Hacettepe Üniversitesi Eğitim Fakültesi Dergisi (H. U. Journal of Education) 39: 271-283 [2010]

ÜNİVERSİTE ÖĞRENCİLERİNİN ÖĞRETİM ELEMANINA İLİŞKİN DEĞERLENDİRMELERİNDE NOTUN ETKİSİ

THE EFFECT OF GRADE ON COLLEGE STUDENTS' EVALUATION OF THE TEACHER

Sevgi ÖZGÜNGÖR*

ÖZET: Bu çalışmanın amacı öğretim elemanının not verirken katı ya da müsamahalı davranmasının öğrencilerin öğretim elemanına ilişkin değerlendirmeleri üzerindeki etkilerinin incelenmesidir. Bu amaçla 144 birinci sınıf öğrencisine öğretim elemanına ilişkin değerlendirilmelerini belirlemek amacıyla Ders Değerlendirme Ölçeği uygulanmıştır. Daha sonra öğrenciler biri zor, diğeri kolay sorulardan oluşan ve deneysel olarak yapılandırılan iki farklı sınavdan birisine tabi tutulmuşlardır. Deneysel işlem sonrası öğrencilerin öğretim elemanına ilişkin değerlendirmeleri tekrar alınmış ve değişim olup olmadığını araştırmak amacıyla tekrarlı ölçümler için ANOVA uygulanmıştır. Bulgular öğrencilerin öğretim elemanına ilişkin ön değerlendirmeleri ile son değerlendirmeleri arasında farklılıklar olduğunu ve ölçeğin iyi öğretmen ve düşünme becerileri boyutlarında son değerlendirmelerin her iki grup içinde daha düşük olduğunu ortaya koymuştur. Bulgular aynı zamanda müsamahalı bir şekilde değerlendirilen gruptaki öğrencilerin öğretim elemanının değerlendirme yöntemini daha adil bulmaya başladıklarını ve doyum düzeylerinin değişmediğini buna karşılık katı bir şekilde değerlendirilen gruptaki öğrencilerin dersten aldıkları doyum düzeyinin düştüğünü ve öğretim elemanının değerlendirme yöntemlerini adil bulmadıklarını göstermiştir.

Anahtar sözcükler: öğretmen değerlendirmeleri, not, üniversite öğrencisi.

ABSTRACT: The purpose of this study was to determine the effect of grading leniency on students' evaluations. For this purpose, 144 freshman students responded to CEQ to evaluate instructor's teaching behaviors. Then, students were randomly assigned to one of the two different experimentally constructed exams. The students in the first group received an easy exam and were graded leniently. On the other hand, the second group had hard questions and were graded strictly. Following experimental treatment, students reevaluated their instructors. Repeated measures ANOVA was used to determine whether students' evaluations changed based on how leniently they were graded or not. According to the findings, students gave lower evaluations points for good teaching, generic skills and evaluations subscales of CEQ after the treatment. Also, students in strictly graded groups reported lower levels of course satisfaction and thought their instructors' evaluations methods were unfair.

Keywords: students' evaluations, grading leniency, college student.

1. GİRİŞ

Resmi olarak ilk kez 1920'lerde Washington Üniversitesi'nde uygulamaya konulduklarından beri (Kulik 2001) tüm dünyada yaygın olarak kullanılmalarına karşın öğrenci değerlendirmeleri pek çok önemli bilimsel tartışmanın da konusu haline gelmiştir. Bu tartışmaların önemli bir nedeni kuramsal olarak öğretim etkinliğiyle alâkası olmayan pek çok değişkenin öğrencilerin öğretim elemanının performansına ilişkin değerlendirmeleri ile pozitif korelasyon göstermeleridir. Söz konusu değişkenler arasında öğrencilerin derse ilişkin ilgi düzeyleri ve derse verilen değer (Heckert, Latier, Ringwald-Burton & Drazen 2006; Marsh & Dunkin 1992), öğretim elemanının cinsiyeti (Feldman 1978; Griffin 2001; McKeachie 1997), öğrencinin yaşı (Grimes, Millea & Woodruff 2004), dersin saati (Koushki & Kuhn 1982), dersin alım nedeni (seçmeli mi, zorunlu mu) (Feldman 1978; Marsh 1987), sınıf mevcudu (Feldman 1978), öğretim elemanının kişiliği (Radmacher & Martin 2001), öğrencinin denetim odağı (Grimes, Millea & Woodruff 2004), dersin zorluk derecesi ve beklenen not (örn., Chambers & Schmitt 2002; Greenwald & Gillmore 1997) gibi değişkenler yer almaktadır.

Bu değişkenlerden dersin zorluk derecesi ve öğrencinin yılsonu notuyla da yakından ilişkili olan müsamahalı not verme (grading leniency) konusu öğrenci değerlendirmelerinin geçerliliğine ilişkin tartışmaların ilgi odağı haline gelmiştir. Bu tartışmanın özünü öğrenci değerlendirmelerinin öğretim elemanının gerçek davranışlarından çok dersin zorluk derecesi ve öğrencilerin notlarına ilişkin beklentileri gibi duygusal etkenler tarafından belirlendiği ve bu yüzden kolay geçiren öğretim

*Yard. Doç. Dr., Pamukkale Üniversitesi, Eğitim Bilimleri Bölümü, sozgungor@hotmail.com

elemanlarının gerçeği yansıtmayan pozitif değerlendirmeler alacağı görüşü oluşturmaktadır. Bu görüşün savunucuları öğrencilerin kendilerini kolay geçiren öğretim elemanlarını ödüllendirmek amacıyla daha pozitif değerlendireceklerini öne sürmekte ve bu yüzden yüksek öğretimde kalitenin düşmesine neden olabilecek iki temel sorundan bahsetmektedirler. İlk olarak, öğretim elemanının pozitif geribildirim karşılığında not verirken esnek davranması gerek öğretim elemanına öğrenci değerlendirmeleri aracılığıyla sunulan gerekse öğrencilere notlarıyla sağlanan objektif ve yapıcı geribildirimin çarpıtılarak öğrenci değerlendirmelerinin yüksek öğretimde kaliteyi artırma işlevini kaybetmesine neden olması tehlikesidir (Baummeistr 1996; DeBoer, Anderson & Elfessi 2007; Walwoord & Anderson, 1998). İkinci ve ilk etkene bağlı olarak, kolay not alabilme öğrencilerin daha az çaba ve öğrenme ile daha yüksek not alabilmelerini sağlayacağından üniversitede eğitim kalitesinin gittikçe azalmasına neden olacaktır (Eiszler 2002). Nitekim Bonesronning'in (1999) çalışması daha zor geçiren öğretim elemanlarının öğrencilerinin öğrenme düzeylerinin daha yüksek olduğunu göstermiştir.

Öğretim elemanlarının not verirken müsamahalı davranmalarının öğrencilerin öğretim elamanına ilişkin değerlendirmeleri üzerindeki etkisine açıklık getirmek amacıyla geçmiş yıllardan günümüze gerek korelasyona dayalı gerekse deneysel pek çok çalışma yapılmıştır. Bu konudaki ilk deneysel çalışmalardan birinde Holmes (1972) öğrencilerin yarısının notlarını bir harf aşağı çekmiştir (A alanların puanları B'ye, B alanların ki C'ye indirilmiştir). Holmes yüksek not alan öğrencilerin öğretim elemanlarını daha pozitif değerlendirdiklerini bildirmiştir. Bu çalışmayı takiben Powell'ın (1977) yarı deneysel çalışmasında aynı öğretim elemanı 5 dönem boyunca farklı sınıflara ders vermiş ancak bazı sınıflarda not verirken daha katı davranırken diğer sınıflarda daha müsamahalı davranmıştır. Sonuçlar öğretim elemanının not verirken daha müsamahalı davrandığı sınıflardaki öğrencilerin dersten daha çok doyum aldıklarını belirttiklerini göstermiştir. Worhington ve Wong (1979) öğrencilerin gerçekte hak ettikleri notları ile kendilerine deneysel olarak verilen sahte notların değerlendirmeleri üzerindeki etkilerini karşılaştırmışlardır. Bu amaçla öğrencileri bir sınava tabi tuttuktan sonra aldıkları gerçek puanlara göre iyi, kötü ve orta olmak üzere 3 gruba ayırmışlardır. Bu gruplar yine öğrencilerin hak ettikleri notlardan bağımsız olarak araştırmacılar tarafından rastlantısal olarak atanan notlara dayalı olarak kötü, yeterli ve iyi olmak üzere 3'er gruba ayrılmışlardır (toplam 9 grup). Bulgular, öğrencilerin değerlendirmelerinde gerçekten hak ettikleri notlar açısından önemli bir fark görülmediğini belirtmişlerdir. Ancak aynı çalışmada öğrencilere rastgele atanan puanlara göre iyi not alan öğrencilerin kötü ve yeterli not alan öğrencilere göre öğretim elemanlarını daha arkadaşça buldukları ve notları ile gerçek yetenekleri arasındaki ilişkinin doğru olduğuna ve yeteneklerinin kaynağının öğretim elemanı olduğunu belirttikleri ortaya çıkmıştır. Çalışmanın çarpıcı bir bulgusu öğretim elemanını değerlendirmeye ilişkin 18 maddeden 17'si sahte not dağılımı sonucunda yeterli puan alanlarca kötü not alanlardan daha pozitif olarak değerlendirilirken, aynı 18 maddeden sadece 5'inin gerçekten iyi not alan öğrenciler tarafından gerçekten kötü not alan öğrencilerden daha pozitif olarak değerlendirilmesidir.

Öğrenci notlarının öğrenci değerlendirmelerini etkilediğine ilişkin bulgular sunan bu öncü çalışmalar daha sonraki çalışmalar tarafından desteklenmediği gibi ağır bir şekilde eleştirilmiştir. Howard ve Maxwell (1980) öğrenci motivasyonunun, öğrencinin performansındaki gelişimine dair algılarının ve beklenen notun öğretim elemanlarına ilişkin değerlendirmeler üzerindeki etkilerini inceledikleri path çalışmalarında beklenen notun değerlendirmeler üzerindeki etkisinin öğrenci motivasyonu ve öğrenci motivasyonuna dayalı performans gelişimine dair algılara kıyasla çok düşük olduğunu bildirmişlerdir.

Öğrenci değerlendirmelerinin en önemli savunucularından biri olan Marsh (1980) her ne kadar beklenen notlar ile öğrenci değerlendirmeleri arasında pozitif bir ilişki olsa da bu ilişkinin önemli bir nedeninin öğrencinin gerçek öğrenme düzeyiyle açıklanabileceğini öne sürmüştür. Marsh ve Roche (1997) deneysel çalışmaların sonuçlarının öğrencilerin notlarına ilişkin yanlış beklentiler oluşturmanın etik olmaması ve sonuçların genellenebilirliğinin sınırlanmasına neden olduğu, öğrencilerin gruplara rastgele atanmadığı ve etki büyüklüğünün hesaplanmaması ya da çok küçük değerlerde olması gibi nedenlerle geçerli sayılamayacağını öne sürmüşlerdir. Centra'nın (2003) öğrenci değerlendirmelerinin öğrencilerin bekledikleri yılsonu başarı puanları ile ilişkisi yanında öğrencilerin öğrenme düzeyleri ve ne kadar öğrendiklerine ilişkin algılarıyla da ilişkilerinin incelendiği ve beklenen not ile

değerlendirmeler arasındaki ilişkilerin ancak .11 iken değerlendirmeler ile öğrenme düzeyleri arasındaki ilişkinin manidar ve güçlü olduğunu gösteren çalışması Marsh'ın görüşlerini destekler niteliktedir. Aynı zamanda deneysel çalışmaların bulgularını sentezleyen pek çok literatür tarama çalışmasında (örn., Cohen 1981, Alemoni 1999; Marsh & Dunkin 1997; Wachel 1998; Feldman 1997) söz konusu ilişkilerin .10 ile .30 arasında değiştiği, bu yüzden gerçek öğrenme ile öğretim elemanının performansı arasındaki ilişkiye kıyasla önemsiz olduğu sonucuna ulaşılmıştır.

Ancak son yıllarda tekrarlanan daha yapılandırılmış ve güçlü çalışmalarda da kolay not vermenin öğrenci değerlendirmelerini etkilediğine dair bulgular elde edilmektedir. Greenwald ve Gillmore (1997) Washinton Üniversitesinde sunulan 200 civarında dersin sorumlularına ilişkin değerlendirmeleri veri olarak kullanarak beklenen başarı düzeyi ile değerlendirmeler arasındaki ilişkileri yapısal eşitlik modeli aracılığıyla açıklamaya yönelik çalışmalarında öğrencilerin daha yüksek notlar aldıkları ve/veya daha kolay olarak tanımladıkları derslerin öğretim elemanlarını daha olumlu yönde değerlendirdiklerini bildirmişlerdir. Ellis, Burke, Lomire ve McCormack (2003) geçmiş çalışmalarda küçük örneklemlerle elde edilen bulguların genellenebilirliğinin ölçülmesine olanak veren 5602 öğrenci değerlendirmesinin analiz edildiği çalışmalarında A ve B alan öğrencilerin sayısının fazla olduğu sınıflarda değerlendirmelerin pozitif yönde artarken, C, D ve F alan öğrencilerin çoğunlukta olduğu sınıflarda olumsuz değerlendirmelerin arttığını ortaya çıkarmışlardır.

Özetle, geçmiş çalışmaların müsamahalı notun öğretim elemanına ilişkin değerlendirmeleri üzerindeki etkisine açıklık getirmekten çok yazın alanını ikiye böldüğü görülmektedir. Geçmiş yıllardaki deneysel çalışmaların bulguları müsamahalı notun öğrenci değerlendirmelerini pozitif yönde değiştireceği görüşünü desteklerken, son yıllardaki korelasyona dayalı çalışmalar öğrenci değerlendirmeleri ile müsamahalı not arasındaki pozitif ilişkiyi etkili öğretim elemanlarının öğrencilerinin konuyu öğrenmelerinin doğal bir göstergesi olduğu şeklinde yorumlama eğilimindedirler. Oliveras (2001) yazın alanındaki bu çelişkinin bir nedeninin daha önceki çalışmalarda not verirken müsamahalı davranmanın öğrencilerin yılsonu not beklentisiyle eş değer tutulmasından kaynaklandığını belirtmektedir. Öğrencinin yılsonu not beklentisinin düşük olması ders sorumlusunun belirlediği katı standartların sonucu olabileceği gibi öğrencinin derse karşı ilgisizliği ya da yetersiz çalışma becerileri gibi farklı nedenlere de bağlı olabilir. Nitekim Oliveras'ın müsamahalı notun etkilerini öğrencilerin öğretim elemanının not verme davranışlarının diğer öğretim elamanlarına kıyasla değerlendirildiği çalışmasında öğrencilerin müsamahalı not algılarının öğretim elemanına ilişkin değerlendirmelerinin manidar bir yordayıcısı olduğu ortaya çıkmıştır. Aynı kaygıları dile getirdiği benzer bir çalışmada Griffin (2004) öğrencilerin öğretim elemanının not verirken ne kadar müsamahalı davrandığına ilişkin algılarının öğretim elemanlarına ilişkin değerlendirmeler üzerinde negatif etkileri olduğunu göstermiştir. Bu yüzden bu konuya açıklık getirmek amacıyla öğretim elemanının not verirken öğrencileri ne kadar zorladığına ilişkin algıların öğrencilerin kişisel ihtiyaçlarından bağımsız objektif yöntemlerle belirlendiği ek çalışmalara ihtiyaç duyulmaktadır.

Yazın alanında öğrenci değerlendirmeleri ile müsamahalı not verme arasında görülen korelasyona ilişkin çelişkili sonuçların bir diğer nedeni, bu korelasyonun öğretim elemanına ilişkin değerlendirmelerinin çok boyutlu yapısının göz ardı edilerek analiz edilmeleri olabilir (Marsh 1983). Müsamahalı notun öğrenci değerlendirmeleriyle ilişkilerini inceleyen çalışmalar genellikle öğrenci doyumu ya da öğrencinin derse verdiği değer ile yılsonu başarı puanı beklentisi arasındaki ilişkilere bakmaktadır. Ancak, öğrenci değerlendirmeleri gerçekten geçerli bir ölçüm aracı ise müsamahalı notun etkisinin öğretim elemanının dersi sunmasındaki başarısıyla ilişkili davranışlarından çok dersi değerlendirme davranışlarıyla ilişkili olması beklenir.

Bu çalışmanın amacı müsamahalı not vermenin öğrenci değerlendirmeleri üzerindeki etkilerini müsamahalı not algısının objektif olarak oluşturulduğu bir sınıf ortamında öğrenci değerlendirmelerinin çok boyutlu yapısını dikkate alarak incelemektir. Bu amaçla öğretim elemanının not vermeye ilişkin standartları manipüle edilerek öğrencilerin öğretim elemanına ilişkin değerlendirmelerindeki değişimler ölçülmüştür. Eğer not ile öğrenci değerlendirmeleri arasındaki ilişki öğrencilerin dersi gerçekten öğrenmeleri sonucu ise öğrenci değerlendirmelerinin not verme algısı manipüle edildikten sonra değişmemesi gerekir. Aynı zamanda öğrenci değerlendirmeleri öğretim elemanının gerçek performansını yansıtan geçerli bir ölçüm aracı ise öğrenci değerlendirmelerinin yılsonu başarı puanlarıyla ilişkili olması beklenir. Buna karşılık eğer yazın alanında gözlenen öğrenci

değerlendirmeleri ile not arasındaki pozitif ilişkilerin nedeni öğrencinin müsamahalı not veren öğretim elemanını daha pozitif değerlendirme eğilimine bağlı ise deneysel işlem öncesi elde edilen öğrenci değerlendirmelerinin deneysel işlemi takiben elde edilen değerlendirmelerden farklılaşması ve yılsonu başarı puanlarının sadece deneysel işlem öncesi elde edilen değerlendirmelerle ilişkili olması beklenir. Geçmiş çalışmalar deneysel çalışmalardaki değişimin öğrencinin haksızlığa uğradığına ilişkin algılarına bağlı olduğunu ifade etmektedirler. Değerlendirmelerin çok boyutlu yapısı göz önünde bulundurulduğunda, eğer deneysel işlem sonucu ortaya çıkan farklılıklar öğrencinin haksızlığa uğradığına ilişkin algılara dayalı ise, öğrenci değerlendirmelerine ilişkin değişimlerin öğretim elemanının dersi anlatma becerisini yansıtan iyi öğretmen ya da düşünme becerileri alt boyutlarında değil, adil değerlendirme alt boyutuyla sınırlı olması gerekir.

Özetle bu çalışmanın amacı aşağıdaki denenceleri test etmektir:

a) Not verilirken katı davranılan grubun öğretim elamanına ilişkin ön ve son değerlendirmeleri ile müsamahalı davranılan grubun ön ve son değerlendirmeleri arasında fark olacaktır.

b) Öğrencilerin öğretim elemanına ilişkin değerlendirmelerindeki değişim sadece adil değerlendirme alt boyutuyla sınırlı olmayıp diğer alt boyutlarda da gözlenecektir.

c) Öğrencilerin öğrenme düzeyini gösteren yılsonu başarı puanları sadece deneysel işlem öncesinde elde edilen ön değerlendirmelerle ilişkili iken deneysel işlem sonucunda elde edilen notlar deneysel işlem sonrasında elde edilen son değerlendirmelerle ilişkili olacaktır.

2. YÖNTEM

2.1. Denekler

Araştırmada, öğrencilerin öğretim elemanı hakkındaki değerlendirmeleri üzerinde öğretim elemanı ile daha önceki deneyimleri gibi notların dışındaki etkenlerin olası etkilerini minimum tutabilmek amacıyla öğretim elemanı ile daha önceki derslerde deneyimleri olan tüm öğrenciler araştırma kapsamı dışında bırakılmıştır. Bu yüzden çalışmanın denekleri 2008-2009 bahar yarıyılında Pamukkale Üniversitesi Eğitim Fakültesinde eğitim yaşamlarına başlayan ve aynı öğretim elemanının verdiği Eğitim Psikolojisi dersini alan 6 farklı sınıftaki toplam 147 birinci sınıf öğrencisinden (106 kız ve 41 erkek) oluşmaktadır.

2.2. Materyal

Ders Deneyim Ölçeği. Öğrencilerin öğretim elemanının ders işleyişi ve öğrencinin öğrenmesini sağlamaya yönelik başarı düzeylerine ilişkin algılarını ölçen Ders Deneyim Ölçeği (The Course Experience Questionnaire (CEQ)) Ramsden (1991) tarafından geliştirilmiş ve Türkçeye uyarlama çalışmaları Özgüngör (2009) tarafından yapılmıştır. Geçerlik ve güvenirlik çalışmaları pek çok araştırma tarafından ortaya konan ve pek çok çalışmada kullanılan (örn., Ainley & Long 1994; Byrne & Flood, 2003; Lawless & Richardson, 2002; Ramsden 1991; Wilson, Lizzio & Ramsden 1997) ölçeğin uyarlama çalışmaları iyi öğretmen, düşünme becerileri, açık beklenti ve standartlar ve adil değerlendirme alt boyutu olmak üzere 4 alt boyut olduğunu ortaya çıkarmıştır. Ölçeğin adil değerlendirme alt boyutu için yüksek puanlar olumsuz değerlendirmeleri, düşük puanlar ise olumlu değerlendirmeyi ifade etmektedir. Diğer alt boyutlarda yüksek puanlar olumlu değerlendirmeyi ifade etmektedir. Beşli Likert tipi ölçeğin aynı zamanda öğrencinin derse ilişkin genel doyum düzeyini ölçen alt boyutu bulunmaktadır. Bu çalışmada alt boyutlara ilişkin iç tutarlılık güvenirlik katsayıları; Düşünme Becerileri için .85, İyi Öğretmen için .83, Adil Değerlendirme için .64 ve Açık Beklenti ve Standartlar için .68 ve genel doyum düzeyi için .89 olarak bulunmuştur.

2.3. İşlem

Öğretim elemanı öğretim yarıyılının ilk dersinde öğrencilere fakültede genel olarak sınavların bir vize bir finalden oluştuğunu, ancak öğrencilerin üniversitedeki sınav sistemine alışmalarını kolaylaştırdığı ve dersin ne kadar anlaşıldığı konusunda yararlı geri bildirim sağladığı için söz konusu derste iki vize bir final uygulaması yapılacağını söylemiştir. Öğrencilerin bu konudaki kaygıları

öğrencilerle tartışılmış ve iki vizenin daha yararlı olacağı sonucuna varılmıştır. Takip eden haftalarda dersler alışılageldiği şekilde işlenmiştir. Vizeden bir hafta önceki dersin başlangıcında araştırmacı tarafından kendisine veri toplamada yardımcı olması rica edilen ancak öğrencilerin kendisi adına veri topladığına inandırıldıkları bir öğretim elemanı, öğrencilere öğretim elemanlarının etkinliği konusunda bir çalışma için değerlendirmelerine ihtiyacı olduğunu ve diğer sınıflarda da veri topladığını belirtmiştir. Öğrencilerden değerlendirmelerini şu anda hangi öğretim elemanı ders veriyorsa O'nun ders içindeki davranışlarını göz önünde bulundurarak yapmalarını istemiştir. Öğrencilere aynı zamanda bu çalışmanın henüz ölçek oluşturma aşamasında bulunduğu, bu yüzden iki hafta sonra ölçeğin test-tekrar güvenirliliğini belirlemek amacıyla tekrar uygulanacağı belirtilmiş ve her ne kadar bireysel olarak değerlendirilme yapılmayacak olsa da iki ölçeğin eşleştirilebilmesi için kâğıt üzerine sakıncası yoksa numaralarını yazmaları istenmiştir. Takip eden haftada öğrenciler vizeye girmişlerdir. Vizede deneysel etkiyi yaratarak müsamahalı notun öğrenci değerlendirmeleri üzerindeki etkilerini inceleyebilmek amacıyla her sınıfta sorulan soruların zorluk derecesi tamamıyla farklı olan iki grup oluşturulmuştur. Bu amaçla her sınıf için bir gruptaki sorular öğretim elemanının geçmiş yıllardaki sınavlarının en zorlarından oluşurken, diğer sınav en kolay soruların derlemesinden oluşturulmuştur. Vize sırasında sınav türleri öğrencilere tamamıyla rastgele bir yöntemle dağıtılmıştır. Bu amaçla ilk sırada oturan öğrenci zor sorulardan oluşan sınavı alırken arkasındaki öğrenci kolay sorulardan oluşan sınavı, bir arkadaki öğrenci zor sorulardan oluşan sınavı alacak şekilde dağıtım yapılmıştır. Hem zor sınav hem de kolay sınav öğrenilen kavramların günlük yaşamdaki örneklerini içeren çoktan seçmeli sorulardan oluşmaktadır (ödeve katkısı hiç olmayan en yakın arkadaşını kırmaktan çekinen Ayla'nın bu davranışı Kohlberg'in ahlak gelişim evrelerinden hangisini örnekler? gibi). Benzer şekilde sınav okunurken de bir gruptaki öğrencilere tam doğru olmayan sorularda puan verilmezken, diğer gruptaki öğrencilerin puanlandırılmasında daha hoşgörülü olunmuştur. Sınav sonuçları takip eden ilk dersin başında açıklanmıştır. Aynı dersin son 20 dakikasında öğrencilere iki hafta önce derse gelen araştırmacının ölçeğinin test-tekrar güvenirliğini belirlemek amacıyla ikinci ölçümü yapacağı belirtilmiş ve öğrencilerin öğretim elemanı hakkındaki değerlendirmeleri tekrar alınmıştır. Veri toplanmasının ardından öğrencilere sınavın asıl amacı ve doğası açıklanmış, notlarının finale etki etmeyeceği, ancak finalde karşılarına çıkabilecek soruların aşırı uçlardaki örnekleri olduğu belirtilmiştir. Öğrencilerin çoğu çok şaşırdıklarını ve böyle bir şeyi hiç beklemediklerini belirtmişlerdir. Öğrencilerden kandırmaca için özür dilenmiş ve sonuçların isteyenlerle paylaşılacağı belirtilmiştir.

3. BULGULAR

Analizin ilk aşamasında deneysel işlem öncesi müsamahalı not verilen grup ile not verilirken katı standartlar uygulanan grubun, öğretim elemanına ilişkin değerlendirmelerine ait ön test puanları açısından farklılıklar olup olmadığını belirlemek amacıyla bağımsız örneklem grupları için uygulanan "t" testi analizi yapılmıştır. Analiz sonuçları Tablo 1'de verilmiştir.

Tablo 1: Gruplara Göre Öğretim Elemanına İlişkin Ön Değerlendirmelerin Aritmetik Ortalama, Standart Sapma ve T Değerleri

Değerlendirmelere Ait Alt Boyutlar	Gruplar	n	$\overline{X}$	Std. sapma	t	p
Düşünme Becerileri	**Kolay**	74	3.79	.61	.68	.49
	Zor	70	3.73	.58		
İyi öğretmen	**Kolay**	74	4.10	.62	.80	.42
	Zor	70	4.01	.72		
Adil Değerlendirme	**Kolay**	74	2.00	.61	-1.13	.25
	Zor	70	2.12	.57		
Açık Beklenti ve Standartlar	**Kolay**	74	3.20	.59	.34	.72
	Zor	70	3.16	.57		
Genel Doyum	**Kolay**	74	4.27	.89	1.01	.31
	Zor	70	4.11	.94		

Tablo 1'de görüldüğü gibi t testi sonuçlarına göre deneysel işlem öncesi alınan ön test puanları açısından iki grup arasında öğretim elemanına ilişkin hiçbir alt boyutta anlamlı bir farklılık yoktur. Dolayısıyla araştırmanın farklı gruplarındaki öğrencilerin deneysel işlem öncesi öğretim elemanına ilişkin algılarının birbirine denk olduğu söylenebilir.

Analizin ikinci aşamasında araştırmanın ilk denencenin sınanması amacıyla öğrencilerin deneysel işlem öncesi ve sonrasındaki puanları arasında farklılık olup olmadığı ve olası farklılıkların gruplara göre değişip değişmediğini belirlemek amacıyla tekrarlı ölçümler için ANOVA testi kullanılmıştır. Bu analize ait aritmetik ortalama ve standart sapmalar Tablo 2'de, ANOVA sonuçları ise Tablo 3'de verilmiştir. Ancak açık beklenti ve standartlar alt boyutuna ilişkin analizlerin hiçbir aşamasında anlamlı sonuçlar elde edilmediğinden çalışmanın bundan sonraki bölümünde bu alt boyuta ilişkin sonuçlara yer verilmemiştir.

Tablo 2: Öğretim Elemanına İlişkin Ön ve Son Değerlendirmelerin Aritmetik Ortalama ve Standart Sapma Değerleri

Değerlendirmelere Ait Alt Boyutlar		Gruplar	$\overline{X}$	Std. sapma	n
Düşünme Becerileri	ÖNTEST	Kolay	3.79	.61	74
		Zor	3.73	.58	70
	SONTEST	Kolay	3.72	.57	74
		Zor	3.58	.63	70
İyi Öğretmen	ÖNTEST	Kolay	4.10	.62	74
		Zor	4.01	.72	70
	SONTEST	Kolay	3.84	.60	74
		Zor	3.61	.68	70
Adil Değerlendirme	ÖNTEST	Kolay	2.00	.61	74
		Zor	2.12	.57	70
	SONTEST	Kolay	1.90	.52	74
		Zor	2.29	.60	70
Doyum	ÖNTEST	Kolay	4.27	.89	74
		Zor	4.11	.94	70
	SONTEST	Kolay	4.26	.88	74
		Zor	3.86	1.02	70

Beşli Likert tipi ölçeğin tüm alt boyutları için alınabilecek minimum puan birken alınabilecek maksimum puan beşti. Tablo 2'deki değerlerden anlaşılacağı üzere, öğretim elemanının tüm alt boyutlardan aldığı puanlar hem ön test, hem son testlerde ortalamanın üzerindedir. Deneysel işlem sonrası gruplar arasında fark olup olmadığını belirlemek amacıyla yapılan ANOVA sonuçlarına göre işlemler (ön test ve son test) arasındaki fark düşünme becerileri, iyi öğretmen ve doyum açısından anlamlı düzeydedir. Gruplar (müsamahalı ve katı not) arasındaki fark sadece adil değerlendirme için anlamlıdır. Aynı zamanda adil değerlendirme ve genel doyum düzeyi için grup işlem ortak etkisi (grupxişlem) anlamlıdır. Etki büyüklüğünü gösteren Eta kare değerlerini incelediğimizde; en büyük değişimin toplam varyansın yaklaşık %30'unu açıklayan iyi öğretmen alt boyutunda olduğunu, iyi öğretmeni adil değerlendirme ve düşünme becerileri genel doyum düzeyinin takip ettiği görülmektedir.

Son olarak, öğrencilerin ön ve son değerlendirmelerinin yılsonu başarı puanları ve deneysel işlem sonrası elde edilen başarı puanlarıyla ilişkilerini belirlemek amacıyla korelasyon analizi uygulanmış ve sonuçlar Tablo 4'de verilmiştir. Tablo 4'de görüldüğü gibi öğrencilerin yılsonu başarı puanları öğretim elemanına ilişkin deneysel işlem öncesine ait değerlendirmelerin iyi öğretmen ve adil değerlendirme alt boyutları ile anlamlı korelasyonlara sahiptir. Ancak yılsonu başarı puanları deneysel işlem sonrası elde edilen değerlendirmelerin hiçbiriyle ilişkili değildir. Buna karşılık, deneysel olarak elde edilen sınav puanları öğretim elemanına ilişkin ön değerlendirmelerden sadece değerlendirme alt boyutuyla ilişkili iken, deneysel işlem sonrası toplanan değerlendirmelerin tüm alt boyutlarıyla anlamlı ilişkilere sahiptir.

Tablo 3: Gruplara Göre Öğretim Elemanına İlişkin Değerlendirmelere Ait Ön Test-Son Test ANOVA Sonuçları

Alt Boyutlar	Varyansın Kaynağı	KT	Sd	KO	F	Eta değeri
Düşünme Becerileri	Grup (D/K)	.78	1	.78	1.22	.00
	Ölçüm (Ön-Son)	.87	1	.87	10.45**	.07
	Ortak etki	.01	1	.01	1.07	.01
	Hata	91.53	142	.64		
İyi öğretmen	Grup (D/K)	1.82	1	1.82	2.46	.01
	Ölçüm (Ön-Son)	7.70	1	7.70	59.71***	.30
	Ortak etki	.34	1	.34	2.64	.01
	Hata	105.37	142	.74		
Adil Değerlendirme	Grup (D/K)	4.62	1	4.62	8.61**	.05
	Ölçüm (Ön-Son)	.10	1	.10	.75	.01
	Ortak etki	1.43	1	1.43	10.70***	.07
	Hata	76.15	142	.53		
Doyum	Grup (D/K)	5.59	1	5.59	3.46	.02
	Ölçüm (Ön-Son)	1.20	1	1.20	8.41**	.06
	Ortak etki	1.08	1	1.08	7.55**	.05
	Hata	229.50	142	1.61		

**p< .005
*** p<.005

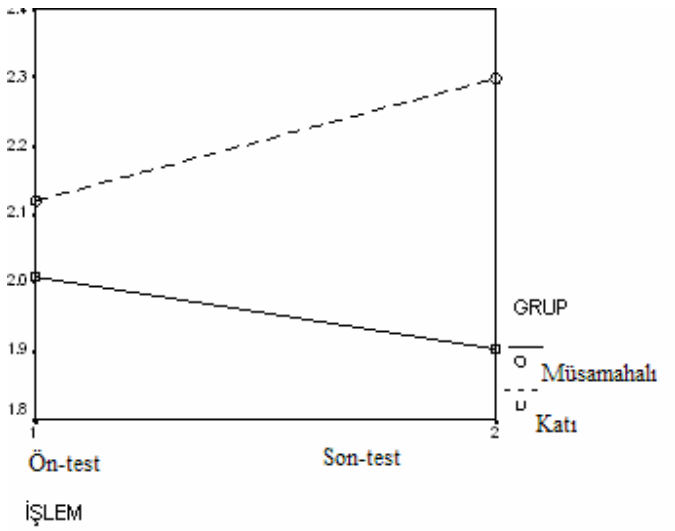
Tablo 4: Öğretim Elemanına İlişkin Ön ve Son Değerlendirmelerin Deneysel Notlarla ve Yılsonu Başarı Puanlarıyla İlişkileri

		Öğretim Elemanına İlişkin Değerlendirmeler			
		Düşünme Becerileri	İyi Öğretmen	Adil Değerlendirme	Doyum
Ortalama	Öntest	.14	.21**	-.12**	.20
	Son test	.12	.123	-.15	.17
Not	Öntest	.15	.150	-.24**	.18
	Son test	.24**	.28***	-.39***	.30***

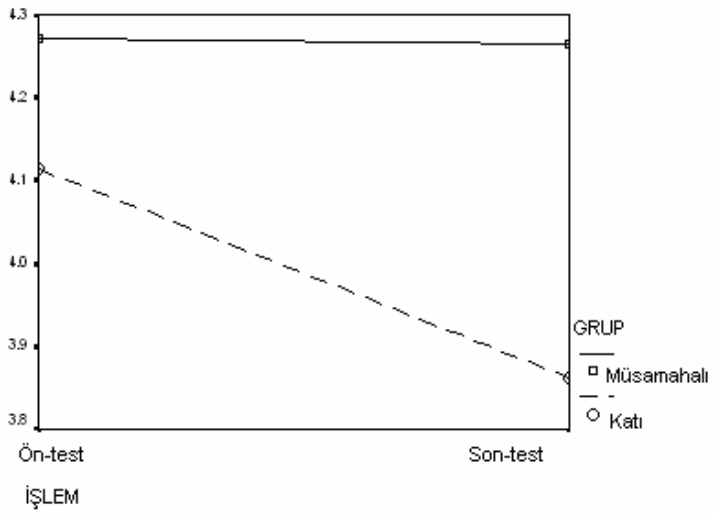
**p< .005
*** p<.005

Ayrıca, etkileşim terimlerine baktığımızda, adil değerlendirme için işlemxgrup etkileşimini temsil eden Şekil 1'de görüldüğü gibi müsamahalı not verilen grubun öğrencilerinin deneysel işlem sonucu öğretim elemanının değerlendirmelerini daha adil bulmaya başladıkları, katı bir şekilde değerlendirmeye tabi tutulan grubun öğrencilerinin ise öğretim elemanının değerlendirmelerini deneysel işlem öncesine nazaran daha az adil olarak algılanmaya başladıkları görülmektedir. Benzer şekilde, genel doyum düzeyi açısından işlemxgrup ortak etkisini gösteren Şekil 2'de görüldüğü gibi

müsamahalı not verilen grubun genel doyum düzeyleri sabit kalırken katı standartlar uygulanarak not verilen grubun öğrenciler deneysel işlem sonrası dersten daha az doyum aldıklarını bildirmişlerdir.



Şekil 1: Adil Değerlendirmeye İlişkin İşlemxgrup Ortak Etkisi



Şekil 2: Dersten Alınan Genel Doyum Düzeyine İlişkin İşlemxgrup Ortak etkisi

4. YORUM / TARTIŞMA

Bu çalışmanın amacı öğretim elemanının not verirken kullandığı standartların öğrencilerin öğretim elemanına ilişkin değerlendirmeleri üzerindeki etkisinin belirlenmesidir. Bulgular bu amaçla deneysel olarak oluşturulan grupların, deneysel işlem öncesi ve sonrasında öğretim elemanına ilişkin değerlendirmelerinde farklılıklar olduğunu ortaya koymuştur. Müsamahalı bir şekilde not verilen gruptaki öğrenciler öğretim elemanının öğrenci başarısını değerlendirmeye ilişkin davranışlarını daha olumlu algılamaya başlarken, katı bir şekilde not verilen grubun öğrencilerinin öğretim elemanının değerlendirmelerinde adil davranmadığına ilişkin algıları artmıştır. Diğer yandan, müsamahalı not verilen grubun öğrencilerinin dersten aldıkları doyum düzeyi değişmezken, katı bir şekilde değerlendirmeye tabi tutulan öğrencilerin deneysel işlem sonrası dersten aldıkları genel doyum düzeyinin düştüğü görülmüştür. Dahası deneysel işlem sadece öğretim elemanının değerlendirme davranışlarına ilişkin değerlendirmeleri değiştirmemiş, öğrencilerin öğretim elemanının ders içeriğini

sunmadaki başarısını ölçen iyi öğretmen alt boyutu ile öğrencilerin düşünme ve karar becerilerini geliştirmedeki başarısını ölçen düşünme becerileri alt boyutlarına ilişkin algılarının da olumsuz yönde değişmesine sebep olmuştur. Daha da çarpıcı olarak en büyük değişim öğretim elemanının öğrenci başarısını değerlendirmesine ilişkin davranışlarıyla hiç ilgisi olmayan iyi öğretmen alt boyutunda gözlenmiş, bu değişimi takiben öğretim elemanının öğrencilerin düşünme yeteneklerini geliştirici davranışlarına ilişkin algıları ve değerlendirmeye ilişkin algıları yer almıştır.

Bu bulgular müsamahalı notun öğrencilerin öğretim elemanlarına ilişkin değerlendirmelerini olumlu yönde etkilediğini gösteren önceki yıllardaki deneysel çalışmaların (örn; Holmes 1972; Powell 1977) ve müsamahalı notun değerlendirmeler üzerindeki etkisinin beklenen not algısı yerine doğrudan müsamahalı not algısının ölçülerek incelendiği daha güncel korelasyona dayalı çalışmaların (örn., Oliveras 2001; Griffin 2004) bulgularını destekler niteliktedir. Bu bulgulara ek destek araştırmanın üçüncü denencesine ait analizlerden gelmektedir. Analizler, öğrencilerin yılsonu başarı puanlarının deneysel işlem öncesi değerlendirmeleriyle ilişkili olduğunu, deneysel işlem sonrası değerlendirmelerin ise öğrencilerin gerçek başarılarıyla değil de deneysel işlem sonucu elde edilen puanlarla ilişkili olduğunu ortaya koymuştur. Bu bulgular Worhington ve Wong'un (1979) deneysel olarak değiştirilen notların değerlendirmelerle olan ilişkilerinin gerçek notlarla ilişkilerinden yüksek olduğuna ilişkin bulgularıyla paraleldir.

Bu çalışmanın bulguları bir bütün olarak ele alındığında öğrenci değerlendirmelerinin bütün dünyada yaygın olarak kullanılmalarına karşılık geçerliliğine ilişkin kaygıların göz ardı edilmemesi gereğini ortaya çıkarmaktadır. Öyle görünüyor ki her ne kadar öğrenci değerlendirmeleri öğretim elemanının akademik davranışları konusunda geri bildirim aracı olarak yararlı olsa da, öğrencilerin bireysel ihtiyaç ve duyguları aracın geçerliliğini tehdit edebilmektedir. Bu sonuçlar kendini mükemmelliğe adamış olan ve öğrencileri için yüksek başarı kriterleri belirlemiş öğretim elemanının ne kadar iyi bir öğretim elemanı olursa olsun öğrencilere kolay sorular soran ve düşük başarı kriterleri belirleyen öğretim elemanlarından daha yetersiz olarak değerlendirilebileceklerini ortaya koymaktadır. Sonuçlar öğretim elemanının maaş ve yükselme kriterlerinin öğrenci değerlendirmelerine dayandırıldığı uygulamalar sonucu öğretim elemanının daha olumlu değerlendirme karşılığı öğrencileri daha kolay geçirerek eğitim kalitesini zedeleyeceğine ilişkin kaygıları (Baummeistr 1996; DeBoer, Anderson & Elfessi 2007; Walwoord & Anderson 1998) desteklemektedir.

Ancak bu çalışmanın bulguları pek çok önemli ve değerli çalışmanın (örn., Marsh & Dunkin 1997; Marsh & Roche 1997) bulgu ve yorumlarıyla da çelişmektedir. Eldeki çalışmanın bulguları gözlenen çelişkilerin önemli bir nedeninin Oliveras'ın (2001) öne sürdüğü gibi geçmiş çalışmalarda öğrencilerin not beklentilerinin "zorluk algısıyla" eş değer tutulmasına bağlı olduğunu ortaya koymaktadır. Her ne kadar öğrenci değerlendirmelerinin geçerli olduğu görüşünü destekleyen korelasyona dayalı çalışmalarda öğrenci değerlendirmeleri ile gerek notlar gerekse öğrenme algıları arasında olumlu ilişkiler tespit edilmiş olsa da söz konusu ilişkiler öğretim elemanının yeteneklerine atfedilebileceği gibi öğrencilerin derse verdikleri değer ya da ilgi ve çalışma davranışların sonucu oluşan öğrenme düzeyleri ve not beklentileriyle ilişkili de olabilir. Bu durumda ortaya çıkan öğrenme ürünü öğretim elemanının değil öğrencilerin özelliklerine atfedilebilir. Oysa eldeki çalışmada oluşturulan not verirken kullanılan standartlara (katı ya da müsamahalı) ilişkin algılar öğrenci özelliklerinden bağımsız deneysel olarak yapılandırıldığından değerlendirmeler ile müsamahalı not algısı arasındaki ilişkiler öğrencilerin özelliklerine dayalı algılarına atfedilemez.

Bu çalışmanın bulguları yazın alanında gözlenen çelişkilerin olası bir nedeninin geçmiş çalışmalarda öğrenci değerlendirmelerinin çok boyutlu yapısının dikkate alınmamasına bağlı olabileceği sayıtlısını da destekler niteliktedir. Geçmiş deneysel çalışmaların bulgularına ilişkin ciddi bir eleştiri söz konusu etkilerin dikkate alınmaya değmeyecek kadar küçük olduğuna ilişkindir. Oysa geçmiş çalışmalarda genellikle müsamahalı notun derse ilişkin doyum düzeyi üzerindeki etkisi incelenmiştir. Ancak bu çalışmanın bulguları bu etkinin en fazla doyum üzerinde değil iyi öğretmen alt boyutunda olduğunu ortaya koymakta, genel doyum üzerindeki etkinin ise çok daha az olduğunu ortaya çıkarmaktadır.

Özetle, bu çalışma, öğretim elemanının not verirken belirlediği ulaşılması zor standartların öğretim elemanına ilişkin değerlendirmeleri etkilemediği ve bu yüzden geçerli bir ölçüm aracı olduğuna ilişkin yorumlara gölge düşürücü niteliktedir. Bu yüzden öğrencilerin değerlendirmeleri ile

öğrencilerin derse verdiği değer arasında bulunan pozitif ilişkileri öğretim elemanının becerilerine dayalı öğrenmenin bir ürünü şeklindeki yorumlara da gölge düşürmektedir. Her ne kadar öğrenci değerlendirmelerinin geçerliliğini savunan geçmiş çalışmalarda öğrenci değerlendirmeleri ile öğrenme arasında olumlu ilişkiler gözlenmiş olsa da bu çalışmalarda derse verilen değer öğretim yılının başında öğretim elemanıyla karşılaşılmadan önce ölçülmemiştir. Bunun yerine değerlendirmelerin toplanması sırasında öğrencilerden dönemin başında bu derse ne kadar değer verdiklerini "hatırlamaları" istenmiştir. Bu tür bir ölçüm yanıltıcı olabileceğinden bu çalışmanın bulguları öğrencilerin derse verdikleri değer ile öğretim elemanına ilişkin değerlendirmeler arasında görülen ilişkilerin öğrenci özelliklerine mi yoksa öğretim elemanının özelliklerine mi bağlı olduğunu belirleyen yeni çalışmaların gerekliliğini ortaya koymaktadır.

5. SONUÇ ve ÖNERİLER

Sonuç olarak bu çalışma öğrencilerin öğretim elemanlarına ilişkin değerlendirmelerinin öğretim elemanının not verirken kullandığı standartlardan etkilenebileceğini ortaya koymaktadır. Ancak dikkate değer bir nokta bu çalışmanın bulgularının öğrenci değerlendirmelerinin öğretim elemanının not verirken kullandığı standartlardan olumsuz yönde etkilenebileceği şeklinde olmasına karşın, sonuçların öğrenci değerlendirmelerinin geçersiz ve kullanışsız olduğu şeklinde yorumlanmaması gereğidir. Deneysel işlemin değerlendirmeler üzerinde anlamlı etkileri olsa da etki değerleri bu etkinin orta düzeyde ya da düşük olduğunu ortaya koymaktadır. Bu bulgular öğrenci değerlendirmelerinin öğretim elemanının başarısını belirlemede yararlı bir geri bildirim aracı olabileceği, ancak bu aracın değerlendirilmesinde öğrencilerin bireysel ihtiyaç ve özellikleri gibi, sürece etki edebilecek değişkenlerin göz önünde bulundurulması gereğini ortaya koymaktadır. Bu uyarı özellikle bu tür değerlendirmelerin ancak son yıllarda uygulamaya koyulduğu ve genellikle belli standartlardan yoksun olduğu ülkemizde (Collins 2002) özellikle önem taşımaktadır. Öğrenci değerlendirmeleri profesyonelce kullanıldığında çok yararlı bir geri bildirim aracı olabilecekken, tek başına ve öğrenci değerlendirmelerine etki edebilecek değişkenler göz önünde bulundurulmaksızın kullanıldığında, son derece yanıltıcı ve bu yüzden zararlı olabilir. Öğrenci değerlendirmelerinin yüksek öğretimde kaliteyi artırıcı önemli bir geri bildirim ve araştırma aracı olarak işlevini devam ettirebilmesi için bu tür araştırmalara devam edilerek, süreci etkileyen potansiyel değişkenlerin tespit edilmesi ve sonrasında bu değişkenler konusunda öğrencilerin değerlendirme öncesinde bilgilendirilmesi büyük önem taşımaktadır.

KAYNAKLAR

Aleamoni, L.M. (1999). Student rating myths versus research facts from 1924 to 1998. *Journal of Personal Evaluation in Education*, *13*(2), 153-166.

Baummeistr, R. F. (1996, Summer). Should schools try to boost self-esteem? *American Educator, 22*, 14-19.

Bonesronning, H. (1999). The variation in teachers' grading practices: Causes and consequences. *Economics of Educational Review*, *18*, 89-105.

Byrne, M., & Flood, B. (2003). Assessing the teaching quality of accounting programmes: An evaluation of the Course Experience Questionnaire. *Assessment and Evaluation in Higher* Education, 28 (2),135-145.

Chambers, B.A., & Schmitt, N. (2002). Inequity in the performance evaluation process: How you rate me affects how I rate you. *Journal of Personnel Evaluation in Education,* 16 (2), 103-112.

Centra, J.A. (2003). Will teachers receive higher student evaluations by giving higher grades and less course work? *Research in Higher Education, 44*, 495-518.

Cohen, P. A. (1981). Student ratings of instruction and student achievement: A meta-analysis of multi-section validity studies. *Review of Educational Research, 51*, 281-309.

Collins, A. B. (2002). Üniversite öğrencileri öğretim elemanlarının başarısını değerlendirebilir mi? İkilemler ve problemler. *Ankara Üniversitesi Eğitim Bilimleri Fakültesi Dergisi, 35*, (1-2), 81-91.

DeBoer, B. V., Anderson, D. M., & Elfessi, A. M. (2007). Grading styles and instructor attitudes. *College Teaching, 55* (2), 57-64.

Ellis, L., Burke, D. M., Lomire, P., & McCormack, D. R. (2003). Student grades and average ratings of instructional quality: The need for adjustment. *The Journal of Educational Research, 97*, (1), 35-40.

Feldman, K. A. (1978). Course characteristics and college students' ratings of their teachers; what we know and what we don't. *Research in Higher Education, 9*, 199-242.

Feldman, K. (1997). Identifying exemplary teachers and teaching: Evidence from student ratings. In R. P. Perry & J. C. Smart (Eds.), *Effective teaching in higher education: research and practice* (pp. 368-395). New York : Agathon.

Greenwald, A. G., & Gillmore, J. M. (1997). No pain, no gain? The importance of measuring course workload in student ratings of instruction. *Journal of Educational Psychology, 89*, 743-751.

Griffin, B. W. (2001). Instructor reputation and student ratings of instruction. *Contemporary Educational Psychology, 26*, 534-552.

Griffin, B. W. (2004). Grading leniency, grade discrepancy, and student ratings of instruction. *Contemporary Educational Psychology, 29*(4), 410-425.

Grimes, P. W., Millea, M. J., & Woodruff, T. W. (2004). Grades; who to blame? Student evaluation of teaching and locus of control. *Journal of Economic Education, 35* (2), 129-147.

Heckert, T. M., Latier, A., & Ringwald-Burton, A., & Drazen, C. (2006). Relations among student effort, perceived class difficulty appropriateness and student evaluations of teaching: Is it possible to buy better evaluations? *College Student Journal, 40* (3), 588-596.

Holmes, D. S. (1972). Effects of grades and disconfirmed grade expectancies on students' evaluations of their instructor. *Journal of Educational Psychology, 63*, 130- 133.

Howard, G., & Maxwell, S. (1980). Correlation between student satisfaction and grades: A case of mistaken causation. *Journal of Educational Psychology, 72*, 810-820.

Koushki, P. A., & Kuhn, H. A. J. (1982) How reliable are student evaluations of teachers? *Engineering Education, 72,* 362-367.

Kulik, J.A. (2001). Student ratings: Validity, utility, and controversy. *New Directions for Institutional Research 109*, 9-25.

Lawless, C.J., & Richardson, J.T.E. (2002). Approaches to studying and perceptions of academic quality in distance education. *Higher Education, 44*, (2), 257-82.

Marsh, H. W. (1980). The influence of student, course and instructor characteristics on evaluations of university teaching. *American Educational Research Journal, 17*, 219-237.

Marsh, H.W. (1983). Multidimensional ratings of teaching effectiveness by students from different academic settings and their relation to student/course/instructor characteristics. *Journal of Educational Psychology, 75,* 150-166.

Marsh, H.W. (1987). Students' evaluation of university teaching: Research findings, methodological issues, and directions for future research. *International Journal of Educational Research, 11*, 253-388.

Marsh, H. W. & Dunkin, M. (1992). Students' evaluations of university teaching: A multidimensional perspective. In J. C. Smart (Ed.), *Higher Education: Handbook of Theory and Research*: Volume 8, (pp. 143-233). New York: Agathon.

Marsh, H. W., & Dunkin, M. (1997). Student evaluation of university teaching: A multidimensional perspective. In Perry, P. R., and Smart, J. C. (Eds.), *Effective Teaching in Higher Education: Research and Practice,* (pp. 241-320). Agathon, New York.

Marsh, H. W., & Roche, L. A. (1997). Making students' evaluations of teaching effectiveness effective. *American Psychologist, 52*, 1187-1197.

McKeachie, W. J. (1997). Student ratings: The validity of use. *American Psychologist, 52*, 1218-1225.

Oliveras, O. J. (2001). Student interest, grading leniency, and teacher ratings: A conceptual analysis. *Contemporary Educational Psychology, 26*, 382, 399.

Özgüngör, S. (2009). The relationships between students' evaluations of teaching behaviors and self efficacy beliefs. *Procedia Social and Behavioral Sciences, 1*, 2687–2691.

Powell, P.W. (1977). Grades, learning, and student evaluation of instructors. *Research in Higher Education, 7*, 193-205

Radmacher, S. A., & D. J. Martin (2001). Identifying significant predictors of student evaluations of faculty through hierarchical regression analysis. *The Journal of Psychology, 135*(3), 259-268.

Ramsden, P. (1991). A performance indicator of teaching quality in higher education: The Course Experience Questionnaire. *Studies in Higher Education, 16,* 129-50.

Wachtel, K. H. (1998). Student evaluation of college teaching effectiveness: A brief review. *Assessment & Evaluation in Higher Education, 23*(2), 191-211.

Walwoord, B., & Anderson, V. J. (1998). *Effective grading: A tool for learning and assessment*. San. Francisco: Jossey-Bass.

Wilson, K., Lizzio, A., & Ramsden. P. (1997). The development, validation and application of the course experience questionnaire. *Studies in Higher Education 22* (1), 33–53.

Worthington, A. G., & Wong, P. T. P. (1979). Effects of earned and assigned grades on student evaluations of an instructor. *Journal of Educational Psychology, 71*, 764-775.

Extended Abstract

Student evaluations have been widely used all over the world as a useful tool to improve higher education since the 1920's (Kulik, 2001). However, they have also been subject to many criticisms regarding their validity since many variables which theoretically bears no relationship with the teaching ability such as instructors' gender (Feldman 1978; Griffin 2001), instructor's personality (Radmacher and Martin 2001), students' locus of control (Grimes, Millea and Woodruff 2004), difficulty of the course and grading leniency (e.g., Greenwald & Gillmore 1997) have been linked to them. Among these variables, students' expected grades and grading leniency issues have become the center of the validity arguments. According to the grading leniency argument, students would assign higher evaluation points to the instructors, not because they are good instructors but just to appreciate or reward their generosity in grading, which in turn would harm higher education quality since easy grading would result less effort and learning by students (Eiszler 2002).

Although a vast amount of research has been conducted to determine the effect of grading leniency over the years, studies appear to adding conflicting results rather than solving the problem at hand. The early experimental studies (e.g., Holmes 1972; Powell 1977) usually produced data supporting leniency effect. However, these studies were not only not supported by follow up studies but also have been critiqued severely on the ground of not assigning subjects randomly, failing to report effect size, misleading students regarding their expected grades and so on (e.g., Howard and Maxwell 1980; Marsh 1980; Marsh and Roche 1997). Furthermore, literature reviews usually concluded against leniency effect (e.g., Cohen 1981, Alemoni 1999; Marsh and Dunkin 1997).

However, more recent and stronger studies continue to support leniency effect (e.g., Ellis, Burke, Lomire and McCormack 2003). Oliveras (2001) argued that the main reason for the inconsistencies on the literature is the fact that many studies mistakenly equated expected grades to leniency. However, expected grades could depend on students' own interest level or effort, which is independent of instructor's real grading habits. As a matter of fact, later studies requiring students to evaluate the instructors' leniency in comparison to other instructors' grading behaviors revealed further findings supporting grading leniency effect (e.g., Oliveras 2001; Griffin 2004).

The inconsistencies existing in the literature could also stem from researchers failure to take into consideration of multidimensional structure of the teaching (Marsh, 1983). Studies against leniency hypothesis usually tested the relationship between expected grades and course satisfaction. However, if students' ratings are not affected by instructors' grading styles, instructors' strict grading would correlated only with evaluation dimension of the teaching, whereas it should bear no or low relationship to other teaching aspects tapping instructors' teaching behaviors such as good teaching or generic skills dimensions.

The purpose of this study was to test leniency hypothesis within a real life classroom context where students' perception of instructors' grading styles was manipulated and data were analyzed by taking in to consideration of multidimensional structure of the ratings. For this purpose 144 students were asked to respond to CEQ to rate their instructor's teaching behaviors one week before the exam. The ratings were collected by a researcher who was helping to the main researcher without the students' awareness. The researcher instructed the students that he was developing a scale to measure instructor's skills and he had to collect the data twice to establish scales' reliability. The following week after the first data collection students took one of the two experimental exams, although the students thought it was a real. The first exam was a collection of instructors' hardest questions harvested from previous years' exams and the students were graded strictly during evaluations in that only completely correct answers received point. The second exam was a collection of the instructors' easiest questions used in previous years' exams and the students who took this exam were graded very leniently. The following week the students received their grades and the researcher who collected the first evaluations asked students to provide the second part of the evaluations. After the collection of the data the students were informed about the true nature of the exam.

In order to determine possible differences between the ratings of the students' in two groups before the treatment, t test was used. According to the results, there was no difference between two groups' ratings in terms of any of the teaching dimensions. The repeated measure of ANOVA was

used to determine whether students' ratings changed after the treatment and whether the change differed according to the groups. According to the results, students' ratings changed on good teaching, generic skills, fair evaluation and general satisfaction subscales. Both groups assigned lower ratings on good teaching and generic skills dimensions. However; students in lenient group gave higher ratings to the instructor on fair evaluation dimension and their satisfaction level remained the same. On the other hand, students in strictly graded group reported lowered levels of course satisfaction and assigned lower points to the instructor on fair evaluation dimension. The effect sizes indicated that the highest decrease was on good teaching dimension, followed by generic skills, fair evaluations and satisfaction. Also, students' achievement scores were significantly correlated to both good teaching and fair evaluation subscales of first evaluations but had no relationships with the any of the second evaluations. Similarly, students' experimental scores had no relationship with any of the first evaluations but had significant relationships with all of the second ratings.

Overall, these results provide further supportive findings for grading leniency hypothesis. The results are consistent with early experimental studies and challenge major literature reviews' conclusions. The main reason for the inconsistency seems to stem from past studies' failure to differentiate leniency and expected scores as Oliveras (2001) argued. The results highlight the importance of continuous studies to determine the main validity threats to students' evaluations. Thus, it would be possible to ensure their objective use as a useful research and educational tool by training students about the potential biases and by statistically controlling confounding variables.