Hacettepe Üniversitesi Eğitim Fakültesi Dergisi (H. U. Journal of Education) 28(2), 308-320 [2013]

Dereceli Puanlama Anahtarıyla, Genel İzlenimle ve İkili Karşılaştırmalar Yöntemiyle Yapılan Değerlendirmelerin Karşılaştırılması *

Comparison of the Evaluations Which Were Done With Rubric, Overall Impression and Paired Comparisons

Seçil ÖMÜR**, Adnan ERKUŞ***

ÖZ: Bu çalışmada, üç farklı şekilde (genel izlenimle, dereceli puanlama anahtarıyla ve Thurstone ikili karşılaştırmalar yöntemiyle) puanlanan ve değerlendirilen kompozisyonlardan elde edilen verilere dayanarak, üç farklı değerlendirmenin karşılaştırılması amaçlanmıştır. Bu amaçla "Kompozisyon Yazmaya İlişkin Dereceli Puanlama Anahtarı" geliştirilmiştir. Dereceleme ölçeğinin davranış ölçütlerinin nasıl bir yapı sergilediğini görebilmek için faktör analizi yapılmış ve hem tüm ölçeğin hem de üç ana bileşenin tek faktörlü olduğu ve Cronbach α katsayılarının oldukça yüksek çıktığı gözlenmiştir. Seçilen iki kompozisyon için dereceli puanlama anahtarı ile yapılan değerlendirmelerin güvenirliklerine test tekrar test yöntemi ile bakılmıştır. Yargıcıların kompozisyonları değerlendirmeleri sonucunda, hangi değerlendirme yapılırsa yapılsın kompozisyonların sıralamalarında bir değişikliğin olmadığı gözlenmiştir. Farklı yargıcıların beş ayrı kompozisyonu, genel izlenimle ve dereceli puanlama anahtarı ile puanlanmasında puanlayıcı güvenirliğinin belirlenmesinde G katsayıları hesaplanmış ve bu katsayılar yüksek bulunmuştur.

Anahtar sözcükler: derecelendirme ölçekleri, dereceli puanlama anahtarı, ikili karşılaştırmalar yöntemi.

ABSTRACT: This study investigates comparison of three different evaluation types based on the data of the compositions which have been going to be rated and evaluated in three different types (overall impression, rubric and Thurstone's method of paired comparison). Firstly, "The Rubric For Writing Composition" was constructed which was going to be used as a tool for evaluating writing skill. It is observed that, one factor explatory solution which is done to group of behavior criteria of rating scale constructed as rubric is accurate for both scale and its three components. Additionally, Cronbach alpha coefficients calculated for major and minor criteria of all composition is observed to be high. The relationship among rates of compositions which are evaluated with the rubric by the same judges found positive and significant. It is observed that there is no change in rank order of compositions for three different evaluation types. For determining rating reliability, G coefficients were calculated for overall impression and rubric evaluation and these coefficients are found as high.

Keywords: rating scales, rubric, paired comparison method.

1. GİRİŞ

Günümüzde, dünyanın birçok ülkesinde olduğu gibi, ülkemizde de toplumsal, kültürel, ekonomik, siyasal, teknolojik vb. alanlarda hızlı değişimler olmaktadır. Bu değişim sürecinde eğitim alanında; öğrencinin "yaparak öğrenmesini" sağlama temeline dayanan bir eğitim felsefesi ve yönteminin uygulanmasına geçilmiştir. Ülkemizde yeni eğitim programının yaşama geçirilmesiyle birlikte, gerek eğitim sürecinin uygulamalarında, gerekse bu süreçte ve süreç sonunda öğrenci kazanımlarının ölçümünde hala bir karmaşa yaşanmaya devam etmektedir. Eğitim uygulamasında başvurulan yöntem ve tekniklerin anlaşılırlığını ve uygulamadaki sorunları bir yana bırakacak olursak, en çok sıkıntının ölçme ve değerlendirme alanında

* Bu çalışma Mersin Üniversitesi Sosyal Bilimler Enstitüsü Eğitimde Ölçme ve Değerlendirme Anabilim Dalı'nda Yüksek Lisans Tezi olarak hazırlanmıştır.

** Arş. Gör., Mersin Üniversitesi, Eğitim Fakültesi, Mersin-Türkiye, secilomur@gmail.com
*** Prof. Dr., Emekli öğretim üyesi, Mersin-Türkiye, adnanerkuspsi@gmail.com

yaşandığı gözlenmektedir. Gelbal ve Kelecioğlu'na (2007) göre, öğrenci değerlendirmesine dayalı yöntemlerin, eğitim sisteminde yaygın olarak kullanılmaması ve bu araçların nasıl kullanılacağına ve sonuçlarının nasıl değerlendirileceğine ilişkin yeterince örneğin bulunmaması öğretmenlerin bu alanda güçlük çekmelerinin nedenleri arasında yer alabilir.

Öğrencinin öğrendiklerini gerçekleştirme sürecinin ve bu süreç sonunda ortaya çıkardığı ürünün (sunu, rapor, elişi vb.) ölçülüp değerlendirilmesinde ölçme işlemi ***dolaylı*** olarak; ya öğrencinin *kendisini üçüncü bir gözle değerlendirmesi* ya da öğrencinin bir *başkası (öğretmen, veli, akran) tarafından gözlenip ölçülmesi* söz konusu olduğundan, bu süreçlerde kullanılan ölçme araçları "değerlendirme ölçekleri" olarak adlandırılmaktadır (Aiken, 1995; Hafner ve Hafner, 2003; Erkuş, 2006).

Bu değerlendirmeler ise bir davranışın veya özelliğin varlığını ya da yokluğunu saptamaya yönelik "kontrol listeleri" veya bir davranışın ya da özelliğin ne kadar sık, yoğun veya nitelikli olduğunu belirlemek için kullanılan "dereceleme ölçekleri" ile yapılmak durumundadır.

Değerlendirme ölçeklerinin de bir ölçme aracı olduğu, diğer ölçme araçlarının geliştirilme süreçlerindeki gibi aynı işlemlerle geliştirilmesi ve bunların güvenilir ve geçerli bir ölçme aracı olması gerektiği açıktır. Çünkü bu araçlarla yapılan ölçme sonuçlarına dayanarak öğrencinin ürünü veya kendisi hakkında ciddi kararlar verilmektedir.

Bu bakımdan ölçülecek performansın veya özelliğin;

(varsa) aşamalarının, kritik davranışsal göstergelerinin, kısaca ölçütlerinin belirlenerek yazılması,

a) puanlama türü ve düzeyleri ile bunlara uygun tepki kategorilerinin yazılması,

b) performansı oluşturan parçaların ve kapsamlarının belirlenmesi,

c) bu parçaların ağırlıklarının belirlenmesi,

d) kapsam uygunluğunun irdelenmesi,

e) değerlendiriciler arası uyumun (puanlama güvenirliğinin) irdelenmesi,

f) değerlendirme ölçeğinin işe yararlılığı (geçerliği) hakkında kanıt toplanması,

gibi bir ölçme aracının geliştirilme aşamalarının görmezden gelinmesi veya basit bir süreç olarak algılanmasına pratikte çokça rastlanmaktadır.

Değerlendirme ölçekleriyle yapılan ölçme sonuçlarına dayanarak, öğrencinin ürünü ya da kendisi hakkında ciddi kararlar verileceğinden, değerlendirmede iyi tanımlanmış, sınırları belirlenmiş bir anlayış içinde farklı derecelerde tanımlamalar yapmaya olanak tanıyacak bir yönteme gereksinim duyulur. Bu bağlamda, dereceli puanlama anahtarlarının, puanlama rehberi olarak kullanılması etkili bir yöntem olacaktır (Atılgan, Kan ve Doğan, 2006).

Popham'a (1997) göre, dereceli puanlama anahtarı; *değerlendirme ölçütleri*, *ölçüt tanımlamaları* ve bir *puanlama stratejisi* olmak üzere üç bölümden oluşur. Değerlendirme ölçütlerinin, öğretime ya da değerlendirmeye konu olan performansın ya da ürüne özgü özelliklerin ve boyutların tanımlanmaları gerekir (Tierney ve Simon, 2004). Ölçüt tanımları, performansın kritik aşamalarına ve düzeylerine ilişkin gözlenebilir özelikleri içeren ifadelerdir ve her bir performans düzeyine ve ölçütüne bağlı olarak ayrı ayrı düzenlenir. Dereceli puanlama anahtarlarıyla, sunulan ürün ve bu ürünün ortaya çıkma süreci de değerlendirilebilir. Popham'a (1997) göre, puanlama, bütünsel (holistic) ya da analitik (analitical) biçimde olabilir.

Değerlendirmenin amacı ve ölçülen nitelik göz önüne alındığında, ölçmeye konu olan özelliğin parçalara, bağımsız öğelere ayrıştırılıp ayrıştırılamayacağına göre hangi tip puanlama yönergesinin kullanılacağına karar verilebilir. Bazı durumlarda gözlenen ya da ölçülen özellik

öğelerine ayrıştırılabilmekte, bazı durumlarda ise ölçülen özelliği belirli bağımsız öğelere ayrıştırmak mümkün olmamaktadır (Atılgan, Kan ve Doğan, 2006). Analitik puanlama anahtarında, ürünü, süreci ya da performansı oluşturan parçalar ayrı ayrı belirlenen ölçütler doğrultusunda birbirinden bağımsız olarak puanlanır (Moskal, 2000).

Bir performansın sergilenmesinde gerekli olan ölçüt davranışların o performansa katkısı çoğunlukla aynı oranda olmayacaktır. Bu nedenle dereceli puanlama anahtarında yer alan ana ölçütler ve bunları oluşturan alt ölçütlerin ağırlıklı puanlanması toplam puanı ve dolayısıyla buna bağlı olarak yapılan değerlendirmeyi etkileyecektir. Dereceleme ölçeğinden alınan toplam puanın bir bileşke (composite) puan olmasının, dereceleme ölçeğinin psikometrik niteliklerini de etkileyeceği unutulmamalıdır. Bu bakımdan, puanlama anahtarı, çoğunlukla ağırlıklı puanlama anahtarı şeklinde olmak durumundadır. Burada tek bir davranış ölçütünün ağırlıklandırılması yapılabileceği gibi, ölçüt davranış gruplarının ağırlıklandırılması da yapılabilir.Ayrıca, az sayıda uyarıcının bir ölçek boyutu üzerinde sıralanmaları Thurstone'un ikili karşılaştırmalar yöntemine göre (Turgut ve Baykul, 1992) de gerçekleştirilebilir.

Günümüzde hızla güncelleşen ve herkes tarafından kabul edilen, öğrencilerin performans sergilemesi ve bu performansa göre değerlendirilmesi anlayışı, yazılı anlatım becerileri için de uygulanmaya başlanmıştır. Yazılı anlatım bir beceri olduğu için, bu konuda bilginin değil, becerinin ölçülmesi ve değerlendirilmesi daha doğru bir yaklaşım olacaktır. Bu becerinin ölçme ve değerlendirme sürecinde kullanılan yöntemlerden bir tanesi ise, öğrencilerin bu konudaki çalışmalarının (kompozisyonları) değerlendirilmesidir. Öğrencilerin duygu ve düşüncelerini dilin kurallarına uygun, etkili bir anlatım kullanarak düzenlemeleri istenen kompozisyonların puanlanması ve değerlendirilmesi, değerlendirmeyi yapan kişiye göre değişir. Bu durum kompozisyonların objektif bir biçimde değerlendirilmemesine neden olacak, geçerlik ve güvenirlik problemleri yaratacaktır. Bu nedenle, yazma becerisinin doğası göz önünde bulundurularak, bu becerinin hangi yaklaşımla nasıl puanlanacağına karar verilmesi, güvenirlik ve geçerlik çalışmalarının irdelenmesi için görgül çalışmaların yapılması gerekmektedir.

Çetin ve Kelecioğlu (2004), kompozisyon sınavlarında, kompozisyonun biçimsel özelliklerinin anahtarla puanlamayı ve genel izlenimle puanlamayı ne derece yordadığı ve en iyi yordayıcıların hangi biçimsel özellikler olduğunu incelemişlerdir. Araştırmanın sonunda, biçimsel özelliklerin anahtarla yapılan puanlamanın 0.52'sini, genel izlenimle yapılan puanlamanın 0.60'ını açıkladığı görülmüştür.

Bu çalışmada, değerlendirme işleminin bir aracı olarak yazılı anlatım becerisi ve bunun özel bir uygulaması olan kompozisyon kullanılmıştır. Burada, üç farklı şekilde (genel izlenimle, dereceli puanlama anahtarıyla ve Thurstone ikili karşılaştırmalar yöntemiyle) puanlanan ve değerlendirilen kompozisyonlardan elde edilen verilere dayanarak yapılan sıralamayla, üç farklı değerlendirme biçiminin karşılaştırılması amaçlanmaktadır.

Araştırma amacına bağlı olarak aşağıdaki sorulara yanıt aranacaktır:

1. İkili karşılaştırmalarla, genel izlenimle ve dereceli puanlama anahtarıyla yapılan değerlendirmelere dayalı kompozisyonların sıralanmaları değişmekte midir?
2. Aynı yargıcının beş ayrı kompozisyon için, genel izlenimle ve dereceli puanlama anahtarıyla
   a) verdikleri puanlar arasında anlamlı bir ilişki var mıdır?
   b) verdikleri puanların ortalamaları arasında anlamlı bir fark var mıdır?
3. Farklı yargıcıların beş ayrı kompozisyonu,
   a) genel izlenimle puanlanmasında puanlayıcı güvenirliği nasıldır?
   b) dereceli puanlama anahtarı ile puanlanmasında puanlayıcı güvenirliği nasıldır?

2. YÖNTEM

Bu araştırma, dereceli puanlama anahtarıyla, genel izlenimle ve ikili karşılaştırmalar yöntemiyle puanlanan ve değerlendirilen kompozisyonların sıralanmalarının karşılaştırılması incelendiğinden betimsel bir çalışmadır.

2.1. Çalışma Grubu

Araştırmada kullanılan ve aynı konu üzerine bir ders kapsamında yazdırılan beş kompozisyonun seçilmesinde iki İlköğretim Türkçe öğretmeninin görüşlerine başvurulmuştur. Dereceli puanlama anahtarının davranışsal ölçütlerinin geliştirilmesi aşamasında çeşitli üniversitelerden Türkçe kompozisyon yazma konusunda uzman 36 öğretim elemanı yargıcı olarak kullanılmıştır. Kompozisyonların dereceli puanlama anahtarıyla, genel izlenimle ve ikili karşılaştırmalar yöntemiyle değerlendirilmesi Mersin ilindeki çeşitli ilköğretim okullarında görev yapan gönüllü 194 Türkçe (Edebiyat) öğretmeniyle, dereceli puanlama anahtarının güvenirlik çalışması ise 21 Türkçe öğretmeniyle gerçekleştirilmiştir.

2.2. Dereceli Puanlama Anahtarının Hazırlanması ve Geliştirilmesi

İlk aşamada, hem yurt dışında hem de yurt içinde yazılı anlatım becerilerini değerlendirmek için hazırlanan çeşitli dereceli puanlama anahtarları araştırmacı tarafından incelenerek, öğrencilerin kompozisyonlarını değerlendirmek amacıyla kullanılması planlanan denemelik bir dereceli puanlama anahtarı hazırlanmıştır. Bunun sonucunda, denemelik dereceli puanlama anahtarında bulunması gereken ana ölçütler, alt ölçütler ve ölçüt göstergeleri belirlenerek, bunlara ilişkin ağırlıklandırma önsel (a priori) olarak yapılmıştır. Hazırlanan denemelik dereceli puanlama anahtarı çeşitli üniversitelerden Türkçe kompozisyon yazma konusunda uzman 36 öğretim görevlisi tarafından değerlendirilmiştir. Uzman görüşleri arasındaki uyuşma oranı (P) hesaplanmıştır (Aiken, 1995). Ana ölçütlerin ve bunlara ait alt ölçütlerin ayrılmasına ilişkin uzmanların uyuşma oranı (P) değerlerinin yüksek çıktığı görülmüştür.

Alt ölçüt göstergelerinin uygunluğu aşamasında ise, uzmanlara dereceletilen davranış göstergelerinin beklenen ortalama değeri (3) gözlenen ortalama değeriyle karşılaştırılmıştır (Erkuş, 2003). "Beklenen ortalama değer < gözlenen ortalama değer" olduğunda davranış göstergesinin uygun olduğu yargısına varılmıştır.

Bu süreçte, Türkçe kompozisyon yazma konusunda uzman olduğu bilinen 36 öğretim elemanının değerlendirmeleri sonucu "Kompozisyon Yazmaya İlişkin Dereceli Puanlama Anahtarı'na" son şekli verilmiştir.

Dereceli puanlama anahtarıyla yapılan dereceleme ölçeğinin davranış ölçütlerinin nasıl bir yapı sergilediğini görebilmek için yapılan faktör analizinde, veri setinin faktör analizi için uygun olup olmadığını değerlendirmek amacıyla Barlett Küresellik ve KMO testleri sonuçları incelenmiştir. KMO test değerlerinin yüksek (>.70) ve Barlett testinin de anlamlı çıktığı ($p< .05$) görülmüştür.

Her kompozisyonun bütününe ya da alt ölçüt puanlarına yapılan döndürmesiz temel bileşenler faktör analizi sonucu elde edilen özdeğeri birden büyük olan faktör ve bunu takip eden en yakın faktöre ilişkin özdeğerler ve bu faktörlerin açıkladıkları varyans yüzdeleri Tablo 1'de gösterilmektedir.

Üç ana alt bileşenin döndürmesiz temel bileşenler faktör analizi sonucunda Tablo 1'de görüldüğü gibi, 1. Faktörlerin açıklama gücü oldukça yüksektir. 1. Faktörü takip eden en yakın faktörün açıklama gücünün oldukça düşük olması alt bileşenlerin tek faktörlü olduğunun göstergesidir (Lord, 1980). Bununla beraber, daha önce de sözü edildiği gibi, öz değerlendirme (self-report) ölçme araçlarından farklı olarak burada yazma becerisi (kompozisyon) ve

değerlendirici şeklinde iki yüzey (facet) bulunmasına rağmen, bu kadar yüksek açıklama yüzdeleri, tüm ana ölçütlerin bu çalışma kapsamında tek boyutlu olduğunun göstergesi olarak kabul edilebilir. Ana ölçütler ve tüm ölçek için hesaplanan iç tutarlık katsayıları incelendiğinde, tüm kompozisyonlar için hesaplanan ana ölçütlerin ve tüm ölçeğin Cronbach α katsayılarının oldukça yüksek çıkması, değerlendirme ölçeğinin tek faktörlü bir yapıda olduğunu desteklemektedir. Bu iç tutarlık katsayıları faktör analizi ile birlikte ele alındığında, *kompozisyon yazma becerisinin*, bir başka açıdan da *değerlendirme sürecinin* aynı bilişsel mekanizmaya dayandığı ileri sürülebilir.

Tablo 1: Temel Bileşenler Faktör Analizi Sonuçları

		I. Kompozisyon		II. Kompozisyon		III. Kompozisyon		IV. Kompozisyon		V. Kompozisyon	
Ölçüt	**Faktör**	**1**	**2**	**1**	**2**	**1**	**2**	**1**	**2**	**1**	**2**
Biçim	Özdeğeri	4.69	1.01	5.35	1.13	5.53	-	5.61	1.12	5.66	-
	Açıklama Yüzdesi	52.06	11.24	59.43	12.59	61.46	-	62.34	12.47	62.85	-
	Cronbach α	.88		.91		.92		.92		.92	
İçerik	Özdeğeri	10.95	1.54	13.42	1.05	13.00	1.03	12.14	1.64	12.80	1.36
	Açıklama Yüzdesi	54.77	7.72	67.08	5.24	65.04	5.15	60.70	8.17	64.01	6.78
	Cronbach α	.95		.97		.97		.96		.97	
Hedef Kitleye Uygunluk	Özdeğeri	4.96	1.10	5.54	-	5.69	-	5.46	-	5.66	-
	Açıklama Yüzdesi	61.95	13.74	69.21	-	71.11	-	68.29	-	70.74	-
	Cronbach α	.91		.93		.94		.93		.94	
Tüm	Özdeğeri	17.91	2.84	22.52	2.01	22.15	1.83	20.77	2.62	21.89	2.34
	Açıklama Yüzdesi	48.41	7.69	60.86	5.42	59.86	4.94	56.13	7.07	59.15	6.33
	Cronbach α	.97		.98		.98		.98		.98	

Yargıcıların dereceli puanlama anahtarıyla yaptıkları puanlamaların kararlılığını (stability) saptamak amacıyla, "en iyi" ve "en kötü" olarak değerlendirilen iki kompozisyonu aynı yargıcılar 7-14 gün zaman aralığıyla iki kez değerlendirmiş, iki değerlendirme arasındaki ilişki Spearman'ın Sıra Farkları Korelasyon Katsayısı ile hesaplanmıştır. Yargıcıların aynı dereceli puanlama anahtarıyla iki kez yapmış oldukları değerlendirmeler arasındaki korelasyon katsayıları "iyi" kompozisyon için 0.819 ve "kötü" kompozisyon için 0.683 olarak bulunmuştur. Burada, yapılan değerlendirmeler arasında pozitif yönde ve anlamlı bir ilişkinin olduğu görülmektedir. Bu durum, yargıcıların "iyi" olan kompozisyonu kararlı bir şekilde puanlandıkları şeklinde yorumlanabilir. Ancak yargıcıların "kötü" kompozisyon üzerinde yapmış oldukları değerlendirmeler arasındaki korelasyonun düşük çıkmasının bu kompozisyon üzerinde çok farklı puanlar vermelerinden kaynaklandığı ileri sürülebilir.

Beş ayrı kompozisyonun, dereceli puanlama anahtarıyla puanlanmasında, alt ölçüt puanlarının birbirleriyle ve toplam puanla olan ilişkilerine ait sonuçlar Tablo 2'de gösterilmektedir.

Tablo 2 incelendiğinde her bir kompozisyon için, içerik alt ölçüt puanları ile toplam puan arasındaki korelasyon değerlerinin diğer alt ölçütlerden elde edilen korelasyon değerlerine göre

en yüksek olduğu görülmektedir. Bu sonuç bize, kompozisyon yazma becerisi için yapılan dereceli puanlama anahtarındaki içerik alt ölçütün diğer alt ölçütlere göre toplam puan ile daha yüksek pozitif ilişkili olduğunu göstermektedir.

Tablo 2: Alt Ölçütlerden Elde Edilen Puanların Birbirleriyle ve Toplam Ölçekten Elde Edilen Puanlarla Olan Korelasyonları

		İçerik	Hedef Kitleye Uygunluk	Toplam
1. Kompozisyon	Biçim	.759	.666	.850
	İçerik	-	.842	.976
	Hedef Kitleye Uygunluk	-	-	.904
		İçerik	Hedef Kitleye Uygunluk	Toplam
2. Kompozisyon	Biçim	.840	.818	.911
	İçerik	-	.928	.984
	Hedef Kitleye Uygunluk	-	-	.956
		İçerik	Hedef Kitleye Uygunluk	Toplam
3. Kompozisyon	Biçim	.833	.791	.902
	İçerik	-	.903	.984
	Hedef Kitleye Uygunluk	-	-	.941
		İçerik	Hedef Kitleye Uygunluk	Toplam
4. Kompozisyon	Biçim	.830	.712	.896
	İçerik	-	.870	.982
	Hedef Kitleye Uygunluk	-	-	.910
		İçerik	Hedef Kitleye Uygunluk	Toplam
5. Kompozisyon	Biçim	.814	.747	.883
	İçerik	-	.912	.985
	Hedef Kitleye Uygunluk	-	-	.942

Alt ölçütlerin kendi aralarındaki korelasyon değerlerine bakıldığında ise, biçim ile hedef kitleye uygunluk alt ölçüt puanları arasındaki korelasyon değerlerinin diğer alt ölçütlerden elde edilen korelasyon değerlerine göre en düşük, içerik ile hedef kitleye uygunluk alt ölçüt puanları arasındaki korelasyon değerlerinin diğer alt ölçütlerden elde edilen korelasyon değerlerine göre en yüksek olduğu görülmektedir.

Tüm kompozisyonlar için, alt ölçütler ile toplam puanlar arasındaki korelasyon katsayılarının oldukça yüksek olmasının (multicolinearity) kompozisyon yazma becerisinin ve değerlendirmenin ortak bir bilişsel sürecin ürünü olduğuna işaret ettiği ileri sürülebilir.

2.3. Verilerin Analizi

Kompozisyonların ikili karşılaştırmalarla sıralanması, Thurstone'nun V. Hal eşitliği yardımıyla tam veri matrisinden ölçekleme kullanılarak yapılmıştır. İkili karşılaştırmalar yönteminde N gözlemciye U_j ve U_K uyarıcılarından hangisinin uyarıcılık değerinin büyük olduğu sorulur. Örneğin, gözlemciler belirtilen iki uyarıcıdan hangisinin "daha büyük", "daha iyi", "daha olumsuz" veya "daha iyi görünüşlü" olduğuna karar verirler. Her gözlemciden, her uyarıcı çiftinden birini diğerine mutlaka tercih etmeleri istenir. Eşitlik veya ayırt etmeme yargılarına izin verilmez (Turgut ve Baykul, 1992). Bu yöntemde yargıcı kararlarının frekansları önce p matrisine, sonra da z matrisine çevrilerek, z ortalamaları uyarıcıların yargı boyutundaki ölçek değeri olarak bulunur.

Genel izlenimle ve dereceli puanlama anahtarıyla kompozisyonların sıralanmaları, öğretmenlerin verdikleri toplam puanların ortalamaları alınarak yapılmıştır. Kompozisyonların ortalamaları arasındaki farkın anlamlı olup olmamasına ise Tek Yönlü ANOVA ile bakılmıştır.

Aynı yargıcı tarafından iki farklı şekilde (genel izlenimle ve dereceli puanlama anahtarıyla) ayrı ayrı puanlanan aynı kompozisyonlardan elde edilen puanlar arasındaki ilişki Pearson Momentler Çarpımı Korelasyon Katsayısıyla hesaplanmıştır. Genel izlenimle ve

dereceli puanlama anahtarıyla elde edilen puanların ortalamaları arasındaki farkın anlamlılığı ise, bağımlı gruplar arası t testi ile test edilmiştir.

Farklı yargıcıların beş ayrı kompozisyonu, genel izlenimle ve dereceli puanlama anahtarı ile puanlamasında puanlayıcı güvenirliğini belirlemek için Genellenebilirlik (G) kuramından yararlanılmıştır. Hesaplanacak olan G katsayısında değişkenlik kaynağı olarak öğretmenler, kompozisyonlar ve bunların birbiriyle etkileşimden meydana gelecek ortak etkiler dikkate alınmıştır.

3. BULGULAR

3.1. Thurstone İkili Karşılaştırmalar Yöntemiyle Yapılan Değerlendirmelere Dayalı Kompozisyonların Sıralamalarına İlişkin Bulgular

Her bir öğretmenin nitelikli bir yazıda bulunması gereken özellikler açısından beş kompozisyonu ikili karşılaştırma yaparak değerlendirme sonucu elde edilen ölçek değerleri ve sıralamalar, Tablo 3'te gösterildiği şekilde oluşturulmuştur.

Tablo 3: Beş Kompozisyona Ait Ölçek Değerleri ve Uyarıcı Sıraları

Kompozisyonlar	Ölçek Değerleri	Uyarıcı Sıraları
1	0.827	3
2	0.690	4
3	1.252	1
4	0.000	5
5	0.929	2

Tablo 3'e göre, nitelikli bir yazıda bulunması gereken özellikler açısından ilk tercih edilen kompozisyonun 3. kompozisyon ve son tercih edilen kompozisyonun ise 4. kompozisyon olduğu görülmektedir.

3.2. Genel İzlenimle Yapılan Değerlendirmelere Dayalı Kompozisyonların Sıralamalarına İlişkin Bulgular

194 öğretmenin beş kompozisyonu genel izlenime dayalı bütüncül puanlamalarına ilişkin istatistikler Tablo 4'te gösterilmektedir.

Tablo 4: Beş Kompozisyonun Genel İzlenime Dayalı Bütüncül Puanlanmalarına İlişkin İstatistikler

Kompozisyonlar	N	Minimum	Maksimum	Ortalama	Standart Sapma
1	194	40,000	100,000	75,531	11,778
2	194	35,000	100,000	73,557	13,349
3	194	40,000	100,000	81,289	12,744
4	194	10,000	90,000	62,242	13,741
5	194	20,000	100,000	76,835	12,016

Tabloda kompozisyonların ortalamalarına bakıldığında en yüksek ortalama değerinin 3. Kompozisyonda (81, 289) ve en düşük ortalama değerinin ise 4. Kompozisyonda (62,242) olduğu görülmektedir. Kompozisyonların ortalamaları arasındaki farkın anlamlı olup olmamasına ise Tek Yönlü ANOVA ile bakılmış ve kompozisyonların ortalamaları arasında anlamlı bir farklılık olduğu bulunmuştur [$F_{(4-965)} = 60{,}248$, $p<0.05$]. Hangi kompozisyonlar

arasında fark olup olmadığına ise Post Hoc Testi ile bakılmıştır. Post Hoc Testi sonuçlarında, genel izlenime dayalı değerlendirilen kompozisyonların tümü için ortalamalar arası farkın anlamlı olduğu görülmektedir. Kompozisyonlar genel izlenimle verilen puanların ortalamasına göre sıralandığında; ilk sırada 3. kompozisyonun, ikinci sırada 5. kompozisyonun, üçüncü sırada 1. kompozisyonun, dördüncü sırada 2. kompozisyonun ve en son sırada ise, 4. kompozisyonun yer aldığı görülmektedir.

3.3. Her Davranış Ölçütü Eşit Ağırlıklandırılan Dereceli Puanlama Anahtarı İle Yapılan Değerlendirmelere Dayalı Kompozisyonların Sıralamalarına İlişkin Bulgular

194 öğretmenin beş kompozisyonu her davranış ölçütü eşit ağırlıklandırılan dereceli puanlama anahtarı ile ayrı ayrı değerlendirmelerine ilişkin istatistikler Tablo 5'te gösterilmektedir.

Tablo 5: Beş Kompozisyonun Her Davranış Ölçütü Eşit Ağırlıklandırılmış Dereceli Puanlama Anahtarı İle Değerlendirilmesine İlişkin İstatistikler

Kompozisyonlar	N	Minimum	Maksimum	Ortalama	Standart Sapma
1	194	59,000	179,000	125,701	26,435
2	194	31,000	175,000	115,680	32,512
3	194	59,000	185,000	142,665	28,694
4	194	20,000	171,000	98,062	29,485
5	194	63,000	185,000	131,263	29,765

Tablo 5'teki kompozisyonların ortalamalarına bakıldığında en yüksek ortalama değerinin 3. Kompozisyonda (142.665) ve en düşük ortalama değerinin ise 4. Kompozisyonda (98.062) olduğu görülmektedir. Kompozisyonların ortalamaları arasındaki farkın anlamlı olup olmamasına ise Tek Yönlü ANOVA ile bakılmıştır. Burada kompozisyonların ortalamaları arasında anlamlı bir farklılık olduğu bulunmuştur [$F_{(4-965)}$=63,623, p<0.05]. Hangi kompozisyonlar arasında fark olup olmadığına ise Post Hoc Testi ile bakılmıştır. Post Hoc Testi sonuçlarına bakıldığında, her davranış ölçütü eşit ağırlıklandırılan dereceli puanlama anahtarıyla değerlendirilen kompozisyonların tümü için ortalamalar arası farkın anlamlı olduğu görülmektedir. Kompozisyonlar her davranış ölçütünün eşit ağırlıklandırılmış dereceli puanlama anahtarıyla verilen puanların ortalamasına göre sıralandığında; ilk sırada 3. kompozisyonun, ikinci sırada 5. kompozisyonun, üçüncü sırada 1. kompozisyonun, dördüncü sırada 2. kompozisyonun ve en son sırada ise, 4. kompozisyonun yer aldığı görülmektedir.

3.4. Davranış Ölçütleri Uzman Kanılarına Dayalı Olarak Ağırlıklandırılan Dereceli Puanlama Anahtarından Elde Edilen Puanlara Dayalı Kompozisyonların Sıralanmalarına İlişkin Bulgular

194 öğretmenin beş kompozisyonu, uzman kanılarına dayalı ağırlıklı dereceli puanlama anahtarı ile değerlendirmelerine ilişkin istatistikler Tablo 6'da gösterilmektedir. Tablo 6'da kompozisyonların ortalamalarına bakıldığında en yüksek ortalama değerinin 3. Kompozisyonda (77.048) ve en düşük ortalama değerinin ise 4. Kompozisyonda (53.911) olduğu görülmektedir. Kompozisyonların ortalamaları arasındaki farkın anlamlı olup olmamasına ise Tek Yönlü ANOVA ile bakılmıştır. Burada kompozisyonların ortalamaları arasında anlamlı bir farklılık olduğu bulunmuştur [$F_{(4-965)}$=64,836, p<0.05]. Hangi kompozisyonlar arasında fark olup olmadığına ise Post Hoc Testi ile bakılmıştır Post Hoc Testi sonuçları incelendiğinde davranış ölçütleri uzman kanılarına dayalı ağırlıklandırılan dereceli puanlama anahtarıyla değerlendirilen kompozisyonların tümünde ortalamalar arası farkın anlamlı olduğu görülmektedir. Kompozisyonlar uzman kanılarına dayalı ağırlıklandırılan dereceli puanlama anahtarıyla verilen puanların ortalamasına göre sıralandığında; ilk sırada 3. kompozisyonun, ikinci sırada 5.

kompozisyonun, üçüncü sırada 1. kompozisyonun, dördüncü sırada 2. kompozisyonun ve en son sırada ise, 4. kompozisyonun yer aldığı görülmektedir

Tablo 6: Beş Kompozisyonun Uzman Kanılarına Dayalı Dereceli Puanlama Anahtarı İle Değerlendirilmesine İlişkin İstatistikler

Kompozisyonlar	N	Minimum	Maksimum	Ortalama	Standart Sapma
1	194	33,27	96,87	68,974	14,010
2	194	31,71	93,11	64,327	14,856
3	194	33,24	100,00	77,048	15,388
4	194	11,66	91,25	53,911	15,845
5	194	33,37	100,00	71,886	15,590

.

Kompozisyonların değerlendirilmelerine ilişkin elde edilen tüm bulgular incelendiğinde, hangi değerlendirme (ikili karşılaştırmalarla, genel izlenimle ve dereceli puanlama anahtarıyla) yapılırsa yapılsın kompozisyonların sıralamalarında bir değişikliğin olmadığı görülmektedir.

3.5. Aynı Yargıcıların Beş Ayrı Kompozisyon İçin, Genel İzlenimle ve Dereceli Puanlama Anahtarıyla Verdikleri Puanlar Arasındaki (Puanlama Geçerliği) İlişkiye Ait Bulgular

Puanlama geçerliğini incelemek için her değerlendirme bir diğerinin ölçütü olarak ele alınmıştır. Aynı yargıcılar tarafından iki farklı şekilde (genel izlenimle ve dereceli puanlama anahtarıyla) ayrı ayrı puanlanan aynı kompozisyonlardan elde edilen puanlar arasındaki korelasyon katsayıları Tablo 7'de gösterilmektedir.

Tablo 7 incelendiğinde, yargıcıların genel izlenimle ve dereceli puanlama anahtarıyla yapmış oldukları değerlendirmeler arasında pozitif yönde ve anlamlı bir ilişkinin olduğu görülmektedir. Burada yapılan değerlendirmeler arasındaki en yüksek korelasyon değeri 5. kompozisyona (0.716) ait iken, en düşük korelasyon değeri ise, 2. kompozisyona (0.586) aittir. Bulunan korelasyon değerlerinin çok yüksek çıkmaması, yargıcıların kompozisyonları genel izlenimle ve dereceli puanlama anahtarıyla değerlendirmeleri arasında farklılık yaratabildiği şeklinde yorumlanabilir.

Tablo 7: Genel İzlenim ve Dereceli Puanlama Anahtarı Puanları Arasındaki Korelasyon Katsayıları

Kompozisyon	N	Genel İzlenim ve Dereceli Puanlama Anahtarı Puanları Arasındaki Korelasyon	p
1	194	0.668	0.000
2	194	0.586	0.000
3	194	0.671	0.000
4	194	0.650	0.000
5	194	0.716	0.000

3.6. Aynı Yargıcıların Beş Ayrı Kompozisyon İçin, Genel İzlenim ile ve Dereceli Puanlama Anahtarı ile Verdikleri Puanların Ortalamaları Arasındaki Farkın Anlamlılığına İlişkin Bulgular

Aynı yargıcılar tarafından iki farklı şekilde (genel izlenimle ve dereceli puanlama anahtarıyla) ayrı ayrı puanlanan aynı kompozisyonlardan elde edilen puanların ortalamaları arasındaki farkın anlamlılığını test etmek amacıyla yapılan bağımlı gruplar arası t testi sonuçları Tablo 8'de gösterilmektedir. Tablo 8'de görüldüğü gibi, yargıcıların iki farklı şekilde aynı kompozisyonlara vermiş oldukları puanların ortalamaları arasında anlamlı bir farklılık olduğu görülmektedir ($p< 0.05$).

Tablo 8: Genel İzlenim ve Dereceli Puanlama Anahtarıyla Elde Edilen Puanlara İlişkin Bağımlı Gruplar Arası t Testi Sonuçları

Kompozisyon	N	Değerlendirme	Ortalama	Standart Sapma	t	p
1	194	Genel İzlenim	75,531	11,778	8.533	0.000
	194	Dereceli Puanlama Anahtarı	68,974	14,010		
2	194	Genel İzlenim	73,557	13,350	9.961	0.000
	194	Dereceli Puanlama Anahtarı	64,327	14,856		
3	194	Genel İzlenim	81,289	12,745	5.064	0.000
	194	Dereceli Puanlama Anahtarı	77,048	15,388		
4	194	Genel İzlenim	62,242	13,742	9.267	0.000
	194	Dereceli Puanlama Anahtarı	53,911	15,845		
5	194	Genel İzlenim	76,835	12,016	6.313	0.000
	194	Dereceli Puanlama Anahtarı	71,886	15,590		

Tabloya bakıldığında, yargıcıların tüm kompozisyonlara genel izlenimle vermiş oldukları puanların ortalamasının, dereceli puanlama anahtarı ile vermiş oldukları puanların ortalamasından büyük olduğu görülmektedir. Başka bir ifadeyle, yargıcıların genel izlenim ile kompozisyonlara daha fazla puan verdikleri görülmektedir. Ayrıca, yargıcıların tüm kompozisyonlara genel izlenimle vermiş oldukları puanların standart sapmasının, dereceli puanlama anahtarı ile vermiş oldukları puanların standart sapmasından küçük olduğu görülmektedir. Bu sonuçlara göre, yargıcıların tümünün genel izlenimle değerlendirmelerine "cömertlik hatası" karıştığını söylemek mümkündür

3.7. Yargıcıların Beş Ayrı Kompozisyonu, Genel İzlenimle Puanlamasında Puanlayıcı Güvenirliği Nasıldır?

Değerlendirme ölçeklerinde, değerlendiriciler ve değerlendirilenler şeklinde iki varyans kaynağı bulunduğundan, puanlama güvenirliği için psikometrik açıdan en uygun yol, bu iki varyans kaynağının da dikkate alındığı genellenebilirlik katsayısının hesaplanmasıdır.

Burada puanlayıcı güvenirliğini hesaplamak için, değişkenlik kaynağı olarak öğretmenler, kompozisyonlar ve bunların birbiriyle etkileşimi göz önünde bulundurulmuş ve G katsayısı hesaplanmıştır. Tablo 9'da kompozisyonların genel izlenimle değerlendirilmelerindeki varyans kestirimlerine ilişkin sonuçlar yer almaktadır.

Tablo 9: Kompozisyonların Genel İzlenimle Değerlendirilmelerindeki Varyans Kestirimlerine İlişkin Sonuçlar

Bileşen	Kestirim
Varyans (Öğretmen)	65,905
Varyans (Kompozisyon)	49,974
Varyans (Öğretmen*Kompozisyon)	96,615

Yargıcıların (n=194) kompozisyonları genel izlenimle puanlamalarından elde edilen G katsayısı 0,99 olarak bulunmuştur. Ayrıca, bir öğretmenin puanlama yapması durumunda, tutarlığın ve/veya genellenebilirliğin ne olacağı da merak edilmiş ve 0,37 bulunmuştur. Bu sonuç, yargıcı sayısının arttırılmasının kompozisyonların genel izlenimle değerlendirilmelerindeki tutarlılığı arttıracağı şeklinde yorumlanabilir.

3.8. Yargıcıların Beş Ayrı Kompozisyonu, Dereceli Puanlama Anahtarı İle Puanlamasında Puanlayıcı Güvenirliği Nasıldır?

Burada puanlayıcı güvenirliğini hesaplamak için, değişkenlik kaynağı olarak öğretmenler, kompozisyonlar ve bunların birbiriyle etkileşiminden meydana gelebilecek varyans göz önünde bulundurulmuş ve G katsayısı hesaplanmıştır. Tablo 10'da kompozisyonların dereceli puanlama anahtarıyla değerlendirilmelerindeki varyans kestirimlerine ilişkin sonuçlar yer almaktadır.

Tablo 10: Kompozisyonların Dereceli Puanlama Anahtarıyla Değerlendirilmelerindeki Varyans Kestirimlerine İlişkin Sonuçlar

Bileşen	Kestirim
Varyans (Öğretmen)	119,732
Varyans (Kompozisyon)	76,166
Varyans (Öğretmen*Kompozisyon)	109,864

Farklı yargıcıların kompozisyonları dereceli puanlama anahtarı ile puanlamalarından elde edilen G katsayısı bir öğretmenin puanlama yapması durumunda .40 bulunurken, 194 öğretmenin puanlama yapması durumda ise .99 olarak bulunmuştur. Bu sonuç, yargıcı sayısının arttırılmasıyla, kompozisyonların dereceli puanlama anahtarıyla değerlendirilmelerindeki tutarlılığı arttıracağı şeklinde yorumlanabilir.

4. TARTIŞMA ve SONUÇ

İkili karşılaştırmalarla, genel izlenimle ve dereceli puanlama anahtarıyla yapılan değerlendirmelere ilişkin analizler sonucunda, hangi değerlendirme yapılırsa yapılsın kompozisyonların sıralamalarında bir değişikliğin olmadığı bulunmuştur. Kompozisyon değerlendirmelerinde sıralama söz konusu olduğu durumlarda, ikili karşılaştırmalar yönteminin kullanılmasının uygun olacağı ileri sürülebilir. Eğer yazma becerisinin ayrıntıları söz konusu değilse, bu konuda yetişmiş uzmanların genel izlenimle kompozisyonları sıraya dizebilecekleri; başka bir ifadeyle "iyi-kötü" kompozisyonları ayırt edebilecekleri görülmektedir. Ancak özellikle ciddi ve titiz değerlendirme ile bireylerin alt yazma becerileri de merak edildiğinde "ayrıntılı dereceli puanlama anahtarıyla" değerlendirme yapılması daha yararlı olacaktır.

Yargıcıların kompozisyonlara iki farklı şekilde (genel izlenimle ve dereceli puanlama anahtarıyla) verdikleri puanlar arasında bulunan korelasyon değerlerinin çok yüksek çıkmaması, değerlendirmeler arasında farklılık olabileceği şeklinde yorumlanmıştır. Genel izlenimle ve dereceli puanlama anahtarıyla elde edilen puanların ortalamaları arasında puanların lehine anlamlı bir fark bulunmuştur. Bu durum değerlendiricilerin genel izlenimle değerlendirmelerinde *cömertlik;* ayrıntılı puanlamada ise *cimrilik* değerlendirme hataları yaptıkları şeklinde yorumlanabilir.

Farklı yargıcıların beş ayrı kompozisyonu, genel izlenimle ve dereceli puanlama anahtarıyla puanlanmasında puanlayıcı güvenirliğinin belirlenmesinde, değişkenlik kaynağı olarak öğretmenler, kompozisyonlar ve bunların birbiriyle etkileşimden meydana gelecek hatalar göz önünde bulundurulmuş ve G katsayıları hesaplanmıştır. Bu katsayıların yüksek çıkması, puanlama güvenirliğinin olduğunu, başka bir ifadeyle, yargıcıların tutarlı bir şekilde puanlama yaptıklarını göstermektedir. Ancak başka çalışmalarda iki düzeye (facet) madde düzeyinin de eklenmesi daha zengin bilgiler verebilir.

Bu çalışmada, beş ayrı kompozisyonun üç farklı yöntemle yapılan sıralamaları incelenmiştir. Üç farklı yönteme göre kompozisyon sıralamalarının değişmemesi, öğrencilerin bireysel puanları arasında fark olmayacağı anlamına gelmemelidir. Bu çalışmayı takiben yapılacak çalışmalarda çok sayıda kompozisyonun (öğrencinin) puanlanmaları arasında ve sıralanmalarında bir fark olup olmadığının incelenmesi yararlı olacaktır.

KAYNAKLAR

Airasian, P. W. (2000). Assesment in the clasroom: A concise approach. Boston: McGraw-Hill.

Aiken, L. R. (1995). *Rating scales and checklists; evaluating behavior, personality, and attitudes.* New York: John Wiley & Sons.

Atılgan, H., Kan A., ve Doğan N. (2006). *Eğitimde ölçme ve değerlendirme,* Ankara: Anı Yayıncılık.

Çetin, B. ve H. Kelecioğlu (2004). Kompozisyon tipi sınavlarda kompozisyonun biçimsel özelliklerinden kestirilen puanların anahtarla ve genel izlenimle elde edilen puanlarla ilişkisi , *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi*, 26, 19-26.

Erkuş, A. (2003). *Psikometri üzerine yazılar*, Ankara: Türk Psikologlar Derneği Yayınları.

Erkuş, A. (2006). *Sınıf öğretmenleri için ölçme ve değerlendirme*, Ankara: Ekinoks Yayıncılık.

Gelbal, S. ve Kelecioğlu H. (2007). Öğretmenlerin ölçme ve değerlendirme yöntemleri hakkındaki yeterlik algıları ve karşılaştıkları sorunlar. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi.* 33.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Hafner, J., & Hafner M. (2003). Quantitative analysis of the rubrics as an assessment tool : An emprical study of student peer- group rating. *İnternational Journal of Science Education.*

Mertler, C. A. (2001). Designing scoring rubrics for your classroom. *Practical Assessment, Research, and Evaluation,*

Moskal, B., M. (2000). Scoring rubrics: what, when, how? *Pratical Assessment, Research and Evaluation,* 8 (14).

Popham, W. J. (1997). What's wrong- and what's right – with rubrics. *Educational Leadership.*

Tierney, R. & Simon, M. (2004). What's still wrong with rubrics : Focusing on the contistency of performance criteria across scale levels. *Pratical Assessment, Research and Evaluation,* 9 (2).

Turgut, F ve Baykul, Y. (1992). *Ölçekleme teknikleri.* Ankara: ÖSYM Yayınları

Extended Abstract

The social, cultural, economic, political, and technological areas are in rapid changes in our country. Assessment of student performance has became a fundamental aspect of teaching and learning. Assessment of performance also provides new challenges for many fields. The use of alternative assessment approaches has been recommended in educational curriculum.

Scoring rubrics are useful to serve performance assessment and currently used by students and teachers in classrooms from kindergarten to college. Rubrics are rating scales as opposed to checklists that are used with performance assessments. They are formally defined as scoring guides, consisting of specific pre-established performance criteria, used in evaluating student work on performance assessments (Mertler, 2001).

According to Popham (1997), a scoring rubric has typically three parts: evaluative criteria, quality definitions and scoring strategy. Evaluative criteria are used to distinguished acceptable responses from unacceptable responses. Quality definitions describe the way that qualitative differences to student's responses are to be judged scoring strategy may be holistic or analytic (Popham, 1997).

Rubrics are typically the specific form of scoring instrument used when evaluating student performances or products resulting from a performance task. There are two types of rubrics: holistic and analytic. A holistic rubric requires the teacher to score the overall process or product as a whole, without judging the component parts separately (Mertler, 2001). An analytic rubric, the teacher scores separate, individual parts of the product or performance first, then sums the individual scores to obtain a total score

(Moskal, 2000). Prior to designing a specific rubric, a teacher must decide whether the performance or product will be scored holistically or analytically (Airasian, 2000).

Unfortunately, many rubrics are still not instructionally useful because of inconsistencies in the descriptions of performance criteria across their scale levels For scoring rubrics to fulfill their educational ideal, they must first be designed to reflect greater consistency in their performance criteria descriptors. Rubric development can be challenging, Rubric's design must be thoughtfully matched to its purpose. Consistency is an important technical requirement that should be considered carefully for all scoring rubrics designed or adapted for classroom use.

In using any evaluation technique, you must make sure that you are assessing student learning using reliable and valid measures. A rubric will be reliable if it yields results that are accurate and stable; and it is valid for a particular purpose, if it measures what it was intended to measure (Moskal, 2000).

This study investigates comparison of three different evaluation types based on the data of the compositions which have been going to be rated and evaluated in three different types (overall impression, rubric and Thurstone's method of paired comparison.

Firstly, "The Rubric For Writing Composition" was constructed which was going to be used as a tool for evaluating writing skill. It is observed that, one factor explatory solution which is done to group of behavior criteria of rating scale constructed as rubric is accurate for both scale and its three components. Additionally, Cronbach alpha coefficients calculated for major and minor criteria of all composition is observed to be high. The relationship among rates of compositions which are evaluated with the rubric by the same judges found positive and significant.

Rankings of compositions by overall impressions and rubrics were conducted according to means of points which was assigned by teachers. The differences between the means of compositions were tested by oneway ANOVA.

The relationship between the gradings of same judges by using overall impressions and rubrics were calculated by Pearson Product-Moment Correlation Coefficient separately. The differences of overall impressions and rubrics mean grades were analyzed by using paired t test.

Generalibility Theory is used to determine the rater reliability of five compositions which was graded by five judges. Teachers, compositions and the interaction of them was used for source of variance for G coefficient.

It has been shown that there is no difference between the ranks of compositions, independent from rating methods. Correlations which for the relationships between the grading of judges according to different grading methods were found moderately. The differences between the gradings have been thought to explain those moderate correlations. G coefficient which was calculated to determine rater reliability was found extremely high and this can be evidence for reliable rater gradings in other words raters graded the compositions consistently.

In this study, rankings of 5 compositions according to 3 grading methods were investigated. Invariance composition rankings shouldn't be interpreted as there is no difference between individual student grades. For the further studies it will be suggested to focus on increasing the numbers of compositions and students for investigating the difference between the gradings and the rankings.

Kaynakça Bilgisi:

Ömür, S. ve Erkuş, A. (2013). Dereceli puanlama anahtarıyla, genel izlenimle ve ikili karşılaştırmalar yöntemiyle yapılan değerlendirmelerin karşılaştırılması. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi [Hacettepe University Journal of Education]*, *28*(2), 308-320.

Citation Information:

Ömür, S., & Erkuş, A. (2013). Comparison of the evaluations which were done with rubric,overall impression and paired comparisons [in Turkish]. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi [Hacettepe University Journal of Education]*, *28*(2), 308-320.