uygunluk İstatistikleri: ÖYSS'ında kullanılmış olan toplam 13 görevin güçlük logit değerleri –0,43 ile 0,55 aralığında değişmekte, ortalaması 0,00 ve standart sapması 0,28'dir. En zor görev 3. (0,55 logit) görevken, en kolay görev 8. (-0,43 logit) görevdir. Görev ayırma indeksi 6,91 ve ayırma güvenirliği 0,98 test edildiğinde, sabit etki (fixed effect) hipotezi, Kay-kare testiyle (χ^2=709,9 ve 12 serbestlik derecesinde, p=0,00) reddedilmektedir. Bu sonuç, 13 görevin güçlükleri arasında istatistiksel olarak anlamlı farklılıklar olduğunu göstermektedir.

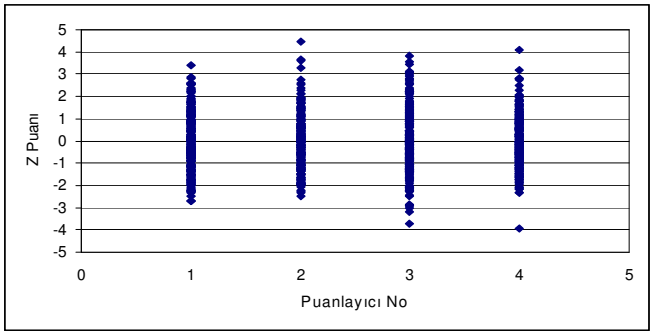
Tablo 2 incelendiğinde, 13 görevin uygunluk kareler ortalaması 0,8 ile 1,5 arasında değerler alıkken, uygunluk standart Z puanları –8 ile 9 arasında değişmektedir. 13 görevden uygunluk standart Z puanının mutlak değerce 2'ye eşit ve büyük olanlarının 12 tane olduğu, bir görevin (5. görev) ise uygunluk Z puanının sıfır olduğu görülmektedir. Uygunluk standart Z puanı 2'den büyük olan 12 görevden 8 tanesinin Z puanları negatif işaretli ve uygunluk kareler ortalamaları birden küçük olduğundan aşırı uygun olarak değerlendirilir. Uygunluk standart Z puanı 2'den büyük olan 13 görevden 4 tanesinin ise Z puanları pozitif işaretli olduğundan uygun olmayan olarak değerlendirilir. Z puanı 2'den büyük olan 12 görevden bir tanesinin (8. görev) uygunluk kareler ortalaması, Myford ve Wolfe (2000) tarafından kalite kontrol aralığı olarak kullanılan 0,7 ile 1,3 aralığının sınırları dışında, iki tanesi sınırda (7. ve 9. görevler) diğer görevlerin uygunluk kareler ortalamaları ise belirtilen aralıktadır. Bu bakımdan 8. görev dışında bütün görevlerin model veri uygunluklarının oldukça iyi olduğu, testin 8. görev dışında çok boyutlu olmadığı söylenebilir. Yine bu 12 görev için, Lynch ve McNamara'nın (1998) uygunluk kareler ortalamaları, "ortalamaya iki standart sapma eklenmesi kuralı" ile yeniden yorumlandığında 8. görev dışında diğer görevlerin (1,0+/-1,96x0,2) elde edilen 0,61 ile 1,39 kabul edilebilirlik aralığının sınırında bulunduğu görülmektedir. Standart Z puanı 2'den mutlak değerce büyük olan 8. görev, hem Myford ve Wolfe'un (2000) hem de Lynch ve McNamara'nın (1998) yaklaşımlarıyla da uygun olmayan olarak bulunmuştur. Bu sonuç, uygun olmayan olarak kabul edilen 8. görevin diğer görevlerin puanlarıyla aynı çizgide olmadığının, test yapısı içinde farklı bir boyut oluşturduğunun (çok boyutluluk) ya da bu görevin kriterlerinin yorumlanmasında puanlayıcılar arasında temel düzeyde uyuşmazlığın olduğunun göstergesidir.

Tablo2.
ÖYSS Görev Ölçüm Raporu

Görev		Görev Güçlüğü		Uygunluk İçi		Uygunluk Dışı	
No	$\overline{X}$	Logit Ölçüsü	S.H.	Kareler $\overline{X}$	Z	Kareler	Z
1	0,70	-0,30	0,07	0,9	-3	0,9	-4
2	0,50	0,40	0,07	0,9	-8	0,9	-8
3	0,50	0,55	0,07	0,9	-8	0,9	-8
4	0,50	0,41	0,07	0,9	-8	0,9	-7
5	17,10	-0,13	0,02	1,0	0	1,1	1
6	16,90	-0,09	0,02	1,1	2	1,2	2
7	12,10	-0,13	0,02	1,3	6	1,4	7
8	8,40	-0,43	0,02	1,5	9	1,9	6
9	6,70	0,01	0,02	1,3	6	1,7	8
10	16,90	-0,04	0,02	0,8	-3	0,8	-3
11	12,40	-0,23	0,02	0,9	-2	0,9	-1
12	14,00	0,01	0,02	0,9	-3	0,8	-3
13	14,50	-0,05	0,02	0,8	-4	0,8	-4
$\overline{X}$	9,3	0,00	0,03	1,0	-1,4	1,1	-1,2
SS	6,6	0,28	0,02	0,2	5,8	0,3	5,7
RMSE (Model)= 0,04 SS: 0,28			Ayırma İndeksi= 6,91 Güvenirlik= 0,98				
Tamamı Aynı Kay-Kare= 709,9			Sd= 12 p= 0,00				
Normal Kay-Kare = 11,8			Sd=11 p=0,38				

Puanlayıcı x Birey Etkileşimi Yanlılıkları: Puanlayıcı x birey analizi, belli bir puanlayıcının bütün bireyler için aynı biçimde puanlama yapıp yapmadığı ya da belli bir puanlayıcının belli bir birey için cömert veya katı puanlama yapıp yamadığını sınamaya olanak sağlar (Lynch ve McNamara 1998; Engelhart ve Myford 2003; Lee ve Kantor 2003). Dört puanlayıcı tarafından puanlanan 249 birey için olası yanlılık sayısı puanlayıcı sayısının birey sayısı birey sayısı ile (4x249=996) çarpımına eşittir. Puanlayıcı x birey yanlılık logit ölçülerinin ortalaması -0,05, standart sapması 0,36 ve -2,17 ile 2,08 arasında değişmektedir. Standart Z puanlarının ise, ortalaması -0,01 ve standart sapması 1,27 olup –3,95 ile 4,49 arasında değişmektedir.

Olası 996 puanlayıcı x birey etkileşimi için hesaplanan Z puanlarının 889 tanesi –2 ile 2 arasında değerler alırken, 107 tanesinin (% 10,74) Z puanları mutlak değerce 2'den büyük ve anlamlı olarak bulunmuştur (Bkz. Şekil 1). Başka bir ifadeyle; puanlayıcı x birey etkileşiminin olası yanlılık sayısı 996 olmakla birlikte, bunların 107 tanesinin (% 10,74) Z puanı mutlak değerce 2'den büyük ve anlamlı bulunmuştur. Mutlak değerce 2'den büyük olan bu Z puanlarının; 45 tanesi negatif işaretliyken, 62 tanesi pozitif işaretlidir. Negatif işaretli Z puanı (ya da logit değeri), gözlenen puanın beklenenden büyük olduğunu ve o puanlayıcının o birey için cömert puanlama yaptığını, pozitif işaretli Z puanı (ya da logit değeri) gözlenen puanın beklenenden küçük olduğunu ve o puanlayıcının o birey için katı puanlama yaptığını gösterir(Lee ve Kantor 2003).

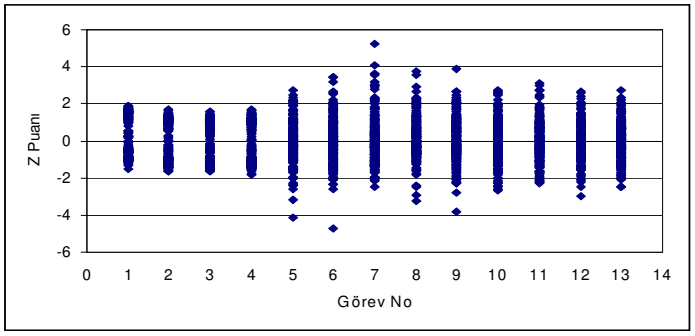


Şekil 2. ÖYSS *Puanlayıcı x Birey* Etkileşimi Z Puanları Grafiği

Şekil 2'de görüldüğü gibi, üçüncü puanlayıcının puanlayıcı-birey eşlemesiyle elde edilen mutlak Z puanlarının 2'den büyük olanlarının sayısı (46 eşleme) en fazladır. Üçüncü puanlayıcının anlamlı olarak belirlenen Z puanlarından 15 tanesi pozitif değerliyken geriye kalan 11 tanesi negatif değerlidir. Bu durum üçüncü puanlayıcının negatif değerli olan iki Z puanı ile ilgili bireyleri cömert puanladığının, pozitif değerli olan Z puanı ile ilgili olan bireyleri katı puanladığını göstermektedir. Birinci puanlayıcıların puanlayıcı x birey eşlemesiyle elde edilen mutlak Z puanlarının 2'den büyük olanlarının 26 tanesinin 11'i negatif işaretliyken (cömert puanlanan), 15'i pozitif işaretlidir (katı puanlanan). İkinci puanlayıcının mutlak değer olarak 2'den büyük olan Z puanlarının sayısı 17 tanedir ve bunların 6 tanesi negatif (cömert puanlanan), 11 tanesi pozitif işaretlidir (katı puanlanan). Dördüncü puanlayıcının anlamlı olarak belirlenen 18 Z puanından 10 tanesi pozitif işaretliyken (katı puanlanan) geriye kalan 8 tanesi negatif işaretlidir (cömert puanlanan).

Görev x Birey Etkileşimi Yanlılıkları: Görev x birey karşılıklı etkisi için yapılan analiz; belli bir görevin bütün bireyler için aynı olup olmadığı ya da belli bir görevin belli bir birey için anlamlı olarak daha zor veya daha kolay olup olmadığı sorusuna yanıt bulmaya olanak sağlar (Lynch ve McNamara 1998; Engelhart ve Myford 2003; Lee ve Kantor 2003). 249 birey 13 görev üzerinden puanlandığından, birey sayısının görev sayı ile çarpımı kadar (249x13=3237) olası yanlılık bulunmaktadır. Görev x birey logit değerlerinin ortalaması –0,01, standart sapması 1,01 olup, logit değerleri –3,01 ile 3,52 arasında değişmektedir. Bununla birlikte, standart Z puanlarının ortalaması 0,03 ve standart sapması 1,08 olup –4,70 ile 5,24 arasında değişmektedir.

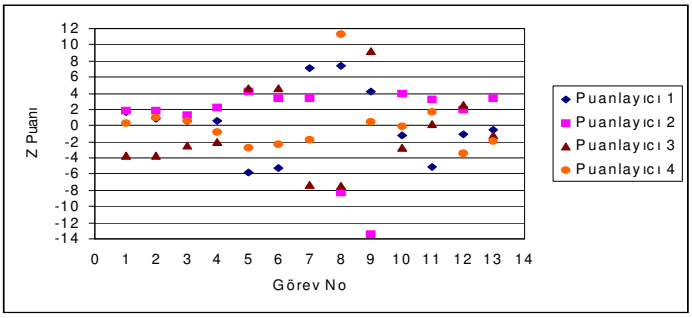
Şekil 3'de görüldüğü gibi, 3237 görev x birey etkileşimi olası yanlılığı için hesaplanan Z puanlarının –2 ile 2 aralığı dışında kalan değerler bulunmaktadır. Görev x birey karşılıklı etkisinin olası yanlılık sayısı 3237 olmakla birlikte, 150 tanesinin (% 4,63) Z puanı mutlak değerce 2'den büyük (anlamlı) bulunmaktadır



Şekil 3. ÖYSS Puanlayıcılar için *Birey x Görev* Etkileşimi Z Puanları Grafiği

Puanlayıcı x Görev Etkileşimi Yanlılıkları: ÇYRM ile yapılan puanlayıcı x görev karşılıklı etkileşimi analizi, belli bir puanlayıcının bütün görevler için aynı biçimde ya da belli bir görev için diğerlerinden farklı davranıp davranmadığını araştırmaya olanak sağlar (Lynch ve McNamara 1998; Engelhart ve Myford 2003; Lee ve Kantor 2003). Sınavda dört puanlayıcının bireyleri 13 görev üzerinden puanlaması nedeniyle (4x13) 52 puanlayıcı x görev yanlılık kombinasyonu bulunmaktadır. Puanlayıcı x görev standart Z puanlarının ortalaması 0,11 ve standart sapması 4,42 olup –13,42 ile 11,29 arasında değişmektedir. Bununla birlikte, puanlayıcı x görev karşılıklı etkisi için hesaplanan logit değerlerinin ortalaması -0,01 ve standart sapması 0,22 olup, -0,62 ile 0,36 arasında değerler almaktadır. 52 puanlayıcı x görev yanlılık kombinasyonun 30 tanesinin (% 57,69) Z puanları mutlak değer olarak 2'den büyüktür. Standart Z puanları anlamlı olan bu 30 puanlayıcı x görev karşılıklı etkisinin 15 tanesi pozitif işaretliyken 15 tanesi de negatif işaretlidir.

Şekil 4. ÖYSS Puanlayıcı x Görev Etkileşimi Z Puanları Grafiği



Z puanları anlamlı olan puanlayıcı x görev karşılıklı etkisinin 30 tanesinin 6 tanesi (3'ü pozitif ve 3'ü negatif işaretli) birinci puanlayıcıya, 9 tanesi (6'sı pozitif, 3'ü negatif işaretli) ikinci puanlayıcıya, 11 tanesi (4'ü pozitif ve 7'si negatif işaretli) üçüncü puanlayıcıya ve 4 tanesi ise (1'i pozitif, 3'ü negatif) dördüncü puanlayıcıya aittir.

Anlamlı olan Z puanları tek tek incelendiğinde; birinci puanlayıcının üç görev (5., 6., ve 11. görevler) için anlamlı Z puanları negatif olduğundan bu görevler için cömert puanlama yaptığı, üç görev (7., 8., ve 9. görevler) için ise anlamlı Z puanlarının pozitif olduğundan katı puanlama yaptığı görülmektedir. İkinci puanlayıcının iki görev (8. ve 9. görevler) için anlamlı Z puanları negatif olduğundan bu görevler için cömert puanlama yaptığı, yedi görev (4., 5., 6., 7., 10., 11. ve 13. görevler) için ise anlamlı Z puanları pozitif olduğundan katı puanlama yaptığı söylenebilir. Üçüncü puanlayıcının yedi görev (1., 2., 3., 4., 7., 8. ve 10. görevler) için anlamlı Z puanları negatif olduğundan bu görevler için cömert puanlama yaptığı, dört görev (5., 6., 9., ve 12. görevler) için ise anlamlı Z puanları pozitif olduğundan katı puanlama yaptığı görülmektedir. Dördüncü puanlayıcının ise üç görev (5., 6. ve 12. görevler) için anlamlı Z puanları negatif olduğundan, bu görevler için cömert puanlama yaptığı, bir görev (8. görev) için ise anlamlı Z puanları pozitif olduğundan katı puanlama yaptığı görülmektedir.

SONUÇ VE TARTIŞMA

2002-2003 öğretim yılı İnönü Üniversitesi Güzel Sanatlar Eğitimi Bölümü Müzik Öğretmenliği Programı Özel Yetenek Seçme Sınavı ikinci aşaması için Çok-Yüzeyli Rasch Modeli (ÇYRM) model-veri uyumu sağlanmıştır.

Bu sınavın yüzeylerine göre varılan sonuçlar ve sonuçlara dayalı öneriler; (a) 249 bireyden olağan olmayan uygunluk istatistiklerine sahip 10 bireyin (% 4,02) bulunduğu, bu sınavla bireylerin istatistiksel olarak anlamlı yaklaşık yedi (6,97) yetenek tabakasına ayrılabilmektedir. Bu yetenek tabaklaması yeterli sayılabilir. (b) Dört puanlayıcının bütün bireyler boyunca verdikleri puanların uygun olduğu ve puanlayıcıların bütün bireyler boyunca yaptıkları puanlamalarda istatistiksel olarak anlamlı düzeyde farklı katılık/cömertlikte puanlama yapmadıkları, dört puanlayıcının bütün bireyler boyunca yaptıkları puanlamaları arasında göreceli olarak çok az farklılık olduğu, özellikle üçüncü puanlayıcının diğer üç puanlayıcıya göre daha katı puanlama yaptığı, fakat bu farklılıkların ılımlı olduğu söylenebilir. Farklılıkların daha da azaltılabilmesi için puanlayıcılar puanlama kriterleri bakımından eğitilebilir. (c) Sekizinci görevin (konuşmada anlaşılırlık) uygun bulunmaması nedeniyle; bu görevin puanlarının diğer görevlerin puanlarıyla aynı çizgide olmadığı, test yapısı içinde farklı bir boyut oluşturduğu, 13 görevin güçlükleri arasında istatistiksel olarak anlamlı farklılık olduğu görülmektedir. Söyleme (şarkı) boyutuyla doğrudan ilgili olmadığı düşünülen ve analizde de bu sonuca varılan bu görev çıkarılmalıdır. (d) 996 puanlayıcı x birey olası yanlılığından 107 tanesinin (% 10,74) yanlı olduğu ve bazı puanlayıcıların bazı bireyleri diğerlerine göre daha katı ya da cömert puanlamışlardır. Puanlayıcıların bu konuda daha objektif davranması veya yanlı bulunan puanların düzeltilerek kullanılması uygun olacaktır. (e) 3237 olası görev x birey yanlılığından 150 tanesinin (% 4,63) yanlı bulunmuştur. Bu sonuç bazı görevlerin bazı bireyler için yanlı olduğunu göstermektedir. Bu bağlamda bu puanların düzeltilerek kullanılması sonuçları daha anlamlı hale getirecektir. (f) 52 olası puanlayıcı x görev yanlılığından 30 tanesinin (% 57,69) yanlı bulunmuş ve puanlayıcıların görevleri puanlamada birbirlerinden farklı katılık/cömertlikte davrandıkları görülmüştür. Puanlayıcıların her bir görevi puanlama kriterlerinin ortak hale getirilmesi için eğitilmesi ya da puanlama kriterlerinin ayrıntılandırılması etkili olabilir.

KAYNAKÇA

Aldrich, D. A. (1978). Rating scale formilation for ordered response categories. *Psychometrika*, 43, 561-573.

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

Engelhard, G. Jr. & Myford, C. M. (2003). Monitoring faculty consultant performance in the advanced placement English literature and composition program with a many-faceted Rasch model. *College Board Research Report*, No. 2003-1, ETS RR-03-01.

Griffin, P. & Gillis, S. (2000). A multi source measurement approach to assessment of higher order competencies. *Paper Presented at the educational Research Association Conference*, Wales, United Kingdom: Cardiff University.

Hambleton, R. K. & Swamitnathan, H. (1985). *Item response theory: principles and applications*. Boston: Kluwer-Nijhoff Publishing.

İnönü Üniversitesi (2003). *Eğitim Fakültesi Güzel Sanatlar Eğitimi Bölümü Müzik Öğretmenliği Programı Özel Yetenek Sınav Yönergesi*. Malatya: İnönü Üniversitesi.

Linacre, J. M., Wrıght, B. D. & Lunz, M. E.. *A Facet Model for Judgmental Scoring*, MESA Memo 61, http://www.rasch.org/memo61.htm (erişim 13 eylül 2004)

Linacre, John M. (1993). Rasch-based generalizability theory. *Rasch Measurement Transaction*, 7:1 p. 283.

Linacre, J. M. (1990). A facet model for judgmental scoring. *MESA Memo 61*, Chicago:MESA Pres,.

Linacre, J. M. (1989). *A many-facet Rasch measurement*. Chicago: MESA Pres,.

Linacre, J. M.(1991-2003). *A user's guide to FACETS: Rasch-model computer programs*. Chicago: Winsteps.

Linacre, J. M.(1991-2003). *A user's guide to FACFORM data formatter for FACETS: Rasch-model computer programs*. Chicago: Winsteps.

Linacre, J. M.(1993). Generalizability theory and rasch measurement. *American Education Research Area*.

Lord, Frederic M. (1980). *Applications of Item Response Theory to Practical Testing Problems*. Hillsdale, New Jersey: Lawrence Erlbaum Associates Publishers.

Lunz, M. E., Wright, B.D. & Linacre, J. M. (1990). Measuring the impact of judge severity on examination scores. *Applied Measurement in Education*, 3(4), 331-345.

Lunz, M., E. & Wrıght, B. D. (1997). Latent trait models for performance examinations, Jurgen Rost, Rolf Langeheine(Eds.), *Application of Latent Trait and Latent Class Models in Social Science*. New York: 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.

Macmillan, P. D. (2000). Classical, generalizability and multi-faceted Rasch detection of interrater variability in large sparse data sets. *Journal of Experimental Education*, Vol. 68, Issue 2, p.167.

Nunnally, J. C. & Bernstein, I. H. (1994). *Psychometric theory*. (3rd Editions) New York: McGraw-Hill Inc..

ÖSYM (2003). 2003-ÖSS Öğrenci Seçme Sınavı Kılavuzu. Ankara: *ÖSYM*.

Rost, J. & Langeheine, R. (1997). A guide through latent structure models for categorical data, Jurgen Rost, Rolf Langeheine(Eds.) *Application of Latent Trait and Latent Class Models in Social Science*. New York: Waxmann.

Schumacker, E. R. (April 25, 2003). Reliability in Rasch measurement: avoiding the rubber ruler. Papere Presented at Annual Meeting of the *American Educational Research Association*, Chicagoi Illinois.

Zhu, W., Ennis, D. C & Chen, A. (1998). Many-faceted Rasch modeling expert judgment in test development. *Measurement in Physical Education and Exercise Science*, 2(1), 21-39.

SUMMARY

In this reserarch, by analyzing Sepecial Ability Sellection Examination for music education department of Education Faculty of İnönü University, it was aimed to make the exam more convenient by determining individual ability measure, severity of raters, task difficulty level and fit statistics, and bias of interaction between raters by individuals, tasks by individuals and raters by tasks.

Method

In this reserarch, the data of 249 candidatres out of 689, who took and passed the first exam and later took the second exam for music education department in the faculty of education of İnönü University during the academic year of 2003-2004, and also these data were from decoding, playing and singing dimensions, 13 tasks in total, and rated by four raters. In analyzing the data, $\log(P_{nggpk}/ P_{nggp(k-1)})=B_n-A_gD_i-C_p-F_k$ MFRM pattern, that is B_n: ability of person n, A_g: the challenge of task g, D_g: the difficultty of task g, C_p: the severity of rater p, F_k: the barrier to being observed in category k relative to category k-1, FACDOS, FACETS and FACFORM computer softwares were used.

Results and Discussion

The findings of the research acording to the reserch data and facets of the exam with fited Many-Facets Rasch Measurement model and data are as follows: (a) 4,02% of individuals have unusual fit statistics and individuals may separeted aproximately seven (6,97) significant ability levels. Such an ability level may be taken as adequate. (b) The four raters had sinificantly differ severity in their rating throught all individuals, especially the third rater can be said to have severe rated when compared the other three raters. To decrese the differences between the raters, the raters can be educated in terms of rating criteria. (c) The eighth task was like a different dimension due to being inappropriate in (comprehensibility in speaking), and between the difficulty levels of 13 tasks, there was significant difference. It was not directly related to singing dimension and was proved so after the analysis; hence it must be taken out. (d) 10,74% of raters by individuals possible interaction bias was partialed. It will be appropriate for the raters to be behave more objectively or the biased scores may be use after correction. (e) 4,63% of tasks by individuals possible interaction bias was partialed. This result indicates that some tasks are bias to some individuals. So, using the corrected scores makes the results more meaningful. (f) 57,69% of raters by tasks possible intraction bias was partialed, and it was seen that the raters had different severity behaviours in rating scores. Bias can be decreased only by making rating criteria common for each tas kor making the criteria detailed