Hacettepe Üniversitesi Eğitim Fakültesi Dergisi [H. U. Journal of Education] 44: 237-248 [2013]

ÖĞRETİM ELEMANI DEĞERLENDİRMESİNDE KULLANILAN FORMATLARIN HALE ETKİSİ, CÖMERTLİK ETKİSİ VE KULLANICI TEPKİLERİ AÇISINDAN KARŞILAŞTIRILMASI

COMPARING STUDENT EVALUATION FORMATS IN TERMS OF HALO EFFECT, LENIENCY EFFECT AND USER REACTIONS

Afife Başak OK[*] H. Canan SÜMER[**] Reyhan BİLGİÇ[***]

ÖZ: Bu çalışmanın amacı öğrencilerin öğretim elemanı performansını değerlendirmede kullanılabilecek iki farklı değerlendirme formatını psikometrik özellikleri (hale ve cömertlik/aşırı olumluluk etkileri) ve kullanıcı tepkileri açısından karşılaştırmaktır. Bu amaçla geliştirilen davranış odaklı değerlendirme ölçeği (DODÖ) ve grafik değerlendirme ölçeği (GDÖ) toplam 270 üniversite öğrencisine (N_{BARS} = 126; N_{GRS} = 144) uygulanmış ve öğrencilerden gerçek hayattaki öğretim elemanlarının ders sırasındaki performansını değerlendirmeleri istenmiştir. Formatlardan hiçbiri psikometrik açıdan diğerinden belirgin bir şekilde üstün bulunmamakla birlikte GDÖ'nün hale etkisine dayanıklılık açısından DODÖ'den biraz daha iyi olduğu bulunmuştur. Ayrıca, GDÖ ile karşılaştırıldığında DODÖ kullanıcılardan daha az olumlu tepkiler almıştır. Davranış odaklı bir değerlendirme ölçeği geliştirmek için gereken farklı kaynaklar düşünüldüğünde GDÖ'nün DODÖ'ye tercih edilebileceği yönünde bulgular elde edilmiştir. Elde edilen bulguların sonuçları geleceğe yönelik önerilerle birlikte tartışılmaktadır.

Anahtar sözcükler: değerlendirme formatları, öğrenci değerlendirmeleri, kullanıcı tepkileri, hale etkisi, cömertlik etkisi

ABSTRACT: The purpose of this study was to compare two different rating scales to be used in student evaluation of instructors in terms of psychometric properties, mainly halo and leniency effects, and user reactions. A behaviourally anchored rating scale (BARS) and a graphic rating scale (GRS) were developed and administered to a total sample of 270 college students (N_{BARS} = 126; N_{GRS} = 144) rating their real-life instructors. Although neither format had a clear psychometric advantage over the other, the GRS format was found to be slightly better than the BARS format in terms of resistance to halo. Furthermore, the BARS format received less favourable user reactions than did the GRS format. Considering the differential resources required for the development of a behaviour-based rating scale, the present results indicated the GRS format as being preferable to the BARS format. The implications of the findings are discussed along with suggestions for future research.

Keywords: rating formats, student evaluations, user reactions, halo, leniency

1. GİRİŞ

Landy ve Farr'ın (1980) değerlendirme ölçeği formatı araştırmalarına ara verilmesi yönündeki çağrılarına rağmen kullanıcı tepkileri ve psikometrik kalite açısından daha iyi formatı arama çalışmaları devam etmektedir (Jelley ve Goffin, 2001; MacDonald ve Sulsky, 2009; Roch, Sternburgh ve Caputo, 2007; Tziner, Joanis ve Murphy, 2000; Tziner ve Kopelman, 2002). 1980'lerde ve 1990'lardaki değerlendirme formatı araştırmaları, farklı değerlendirme ölçeklerini psikometrik özellikleri (Tziner, 1984), performans hedeflerinin netliği ve kullanıcılar tarafından kabulü (Tziner ve Kopelman, 1988; Tziner, Kopelman ve Livneh, 1993), kullanıcı memnuniyeti (Tziner, Joanis ve Murphy, 2000; Tziner, Kopelman ve Joanis, 1997; Tziner, Kopelman ve Livneh, 1993), geribildirim doğruluğu (Jelley ve Goffin, 2001), adalet ve kurumsal adalet algıları (Roch ve ark., 2007) ve performanstaki izleyen değişimleri (Tziner, Kopelman ve Livneh, 1993) açısından karşılaştırmaya devam etmiştir.

[*] Yrd. Doç. Dr. A. Basak Ok., Istanbul Kemeburgaz Univ. Psikoloji Bol., okbasak@yahoo.com
[**]Prof. Dr. H. Canan Sümer, ODTÜ Psikoloji Böl., hcanan@metu.edu.tr
[***]Prof. Dr. Reyhan Bilgiç. ODTÜ Psikoloji Böl., rey@metu.edu.tr

MacDonald ve Sulsky'ye göre (2009), performans değerlendirmede kullanılan format; yapılan değerlendirmelerin kalitesi, değerlendirmelerin geribildirim ve gelişim amaçlı kullanılabilirliği, değerlendirici tepkileri ve değerlendirme hataları üzerinde etkilidir. Bu tespitler, değerlendirme formatları üzerinde halen niye çalışılması gerektiğinin anlaşılması açısından önemlidir. *Grafik Değerlendirme Ölçekleri* – GDÖ (*Graphic Rating Scales*) ve *Davranış Odaklı Değerlendirme Ölçekleri* – DODÖ (*Behavioral Anchored Rating Scales*) en yaygın olarak çalışılan formatlar arasındadır (bkz. Tziner ve ark., 2000).

1.1. Grafik Değerlendirme ve Davranış Odaklı Değerlendirme Ölçekleri

1920'lerde Donald Paterson tarafından tanıtılan GDÖ, en eski ve en yaygın olarak kullanılan değerlendirme formatıdır (Landy ve Farr, 1983). GDÖ'de, değerlendiricilerden değerlendirdikleri kişileri grafiksel/görüntüsel olarak üzerinde değerlendirmelerin yapıldığı yatay bir düzlemde birçok özellik ya da genel beceri boyutlarında değerlendirmeleri istenmektedir. Bu formatlarda/ölçeklerde, değerlendiricilerden ilgili özelliğin, sözkonusu olan kişide ne ölçüde bulunduğunu temsil eden rakamı daire içine alarak ya da kutucuğu işaretleyerek değerlendirmesi istenmektedir. Krzystofiak, Cardy ve Newman'a (1988) göre, GDÖ'nün yaygın olarak kullanılmasının nedeni kişilerin performans değerlendirirken belirgin davranışlardan çok geniş/genel özellikler ya da performans boyutları açısından düşünmeyi tercih ediyor olmaları olabilir. Ancak, Kline ve Sulsky'nin de (2009) belirttikleri gibi; GDÖ'de kullanılan ve soyut özellikler temelinde yapılan değerlendirmeler sözkonusu özelliğin iş davranışından çıkarım yapılmasını gerektirmektedir. Bu durum hem değerlendirmelerle çalışan davranışları arasında ilişki kurulmasını zorlaştırmakta (MacDonald ve Sulsky, 2009) hem de özellikle Kuzey Amerika'da yasal açıdan da birtakım sorunları beraberinde getirmektedir (Kline ve Sulsky, 2009). Değerlendirilecek boyutların ve ölçek noktalarının belirsizliği kullanıcıları GDÖ'den özellikle psikometrik açıdan daha üstün olması beklenen alternatif formatlara yöneltmiştir.

Smith ve Kendall (1963) tarafından geliştirilen DODÖ, geleneksel GDÖ'den birçok yönden farklıdır. Birinci olarak, DODÖ'de, belirsizliği azaltmak amacıyla değerlendirilecek boyutlar davranışsal olarak tanımlanmaktadır. İkinci olarak, sözkonusu boyut üzerinde farklı performans düzeylerini temsil eden ölçek noktalarında sıfatlar yerine spesifik davranışlar kullanılmaktadır. Üçüncü olarak, değerlendirilecek boyutları temsil eden ölçekler dikey olarak sunulmaktadır. Son olarak, davranış odaklı bir değerlendirme ölçeğinin geliştirilme sürecinde başlıca ilgili tarafların (olası değerlendirici ve değerlendirilenlerin) aktif katılımı gerekmektedir (Bernardin, 2005).

İlk araştırmalar, DODÖ'nün GDÖ'ye kıyasla, daha objektif ve doğru değerlendirmeler sağladığına (Borman ve Dunnette, 1975) ve hem hale hem de cömertlik etkisine daha az maruz olduğuna işaret etmektedir (Tziner, 1984). Bununla birlikte, izleyen araştırmalar, DODÖ'nün performans değerlendirmelerinde kullanılan diğer formatlardan tutarlı bir şekilde, psikometrik açıdan üstün olduğuna ya da daha doğru değerlendirmeler sağladığına yönelik net bilgiler vermemektedir. Ayrıca, görgül kanıtlar, DODÖ'ye dayalı performans değerlendirme sistemlerinin diğer formatlara dayalı sistemler ile karşılaştırıldığında her zaman daha olumlu kullanıcı tepkilerine neden olmadığını da göstermektedir (Tziner ve ark., 2002).

Tziner ve Kopelman (2002) üç değerlendirme formatının (Davranışsal Gözlem Ölçeği – DGÖ, DODÖ ve GDÖ) farklı kombinasyonlarla karşılaştırıldığı altı bağımsız çalışmayı gözden geçirmişlerdir. Bu araştırmacıların çalışması DGÖ'nin özellikle kullanıcı tepkileri, hedef netliği, hedef kabulü ve hedef/amaç bağlılığı açılarından diğer iki formattan üstünlüğünü göstermektedir. Ayrıca, bu yazarlar DGÖ ile GDÖ arasındaki farkın özellikle değerlendirici memnuniyeti ve hedef netliği gibi değişkenler açısından küçük olduğunu belirtmektedirler. Bu gözden geçirme, DODÖ'nün, özellikle GDÖ'ye kıyasla psikometrik açıdan bazı avantajlara (hale ve cömertlik etkisine daha az maruz olma gibi) sahip olmakla birlikte, beklentilerin tersine üç format arasında en az tercih edilen format olduğunu göstermiştir. Özellikle kullanıcı tepkileri açısından, GDÖ'nün DODÖ'den daha avantajlı olduğuna işaret eden bulgular mevcuttur. Tziner ve Kopelman okuyucuları kullanımı daha kolay ve geliştirilmesi daha az masraflı ve en azından kullanıcı memnuniyeti açısından daha olumlu sonuçlar veriyor gibi görünen GDÖ'ye karşı otomatik muhalefete karşı uyarmaktadırlar.

Bu çalışmada GDÖ ve DODÖ formatları, psikometrik özellikleri (hale ve cömertliğe açıklık) ve kullanıcı tepkileri açısından, araştırma amaçlı olarak, büyük bir devlet üniversitesinde öğrencilerin öğretim elemanı değerlendirmesinde karşılaştırılmıştır.

1.2. Öğretim Etkililiğinin Öğrenciler Tarafından Değerlendirilmesi

Öğrencilerin öğretme etkililiği değerlendirmesi ile ilgili olan yazın geleneksel olarak iki konuya odaklanmaktadır: öğrenci değerlendirmelerinin psikometrik kalitesi ve öğrenci değerlendirmelerinin boyutsallığı ("dimensionality"). Çalışmalar genel olarak öğrencilerin öğretim değerlendirmelerinin güvenirlik, tutarlık ve geçerliğini desteklemektedir (d'Appollonia ve Abrami, 1997; Madesen ve Napoles, 2006; Marsh ve Roche, 1997; Milanowski, 2004). Marsh ve Roche (1997) gözden geçirme çalışmalarında "Bir sınıfta yeterli sayıda öğrenci olması durumunda (ya da, belki, [aynı öğretim elemanının girdiği] farklı sınıfların ortalaması alındığında), öğrenci değerlendirmelerinin ortalamalarının güvenirliğinin, objektif ölçümlerle mukayese edilecek kadar iyi olduğunu" (s.1188) belirtmişlerdir. Benzer şekilde, d'Apollonia ve Abrami (1997) bu değerlendirmelerin geçerliği ile ilgili olarak kanıt sunmuşlardır.

Öğrenci değerlendirmelerinin psikometrik kalitesi üzerindeki çalışmalardan farklı olarak, değerlendirmelerin boyutsallığı konusunda bir fikir birliği bulunmamaktadır. Bazı araştırmacılar, eğitim becerilerini temsil eden genel, global bir faktörün varlığını tartışırken (örn., Abrami ve d'Appollonia, 1990; d'Appollonia ve Abrami, 1997), diğerleri değerlendirmelerin çok boyutluluğunu destekler yönde kanıt sunmaktadır (örn., Heckert, Latier, Ringwald ve Silvey, 2006; Marsh ve Roche, 1997). Şimdiye kadar yapılan çalışmalarda öğrenci değerlendirmelerinde format karşılaştırmasının genel olarak yapılmadığı ve özellikle de kullanıcı memnuniyeti üzerine odaklanılmadığı görülmektedir.

Bu çalışmanın amacı öğretim etkililiğinin boyutsallığını bir Türk üniversitesinde incelemek ve aynı zamanda iki değerlendirme formatını (DODÖ ve GDÖ) psikometrik özellikleri ve kullanıcı tepkileri açısından karşılaştırmaktır. Öğrencilerin öğretim elemanı değerlendirmelerindeki çok boyutlu yapının ortaya çıkmasının, kısmen kullanılan formata bağlı olduğunu düşünmekteyiz. Örneğin, DODÖ gibi davranış yönelimli bir format öğretim etkililiğinin boyutlarının ayrışmasını teşvik ederken, GDÖ gibi bir format değerlendirmelerin global performans boyutları üzerinde yapılması nedeniyle öğretim etkililiği ile ilgili tek bir faktörün ortaya çıkmasına neden olabilir. Tek boyutlu bir öğretim etkililiği faktörünün ortaya çıkması performansın kavramsal olarak farklı boyutlarını ayrıştırma yeteneğinin olmaması şeklinde tanımlanan hale etkisinin olduğunu gösteriyor olabilir (bkz. Cooper, 1981). Bu nedenle, DODÖ formatının hale[1] etkisine GDÖ formatından daha az açık olması beklenmektedir. Benzer şekilde, incelenen yazın temelinde, değerlendirmelerin spesifik davranışlar temelinde yapıldığı bir format olan DODÖ formatının, daha global değerlendirmeleri içeren GDÖ formatına kıyasla cömertlik etkisine (performansın hak ettiğinden çok daha olumlu olarak değerlendirilmesi eğilimi) daha az maruz olacağı da beklenmektedir.

Hale ve cömertlik etkilerine daha fazla maruz olmasını beklediğimiz GDÖ formatının, kullanıcı tepkileri açısından ise avantajlı olacağı düşünülmektedir.

Özetle, incelenen yazına dayanarak bu çalışmada iki hipotez öne sürülmüştür:

Hipotez 1: DODÖ formatı GDÖ formatından görece daha az hale ve cömertlik etkisine neden olacaktır

Hipotez 2: GDÖ formatı DODÖ formatına kıyasla, kullanıcılardan daha olumlu tepkiler alacaktır

2. YÖNTEM

Hipotezleri test etmek amacıyla iki çalışma yapılmıştır. İlk çalışma kritik öğretim elemanı davranışlarının ve performans boyutlarının tanımlanmasını ve öğrencilerin öğretim elemanı performansı değerlendirmesinde kullanacakları DODÖ ve GDÖ ölçeklerinin geliştirilmesini

[1] Bu noktada, "hale" etkisi ile kastedilen "*gerçek hale*" değil, performans boyutları arasında olmayan korelasyonu yansıtan "*illüzyon hale*"dir.

kapsamaktadır. İkinci çalışma ise bu iki ölçeğin psikometrik özellikleri ve kullanıcı tepkileri açısından karşılaştırmasını içermektedir.

2.1. Çalışma I

2.1.1. DODÖ'nün Geliştirilmesi

Geleneksel DODÖ geliştirme yaklaşımıyla tutarlı olarak DODÖ formatında yer alacak performans boyutlarını ve kritik öğretim elemanı davranışlarını belirlemek amacıyla kritik olaylar tekniği (Flanagan, 1954) kullanılmıştır. Bu süreçte üç grup katılımcı yer almıştır. Yaş ortalaması 21 olan ilk grupta (Grup I), sosyal bilimler ve mühendislik bölümlerinden seçkisiz olarak seçilen 252 üniversite öğrencisi yer almıştır (%36.51'i kadın ve %63.49'u erkek). Grup I, kendi deneyimlerine dayanarak olumlu ve olumsuz kritik öğretim elemanı davranışları örnekleri sağlamıştır. Bu süreç sonunda toplam 436 kritik olay toplanmıştır. Kritik olayların toplanmasını takiben bu olaylar araştırmacılar tarafından, farklı öğrenci değerlendirme formlarını da dikkate alarak, etkili öğretme becerilerini temsil ettiği düşünülen 10 performans boyutu (*konusuna hakimiyet, öğretmeye istekli olmak ve öğretmeyi sevmek, değerlendirmede adil olmak, derse hazırlıklı olmak, öğrencilerle iletişim becerileri, öğrenciyi yönlendirici, yardımcı ve destekleyici tutum sergilemek, öğrenciye ayrıntılı geribildirim vermek, ders anlatırken düzenli ve organize olmak, sorulara doyurucu cevap verebilmek* ve *iş ahlakı*) altında gruplanmıştır. Bir olayı belirlenen performans boyutlarından birinin altında tutmak için araştırmacıların hepsinin aynı fikirde olması kriteri kullanılmış ve bu kriteri sağlamayan olaylar elenmiştir.

Kritik öğretim elemanı davranışlarının ve performans boyutlarının belirlenmesini takiben iki psikoloji bölümü öğretim elemanı ve dört psikoloji bölümü yüksek lisans öğrencisi olmak üzere toplam altı değerlendirici (Grup II) kritik olayları belirlenen 10 performans boyutu altına yeniden yerleştirmiştir. Kritik olayların performans kategorileri içine yeniden yerleştirilmesi sırasında üzerinde %50'ın altında uzlaşma sağlanan 250 olay elenmiş ve geriye 180 olay kalmıştır.

Sonraki aşamada 30 psikoloji yüksek lisans öğrencisi (Grup III) her bir kritik olayın yer aldığı boyut altında temsil ettiği performans derecesini 7-basamaklı Likert tipi bir ölçek (1 = Çok düşük performans, 7 = Mükemmel performans) kullanarak değerlendirmiştir. Bu değerlendirmeleri takiben her bir performans boyutu altında yer alan her bir olayın ortalama ve standart sapması hesaplanmış ve standart sapması 1.4'ün üstünde olan olaylar elenmiştir[2] ve bu elemeler sonrası 63 olay muhafaza edilmiştir. Bu aşamaya ulaşan olayların incelenmesi sonunda bazı performans boyutları için (örn., *öğrenciyi yönlendirici, yardımcı ve destekleyici tutum sergilemek*), orta düzey performansı gösteren örneklerin olmadığı gözlenmiştir. Bu nedenle, araştırmacılar bu boyutlarda orta düzey performansı temsil eden (örn., "Öğrencilere karşı ılımlı ve hoşgörülü bir yaklaşımı vardır") yeni kritik öğretim elemanı davranışları geliştirmiş ve bunların görünüş geçerliğini sağlamıştır. Son olarak, her bir boyut altında kalan kritik olayların ortalamalarına bakılarak aynı 7-basamaklı ölçek kullanılarak kritik olaylar farklı performans düzeylerini temsil edecek şekilde boyutların altına yerleştirilmiştir. Aynı boyut altında ortalamaları çok yakın olan olaylardan sadece bir tanesi kullanılmıştır. Bu yerleştirme sonunda her bir boyut altında yer alan kritik olayların sayısı 4 ile 7 arasında değişmiştir.

2.1.2. GDÖ'nün Geliştirilmesi

GDÖ'nün geliştirilmesinde DODÖ'nün geliştirilmesi sırasında belirlenen aynı 10 performans boyutu ve tanımları kullanılmıştır. GDÖ'de, katılımcılar her boyutun tanımını okuyarak öğretim elemanını 7-basamaklı bir ölçek üzerinde (1 = Oldukça düşük performans sergiler, 4 = Orta düzey performans sergiler ve 7 = Mükemmel performans sergiler) değerlendirmişlerdir.

[2] Bu çalışmada kesme noktası olarak kullanılan 1.4 değeri orijinal olarak Schwab, Heneman ve DeCotiis (1975) tarafından önerilen ve 1.5 olarak belirlenen kesme noktasından daha katıdır.

2.2. Çalışma II

2.2.1. Katılımcılar

Çalışmanın örneklemini Ankara'da bir devlet üniversitesinin farklı bölümlerinde okuyan ve öğretim elemanı değerlendirmesine aşina olan 270 öğrenci oluşturmaktadır. Bu 270 öğrencinin 136'sı (%50.4) kadın, yaş ortalamaları 20.55 yıldır.

2.2.2. Ölçekler

GDÖ formatı DODÖ formatında bulunan 10 performans boyutunu içermektedir. Boyutlar ölçek noktalarının rakamlarla gösterildiği ve üç performans tanımlayıcısının (Kötü performans, Orta performans ve Çok iyi performans) yer aldığı 7-basamaklı yatay bir ölçek üzerinde değerlendirilmektedir.

Hem GDÖ hem de DODÖ'nün sonunda değerlendiricilerden kullandıkları formatı 10-basamaklı bir ölçek kullanarak kullanışlılık, uygunluk ve açıklamaların anlaşılırlığı/netliği açısından değerlendirmeleri istenmektedir. Kullanıcı tepkileri soruları şunlardır: Sizce değerlendirmede kullandığınız bu ölçek "Ne kadar kullanışlıydı?" "Amacına ne kadar hizmet ediyordu?" ve "Açıklamalar ne kadar yeterliydi?"

Aynı 10 performans boyutunu içeren DODÖ ve GDÖ, toplam 270 öğrenciye (DODÖ = 126; GDÖ = 144) ders saatlerinde uygulanmıştır. Katılım gönüllü olup katılımcılara değerlendirmelerinin gizli tutulacağı yönünde garanti verilmiştir.

3. BULGULAR

İki ölçekten alınan puanlara yönelik betimleyici istatistikler Tablo 1'de sunulmuştur. DODÖ ve GDÖ formatlarındaki performans boyutları arasındaki korelasyonlar ise sırasıyla Tablo 2 ve Tablo 3'te sunulmaktadır.

Tablo 1: DODÖ ve GDÖ Formatlarının Boyut Ortalama Değerlerine İlişkin Betimleyici İstatistikler

	DODÖ (N = 126)	GDÖ (N = 144)
Ort.	5.30	5.43
Medyan	5.30	5.50
S	.93	1.02
Ranj	4.80	5.50
Uca Kayma Eğilimi	-.57	-1.00

Not. Değerlendirmeler 7-dereceli bir ölçek kullanılarak yapılmıştır.

Tablo 2: DODÖ Boyutları Arasındaki Korelasyonlar

Boyutlar	1	2	3	4	5	6	7	8	9	10
1	1.000									
2	.559**	1.000								
3	.383**	.524**	1.000							
4	.520**	.556**	.517**	1.000						
5	.538**	.461**	.423**	.529**	1.000					
6	.490**	.428**	.274**	.412**	.623**	1.000				
7	.271**	.446**	.377**	.467**	.485**	.491**	1.000			
8	.266**	.489**	.322**	.380**	.245**	.366**	.461**	1.000		
9	.583**	.457**	.371**	.528**	.463**	.547**	.375**	.270**	1.000	
10	.274**	.238**	.390**	.360**	.211**	.296**	.320**	.383**	.347**	1.000

Not. (1) 1= Konusuna Hakimiyet, 2 = Derse Hazırlıklı Olmak, 3 = Değerlendirmede Adil Olmak, 4 = Öğretmeye İstekli Olmak ve Öğretmeyi Sevmek, 5 = Öğrenciyi Yönlendirici, Yardımcı ve Destekleyici Tutum Sergilemek, 6 = Öğrencilerle İletişim Becerileri, 7 = Öğrenciye Ayrıntılı Geribildirim Vermek, 8 = Ders Anlatırken Düzenli ve Organize Olmak, 9 = Sorulara Doyurucu Cevap Verebilmek, 10 = İş Ahlakı. *(2)* ** $p < .01$.

Tablo 3: GDÖ Boyutları Arasındaki Korelasyonlar

Boyutlar	1	2	3	4	5	6	7	8	9	10
1	1.000									
2	.644**	1.000								
3	.187*	.349**	1.000							
4	.284**	.365**	.481**	1.000						
5	.340**	.352**	.411**	.595**	1.000					
6	.359**	.349**	.300**	.520**	.581**	1.000				
7	.268**	.281**	.373**	.444**	.470**	.454**	1.000			
8	.357**	.396**	.287**	.376**	.305**	.403**	.449**	1.000		
9	.493**	.611**	.479**	.434**	.460**	.391**	.355**	.333**	1.000	
10	.411**	.444**	.499**	.443**	.381**	.304**	.312**	.249**	.536**	1.000

Not. (1) 1= Konusuna Hakimiyet, 2 = Derse Hazırlıklı Olmak, 3 = Değerlendirmede Adil Olmak, 4 = Öğretmeye İstekli Olmak ve Öğretmeyi Sevmek, 5 = Öğrenciyi Yönlendirici, Yardımcı ve Destekleyici Tutum Sergilemek, 6 = Öğrencilerle İletişim Becerileri, 7 = Öğrenciye Ayrıntılı Geribildirim Vermek, 8 = Ders Anlatırken Düzenli ve Organize Olmak, 9 = Sorulara Doyurucu Cevap Verebilmek, 10 = İş Ahlakı. *(2)* ** $p < .01$. * $p < .05$.

3.1. Hale Etkisi İçin Yapılan Analizler

Çalışmanın ilk hipotezi DODÖ formatı kullanılarak yapılan değerlendirmelerin GDÖ formatıyla yapılan değerlendirmelere görece olarak daha az hale ve cömertlik etkisine neden olacağı yönündeydi. Hale etkisini ölçmek için iki farklı yöntem kullanılmıştır: (1) faktör analizi yaklaşımı ve (2) boyutlar arası standart sapmaların ortalamalarının hesaplanması yaklaşımı.

Az sayıda faktör ya da bileşenin ortaya çıkması hale etkisinin varlığını akla getirmektedir (Saal, Downey ve Lahey, 1980). Bu çalışmada her iki format üzerinde de varimax rotasyonlu temel bileşenler analizi yapılmıştır. Bu analizlerin sonuçları Tablo 4 ve Tablo 5'te sunulmaktadır. Tablolardan da görülebileceği gibi her iki format için de Kaiser kriteri kullanıldığında iki faktör ortaya çıkmıştır. Ancak, her iki analizde de ilk faktör değerlendirmelerdeki varyansın önemli bir kısmını açıklamaktadır (DODÖ: 34.36% ve GDÖ: 32.39%). Her iki analizde de maddeler anlamlı bir şekilde gruplanmamış ve birçok madde her iki boyutta da yer almıştır. Her iki formatta da iki faktör ortaya çıkmasına karşın, formatların her ikisinde de değerlendirmelerin altında tek bir faktör var gibi görünmesi öğretme performansının tek boyutluluğunu desteklemektedir. İki formatın güvenirlik değerlerinin DODÖ ve GDÖ formatları için sırasıyla .88 ve .87 olduğu bulunmuştur.

Kısaca, faktör analizi sonuçlarına dayanarak her iki format da tek boyutluluk göstermektedir. Bu nedenle, faktör analizi yaklaşımı kullanıldığında her iki değerlendirme formatının da eşit derecede hale etkisine açık olduğunu söyleyebiliriz.

Tablo 4: DODÖ İçin Yapılan Temel Bileşenler Faktör Analizi Sonuçları

	Bileşen	
	1	2
Açıklanan Varyans	%34.36	%24.89
Maddeler		
Konusuna Hakimiyet	.819	
Öğrenciyi Yönlendirici, Yardımcı ve Destekleyici Tutum Sergilemek	.792	
Sorulara Doyurucu Cevap Verebilmek	.750	
Öğrencilerle İletişim Becerileri	.710	
Öğretmeye İstekli Olmak ve Öğretmeyi Sevmek	.615	.472
Derse Hazırlıklı Olmak	.598	.476
Ders Anlatırken Düzenli ve Organize Olmak		.799
İş Ahlakı		.718
Öğrenciye Ayrıntılı Geribildirim Vermek	.370	.636
Değerlendirmede Adil Olmak	.408	.554

Tablo 5: GDÖ İçin Yapılan Temel Bileşenler Faktör Analizi Sonuçları

	Bileşen	
	1	2
Açıklanan Varyans	%32.39	%26.03
Maddeler		
Öğretmeye İstekli Olmak ve Öğretmeyi Sevmek	.774	
Öğrenciyi Yönlendirici, Yardımcı ve Destekleyici Tutum Sergilemek	.749	
Öğrenciye Ayrıntılı Geribildirim Vermek	.735	
Öğrencilerle İletişim Becerileri	.722	
Değerlendirmede Adil Olmak	.592	
Ders Anlatırken Düzenli ve Organize Olmak	.540	
Derse Hazırlıklı Olmak		.845
Konusuna Hakimiyet		.832
Sorulara Doyurucu Cevap Verebilmek	.413	.701
İş Ahlakı	.379	.596

Hale etkisini incelemek için yapılan ikinci grup analizler her bir değerlendirici için boyutlar arası standart sapmaların hesaplanmasını ve bu standart sapmaların iki formattan birini kullanan değerlendiriciler üzerinden ortalamalarının alınmasını içermektedir. Başka bir deyişle, hale etkisinin bu şekilde ölçülmesinde, bir değerlendiricinin belli bir değerlendirilenin farklı performans boyutlarındaki değerlendirmelerinin standart sapmaları kullanılır (boyut değerlendirmeleri arasındaki fark ne kadar az ise hale etkisi o kadar fazladır). Bu yaklaşımda büyük standart sapmalar değerlendiricilerin farklı performans boyutları arasında ayrım yapabildiğini göstermektedir. Her bir değerlendirici için boyutlara ait değerlendirmelerin standart sapmalarını hesaplayabilmek amacı ile SPSS programı kullanılarak değişkenler (10 performans boyutu) ve katılımcılar yer değiştirmiştir. Her bir değerlendirici-değerlendirilen kombinasyonu için boyut değerlendirmelerinin standart sapmaları hesaplanarak iki değerlendirme formatından birini kullanan değerlendiriciler için bu standart sapmaların ortalaması hesaplanmıştır. Ortalama standart sapma değerleri GDÖ (1.13) için DODÖ'ye kıyasla (1.01) daha yüksek bulunmuştur. Bu bulgu, beklenenin tersine, GDÖ kullanıcılarının genel olarak boyutlar arası daha fazla ayırım yapabildiğine işaret etmektedir. Gözlenen bu farkın, istatistiksel olarak anlamlı olup olmadığını test edebilmek amacıyla bir tek yönlü varyans analizi (one-way ANOVA) yapılmıştır. Bu analizde veri noktaları olarak her bir değerlendirme için hesaplanan SS'lar kullanılmıştır. Bu analiz sonunda, beklenenin tersine, GDÖ formatının, boyutlar arasında daha değişken değerlendirmeler yapılmasına izin verdiğini gösterir bir biçimde anlamlı bir fark elde edilmiştir [F (1, 268) = 5.19, p < .02].

Hale etkisini değerlendirmek için yapılan analizler sonucunda bulgular çalışmanın birinci hipotezini desteklememiştir. Yani, hale etkisine dayanıklılık açısından DODÖ'nün GDÖ'den daha iyi olduğu bulunmamıştır. Aksine, boyutlar arası standart sapmaların ortalamaları üzerine yapılan analizler DODÖ ile karşılaştırıldığında GDÖ'nün hale etkisine daha az maruz olduğunu göstermiştir.

3.2. Cömertlik Etkisi İçin Yapılan Analizler

İdeal olarak cömertlik etkisini ölçmek için birden fazla sayıda değerlendirileni değerlendiren bir tane değerlendirici olmalıdır (Murphy ve Cleveland, 1995). Ancak, bu çalışmadaki her uygulamada aynı kişinin (öğretim elemanı) farklı değerlendiriciler (öğrenciler) tarafından değerlendirilmesi söz konusu olduğu için cömertlik etkisini doğrudan test etmek mümkün olmamıştır. Cömertlik etkisinin dolaylı bir ölçümü olarak öncelikle her iki formatın ortalama değerleri incelenmiş ve değerlendirmelerdeki format etkilerini incelemek amacıyla tek yönlü bir varyans analizi yapılmıştır. İkinci olarak, DODÖ ve GDÖ'nün değerlendirme dağılımlarındaki uca kayma eğilimi ("skewness") incelenmiştir. Tablo 1'de görüldüğü gibi, DODÖ için ölçek ortalama değeri 5.30, GDÖ için ise 5.43'tür. Değerlendirmeler üzerinde yapılan tek yönlü varyans analizi sonunda format temel etkisinin anlamlı olmadığı bulunmuştur, F (1, 268) = 1.04, p > .05.

Saal ve arkadaşlarına (1980) göre, bir değerlendirici tarafından farklı kişiler için yapılan değerlendirmelerin, dağılımın olumlu ucuna yığılması cömertlik etkisini yansıtmakta, değerlendirmelerin dağılımın olumsuz ucuna yığılması da katılık etkisini ("severity") göstermektedir. Bu çalışmadaki uçlara yığılma değerleri, aynı kişiyi aynı formatı kullanarak değerlendiren değerlendiricilerin değerlendirmeleri için elde edilmiş olup DODÖ için –.57 ve GDÖ için –1.00 olarak bulunmuştur. Bu bulgu, DODÖ formatı ile karşılaştırıldığında GDÖ formatının daha fazla cömertlik etkisine açık olduğu şeklinde yorumlanabilir. Ancak, bu çalışmada cömertlik etkisinin alışılageldik/geleneksel işevuruk tanımları kullanılamadığı için cömertlik etkisi için yapılan analiz sonuçlarının temkinli yorumlanması gerekmektedir.

Özet olarak, DODÖ formatı özellikle hale etkisi açısından GDÖ'den psikometrik olarak daha üstün değerlendirmelere neden olmamış, ancak cömertlik etkisinin bir indeksi açısından daha avantajlı olduğu görülmüştür.

Hipotez 2, DODÖ formatı ile karşılaştırıldığında GDÖ formatının değerlendiriciler tarafından daha fazla kabul göreceği yönündedir. Kullanıcı tepkileri maddeleri arasında görece yüksek korelasyonların olması (.42 ile .72 arasında değişmekte) kullanıcı tepkilerini ölçen üç tutum sorusunun aynı yapıyı ölçtüğünü göstermektedir. Bu nedenle, bu üç sorunun ortalaması alınarak genel bir "kullanıcı tepkisi" ya da "kullanıcı tutumu" değişkeni oluşturulmuştur. Oluşturulan bu kullanıcı tutumu değişkeni üzerine yapılan varyans analizi anlamlı bir ölçek etkisi olduğunu göstermektedir [F (1, 268) = 18.31, p < .001]. Bu analizler, Hipotez 2'yi desteklemekte olup GDÖ'nün (Ort = 5.43) DODÖ'den (Ort = 5.30) daha olumlu kullanıcı tepkileri aldığını göstermektedir.

4. TARTIŞMA ve SONUÇ

Bu çalışmada farklı formatlardaki iki değerlendirme ölçeği psikometrik özellikleri ve kullanıcı tepkileri temelinde incelenmiştir. Psikometrik özellikler açısından formatlardan hiçbirinin tutarlı olarak diğerinden üstün olmadığı bulunmuştur. Hale etkisi açısından GDÖ formatı DODÖ formatına nazaran biraz daha avantajlı olmakla birlikte, cömertlik etkisi açısından DODÖ formatının GDÖ formatından daha avantajlı olduğu görülmüştür. Kullanıcı (değerlendirici) tepkileri açısından ise beklendiği gibi, GDÖ daha olumlu tepkiler almıştır.

Çalışmanın sonuçları genel olarak DODÖ formatına nazaran GDÖ formatı ile ilgili olarak daha olumludur. Daha fazla cömertlik etkisine sahip olmakla birlikte GDÖ ile yapılan değerlendirmelerin hale etkisine daha az açık olduğu görülmektedir. Bunun bir nedeni bu çalışmadaki GDÖ'nün geliştirilme şekliyle ilgili olabilir. GDÖ'ye yöneltilen en önemli eleştirilerden biri genellikle boyutların ve ölçek noktalarının açık/net bir şekilde tanımlanmaması, soyut ve muğlak olmasıdır (bkz. Murphy ve Cleveland, 1995). Bu çalışmada kullanılan GDÖ ise, genel olarak bu formata atfedilen kısıtlılıklardan bazıları azaltılarak geliştirilmiştir. Örneğin, DODÖ formatında yer alan ayrıntılı boyut tanımlarının aynısı GDÖ formatında da kullanılmıştır. Diğer bir deyişle, bu çalışmada kullanılan GDÖ tipik, geleneksel bir GDÖ değildir; performans boyutlarının tanımları soyut değil, davranış temellidir. Boyutların görece açık/net bir şekilde tanımlanması, değerlendiricilerin değerlendirmelerini yaparken kendi "örtük kuramlarına" (bkz. Krzystofiak ve ark., 1988) daha az başvurmalarına neden olmuş olabilir. GDÖ'nün hale etkisine karşı gözlenen direnci bu şekilde açıklanabilir.

Çalışmanın ikinci hipotezi ve de Tziner ve Kopelman'ın (2002) bulguları ile tutarlı olarak, GDÖ formatı değerlendiricilerden daha olumlu kullanıcı tepkileri almıştır. GDÖ'ye yönelik olarak tespit edilen olumlu tepkilerin olası bir açıklaması bu formatın basitliği ve gelenekselliği olabilir. Değerlendiricilerin alternatif ölçek formatlarına aşinalığının önemli bir konu olduğu belirtilmektedir (Kingstrom ve Bass, 1981). DODÖ'nün değerlendiricilerden daha az olumlu tepkiler almasının bir nedeni, değerlendiricilerin bu formata yabancı olması olabilir. Bu çalışmada öğrenci olan değerlendiriciler, öğretim elemanlarını değerlendirmeye alışkın oldukları halde, dikey olarak sunulan ve aslında GDÖ'den (7 sayfa) daha uzun olan DODÖ (13 sayfa) formatı ile muhtemelen ilk kez karşılaşmışlardı. GDÖ formatı daha aşina bir format olup muhtemelen daha az bunaltıcı ve bilişsel olarak daha az çaba gerektiriyor gibi görünmüş olabilir.

Özetlemek gerekirse, ilgili performans boyutlarının açık ve net bir şekilde tanımlandığı bir GDÖ'nün, daha çok zaman, uzmanlık ve enerji gerektiren DODÖ'den (kullanıcı tepkileri ve hale etkisine dayanıklılık açısından) daha etkin bir değerlendirme aracı olduğu bulunmuştur.

Bu çalışmadaki öncelikli amaç iki değerlendirme formatını psikometrik kriterler ve kullanıcı tepkileri açısından karşılaştırmak olmasına rağmen, bulguların öğretim elemanı değerlendirmesi yazını için de bazı doğurguları bulunmaktadır. Giriş bölümünde tartışıldığı gibi öğretim etkililiği boyutsallığı ile ilgili olarak karşıt görüşler mevcuttur. Öğretim etkililiğinin hem tek boyutluluğunu (d'Apollonia ve Abrami, 1997) hem de çok boyutluluğunu (Heckert ve ark., 2006; Lowman, 1994; Marsh ve Roche, 1997) destekler nitelikte kanıtlar bulunmaktadır. Bu çalışmada, her iki format kullanılarak elde edilen değerlendirmeler üzerinde yapılan temel bileşenler analizi öğretim etkililiğinin tek boyutlu olduğunu desteklemektedir. Her iki formda da yer alan boyutların kavramsal olarak farklı olması amaçlanmış olduğundan bu bulgu hale etkisinin bir göstergesi olarak yorumlanmıştır. Ancak, alternatif olarak, öğretim elemanı değerlendirmelerinin tek boyutlu olması, değerlendirmelerde psikometrik açıdan bir yanlılığa işaret etmeyebilir. Gözlenen tek boyutluluk; yanlılığı temsil eden hale etkisi ("illusory halo") yerine, boyutlar arası gerçekte var olabilecek korelasyonları yansıtan gerçek halenin ("true halo") bir göstergesi olabilir (bkz. Cooper, 1981).

Gerçek haleyi, *hata* ya da *yanlılığı* yansıtan haleden ayrıştırmanın bir yolu, konu ile ilgili iş uzmanlarından uygun koşullarda elde edilen gerçek performans puan tahminlerini ("true score estimates") kullanmaktır. Gerçek puan tahminleri, gözlemlenen değerlendirmelerin psikometrik kalitesinin, yanlılığa maruz olma derecesinin ve de doğruluğunun tespitinde bir referans noktası olarak kullanılmaktadır (bkz. Murphy ve Cleveland, 1995). Gerçek puan tahminlerinin yansıttığı boyutlar arası korelasyonlar ile gözlenen değerlendirmelerin yansıttığı boyutlar arası korelasyonlar, yapılan değerlendirmelerdeki hatayı yansıtan hale yanlılığını tespit etmede kullanılabilir. Örneğin, eğer gerçek puanların analizi kullanılan format için çok boyutlu bir faktör yapısına işaret ediyor, gözlenen değerlendirmeler ise tek boyutlu bir yapıya işaret ediyorsa, bu bulgu o formatta yapılan değerlendirmelerde hale hatasının varlığını işaret ediyor olabilir. Benzer şekilde, gerçek puan tahminleri değerlendirmelerdeki cömertlik etkisi miktarını değerlendirmede de kullanılabilir. Özetle, gerçek puan tahminlerinin kullanımının öğrenci değerlendirmelerindeki hale etkisi, boyutluluk ve cömertlik etkisinin daha iyi test edilmesine imkan tanıdığı bir gerçektir. Ancak, gerçek puan tahminlerinin elde edilmesi, doğal ortamdan uzak, video ya da kâğıt ortamında yaratılan sanal kişilerin iş uzmanlarınca değerlendirildiği katı laboratuar koşullarını zorunlu kılmaktadır. Alanda yapılan bu çalışmanın deseni değerlendirme kalitesini ölçmede gerçek puan tahminlerinin kullanılmasına izin vermemiştir.

Çalışmanın bir diğer sınırlılığı cömertlik etkisini test etmede daha uygun analizlerin kullanılamamış olmasıdır. Cömertlik etkisinin uygun şekilde sınanması için, aynı değerlendirici(ler) tarafından değerlendirilen çok sayıda değerlendirilenin olması gerekmektedir. Bu çalışmada ise aynı öğretim elemanını (değerlendirileni) değerlendiren birden çok öğrenci (değerlendirici) bulunmaktadır. Bu nedenle, cömertlik etkisinin yeterli ya da doğrudan ölçülmediği söylenebilir. Aynı değerlendiricilerin birden çok kişiyi değerlendirdikleri gelecek çalışmalarda, cömertlik etkisinin daha doğrudan ölçülmesi mümkün olabilecektir.

Son olarak, çalışmanın vurgulanması gereken birtakım güçlü yanları mevcuttur. İlk olarak, halen yaygın olarak kullanılan iki değerlendirme formatının karşılaştırıldığı bu çalışmanın, değerlendirme formatları yazınına katkıda bulunması beklenmektedir. İkinci olarak, bu çalışmanın yöntemsel olarak sağlam bazı yönleri olduğu düşünülmektedir. Örneğin, hem DODÖ hem de GDÖ geliştirilme süreçleri son derece sistematik ve titizlikle yürütülmüştür. Ayrıca, değerlendirme hatalarının birden çok indeks kullanılarak ölçülmesinin yöntemsel bir avantaj sağladığı düşünülmektedir. Son olarak, bu çalışmaya gerçek öğretim elemanlarından ders alan oldukça fazla sayıda gerçek öğrenci katıldığı için ve her bir öğretim elemanı için her iki değerlendirme formatı seçkisiz olarak dağıtıldığı için, çalışmanın sonuçlarının küçümsenmeyecek bir genellenebilirliğe sahip olduğu söylenebilir.

Çalışmanın Kayıt Tarihi : 22.03.2010
Yayına Kabul Edildiği Tarih : 09.08.2012

5. KAYNAKLAR

Abrami, P. C., & d'Apollonia, S. (1990). The dimensionality of ratings and their use in personnel decisions. In M. Theall & J. Franklin (Eds.), *Student ratings of instruction: Issues for improving practice* (pp. 97-111). San Francisco: Jossey-Bass.

Bernardin, H. J. (2005). Behaviorally anchored rating scales. In *Blackwell Encyclopedic Dictionary of Human Resource Management* (pp. 22-23). Oxford, England: Blackwell Publishing.

Borman, W. C., & Dunnette, M. D. (1975). Behavior-based versus trait-oriented performance ratings: An empirical study. *Journal of Applied Psychology, 60(5)*, 561-565.

Cooper, W. (1981). Ubiquitous halo. *Psychological Bulletin, 90*, 218-244.

D'Apollonia, S. & Abrami, P. C. (1997). Navigating student ratings of instruction. *American Psychologist, 52(11)*, 1198-1208.

Flanagan, J. C. (1954). The critical incident technique. *Psychological Bulletin, 51*, 327-358.

Heckert, T. M., Latier, A., Ringwald, A., & Silvey, B. (2006). Relation of course, instructor, and student characteristics to dimensions of student ratings of teaching effectiveness. *College Student Journal, 40(1)*, 195-203.

Jelley, R. B. & Goffin, R. D. (2001). Can performance-feedback accuracy be improved? Effects of rater priming and rating-scale format on rating accuracy. *Journal of Applied Psychology, 86(1)*, 134-144.

Kingstrom, P. O., & Bass, A. R. (1981). A critical analysis of studies comparing behaviorally anchored rating scales (BARS) and other rating formats. *Personnel Psychology, 34*, 263-289.

Kline, T. J. B., & Sulsky, L. M. (2009). Measurement and assessment issues in performance appraisal. *Canadian Psychology, 50(3)*, 161-171.

Krzystofiak, F., Cardy, R., & Newman, J. (1988). Implicit personality and performance appraisal: The influence of trait inferences on evaluations of behavior. *Journal of Applied Psychology, 73(3)*, 515-521.

Landy, F. J., & Farr, J. L. (1980). Performance rating. *Psychological Bulletin, 87(1)*, 72-107.

Landy, F. J., & Farr, J. L. (1983). *The measurement of work performance: Methods, theory, and applications.* New York: Academic Press.

Lowman, J. (1994). Professors as performers and motivators. *College Teaching, 42(4)*, 137-141.

Madesen, C. K. & Napoles, J. (2006). A 30-year follow-up study of perceptions of students' ratings of former instructors. *UPDATE: Applications of Research in Music Education, 24(2)*, 45-53.

Marsh, H. W. & Roche, L. A. (1997). Making students' evaluations of teaching effectiveness effective: The critical issues of validity, bias, and utility. *American Psychologist, 52(11)*, 1187-1197.

MacDonald, H. A., & Sulsky, L. M. (2009). Rating formats and rater training redux: A context-specific approach for enhancing the effectiveness of performance management. *Canadian Journal of Behavioural Science, 41(4)*, 227-240.

Milanowski, A. (2004). The relationship between teacher performance evaluation scores and student achievement: Evidence from Cincinnati. *Peabody Journal of Education, 79(4)*, 33-53.

Murphy, K. R., & Cleveland, J. N. (1995). *Understanding performance appraisal: Social, organizational, and goal-based perspectives.* Thousand Oaks, CA: SAGE Publications.

Roch, S. G., Sternburgh, A. M., Caputo, P. M. (2007). Absolute vs. relative performance rating formats: Implications for fairness and organizational justice. *International Journal of Selection and Assessment, 15(3)*, 302-316.

Saal, F. E., Downey, R. G., & Lahey, M. A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin, 88(2)*, 413-428.

Schwab, D. P., Heneman III, H. G., & DeCotiis, T. A. (1975). Behaviorally anchored rating scales: A review of the literature. *Personnel Psychology, 28*, 549-562.

Smith, P. C., & Kendall, L. M. (1963). Retranslation of expectations: An approach to the construction of unambiguous anchors for rating scales. *Journal of Applied Psychology, 47(2)*, 149-155.

Tziner, A. (1984). A fairer examination of rating scales when used for performance appraisal in a real organizational setting. *Journal of Occupational Behaviour, 5*, 103-112.

Tziner, A., Joanis, C., & Murphy, K. R. (2000). A comparison of three methods of performance appraisal with regard to goal properties, goal perceptions and ratee satisfaction. *Group and Organization Management, 25*, 175-190.

Tziner, A., & Kopelman, R. E. (2002). Is there a preferred performance rating format? A non-psychometric perspective. *Applied Psychology: An International Review, 51(3)*, 479-503.

Tziner, A. & Kopelman, R. E. (1988). Effects of rating format on goal-setting dimensions: A field experiment. *Journal of Applied Psychology, 73*, 323-326.

Tziner, A., Kopelman, R. E., & Joanis, C. (1997). Investigation of raters' and ratees' reactions to three methods of performance appraisal: BOS, BARS, and GRS. *Canadian Journal of Administrative Sciences, 14*, 396-404.

Tziner, A., Kopelman, R. E., & Livneh, N. (1993). Effects of performance appraisal format on perceived goal characteristics, appraisal process satisfaction, and changes in rated job performance: A field experiment. *Journal of Psychology, 127*, 281-292.

Extended Abstract

Despite the famous call for a moratorium on rating scale format research by Landy and Farr (1980), the search for a better format seems not to have ceased (e.g., Jelley & Goffin, 2001; MacDonald & Sulsky, 2009; Roch, Sternburgh, & Caputo, 2007; Tziner & Kopelman, 2002). According to MacDonald and Sulsky (2009), the rating format used in performance appraisal is influential on the quality of ratings, acceptance of feedback, rating errors, and rater and ratee reactions. These points are important for understanding the reasons behind the continuing emphasis on rating formats.

Rating format comparisons have not consistently shown superiority of one format over the others despite some early evidence suggesting that the behaviourally anchored rating scale (BARS) format provided more objective and accurate ratings than the graphic rating scale (GRS) format (e.g., Borman & Dunnette, 1975), and that the BARS format was less prone to both halo and leniency errors (Tziner, 1984). Furthermore, empirical evidence indicates that a BARS-based appraisal system does not necessarily lead to more favourable user reactions than the systems based on the other formats (Tziner et al., 2002).

The literature on student evaluation of teaching effectiveness has traditionally focused on two issues: psychometric quality of student evaluations and dimensionality of student ratings. Studies in general support the reliability, stability, and validity of the students' evaluations of teaching (SET) ratings (d'Appollonia & Abrami, 1997; Madesen & Napoles, 2006; Marsh & Roche, 1997; Milanowski, 2004). However, there seems to be a disagreement concerning the dimensionality of the student ratings.

Our aim in the present study was to compare two rating formats (BARS and GRS) used in evaluating teaching effectiveness in terms of psychometric qualities and user (student) reactions. Hence, based on the reviewed literature, we hypothesized that: (1) the BARS format leads to relatively less halo and leniency in ratings than the GRS format; (2) the GRS format is likely to receive more favourable reactions from the raters than the BARS format.

Two studies were conducted to test the hypotheses. The first study consisted of identification of critical instructor behaviors and performance dimensions and the development of a BARS and a GRS to be used in university students' evaluations of course instructors. The second study involved administering these two scale formats and comparing them in terms of their psychometric properties and user reactions.

The BARS format consisted of 10 items each representing a performance dimension. The performance dimensions were: Knowledge in the Subject Matter, Motivation to Teach, Fairness in Evaluating Students, Being Prepared for the Lecture, Communication Skills, Relations with Students, Feedback Giving, Planning and Organization, Answering Student Questions, and Work Ethics. Each dimension included a 7-point vertically presented scale anchored by numbers and critical incidents representing examples of poor, average, and excellent teaching performance on that dimension.

The GRS format included the same 10 performance dimensions included in the BARS format. The dimensions were rated on a 7-point horizontally presented scale anchored by numbers and three performance descriptors (poor, average, and excellent performance). At the end of both formats, raters were asked to evaluate the format that they had used via three questions measuring user-friendliness, suitability, and the clarity of instructions, using a 10-point scale.

The BARS and the GRS, each including the same 10 performance dimensions, were administered to 270 students (BARS = 126; GRS = 144) during regular class meetings. The results of the present study in general were more favourable towards the GRS than the BARS. Although more lenient, the ratings done with the GRS were relatively less prone to halo. Furthermore, as expected and similar to the findings of Tziner and Kopelman (2002), raters/students seemed to be more satisfied with the GRS format than the BARS format. One plausible explanation for the observed positive reactions towards the GRS could be its simplicity and conventionality. The GRS format was simply a more familiar format, which probably looked less overwhelming and cognitively demanding.

To summarize, considering the time, energy, and effort required to develop a BARS, the present results indicate that a GRS with clearly defined dimensions capturing the domain of interest is likely to generate more positive results. The findings of our study have some implications for student evaluations of teaching effectiveness literature as well. The principal component analyses conducted on the ratings obtained using both formats provided support for the unidimensionality of teaching effectiveness. There are two limitations of the present study. Design of our study did not allow us to use true score estimates in assessing rating quality. Also, in the present study, leniency was not measured adequately or directly.

The findings of the present study are expected to contribute to the rating format literature as we examined two of the commonly used rating formats. Furthermore, since a substantial number of real life students taking courses from real life instructors participated in the present study and the two scale formats were distributed randomly in each group (for each instructor), the results of this study are believed to be somewhat generalizable.