ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2017; 8(4); 469-489

Akran Değerlendirmesinde Puanlayıcı Katılığı Kayması

Rater Severity Drift in Peer Assessment

Bengü BÖRKAN*

Öz

Akran değerlendirmesinde elde edilen puanların geçerliği ve güvenirliği hakkında sağlam psikometrik dayanağı olan ve özellikle puanlayıcı etkisine değinen yeteri kadar çalışma bulunmamaktadır. Bu çalışmada puanlayıcı etkilerinden olan puanlayıcı katılık kaymasının (rater severity drift), akran değerlendirmede ne derece görüldüğü araştırılmıştır. Eğitim fakültesindeki bir ders kapsamında öğrenciler tarafından gerçekleştirilen sözlü sunum performansları aynı dersi alan 29 akran tarafından dereceli puanlama anahtarı kullanılarak puanlanmıştır. İlk üç gün iki sunum, dördüncü gün üç sunum olmak üzere toplam dokuz sunum dört ayrı günde gerçekleştirilmiştir. Puanlayıcı kayması iki farklı çok yüzeyli Rasch ölçme modeli (ayrı modeller ve kukla zaman) yardımıyla incelenmiştir. Her gün için hesaplanan puanlayıcı kestirimlerinden standartlaştırılmış farklar indeksi ve kukla zaman modelinden etkileşim terimleri hesaplanmıştır. Puanlayıcı kayması analizinde, Gün-1 temel gün alınmış, Gün-1'den diğer günlere (Gün-2, 3 ve 4) değişimler incelenmiştir. Analizler genel olarak akran puanlayıcıların arkadaşlarını oldukça cömert bir biçimde puanladıklarını göstermiştir. Puanlayıcılar kendi aralarında kıyaslandığında ise katılık/cömertlik seviyelerinin birbirlerinden farklı olduğu görülmüştür. Sunumlar puanlayıcılar tarafından tutarlı bir şekilde niteliklerine göre sıralandırılmıştır. Puanlayıcı kaymasını incelemek için kullanılan iki yöntem benzer sonuçlar vermiştir. Gün-1 ve 2 arasında puanlayıcı kestirimlerinde bir farklılık görülmemektedir. Her ne kadar ortalamada puanlayıcılar daha cömert puanlama yapsa da, kaymalar istatistiksel olarak anlamlı değildir. Gün-1 ve 3 arasında puanlayıcıların kestirimlerinde önemli kaymaların olduğu puanlayıcıların oranı %38,10'dur. İki yönteme göre de puanlayıcılar ortalamada yaklaşık 0,14 logit kayma gösterip daha katı puanlama davranışı sergilemiştir. Gün-1 ve 4 arasında puanlayıcıların kestirimlerinde önemli kaymaların olduğu puanlayıcıların sayısı standartlaştırılmış farklar yöntemiyle üçgen, etkileşim terimi yöntemiyle birdir. Ortalamada iki yöntemle de puanlayıcılar daha katılaşmıştır. Ortalamada kaymanın en yüksek olduğu Gün-4'tür.

Anahtar Kelimeler: Akran değerlendirme, geçerlik, puanlayıcı kayması, çok yüzeyli Rasch ölçme modeli

Abstract

There are not enough psychometrically sound studies about the validity and reliability of the scores obtained from peer assessment. This study examined degree of rater severity drift, a rater effect, in peer assessment. The college students' presentations were scored by 29 peers in the class using a rating scale. Nine presentations were held on four separate days, two presentations on each of the first three days and three presentations on the fourth day. Drift was investigated with two many-facet Rasch measurement models (separate models and dummy time MFRM). Standardized differences were calculated from the estimates obtained with separate models and interaction terms were calculated with the dummy time MFRM. In drift analysis, shifts in estimations were examined from Day-1 which is a baseline to other three days. Results showed that peer raters varied according to their level of severity and they tend to be lenient. Statistics showed that the quality of the scale was acceptable and its items behaved as expected. In drift analysis, standardized differences and interaction term provided very similar results. Between Day-1 and 2, there was no statistically significant difference in the estimates of the rater severity. Between Day-1 and 3, the percentage of scorers with significant drift in the estimates was 38.10%. The raters' severity shifts on the average of about 0.14 logit and they displayed more severe scoring behavior. Between Day-1 and 4, the number of raters who had significant shifts in their estimates was three according to the standardized difference method, while one according to interaction method. On the average, the raters became more severe. Among three comparison, Day-4 had the

*Yard. Doç. Dr., Boğaziçi Üniversitesi, Istanbul-Turkiye, e-posta: bengu.borkan@boun.edu.tr, ORCID ID: https://orcid.org/0000-0003-1414-1528

Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi, Cilt 8, Sayı 4, Kış 2017, 469-488.
Journal of Measurement and Evaluation in Education and Psychology, Vol. 8, Issue 4, Winter 2017, 469-488.
DOI: 10.21031/epod.328119
Geliş Tarihi: 12.07.2017
Kabul Tarihi: 21.11.2017

largest rater severity drift on the average; Day 3, however, has the highest number of raters with rater severity drift.

Keywords: Peer assessment, validity, rater drift, Many-Facets Rasch Measurement

GİRİŞ

1774-1826 yılları arasında Glasgow Üniversitesi'nde profesör olan George Jardine, yazma becerisini geliştirmede akran değerlendirme yöntemini ve avantajlarını anlatmıştır (Gaillet, 1992 aktaran Topping, 2009). Okul bağlamında akran değerlendirme, öğrencilerin akranlarının performanslarını ya da oluşturdukları bir ürünün seviyesini, değerini veya niteliğini değerlendirmesi olarak tanımlanır (Falchikov, 1995; Topping, 2009). Akran değerlendirmesi, farklı ortamlarda uygulanabilmektedir. Örneğin yüz yüze yapılabildiği gibi son zamanlarda teknolojinin sağladığı olanaklarla çevrimiçi ortamlarda (ör. Demirbilek, 2015, Tseng & Tsai, 2007, Yang & Tsai, 2010) da kullanılabilmektedir (Topping, 2009). Akran değerlendirme tek taraflı olduğu gibi değerlendiriciler arasında karşılıklı yapılabilmekte, sınıf içinde ya da dışında gerçekleştirilebilmektedir (Topping, 2009).

Son yıllarda, özellikle de biçimlendirici değerlendirme açısından akran değerlendirmesine olan ilgi artmıştır. Akran değerlendirmesi, özel eğitime ihtiyaç duyan öğrenciler de dâhil olmak üzere her yaş grubunda ve tüm eğitim kademelerinde başarıyla kullanılmaktadır (Scruggs & Mastropieri, 1998). Akran değerlendirmesi niteliksel ve/veya niceliksel değerlendirme sonuçları sağlayabilmekte ve bu durum bir sürekliliğin iki ucu olarak düşünülebilmektedir. Bir uçta akran değerlendirme sadece geri bildirim amaçlı niteliksel veri/bilgi sağlarken diğer uçta niceliksel bir puanlama yer alır. Akran değerlendirme bu sürekliliğin herhangi bir noktasında yer alabilir. Benzer şekilde, akran değerlendirmesinin amacı biçimlendirici ve/veya düzey belirleme amaçlı olabilir (Somervell, 1993). Akran değerlendirmenin öğrenme etkinliğinde ve niteliğinde iyileşmeler sağlayabileceğine dair önemli kanıtlar bulunmaktadır (Weaver & Cotrell, 1986). Özellikle yazma becerisinin geliştirilmesinde, öğretmen değerlendirmesi kadar etkili olabilmektedir (Topping, 2009).

Akran değerlendirmesinde farklı yöntemler bulunmaktadır. Bunlardan en az kullanılanı akran sıralamadır (*peer ranking*). Bu yöntemde her bir grup üyesi gruptaki bütün akranlarını ölçülen niteliğe göre en başarılıdan en başarısıza doğru sıralar. Bir diğer yöntemde ise her bir grup üyesi, gruptaki en iyi performansı gösteren akranı ya da akranları aday gösterir (*peer nomination*) (Docy, Segers, & Sluijsmans, 1999; Kane & Lawyer, 1978; Love, 1981). Benzer bir şekilde en başarısız performansı gösteren akran ya da akranlar da aday gösterilebilir (Heyman & Sailors, 2011). Üzerine en çok akademik çalışma yapılan yöntem ise akran derecelendirmesidir (*peer rating*). Bu yöntemde kontrol listesi, derecelendirme anahtarı ya da rubrik yardımıyla her bir grup üyesi gruptaki bütün akranları, belirlenmiş bir dizi ölçüte göre puanlar (Docy, Segers, & Sluijsmans, 1999; Kane & Lawyer, 1978; Love, 1981).

Akran değerlendirme, değerlendirme sürecinde bulunan her iki taraf için de kazanç sağlamaktadır (Topping, 2005; Topping & Ehly, 1998). Bu değerlendirme türüyle farklı öğretim programlarındaki yazma becerisi ve sözlü sunum gibi çeşitli ürün veya davranışlar değerlendirilebilmektedir. Akran değerlendirme biçimlendirici değerlendirme olarak kullanıldığında akılcı sorgulamayı, farkındalığı (sels-disclosure), dolayısıyla anlamanın değerlendirmesini sağlar. Buna ek olarak hataların ve kavram yanılgılarının fark edilmesine olanak sağlayarak bilgilerdeki boşlukların kapatılmasını mümkün kılar ve değerlendirme öncesi, sırası ve sonrasında bilişsel ve üstbilişsel yararlar oluşturur (Topping, 2009)

Güvenirlik ve Geçerlik

Her ölçme işleminde olduğu gibi akran değerlendirme sürecinde elde edilen ölçme sonuçlarının geçerliği ve güvenirliğinin tartışılması gerekmektedir. Alan yazınında kullanılan akran değerlendirmesinin güvenirlik ve geçerliği tabiriyle aslında akran değerlendirmeden elde edilen ölçümlerin geçerliliği ve güvenirliğinden bahsedilmektedir. Alan yazınında akran değerlendirmede

ulaşılan ölçümlerin güvenirlik ve geçerliğini belirlemek için, genel olarak öğrencilerin değerlendirme sonuçlarıyla öğretmenin değerlendirme sonuçları arasındaki uyuma bakılmış ancak puanlayıcılar arasındaki uyum ya da aynı puanlayıcının zaman içindeki kendisiyle olan tutarlılığına (puanlayıcı içi güvenirlik) değinilmemiştir. Bu tür çalışmalar öğretmen/uzman değerlendirmen puanlarını kendi başına oldukça güvenilir ve geçerli olduğu varsayılarak yapılmıştır. Bu durum, bazı bağlamlarda şüpheli bir varsayım olduğundan söz konusu çalışmaların güvenirlik veya geçerlik ya da her ikisi için yapılmış çalışmalar olup olmadığı tartışmalıdır (Topping, 2003; Topping, 2009). Bu çalışmalarda genellikle doğruluk (accuracy), geçerlik ve güvenirlik terimleri birbirleri yerine kullanılmıştır.

Akran değerlendirmenin güvenirliği ve geçerliği üzerine yapılan araştırmalar, çoğunlukla yükseköğretim düzeyinde yapılmıştır. (Falchikov, 2001; Topping, 2003). Çalışmaların çoğunluğu güvenirlik ve geçerlik derecesini yeterli bulmaktadır (Sadler & Good, 2006); bazı çalışmalarda ise farklı sonuçlar raporlanmıştır (Falchikov & Goldfinch, 2000; Topping, 1998). İlköğretim ve ortaöğretim düzeyinde yapılan çalışma sonuçları da yükseköğretimde yapılan çalışmalara benzer sonuçlar vermiştir (Toppings, 2003).

Topping (1998) üniversite öğrencileriyle yapılmış 31 çalışmada öğrencilerin akran değerlendirme sonuçlarıyla öğretim elemanı gibi uzman kişiler tarafından yapılan değerlendirme sonuçlarının uyumunu incelemiş ve bu değerlendirme türünün güvenirlik ve geçerlik açısından yüksek ölçümler verdiği sonucuna varmıştır. Benzer şekilde Falchikov ve Goldfinch (2000) tarafından gerçekleştirilen meta analiz çalışmasında 56 deneysel çalışma incelenmiştir. Tüm çalışmalarda ortalama korelasyon katsayısı 0,69 olarak bulunmuş, ortalama olarak akran ve öğretmen puanları arasında uzlaşma sağlandığına dair kanıt sunulmuştur. İstatistiksel olarak anlamlı olmayan etki değeri de öğretmenlerin puanlamasıyla öğrencilerin puanlaması arasında uyum olduğunu göstermektedir. Bu iki meta analiz çalışmasına göre daha yeni olan bir araştırma (Hafner & Hafner, 2003) akran değerlendirmesinde kullanılan rubriğin güvenirliği ve geçerliği üzerine odaklanmıştır. Üç yıllık bir periyotta yürütülen çalışma, biyoloji bölümüne kayıtlı 107 lisans öğrencisinin katılımıyla gerçekleşmiştir. Çalışma kapsamında toplam 1577 akran-grup sözlü sunum puanlaması yapılmıştır. Araştırma sonucunda öğretim elemanı puanlaması ile öğrencilerin puanlaması arasında mükemmel bir ilişki raporlanmıştır (r=1,0). Genellenebilirlik çalışmasına göre de yıllar boyunca akranlar arası güvenirlik (inter-rater reliability) orta seviyede bulunmuştur. Hafner ve Hafner çalışmanın sonucunda akran değerlendirilmesinin üç yıl süresinde tutarlı bir şekilde kullanıldığı sonucuna varmıştır. Sadler ve Good (2006) ortaöğretim fizik dersinde gerçekleştirdikleri çalışmada öğretmen ile akran puanlanması karşılaştırılmış ve öğrencilerin rubrik kullanarak yaptıkları puanlamalar ile öğretmen tarafından verilen puanlar arasında 0,90'ın üzerinde güçlü bir ilişki bulunmuştur. Fakat öğrencilerin akranlarını değerlendirirken yanlı davranış sergiledikleri saptanmış, en iyi performans gösteren akranlara verilen puanların, öğretmen tarafından verilen puanlardan daha düşük olduğu gözlenmiştir.

Puanlayıcı Kayması

Açık uçlu sınavlar ve sunum gibi karmaşık performans görevleri puanlayıcılar tarafından puanlanırken puanlayıcının kendi yargısı ölçme sonuçlarını etkileyerek ölçümlerin geçerliğini düşürebilmektedir. Alan yazınında bu durum puanlayıcı etkisi ve puanlayıcı hatası gibi farklı terminolojilerle ele alınmaktadır; 'puanlayıcı etkisi', 'puanlayıcı yanlılığı' ya da 'puanlayıcı hatası'. Bu terminolojilerin tam olarak tanımları yapılamadığından terimler birbirlerinden ayrılamamaktadır ve birbirleri yerine yanlış kullanılmaktadır (Myford & Wolfe, 2003). Bu çalışmada, Scullen, Mount ve Goff (2000) tanımından yola çıkarak 'puanlayıcı etkisi' terimi kullanılacaktır. Puanlayıcı etkisi, puanlanan bireyin performans puanında bu bireyin gerçek performasından ziyade puanlayıcıdan kaynaklanan, sistematik farklılığa yol açan geniş bir etki kategorisidir. Performans değerlendirmede ölçme sonuçlarında hataya yol açan en büyük etki puanlayıcının kendisidir (Engelhard,1994; Gabrielson, Gordon & Engelhard, 1995; McNamara, 1996). Puanlayıcıdan kaynaklanan herhangi bir hata ölçmek istediğimiz yapının dışındaki nedenlerden dolayı puanların varyansına yansıyabilir (Messick, 1994).

Geleneksel olarak puanlayıcı hataları puanlayıcı katılığı ve cömertliği, halo etkisi, merkezi eğilim ve ranj sınırlaması olarak dört başlıkta incelenir. Fakat bunların yanı sıra daha az bahsedilen hatalılık (inaccuracy), yanlılık (değişen puanlayıcı katılığı/cömertliği), sıralama, puanlayıcı kayması gibi diğer hata türleri de mevcuttur (Myford & Wolfe, 2003). Yaygın olarak tanımlanan puanlayıcı etkileri çok sayıda çalışmada belgelenmiş ve bunun üzerine bu etkilerin ortadan kaldırılabilmesi için farklı işlemler tartışılmıştır (ör. Braun & Wainer, 1989, Engelhard, 1996, Myford & Wolfe, 2003, 2004). Son zamanlarda uygulaması ve puanlaması zamana yayılmış geniş ölçekli testlerde puanlayıcı davranışlarının ölçme hatasına önemli ölçüde yol açabileceği vurgulanmaktadır (Harik ve diğerleri., 2009).

Araştırmacılar geçmişte puanlayıcı etkisini statik etki (puanlayıcı etkisi her öğrencinin performans puanını tam olarak aynı şekilde etkiler) olarak tanımlarken daha yeni çalışmalarda her bir puanlayıcının davranışının zamanla değişebildiği görülmüştür (Myford & Wolfe, 2009). Puanlayıcı kayması (DRIFT—differential rater functioning over time) bir performansın farklı zamanlarda yapılan puanlamalarında, puanlayıcı davranışlarındaki değişikliklerin meydana gelmesi olarak tanımlanmaktadır (Park, 2011; Wolfe, Moulder, & Myford 2001). Puanlayıcı kaymasını inceleyen araştırmalar çoğunlukla İngilizce yeterlilik sınavı puanlamalarıyla yapılan çalışmalarla karşımıza çıkmaktadır (ör., Englehard & Myford, 2003, Yang, 2010). Bunun yanı sıra similasyon verisi kullanılarak gerçekleştirilmiş çalışmalar da bulunmaktadır (ör., Park, 2011; Wolfe, Moulder, & Myford, 2001). Bu çalışmaların tamamında, zamana bağlı olarak puanlayıcı kayması görülmüştür. (ör., Braun, 1988, Casabianca, Lockwood & McCaffrey, 2015, Congdon & McQueen, 2000, Englehard & Myford, 2003, Harik vd.., 2009, Hoskens & Wilson, 2001, Lunz & Stahl, 1990, Lumley & McNamara, 1995, McQueen & Congdon, 1997, Myford, 1991, Myford & Wolfe, 2009, Wolfe, Moulder & Myford, 2001, Wilson & Case, 2000). Puanlayıcı kaymasının sebebi, zamanla puanlayıcıların tecrübe edinmesi veya yorulmalarından kaynaklanabilir. Puanlayıcı kaymasını engellemek için sürekli puanlayıcı eğitimleri önerilmektedir fakat bu eğitimler tam olarak puanlayıcıların davranışlarındaki değişimleri engelleyememektedir (Congdon & McQueen, 2000; McKinley & Boulet, 2004).

Ölçme sonuçlarıyla bireyleri birbirleriyle kıyaslayabilmek için madde kalibrasyonlarının gruptan gruba ve zamandan zamana sabit kalması gerekmektedir (Wright ve Masters, 1982). Benzer bir şekilde birden fazla puanlayıcının kullanıldığı puanlamalarda puanlayıcı kalibrasyonunun puanlanan bireyden bireye ve zamandan zamana değişmemesi beklenir. Ölçmede puanlayıcı kayması üç ayrı şekilde gözlenebilir; 1) zamanla puanlayıcı daha katı ya da daha cömert davranış sergileyebilir (differential severity), 2) puanlayıcının puanlamada yaptığı hata miktarı farklı zamanlarda farklı olabilir (differential accuracy) ve 3) puanlayıcı zamanla ölçekteki kategori kullanımda farklı eğilimler gösterebilir (differential category use). Alan yazınında puanlayıcı kayması çalışan araştırmacıların büyük çoğunluğu ilk kayma türünü incelemiştir. Bunlardan bazıları kronolojik sıraya göre aşağıda özetlenmiştir.

Lunz ve Stahl (1990) çok yüzeyli Rasch ölçme modeli (ÇYRÖM) kullanarak üç farklı sınavda (İngiliz Edebiyatı kompozisyon sınavı, klinik sınavı ve sağlık meslek sözlü sınavı) puanlayıcıların katılık/cömertlik seviyelerinin üç ya da dört gün süren puanlamalarda sabit kalıp kalmadığını incelemişler ve iki sınav türünde (kompozisyon ve klinik) puanlayıcıların katılık düzeylerinde belirgin farklılaşmalar olduğunu bulmuşlardır. Myford (1991) da ÇYRÖM kullanarak gerçekleştirdiği çalışmasında drama performanslarını değerlendiren farklı tecrübeye sahip hakemlerin bir ay süresince yaptıkları puanlamalarda açık bir puanlayıcı katılığı kayması olduğunu bulmuştur. Bir başka çalışmada Lumley and McNamara (1995) puanlaması 20 ay süren İngilizce konuşma testinde puanlayıcı davranışlarını incelemişlerdir. Puanlayıcı ana terimi ve puanlayıcı-zaman etkileşim terimleri için puanlayıcı katılığında önemli değişiklikler bulmuşlardır. Wilson ve Case (1997) ise sekizinci sınıf matematik sınavı için iki oturumda gerçekleşen puanlamada puanlayıcı katılığı kaymasını araştırmışlardır. Puanlayıcıların katılık seviyelerinde bir zaman diliminden diğerine istatiksel olarak anlamlı kaymalar olduğunu bulmuşlardır. Congdon ve McQueen (2000), yedi iş gününe (arada bir hafta sonu olmak üzere) yayılan ilköğretimde yazma performanslarının bütüncül rubrik kullanılarak yapılan değerlendirilmelerinde puanlayıcı kaymasını

araştırmışlardır. Araştırmacılar ÇYRÖM kullanarak her bir gün için 16 puanlayıcının göreceli katılık kestirimlerini hesaplamışlardır. Analiz sonuçları, puanlayıcı katılığının günden güne çoğu puanlayıcı için değiştiğini fakat bu değişimin genel bir deseni olmadığını göstermiştir.

Yukarıdaki çalışmalardan farklı olarak Wolfe, Moulder ve Myford (2001) simülasyon verisi kullanarak puanlayıcı katılığı kaymasını çalışmışlardır. ÇYRÖM kullanan araştırmacılar, çeşitli puanlayıcı kayması türlerini tespit etmişlerdir. Fakat Harik ve diğerleri (2009) bu çalışmanın sonuçlarının sadece simülasyon verisine dayalı olduğu için genellenebilirliğinin sorgulanması gerektiğini, çalışmadaki simüle edilmiş koşulların, her bir denemede farklı bir kayma türünü temsil ettiğini, oysa gerçek verilerde, zamana bağlı puanlayıcı kayma türlerinin farklı kombinasyonlarda ortaya çıkabileceği belirtmişlerdir.

McLaughlin, Ainslie , Coderre, Wright ve Violato (2009) 10-12 dakikalık istasyonlarda farklı zaman dilimlerinde yapılan tıp sınavında puanlayıcı kaymasını incelemişlerdir. Zamanla puanlayıcıların daha katı puanlama davranışı gösterdiklerini bulmuşlar ve bunun sebebini yorgunluk olarak tanımlamışlardır. Puanlayıcı kayması gibi sistematik yanlılığın, testin geçerliğini tehlikeye attığını belirtmişlerdir. Az da olsa çalışmaların bir kısmında da genellenebilirlik kuramı kullanılmıştır. G kuramı kullanılarak gerçekleştirilen çalışmaların birinde Casabianca, Lockwood ve McCaffrey (2015) puanlayıcı kaymasını çalışmışlardır. Çalışmalarında 458 matematik ve İngilizce ortaöğretim öğretmeninin öğrencileri tarafından değerlendirilmelerinde puanlayıcı davranışlarının zaman içinde değişip değişmediğini araştırmışlardır. Toplanan veriye göre eğitim kalitesindeki değişimin çok küçük olduğunu fakat gözlemlerin başında puanlayıcı kaymasının çok büyük olduğunu ve bu kaymanın iki yıl boyunca devam ettiğini raporlamışlardır.

Araştırmanın Önemi ve Amacı

Giriş kısmında alan yazını incelemesinde ele alındığı gibi akran değerlendirmesiyle elde edilen puanların geçerliliğine dair kanıtlar sınırlıdır. Bu nedenle psikometrik ağırlığı olan çalışmaların yapılması oldukça gereklidir. Bu çalışmada olduğu gibi özellikle akran değerlendirmesi sürece yayılıyorsa, performans puanlamada puanlayıcı etkisinin görülme olasılığı daha da artmaktadır. Üniversite öğrencileriyle yapılan bu çalışmada dört ayrı güne yayılan dereceli puanlama anahtarı kullanılarak sunum performansları puanlanma sürecinde, akran puanlayıcıların davranışları ve ölçme aracının niteliği araştırılmıştır ve özel olarak aşağıdaki iki ana araştırma sorusuna cevap aranmıştır.

1. Dört değerlendirme günü boyunca göreceli katılık seviyesinde ilk puanlama gününe göre değişen herhangi bir puanlayıcı var mıdır? Bir başka deyişle değişen katılık ya da değişen cömertlik gösteren puanlayıcı var mıdır?

2. Değişen katılığı/esnekliği tespit etmek için kullanılan farklı yaklaşımlar benzer sonuçlar veriyor mu? Farklı yaklaşımlar değişen katılık/esnekliğe sahip benzer puanlayıcıları ortaya çıkarıyor mu?

YÖNTEM

Bu çalışma, akran değerlendirmesinde puanlayıcı katılığında zamana bağlı kaymayı inceleyen betimsel bir çalışmadır. Dereceli puanlama anahtarı kullanılarak dört ayrı günde puanlanan öğrenci sunum performanslarında akran puanlayıcı katılığı kayması çok yüzeyli Rasch ölçme modeli (ÇYRÖM) (Linecra, 1989) yardımıyla incelenmiştir.

Çalışma Grubu

Bu çalışma orta büyüklükte olan bir üniversitenin eğitim fakültesinde 2017 Bahar döneminde verilen eğitimdeki ölçme ve değerlendirme dersinde gerçekleştirilmiştir. Öğrenciler dersin öğretim elemanının danışmanlığında belirlenen PISA, test yanlılığı gibi eğitimde ölçme ve değerlendirme konularını kapsayan kendi hazırladıkları performans görevini dönem sonunda sınıf arkadaşlarına sözlü olarak sunmuşlardır. Grup çalışması olarak gerçekleştirilen bu çalışmada grup büyüklükleri iki ila dört kişi arasında değişiklik göstermektedir. Sunumlar yaklaşık 45 dakika sürmüştür. Sunumun

sonunda sunumu yapan grubun sunum performansı, sınıftaki akranları tarafından dereceli puanlama anahtarıyla değerlendirilmiştir. Derse devam sağlayan 29 öğrenci bu akran değerlendirme çalışmasına katılmıştır. Fakat değerlendirmenin yapıldığı her bir günde sınıfta tam katılım sağlanamamıştır. Dolayısıyla bu durum az da olsa veri kaybına yol açmıştır. Puanlama yapılmadan önce sınıfta puanlama anahtarı üzerinden gidilmiştir ve açıklama yapılmıştır. Fakat olması gerektiği gibi gerçek bir sunum üzerinden öğrenciler anahtar kullanımı üzerine eğitilmemiştir.

İlk üç gün iki sunum, dördüncü gün üç sunum olmak üzere toplam dokuz sunum dört ayrı günde gerçekleştirilmiştir. Her bir sunum puanlamasına katılan akran sayısı 19 ila 25 arasında değişmektedir. Her sunum grubu, sunumu yapan grup ve derse katılım sağlamamış öğrenciler dışında her akran tarafından puanlanmıştır. Fakat sunum performansları günlerin içinde yuvalanmış olduğu için, örneğin ilk iki sunum ilk gün, sonraki iki sunum ikinci gün gibi, günler arasında bağlantı yoktur. Her sunum, bir puanlayıcı tarafından bir kere sunumun yapıldığı gün puanlanmıştır. Bu nedenle ÇYRÖM'ye gün değişkeni kukla değişken olarak eklenebilmektedir.

Veri Toplama Aracı

Öğrenci sunumlarının akranlar tarafından değerlendirilmesi için araştırmacı, dereceli puanlama anahtarı geliştirmiştir. Bu puanlama anahtarı son beş yıldır aynı derste kullanılmaktır. Geçmiş yıllarda derse kayıtlı ve akran değerlendirmesini kullanan öğrencilerin geri bildirimleriyle son halini almıştır. Anahtarda toplam 10 ölçüt (madde) bulunmaktadır ve her bir ölçüt üç dereceli ölçekle puanlanmıştır. Anahtar örnek maddeler şunlardır: "Sunum sonunda konuyu anladım," "Örnekler vererek fikirlerini açığa kavuşturdu," "Sunum ilginçti[1]."

Ölçme Modeli

Bu çalışmada toplanan verilerin analizinde ÇYRÖM kullanılmış ve analizler FACET v3.71.4 (1987-2014) yazılımıyla gerçekleştirilmiştir. Tüm terimlerin ve öğelerin kestirimleri ortak bir metrikte (logit) ve yaygın olarak kullanılan çeşitli uyum istatistikleri bu yazılım yardımıyla hesaplanmıştır. ÇYRÖM, Rasch ailesinin üyesi olduğu için, temel Rasch modelinin tüm özelliklerine sahip olmakla birlikte daha kullanışlı bir modeldir. ÇYRÖM, birden fazla puanlayıcı olan performans ölçümlerinde çeşitli değişkenlik kaynaklarının incelenmesini sağlar. Örneğin değerlendirilen bireyin becerisi, görev zorluğu, puanlayıcı katılığı ve bu değişkenlerin birbirleriyle olan etkileşimi. Bu çalışmada değişen katılık/esneklik ölçüsü olarak ÇYRÖM kullanılarak hesaplanan iki indeks kullanılmıştır – ***standartlaştırılmış farklar*** ve ***etkileşim terimi***.

Standartlaştırılmış Farklar

Her bir gün için puanlayıcı katılık/cömertlik kestirimleri standart puana çevrilerek puanlayıcı kayması araştırılmıştır. Kestirimlerini elde etmek için üç yüzeyli ÇYRÖM (Denklem 1) kullanılmıştır; yüzeyler sunum grubu, puanlayıcı ve ölçüttür. Tipik olarak böyle bir modelde β_n terimi, ortalaması sıfır olacak şekilde eklenir. Fakat bu tür modelleme, dört gün için ortalaması sıfır olmayan (non-centered) dört ayrı puanlayıcı katılığı kestirimini sağlar. Bu kestirimler her gün değerlendirilen sunum performanslarının nitelik seviyelerine göre değişebilir. Farz edelim ki birinci gün puanlanan sunum, ikinci gün puanlanan sunuma kıyasla niteliği daha yüksektir. Her bir gün için sunum niteliğinin kestirimlerinin ortalaması 0,00 logite sabitlendiğinden puanlayıcı kestirimlerinde, ikinci gün kestirimleri ilk güne göre daha düşük olmasına yol açacaktır. Bir başka değişle puanlayıcılar ilk gün daha cömert gibi algılanacaktır. Bu etkiyi ortadan kaldırmak için, her bir kalibrasyonda puanlayıcı kestirimlerinin ortalaması 0,00 logit olarak alınmalıdır. Bu düzenleme,

[1]Çalışmanın yapıldığı kurumun öğretim dili İngilizce olduğundan maddelerin orijinal dili İngilizcedir. Bu makalede çevirisi kullanılmıştır.

farklı zamanlarda gerçekleştirilen puanlamalarda puanlayıcıların *göreceli katılık seviyelerindeki* değişimi takip etmemizi sağlayacaktır.

$$\ln\left(\frac{\text{Pnrik}}{\text{Pnri(k-1)}}\right) = \beta_n - \gamma_r - \delta_i - \tau_k \qquad \text{(1, Ayrı Model (Seperate Model))}$$

P_{nrik}, = puanlanan sunum n'nin 'i' ölçütünde gösterdiği performansın 'r' puanlayıcı tarafından 'k' kategorisinde puanlanma olasılığı

$P_{nri(k-1)}$ = puanlanan sunum 'n'nin 'i'ölçütünde gösterdiği performansın 'r' puanlayıcısı tarafından 'k-1' kategorisinde puanlanma olasılığı

β_n = Sunum n'nin niteliği

γ_r = Puanlayıcı r'nin katılığı

δ_i = Ölçüt i'nin güçlüğü

Gün 1 temel zaman alınarak sırasıyla diğer günler için sapmalar (SAI_{rc}) hesaplanır.

$$SAI_{rc} = S_{rc} - S_{rb}, \quad (2)$$

SAI_{rc} = Zaman c'yi temel zamanla (baseline) karşılaştıran puanlayıcı SAI (Signed Area Index) indeksi

c = karşılaştırılan zaman

b = temel zaman

S_{rj} = Okuyucu r'nin zaman j'de katılık ölçüsü.

Raju (1988, 1990 Wolfe, Myford, Engelhard and Manalo'da alıntılandığı gibi, 2007) farkların nasıl standardize edileceğini aşağıdaki gibi açıklamıştır.

$$Z_{SAIrc} = \frac{SAI_{rc}}{\sqrt{SE^2_{Src} + SE^2_{Srb}}} \quad (3)$$

Z_{SAIrc}= Okuyu *r* için standartlaştırılmış fark indeksi, temel katılık ölçümüne kıyasla zaman c'deki katılığı

SE^2_{Srj} = *j* zamanda okuyucu *r*'nin katılık kestirimindeki standart hatası

Okucunun temel zamandaki katılık seviyesiyle c zamanındaki katılık seviyesi arasında fark olmadığını belirten sıfır (null) hipotezini test etmek için yukarıdaki formülle hesaplanan z puanı standart normal dağılımla karşılaştırılır. Bununla birlikte SAI değeri bir etki değeri ölçütü olarak da kullanılabilir; 0,50'den büyük değerler anlamlı bir farkın olduğunu göstermektedir (Draba, 1977; Swaminathan & Rogers, 1990). Bu çalışmada pozitif Z_{SAIrc} değerleri, puanlayıcının zamanla daha katılaştığını gösteriken negatif Z_{SAIrc} değeri puanlayıcının zamanla daha az katı hale geldiğini göstermektedir.

Etkileşim Terimi

İkinci değişen katılık/esneklik ölçüsü olarak Denklem 4'de verilen zaman kukla etkileşim modelinden elde edilen etkileşim indeksi (I_{rc}) kullanılmıştır (Linacre, M. kişisel iletişim, Haziran 2017; Wolfe ve diğerleri, 2007). Bu modelde zaman terimi kukla değişken olarak eklenmiştir.

$$\ln\left(\frac{\text{Pnritk}}{\text{Pnrit(k-1)}}\right) = \beta_n - \gamma_r - \delta_i - \pi_t - \tau_k \qquad \text{[4, Zaman Kukla Etkileşim Facet Modeli (Time Dummy Interaction Facet Model)]}$$

P_{nritk}, = puanlanan sunum *n*'nin 'i' ölçütünde gösterdiği performansın 'r' puanlayıcı tarafından 't' zamanında, 'k' kategorisinde puanlanma olasılığı

$P_{nrit(k-1)}$ = puanlanan sunum 'n'in 'i'ölçütünde gösterdiği performansın 'r' puanlayıcısı tarafından 't' zamanında, 'k-1' kategorisinde puanlanma olasılığı

β_n = Sunum n'nin niteliği

γ_r = Puanlayıcı r'nin katılığı

δ_i = Ölçüt i'nin güçlüğü

π_t= t zamanında gözlenen performans azalması

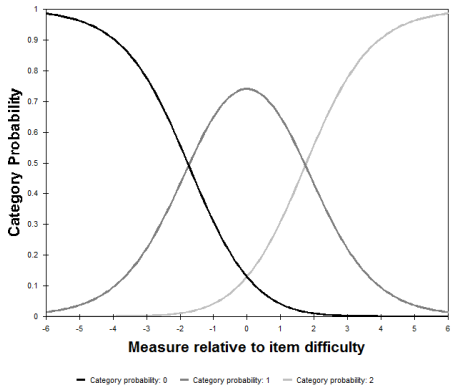
τ_k = Kategori 'k-1'den kategori k'ya geçiş güçlüğü

I_{rc}, standart hataya bölünerek Welch t-test'de kullanılan istatistik hesaplanmış ve bu istatistik I_{rc}'nin sıfırdan farklı olup olmaadığını test etmek için kullanılmıştır.

BULGULAR

Ön Analizler

Bu çalışmada kullanılacak verinin ÇYRÖM modeli ile uyumu ve ölçme aracının tek boyutlu olması beklenmektedir (Bond & Fox, 2015; Eckes, 2011). Model-veri uyumunu değerlendirmek için standartlaştırılmış artık değerler (standardized residuals) kullanılmıştır. Veri-model uyumunun kabul edilebilmesi için bu değerlerin en fazla %1'nin mutlak değerce üçe eşit ya da daha büyük olması ve gene en fazla %3'nün mutlak değerce ikiye eşit ya da ikiden büyük olması gerekmektedir (Linacre, 2002). Analizde kullanılan 1964 veriye ait 20 (%1,02) standartlaştırılmış artık değer |3|'ten büyükken 103 (%5.24) değer ise |2|'den büyüktür. Rasch Modelini kullanabilmek için diğer gereklilik ise kullanılan ölçme aracının tek boyutlu olmasıdır. Ölçekte bulunan maddelere ait uygunluk içi ve dışı indeksler en düşük 0,8 değerini alırken en yüksek 1,12 değerini almaktadır. Bu indeks değerleri ideal olarak 0,6 ile 1,4 aralığında beklenmektedir (Wright & Linacre, 1994). Maddelere ait uyum indeksleri tek boyutluluğa kanıt sağlarken standartlaştırılmış artık değerler ise model-veri uyumunun kabul edilebilir olduğunu göstermektedir. Verinin uygunluğuna karar verildikten sonra kontrol edilmesi gereken bir diğer nokta ise ölçme aracında kullanılan üç puanlı dereceli puanlama ölçeğinin beklendiği gibi kullanılıp kullanılmadığıdır. Şekil 1'de ölçekteki kategorilerin olasılık eğrilerini birbirlerinden net bir şekilde ayrıldığı, beklenen thresholdların beklenen sırada ve yönde arttığı görülmektedir. Threshold değerleri -1.74 ve 1.74'tür. Her ne kadar öğrenciler ikinci ve üçüncü kategoriyi kullanma eğitimde olsalar da ölçme aracında kullanılan üç puanlı dereceli puanlama ölçeğinin beklendiği gibi kullanıldığı sonucuna varabiliriz.

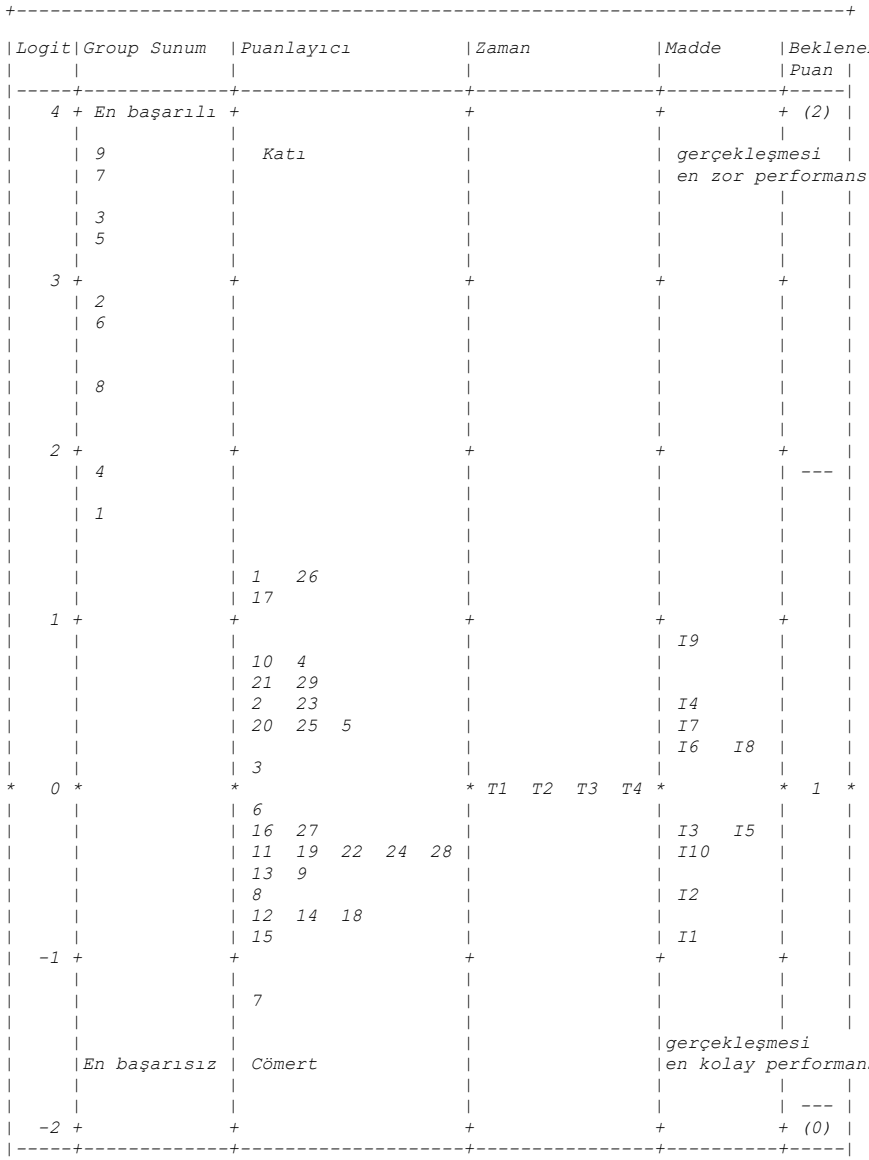


Şekil 1. Kategori Olasılık Eğrisi

Kukla Zaman FACET Modeli

Bu iki araştırmanın sorularına cevap aranmadan önce kukla zaman Facet modelinin sonuçları bu bölümde verilmiştir. Şekil 2, ÇYRÖM analizinde bulunan farklı yüzeylere ait parametre kestirimlerinin sonuçlarını görsel olarak sunmaktadır. Şekildeki birinci sütun bütün yüzeylere ait kestirimlerin gösterildiği ortak logit ölçüsüdür. İkinci sütunda 9 numaralı grubun sunumunun en başarılı, 1 numaralı grubun sunumunun en başarısız bulunduğu görülmektedir. Sunumların logit değerleri 1,56 ile 3,75 arasında değişiklik göstermektedir (daha fazla bilgi için Ek-Tablo 1). Ayırma

indeksi ve güvenirlik katsayısı sırasıyla 4,62 ve 0,96'dır. 'Sunumların öğrenciler tarafından belirlenen nitelikleri arasında bir fark yoktur' hipotezi χ^2 ile test edilmiş ve yokluk hipotezi reddedilmiştir [$\chi^2_{(8)}$=197.0, p<.000]. Tüm bu istatistikler, Şekil 2'de görüldüğü gibi sunumların performans niteliklerine göre anlamlı olarak birbirlerinden ayrıldığı gösterilmektedir.



Şekil 2. Kalibrasyon Haritası

Üçüncü sütunda puanlayıcı olarak akran puanlayıcılara ait değerler yer almaktadır. Akran puanlayıcılar katılık seviyelerine göre farklılık göstermektedir. En katı olan puanlayıcının logit değeri 1,30'ken en cömert puanlayıcının puanı -1,27'dir. Ayırma indeksi ve güvenirlik katsayısı sırasıyla 1,88 ve 0,78'dir. Yokluk hipotezi 'Akran puanlayıcıların katılık dereceleri arasında anlamlı bir fark yoktur' hipotezi χ^2 ile test edilmiş ve hipotez reddedilmiştir [$\chi^2_{(28)}$=160,9, p<.000]. Bu sonuçlar öğrencilerin sunumları değerlendirirken birbirlerinden farklı katılık/cömertlik davranışları sergilediğini göstermektedir. Bu yüzey için düşük ayırma indeksi veya güvenirlik katsayısı beklenirken yokluk hipotezinin de reddedilmemesi gerekmektedir. Geçerli bir puanlama için puanlayıcıların benzer davranışlar sergilenmesi beklenmektedir. Bir başka deyişle bir performansın kim tarafından puanlandığından bağımsız olarak aynı ölçütler göz önüne alındığında her puanlamada aynı değer alınmalıdır. Her bir puanlayıcıyı ayrı ayrı incelemek için de uygunluk indeksleri

kullanılmaktadır. Bu indeksler değerlendirilirken Linacre and Wright'ın (2002) önerdiği gibi 0,5-1,5 aralığı kullanılmıştır. Buna göre 29 puanlayıcı içinde sadece dokuz numaralı puanlayıcının 'uygunluk dışı' indeks değeri, istenilen sınırların biraz dışındadır (Ek-Tablo 2). Uygunluk içi indekslerin işaret ettiği probleme kıyasla uygunluk dışı indekslerin işaret ettiği problem, ölçme işlemi için daha az ciddi bir problemdir ve dolayısıyla başa çıkılması daha kolaydır (Linacre, 2002). Dokuz numaralı puanlayıcının uygunluk içi indeks değerinde bir problem olmadığından uygunluk dışı indeksinin kabul edilebilir aralığı biraz aşmasında çok önemli bir sakınca bulunmamıştır. Sunum ve puanlayıcı kestirimleri karşılaştırıldığında puanlayıcıların cömert davrandıkları görülmektedir.

Puanlamada kullanılan dereceli değerlendirme ölçeğin on maddeden oluşmaktadır. Madde 1, puanlayıcıların kolaylıkla yüksek puan verdikleri, bir başka deyişle sunum sırasında en kolay yerine getirilebilen maddedir. Bunun tam tersi Madde 9 sunumu yapmak gruplar için gerçekleştirilmesi en zor davranışı göstermektedir. Madde logit değerleri -0,84 ile 0,92 arasında değişmektedir. Kestirimde ortalama standart hata 0,17 (SS=0,02)'dir. Puanlamada kullanılan ölçeğin güvenirliğinin (0,90) iyi olduğu söylenebilir. Maddelerin uyum değerleri incelendiğinde hepsinin modele uygun davrandığı, çoğunun değerinin istenen değer olan '1' etrafında olduğu görülmektedir (Ek-Tablo 3). Madde ayırma indeksi 3.01, güvenirlik katsayısı 0,90'dır. Sonuç olarak elde edilen istatistikler, ölçeğin niteliğinin kabul edilebilir olduğunu ve ölçekteki her bir maddenin de beklendiği gibi davrandığını, puanlayıcıların katılık/cömertlik seviyelerinde farklılık olduğunu göstermektedir.

Puanlayıcı Kayması Analizi

Standartlaştırılmış Farklar, Z_{SAIrc}

Puanlayıcı katılığındaki kaymayı belirlemede kullanılan yaklaşım, makalenin yöntem kısmında anlatılmıştır. Bunlardan ilki olan Ayrı Model kullanılarak elde edilen puanlayıcı kestirimine ait betimsel istatistikler aşağıdaki tabloda verilmiştir.

Tablo 1: Ayrı Model Puanlayıcı Katılığı Kestirimleri

	Gün 1	Gün 2	Gün 3	Gün 4	Gün 4*
Ortalama	0,00	0,00	0,00	0,00	0,26
S.S.	0,89	0,99	1,22	1,31	1,44

*3 maximum puanlama yapan puanlayıcı dahil

Ayrı Model kullanılarak elde edilen puanlayıcı kestirimleriyle hesaplanan SAI_{rc} indeksine[2] ait istatistikler Tablo 2'de verilmiştir. SAI_{rc} indeksi yani temel alınan zamandan (Gün 1'den diğer bir güne) bir okuyucunun göreceli katılığındaki değişimi göstermektedir. Toplam 21 puanlayıcıya ait katılık seviyesinde Gün 1'den Gün 2'ye ortalama 0,048 puan düşme gözlenmiştir. Ortalamada puanlayıcıların daha cömert hale geldiği söylenebilir. Puanlayıcı katılık seviyesinde artış olan puanlayıcılarda (%38) gözlenen en yüksek artış 1,48 puanken ikinci günde daha cömert olan puanlayıcılar arasında en fazla kaymayı gösteren puanlayıcı 1,28 logit puan daha cömert olmuştur. Fakat kaymalardaki bu artış ve düşüşlerden hiçbiri istatistiksel olarak anlamlı değildir.

[2] Ayrı modele dayalı puanlayıcı kestirimleri ve standartlaştırılmış farklar Ek-Tablo 4' de verilmiştir.

Tablo 2: SAI_{rc} Katılık Seviyesi Karşılaştırması

	Gün 1 vs. 2	Gün 1 vs. 3	Gün 1 vs. 4
SAI_{rc} Ortalama	-0,048	0,150	0,339
SAI_{rc} S.S.	0,744	1,571	1,299
SAI_{rc} En düşük	-1,28	-2,99	-1,49
SAI_{rc} En yüksek	1,48	2,60	3,79
Daha çok katı olanların %	38,10	52,38	42,86
Z_{SAIrc} > |1,96| %	0	38,10	15,00

Gün 1'den Gün 3'e 21 puanlayıcının göreceli katılık seviyesinde ortalama 0,15 logit puan artış gözlenmiştir. Bunun sebebi, ortalamada puanlayıcıların daha katı puanlama eğilimde olmalarıdır. Göreceli katılık düzeyinde artış olan puanlayıcılarda (%52,38) gözlenen en yuksek artış 2,60 puanken üçüncü gün daha cömert olan puanlayıclar arasında göreceli katılık seviyesinde en fazla düşüş 2,99 logit puandır. Bu katılık düzeyindeki kaymalardan yaklaşık %38'i istatistiksel olarak anlamlıdır. Beş puanlayıcı daha katı puanlama yaparken üç puanlayıcı daha cömert puanlama yapmıştır. Son karşılaştırma da Gün 1 ve Gün 4 arasında yapılmıştır. Sonuçlar bir önceki karşılaştırmaya benzemekle birlikte burada puanlayıcı katılığındaki artış daha yüksektir. Bu katılık düzeyindeki kaymalardan %15'i istatistiksel olarak anlamlıdır. İki puanlayıcı daha katı puanlama yaparken bir puanlayıcı daha cömert puanlama yapmıştır. Her zaman dilimi için ve kaymayı gösteren puanlayıcı logit değerleri ve kaymayı gösteren Z_{SAIrc}değerleri ayrıntılı olarak Ek-Tablo 4'te verilmiştir.

Etkileşim Terimi, I_{rc}

Puanlayıcı kaymasını tespit etmek için kukla değişkenli zaman Facet modeli (Dummy Time-Facet Model) kullanılarak yapılan Rasch analizi, puanlayı ve gün yüzeyi arasında etkileşim olduğunu göstermektedir [Chi-square (99)=146.6, p<.00]. Bu analizden elde edilen etkileşim terimi (I_{rc}) indeksin istatistikleri Tablo 3'de verilmiştir[3]. Tabloda ilk sütun ilgili istatistiğin açıklamasını vermektedir. Son üç sütun ise ikinci, üçüncü ve dördündü gün için elde edilen kestirimlerin temel alınan birinci gün ile karşılaştırılmalarıyla elde edilen indekslerin istatistiklerini vermektedir. İkinci sütunda iki logit arasındaki farkı gösteren I_{rc} indeksinin ortalaması -0,058'dir. Bu durum puanlayıcıların gün 1'e kıyasla gün 2'de daha cömert olarak sunumları puanladıklarını gösterir. Bazı puanlayıcılar daha katı olurken diğerleri daha cömert puanlama davranışı göstermiştir. Üçüncü ve dördüncü sütundaki ortalama I_{rc} ise ilk güne kıyasla puanlayıcıların üçüncü ve dördüncü günde daha katı olma eğiliminde olduklarını göstermektedir. Özellikle son günde ortalama katılık seviyesindeki artış yaklaşık 0,3 logittir. Bu indeksin standart sapmasında yaşanan en düşük ve en yüksek değeri sırasıyla ikinci, üçüncü ve dördündü satırda verilmiştir.

Tablo3: Puanlayıcı-Zaman Arasındaki Etkileşime Ait Özet Değerler

	Gün 1 vs. 2	Gün 1 vs. 3	Gün 1 vs. 4
Ortalama I_{rc}	-0,058	0,126	0,289
SS I_{rc}	0,721	1,538	1,221
En düşük I_{rc}	-1,16	-2,88	-1,23
En yüksek I_{rc}	1,41	2,53	3,44
Anlamlı Welch t-test p<0,05 %	0	38,10	5,00
Etki büyüklüğü |Irc| > 0,50 %	66,7	85,7	65,0

[3] Etkileşim raporu Ek-Şekil 1'de verilmiştir.

Beşinci sıra, istatistiksel olarak sıfırdan farklı olan olan I_{rc} indeksinin yüzdesini vermektedir. Welch t-test'e göre ilk karşılaştırmada (ikinci sütün) katılık seviyesinde gözlenen bu kaymalar, istatistiksel olarak anlamlı değildir. Fakat diğer günlerdeki kaymaların küçük bir kısmı istatistiksel olarak anlamlıdır. İlk güne göre üçüncü gün kaymaların % 38,10'u, ve dördüncü gün kaymalarının % 5'i istatistiksel olarak anlamlıdır.

Tablonun en alt sırasında, Zaman 1'e göre her zaman döneminde anlamlı olabilecek kadar büyük katılık düzeyinde bir değişiklik gösteren puanlayıcıların yüzdesi verilmektedir. Tabloda görüldüğü üzere iki gün arasındaki kaymanın 0,50 logitten büyük olan değişimlerinin yüzdesi oldukça yüksektir. Bu durum, kaymaların büyüklüğünün istatistiksel olarak anlamlı olmasa da büyüklük olarak önemli olduğunu gösteriyor. Bu büyüklüklerin istatistiksel olarak anlamlı olamamasının nedeni, ölçmedeki hata payının yeteri kadar küçük olmamasıdır.

SONUÇLAR ve TARTIŞMA

Zaman içerisinde daha da popülerleşen akran değerlendirmesi yöntemi, hem biçimlendirmeye hem de düzey belirlemeye yönelik bir değerlendirme amacıyla kullanılabilen eğitsel bir araçtır. Buna rağmen diğer değerlendirme yöntemlerine göre daha az sayıda akademik çalışmaya konu olmuştur. Bir değerlendirme yöntemi olarak daha az dikkat çekmesinin sebebi, öğretmenlerin ya da öğretim elemanlarının öğrencilerin güvenilir ve nitelikli bir puanlayıcı/değerlendirici olduklarına inanmamaları olabilir. Alan yazınında akran değerlendirmesiyle elde edilen puanların geçerliliği ve güvenirliği üzerine sınırlı sayıda çalışma vardır ve var olan çalışmalarda genelde öğrenci-öğretmen puanlama ilişkisine bakılmıştır. Yüksek korelasyon değerleri, akran değerlendirmesiyle elde edilen puanların güvenirliği ve/veya geçerliği olarak kabul edilmiştir. Fakat tam olarak geçerliği belli olmayan öğretmen puanlamasını ölçüt olarak kabul edilen bir yöntemle geçerlik kanıtı sağlandığının iddia edilmesi soru işareti oluşturmaktadır.

Üniversite öğrencileriyle yapılan ve dört güne yayılan bu akran değerlendirme çalışmasında dereceli puanlama anahtarı kullanılarak puanlanan sunum performansları sürecinde genel olarak puanlayıcıların davranışları incelenmiş ve özel olarak puanlayıcı etkisi çeşitlerinden olan puanlayıcı katılığı kaymasının olup olmadığı ÇYRÖM analiziyle araştırılmıştır. Bu çalışmada ölçme sürecinin bir öğesi olarak akran puanlayıcı, puanlanan grup, puanlama için kullanılan dereceli puanlama anahtarı ayrı ayrı incelenmiştir. Analizler genel olarak akran puanlayıcıların arkadaşlarını oldukça cömert bir biçimde puanladıklarını göstermiştir. Puanlayıcılar kendi aralarında kıyaslandığında ise katılık/cömetlik seviyelerinin birbirlerinden farklı oldukları bulunmuştur. Puanlayıcılar katılık seviyelerine göre ± 1 logit ranjında dağılım göstermektedirler. Puanlayıcılar dereceli puanlama anahtarını kullanarak 9 sunumu niteliklerine göre birbirlerinden ayırt edebilmektedirler. Puanlayıcıların uyum indekslerine bakıldığında çok az sayıda puanlayıcının modele göre beklenenden farklı davranış sergilediği görülmektedir. Uyum istatistiklerine bakarak sunumların puanlayıcılar tarafından tutarlı bir şekilde niteliklerine göre sıralandıkları sonucuna varılabilir. Güvenirlik katsayısı 0,90 olması ve uyum indekslerinin istenen aralıkta olması, kullanılan dereceli puanlama anahtarının niteliğinin iyi olduğunu göstermektedir. Puanlama anahtarındaki madde kestirimlerine bakıldığında öğrenci performanslarında yerine getirilebilen üç ölçüt, zorluk sırasına göre şu şekilde sıralanmıştır; Madde 9 "sunumu yapan kişi göz temasını ve yüz ifadesini iyi kullandı," Madde 4 "Sunum ilgi çekiciydi" ve Madde 7 "Sunum akıcıydı". Bir başka deyişle, akran puanlayıcıların sunumlarda en zayıf buldukları ölçütler bunlardır. Sunumların en başarılı bulundukları ölçütler ise sırasıyla Madde 1 "Sunum detaylıydı," Madde 2 "Sunum sonunda konuyu anladım," Madde 10 "Slaytlar kalabalık değildi, takibi kolaydı".

Bu çalışmada puanlamanın ilk günü temel alınarak daha sonraki 3 puanlama gününde ilk güne göre puanlayıcı katılık/cömertlik seviyesinde bir değişik olup olmadığı incelenmiştir. Puanlayıcı katılığı kayması iki farklı yöntemle araştırılmıştır: standartlaştırılmış farklar (Z_{SAIrc}) ve etkileşim terimi (I_{rc}). İki yöntemle de oldukça benzer sonuçlara ulaşılmıştır. Gün-1 ve 2 arasında puanlayıcıların kestirimlerinde bir farklılık görülmemektedir. Her ne kadar ortalamada puanlayıcılar daha cömert puanlama yapsa da kaymalar istatistiksel olarak anlamlı değildir. Fakat kaymalardaki logit cinsindeki farklara bakıldığında etki büyüklüğü olarak kaymaların %60'tan fazlasında görülmesi beklenen

minumum farktan daha büyük kayma gösterdiği tespit edilmiştir. Bu kadar kayma, daha fazla puanlamanın yapıldığı ve önemli kararların alındığı sınav sonuçlarında önemli sorunlara yol açabilir. Gün-1 ve 3 arasında puanlayıcıların kestirimlerinde önemli kaymaların olduğu puanlayıcıların oranı %38,10'dur. İki yönteme göre puanlayıcılar ortalamada yaklaşık 0,14 logit kayma gösterip daha katı puanlama davranışı sergilemişlerdir. Puanlayıcılara ayrı ayrı bakıldığında beş puanlayıcının daha katı olduğu, üç puanlayıcının ise daha cömert olduğu görülmektedir. Gün-1 ve 4 arasında puanlayıcıların kestirimlerinde önemli kaymaların olduğu puanlayıcıların sayısı standartlaştırılmış farklar yöntemiyle üçgen, etkileşim terimi yöntemiyle birdir. Ortalamada iki yöntemle de puanlayıcılar daha katılaşmıştır. Ortalamada kaymanın en yüksek olduğu Gün-4'dür. Gün-3 ise kayma gösteren en yüksek sayıda puanlayıcının olduğu gündür.

Alan yazınında bu çalışmanın kapsamında yapılmış çalışma az olduğu için sonuçları karşılaştırabileceğimiz bulgular çok azdır. Giriş bölümünde bahsedilen meta analiz çalışmaları, çalışmaların büyük çoğunluğunda akran değerlendirmesinden elde edilen puanların geçerli ve güvenilir olduğunu raporlamaktadırlar. Bilgimiz dahilinde Casabianca, Lockwood ve McCaffrey'nin çalışması (2015) akran değerlendirmede puanlayıcı etkisini incelemiş tek çalışmadır ve bu çalışmanın sonuçlarıyla paralel olarak puanlayıcılar arasında katılık/cömertlik seviyeleri bakımından önemli farklılıklar görülmüştür. Bu çalışmada Casabianca ve diğerlerine ek olarak bazı puanlayıcıların birinci günden sonra katılık seviyelerini sınıf arkadaşlarına kıyasla sabit tutamadıkları görülmüştür. Bazıları daha katı davranırken bazıları da daha cömert davranmışlardır. Ortalamaya bakılarak sonuca varıldığında sonraki günlerde sunum yapan gruplar dezavantajlı olmuşlardır. Bu nedenle literatürde bahsedildiği gibi (ör. Congdon & McQueen, 2000 ve McKinley & Boulet, 2004) puanlama sürecinde puanlayıcı eğitimleri, puanlayıcı kaymasını tam olarak ortadan kaldırmasa da gereklidir.

KAYNAKÇA

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences*. Mahwah, NJ: Lawrence Erlbaum.

Eckes, T. (2011). *Introduction to many-facet Rasch measurement. Analyzing and evaluating rater-mediated assessments*. Frankfurt am Main: Peter Lang.

Braun, H. I. (1988). Understanding scoring reliability: Experiments in calibrating essay readers. *Journal of Educational Statistics*, *13*(1), 1–18.

Braun, H. I., & Wainer, H. (1989). Making essay test scores fairer with statistics. In J. Tanur, F. Mosteller, W. H. Kruskal, E. L. Lehmann, R. F. Link, R. S. Pieters & G. S. Rising (Eds.), *Statistics: A guide to the unknown* (3rd ed., pp. 178–188). Pacific Grove, CA: Wadsworth.

Casabianca, J. M., Lockwood, J. R., & McCaffrey, D. F. (2015). Trends in classroom observation scores. *Educational and Psychological Measurement, 75*(2), 311–337. doi: 10.1177/0013164414539163

Congdon, P. J., & McQueen, J. (2000). The stability of rater severity in large-scale assessment programs. *Journal of Educational Measurement, 37*(2), 163-178.

Demirbilek, M. (2015). Social media and peer feedback: What do students really think about using Wiki and Facebook as platforms for peer feedback? *Active Learning in Higher Education, 16*(3) 211–224. doi: 10.1177/1469787415589530

Dochy, F., Segers, M., & Sluijsmans, D. (1999). The use of self-, peer and coassessment in higher education: A review. *Studies in Higher Education, 24*(3), 331-350.

Engelhard, G. (1994). Examining rater errors in the assessment of written composition with a many-faceted Rasch model. *Journal of Educational Measurement 31*(2), 93-112.

Engelhard, G. (1996). Evaluating rater accuracy in performance assessments. *Journal of Educational Measurement, 33*(1), 56–70.

Engelhard, G., & Myford, C. M. (2003). *Monitoring faculty consultant performance in the Advanced Placement English Literature and Composition Program with a many-faceted Rasch model* (Research Rep. 03-01). Princeton, NJ: Educational Testing Service

Falchikov, N. (1995) Peer feedback marking: Developing peer assessment. *Innovations in Education and Training International, 32*(2), 175-187.

Falchikov, N. (2001). *Learning together: Peer tutoring in higher education*. London: Routledge Falmer.

Falchikov, N., & Goldfinch, J. (2000). Student peer assessment in higher education: A meta-analysis comparing peer and teacher marks. *Review of Educational Research, 70*(3), 287–322.

Gabrielson, S., Gordon, B., & Engelhard, G. (1995). The effects of task choice on the quality of writing obtained in a statewide assessment. *Applied Measurement in Education 8*(4), 273-290.

Heyman J. E., & Sailors J. J. (2011). Peer assessment of class participation: Applying peer nomination to overcome rating inflation. *Assessment & Evaluation in Higher Education*, *36*(5), 605-618. doi: 10.1080/02602931003632365

Hafner, J. C., & Hafner, P. M. (2003). Quantitative analysis of the rubric as an assessment tool: An empirical study of student peer-group rating. *International Journal of Science Education, 25*(12), 1509–1528. doi: 10.1080/0950069022000038268

Harik, P., Clauser, B. E., Grabovsky, I., Nungester, R. J., Swanson, D., & Nandakumar, R. (2009). An examination of rater drift within a generalizability theory framework. *Journal of Educational Measurement, 46*(1), 43-58. doi: 10.1111/j.1745-3984.2009.01068.x

Hoskens, M., & Wilson, M. (2001). Real-time feedback on rater drift in constructed response items: An example from the Golden State Examination. *Journal of Educational Measurement, 38*(2), 121–146.

Kane, J. S., & Lawler, E. E (1978). Methods of peer assessment. *Psychological Bulletin*, *85*, 555-586.

Linacre, J. M. (2002). What do infit and outfit, mean-square and standardized mean?. *Rasch Measurement Transaction*, *16*(2), 878.

Love, K. G. (1981). Comparison of peer assessment methods: Reliability, validity, friendship bias, and user reaction. *Journal of Applied Psychology*, *66*(4),451-457.

Lumley, T., & McNamara, T. E (1995). Rater characteristics and rater bias: Implications for training. *Language Testing, 12*(1), 54-71.

Lunz, M. E., & Stahl, J. A. (1990). Judge consistency and severity across grading periods. *Evaluation and the Health Professions*, *13*, 425-444.

McLaughlin, K., Ainslie, M., Coderre, S., Wright, B., & Violato, C. (2009). The effect of differential rater function over time (DRIFT) on objective structured clinical examination ratings. Medical Education, 43, 989–992. doi:10.1111/j.1365-2923.2009.03438.x

McNamara, T. F. (1996). *Measuring second language performance*. Harlow, UK: Addison Wesley Longman Limited.

McQueen, J., & Congdon, P. J. (April, 1997). *Rater severity in large-scale assessment: Is it invariant?* Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Messick, S. (1994). Alternatıve modes of assessment, uniform standards of validity. ETS Research Report Series, 2,1-22.

Myford, C. M. (1991). *Judging acting ability: The transition from notice to expert*. Paper presented at the American Educational Research Association, Chicago IL.

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

Myford, C. M., & Wolfe, E. M. (2004). Detecting and measuring rater effects using many-facet Rasch measurement: Part II. *Journal of Applied Measurement, 5*(2), 189-227.

Myford, C. M., & Wolfe, E. M. (2009). Monitoring rater performance over time: A framework for detecting differential accuracy and differential scale use. *Journal of Educational Measurement, 46*(4), 371-389. doi: 10.1111/j.1745-3984.2009.00088.x

Park, Y. S. (2011). *Rater drift in constructed response scoring via latent class signal detection theory and item response theory* (Unpublished doctoral dissertation). University of Columbia, NY.

Rowan, B., Harrison, D. M., & Hayes, A. (2004). Using instructional logs to study elementary school mathematics: A close look at curriculum and teaching in the early grades. *Elementary School Journal, 105*, 103-127.

Sadler, P. M., & Good, E. (2006). The impact of self and peer-grading on student learning. *Educational Assessment, 11*, 1–31. doi:10.4103/2229-516X.186961

Scruggs, T. E., & Mastropieri, M. A. (1998). Tutoring and students with special needs. In K. J. Topping & S. Ehly (Eds.), *Peer-assisted learning* (pp. 165–182). Mahwah, NJ: Lawrence Erlbaum Associates.

Scullen, S. E., Mount, M. K., & Goff, M. (2000). Understanding the latent structure of job performance ratings. *Journal of Applied Psychology*, *85*(6), 956-970.

Somervell, H. (1993). Issues in assessment, enterprise and higher education: The case for self-, peer and collaborative assessment. *Assessment & Evaluation in Higher Education*, *18*(3), 221-233.

Topping, K. J. (1998). Peer assessment between students in college and university. *Review of Educational Research, 68*(3), 249–276.

Topping, K. (2003). *Self and peer assessment in school and university: Reliability, validity and utility. in optimising new modes of assessment: In search of qualities and standards*. M. S. Segers, Dochy,R., and E. C. Cascallar (Ed.). Netherlands.

Topping, K. J. (2005). Trends in peer learning. *Educational Psychology*, *25*, 631–645. doi: 10.1080/01443410500345172

Topping, K. J. (2009). Peer assessment. *Theory Into Practice*, *48*(1), 20-27. doi: 10.1080/00405840802577569

Topping, K. J., & Ehly, S. (Eds.). (1998). *Peer assisted learning*. Mahwah, NJ: Lawrence Erlbaum Associates

Tseng, S. C., & Tsai, C.-C. (2007). On-line peer assessment and the role of the peer feedback: A study of high school computer course. *Computers & Education, 49*(4), 1161-1174. https://doi.org/10.1016/j.compedu.2006.01.007

Weaver II, R., & Cotrell, H. W. (1986). Peer evaluation: A case study. *Innovative Higher Education, 11*(1), 25-39.

Wilson, M., & Case, H. (2000). An examination of variation in rater severity over time: A study of rater drift. *Objective measurement: Theory into practice, 5*, 113-134.

Wolfe, E. W., Moulder, B. C., & Myford, C. M. (2001). Detecting differential rater functioning over time (DRIFT) using a Rasch multi-faceted rating scale model. *Journal of Applied Measurement, 2*(3), 256–80.

Wolfe, E. W., Myford, C. M., Engelhard Jr. G., & Manalo, J. R. (2007). *Monitoring reader performance and DRIFT in the AP® English literature and composition examination using benchmark essays* (Research Report No. 2007-2). Retrieved from https://research.collegeboard.org/publications/content/2012/05/monitoring-reader-performance-and-drift-ap-english-literature-and

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

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

Yang, R. (2010). *A many-facet Rasch analysis of rater effects on an oral English Proficiency Test* (Unpublished doctoral dissertation).Purdue University. IN.

Yang, Y. & Tsai, C. C. (2010). Conceptions of and approaches to learning through online peer assessment. *Learning and Instruction, 20*(1), 72-83. doi: 10.1016/j.learninstruc.2009.01.003

EXTENDED ABSTRACT

Introduction

As in each measurement process, the validity of the scores obtained from the peer assessment needs to be examined. Reliability and validity in the peer assessment literature are generally regarded as the agreement between students' scores and experts' score such as teachers score, not as agreement among peer raters or the consistency of the same peer raters across different time points. Such studies are based on the assumption that teacher scoring are highly reliable and valid. Since this is a suspicious assumption in some contexts, it should be debated whether these studies provide evidence for reliability or validity, or both (Topping, 2003; Topping, 2009). In these studies, the terms accuracy, validity and reliability are often used interchangeably.

Research on the reliability and validity of peer assessment is often included in higher education studies (Falchikov, 2001, Topping, 2003). The majority of the studies have reported high validity (Sadler & Good, 2006); some studies have reported different results (Falchikov & Goldfinch, 2000; Topping, 1998). The results of the studies conducted in the school settings have provided similar results to the studies conducted in the higher education (Toppings, 2003).

In this study, the behavior of peer raters in scoring presentation performance by using the rating scale during four different days will be examined and specifically the following two main research questions will be sought.

1. Are there any rater whose relative severity level drift during four days? In other words, are there any peer rater exhibiting differential severity?

2. Do the different approaches used to detect drift provide similar results?

Method

Participants

The data of this study were collected at the measurement and evaluation course offered in the college of education at the medium sized university in the spring semester of 2017. At the end of the presentation, the presentation was assessed by their peers. Twenty nine students attending the course participated in this study; however, on the day of each assessment, there was no full participation in the classroom. This has led to some data loss.

The total of nine presentations were held on four separate days, two presentations on each of the first three days and three presentations on the fourth day. The number of peer raters in each presentation scoring ranged from 19 to 25. Each presentation performance was rated by each peer, except for the presenter group and the peers who did not attend the class on the particular day. However, since presentation performances are nested within days, there is no link between days. The first two presentations are nested within the first day, and the next two presentations are nested within the second day and so on. Each presentation was rated on the day the presentation had been made. For this reason, time (day) facet can be added the Many-Facet Rasch Measurement model as a dummy variable.

Data Collection Tool

The researcher of this study developed a scoring key for peer raters. This scoring key was developed to be used in the same course and had been finalized with the contribution of the students who were taking the same course one semester earlier. There were a total of 10 criteria (items) in the key, and each criterion was scored on a three-point rating scale. Sample items were: I understand the topic as a result; Elaborated upon ideas by giving examples/reasons, explanations; Presentation was engaging.

Measurement Model

A many-Facet Rasch Measurement (MFRM) (Linacre, 1989) model was used for analyses and they were performed by FACET v3.71.4 (1987-2014) software. In order to examine rater drift, two indexes were calculated using MFRM - *standardized differences* and *interaction term*.

Three Facet Rasch model (facets are presenter group, rater and the rating scale) was used to estimate rater severity for each day separately. Average of rater estimates was set to 0.00 logit. This arrangement allows us to follow the drift in the relative severity/leniency levels of peer rater at different times. Rater drift was calculated between baseline day (Day-1) and other days (Day-2, 3 and 4). The difference (SAI) between two days relative severity estimates were standardized (Z_{SAIrc}). Z values were used to test the null hypothesis that there is no difference between the level of severity at baseline day and the level of severity at time c.

The second index for rater drift, the interaction index (I_{rc}) obtained from the time dummy model was used (Linacre, M. personal communication, June 2017, Wolfe et al., 2007). In this Rasch model time was added as a dummy variable. Interaction between time and rater facet were examined

Results and Discussion

The estimates of the presentations vary between 1.56 and 3.75 logit (See Annex for more information). The separation index and reliability coefficient are 4.62 and .96 respectively [$\chi^2(8)=197.0$, $p <.000$]. All these statistics show that quality of the presentations are significantly different from each other. Peer raters vary according to their level of severity. The logit value of the most severe rater is 1.30, while the logit estimate of the most lenient rater is -1.27. The separation index and reliability coefficient are 1.88 and .78 respectively [χ^2 (28) = 160.9, $p <.0001$]. These results show that peers exhibit different severity behaviors when evaluating presentations. The rating scale used in scoring has nine items. It can be said that the reliability of the scale used in the scoring

(0.90) is good. It is seen that the majority of the fit statistics of items are around the desired value '1'. The item separation index is 3.01, and the reliability coefficient is 0.90. As a result, statistics show that the quality of the scale is acceptable and its each item behaves as expected, and that the raters are a different in their severity/lenience.

Relative rater severity drift from Day-1 to 4 was examined. Standardized differences and interaction term provided very similar results. Between Day-1 and 2, there is no statistically significant difference in the estimates of the rater severity. Between Day-1 and 3, the percentage of scorers with significant drift in the estimates of the raters is 38.10%. According to the two methods, the raters show an average of about 0.14 logit shift and display a more severe scoring behavior. While five raters are more severe, three raters are more lenient. Between Day-1 and 4, the number of raters who had significant shifts in their estimates is three according to the standardized difference method, while one according to interaction method. In the average, the raters became more severe. Among three comparison, Day-4 has the largest rater severity drift on the average; Day-3, however, has the highest number of raters with rater severity drift.

Ekler

Tablo 1: Sunum Performans İstatistikleri

Presentation	FairMAvge	Measure	S.E.	InfitMS	InfitZ	OutfitMS	OutfitZ
1	1.43	1.56	0.14	1.05	0.62	1.04	0.54
2	1.76	2.92	0.16	1.13	1.36	1.08	0.64
3	1.83	3.38	0.17	0.94	-0.50	0.75	-1.71
4	1.50	1.82	0.13	0.95	-0.72	0.97	-0.32
5	1.82	3.26	0.17	1.08	0.80	1.19	1.37
6	1.72	2.74	0.14	1.09	1.04	1.10	0.94
7	1.87	3.67	0.20	0.96	-0.23	0.80	-1.01
8	1.65	2.40	0.16	0.93	-0.70	0.94	-0.49
9	1.88	3.75	0.20	0.89	-0.75	0.74	-1.30
		2.83	.16	1.00	.1	.96	-.1
		.78	.02	.09	.8	.16	1.1
RMSE:0.17	S.D.: 0.76	Ayırma İndeksi:4.62		Güvenirlik:0.96			
χ^2 (8): 197.0	*p*: 0.00						

Tablo 2: Akran Puanlayıcı İstatistikleri

Raters	Fair MAvge	Measure	S.E.	Infit MS	Infit Z	Outfit MS	Outfit Z
9	1.84	-0.56	0.48	1.43	1.35	***1.80***	1.63
6	1.76	-0.09	0.26	1.27	1.62	1.49	2.03
26	1.43	1.26	0.28	1.42	2.25	1.42	2.25
20	1.67	0.35	0.25	1.34	2.15	1.41	2.14
23	1.64	0.46	0.23	1.32	2.27	1.40	2.40
4	1.58	0.71	0.25	1.28	1.89	1.35	2.04
19	1.80	-0.31	0.28	1.04	0.31	1.19	0.81
21	1.61	0.58	0.24	1.04	0.34	1.12	0.80
17	1.46	1.14	0.22	1.08	0.64	1.06	0.49
22	1.80	-0.34	0.24	1.00	0.04	1.05	0.30
2	1.62	0.54	0.26	0.94	-0.37	1.02	0.19
18	1.86	-0.72	0.30	1.05	0.32	0.99	0.05
3	1.73	0.08	0.32	0.94	-0.28	0.94	-0.21
10	1.58	0.71	0.30	0.93	-0.37	0.93	-0.30
16	1.80	-0.30	0.44	0.84	-0.54	0.87	-0.21
25	1.66	0.38	0.23	0.88	-0.90	0.84	-1.00
29	1.61	0.58	0.33	0.88	-0.59	0.82	-0.83
1	1.42	1.30	0.24	0.83	-1.25	0.81	-1.41
5	1.66	0.41	0.35	0.79	-0.97	0.81	-0.71
8	1.84	-0.58	0.35	0.97	-0.05	0.80	-0.54

Tablo 2: Akran Puanlayıcı İstatistikleri – devam ediyor

Raters	Fair MAvge	Measure	S.E.	Infit MS	Infit Z	Outfit MS	Outfit Z
13	1.83	-0.48	0.28	0.88	-0.66	0.76	-0.96
28	1.81	-0.37	0.28	0.87	-0.74	0.73	-1.15
24	1.81	-0.37	0.28	0.81	-1.13	0.71	-1.25
27	1.79	-0.25	0.57	0.83	-0.38	0.70	-0.62
11	1.81	-0.37	0.28	0.78	-1.34	0.66	-1.47
12	1.87	-0.80	0.31	0.86	-0.64	0.61	-1.35
14	1.87	-0.80	0.31	0.81	-0.91	0.58	-1.49
7	1.91	-1.27	0.41	0.84	-0.44	0.54	-1.05
15	1.88	-0.89	0.32	0.77	-1.12	0.50	-1.77
Ortalama	1.70	0.00	0.31	0.99	0.00	0.95	0.00
SS	0.13	0.68	0.08	0.20	1.10	0.32	1.30
RMSE:0,32	S.D.: 0,61	Ayırma İndeksi:1,92		Güvenirlik:0,76			
χ^2 (28): 160,9	*p*: 0.00						

Tablo 3: Madde İstatistikleri

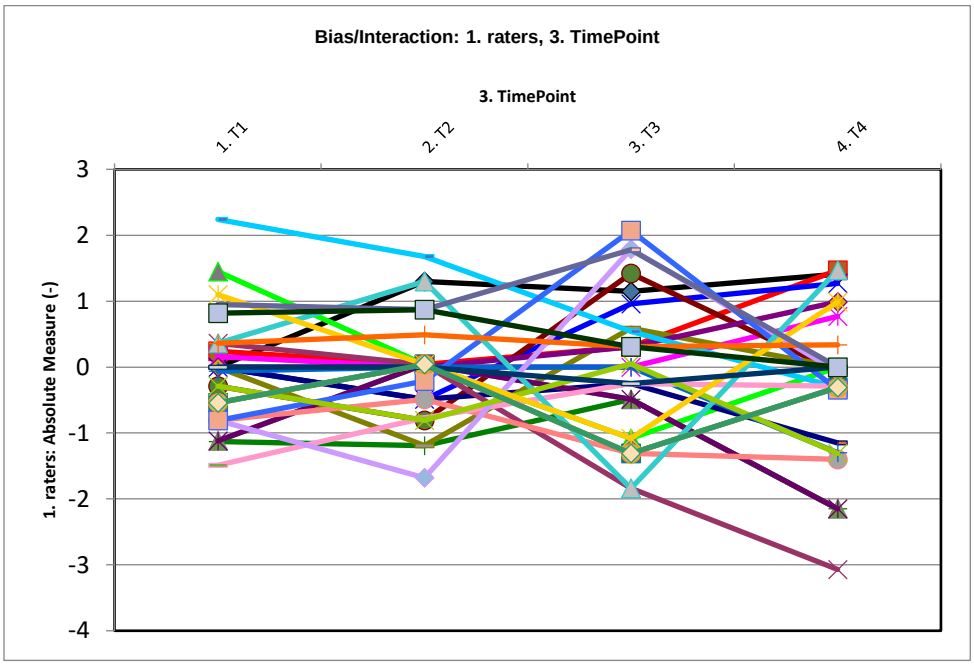
Maddeler	FairMAvge	Measure	S.E.	InfitMS	InfitZ	OutfitMS	OutfitZ
I1	1.87	-0.84	0.2	1.2	1.39	1.05	0.31
I2	1.84	-0.59	0.19	1.1	0.82	1	0.05
I3	1.78	-0.19	0.17	1.05	0.48	1.06	0.45
I4	1.62	0.54	0.15	1.04	0.44	1.02	0.22
I5	1.78	-0.21	0.17	1.06	0.62	0.96	-0.19
I6	1.7	0.19	0.16	1.12	1.29	1.02	0.24
I7	1.67	0.33	0.16	0.91	-1	0.87	-1.13
I8	1.69	0.25	0.16	0.95	-0.56	0.87	-1.13
I9	1.53	0.92	0.15	0.8	-2.47	0.78	-2.57
I10	1.81	-0.4	0.18	1.04	0.4	0.92	-0.42
		0.00	0.17	1.03	0.1	0.96	-0.40
		0.54	0.02	0.11	1.2	0.09	0.90
RMSE:0.17	S.D.: 0.51	Ayırma İndeksi:3,01		Güvenirlik:0.90			
χ^2 (9): 91,6	*p*: 0.00						

Tablo 4: Ayrı Modele Dayalı Puanlayıcı Kestirimleri ve Standartlaştırılmış Farklar

Puanlayıcı	Gün 1		Gün 2		Gün 3		Gün 4		Z_SAIrc		
	Ölçüm	S.H.	Ölçüm	S.H.	Ölçüm	S.H.	Ölçüm	S.H.	Gün 2-1	Gün 3-1	Gün 4-1
1			-1.7	0.51	-1.3	0.45	-1.75	0.41			
2	-0.37	0.69	-0.07	0.54	-0.41	0.51	-1.86	0.49	0.34	-0.05	-1.76
3	-1.55	0.59	-0.07	0.54	1.05	0.76			1.85	***2.70***	
4			0.56	0.58	-1.1	0.46	-1.58	0.42			
5	-0.22	0.46					-1.09	0.55			-1.21
6	0.22	0.48	0.92	0.62	-1.81	0.63	0	0.51	0.89	***-2.56***	-0.31
7	0.99	1.06	1.35	0.69	0.61	0.81	2.25	1.04	0.28	-0.28	0.85
8			0.56	0.58	0.18	0.58	1.26	0.79			
9			1.35	0.69	-0.95	0.71					
10					-0.41	0.51	-1.23	0.42			
11	0.46	0.51	-0.07	0.54	1.45	1.07	0.28	0.54	-0.71	0.84	-0.24
12	1.05	0.58	-0.07	0.54	0.61	0.81	2.25	1.04	-1.41	-0.44	1.01
13	-0.42	0.45	-0.07	0.54	1.83	1.04	3.37	1.84	0.50	***1.99***	***2.00***
14	1.05	0.58	-0.07	0.54	0.61	0.81	2.25	1.04	-1.41	-0.44	1.01
15	0.74	0.54	0.56	0.58	1.45	1.07	1.46	0.76	-0.23	0.59	0.77
16	-0.01	0.47					1.5	1.88			0.78
17	-2.21	0.41	-2.24	0.53	-0.65	0.49	0.35	0.61	-0.04	***2.44***	***3.48***
18	1.42	0.64	0.87	0.63	0.18	0.58	0.3	0.62	-0.61	-1.44	-1.26
19	0.74	0.54	1.89	0.8	-2.19	0.61	0	0.51	1.19	***-3.60***	-1.00
20	0.74	0.54	0.23	0.56	-2.25	0.43	0.1	0.77	-0.66	***-4.33***	-0.68
21	-0.42	0.45	-1.7	0.51	1.83	1.04	-1.86	0.49	-1.88	***1.99***	***-2.16***
22	0.22	0.34	0.92	0.62	-0.07	0.43	1.33	1.88	0.99	-0.53	0.58
23	-1.1	0.44	-0.07	0.54	1.05	0.76	-1.23	0.42	1.48	***2.45***	-0.21
24	0.46	0.51	-0.07	0.54	1.45	1.07	0.28	0.54	-0.71	0.84	-0.24
25	-0.42	0.45	-0.63	0.52	-0.41	0.51	-0.46	0.46	-0.31	0.01	-0.06
26	-0.99	0.43	-1.14	0.76	-1.97	0.46			-0.17	-1.56	
27					0.18	0.58					
28	0.46	0.51	-0.07	0.54	1.45	1.07	0.28	0.54	-0.71	0.84	-0.24
29	-0.81	0.59	-1.14	0.76	-0.41	0.51			-0.34	0.51	

Target Num	Target ra	Target Measr	S.E.	Obs-Exp Average	N	Context TimeP	Target Measr	S.E.	Obs-Exp Average	N	Context TimeP	Target Contrast	Joint S.E.	Welch t	d.f.	Prob.
100120	17	2.24	.43	-.30	1	T1	-.31	.58	.30	4	T4	2.55	.72	3.54	37	.0011
100120	17	2.24	.43	-.30	1	T1	.54	.47	.15	3	T3	1.70	.64	2.67	37	.0113
12276	3	1.45	.60	-.33	1	T1	-1.08	.75	.16	3	T3	2.53	.97	2.62	26	.0146
100165	23	1.10	.44	-.17	1	T1	-1.08	.75	.23	3	T3	2.18	.87	2.49	35	.0176
100084	13	.36	.45	-.19	1	T1	-1.84	1.03	.12	3	T3	2.20	1.13	1.95	32	.0596
100156	21	.36	.45	.06	1	T1	-1.84	1.03	.31	3	T3	2.20	1.13	1.95	32	.0596
100018	6	-.29	.49	.04	1	T1	1.42	.62	-.32	3	T3	-1.71	.78	-2.18	22	.0401
100135	19	-.81	.54	.09	1	T1	1.79	.60	-.45	3	T3	-2.60	.81	-3.22	23	.0038
100150	20	-.81	.54	.25	1	T1	2.07	.43	-.44	3	T3	-2.88	.69	-4.20	37	.0002

Şekil 1. Etkileşim Raporu (Interaction Pairwise Report)



Şekil 2. Yanlılık/Etkileşim: Puanlayıcı vs. Zaman