Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
ISSN: 1309 - 6575

Puanlayıcı Niteliklerinin Kesme Puanlarının Belirlenmesine Etkisinin Genellenebilirlik Kuramı'yla İncelenmesi*

Investigation the Effects of the Raters' Qualifications on Determining Cutoff Scores with Generalizability Theory

Yusuf KARA ** Hülya KELECİOĞLU ***

Öz
Standart belirleme süreçlerinde en önemli unsur, kesme puanları hakkında yargıda bulunan puanlayıcı grubudur. Puanlayıcı yargılarının tutarlı olması, belirlenecek kesme puanı veya puanlarının güvenirliğini doğrudan etkileyecektir. Bu çalışmada, okul öğretmenleri ve alan uzmanları olmak üzere farklı nitelikteki iki puanlayıcı grubunun standart belirleme sürecindeki performansları Genellenebilirlik Kuramı çerçevesinde karşılaştırılmış; böylece puanlayıcı niteliklerinin kesme puanları üzerindeki etkisinin ortaya konulması amaçlanmıştır. Çalışmada test merkezli standart belirleme yöntemlerinden 1-0 ve Nedelsky yöntemleri ele alınmış ve elde edilen veriler tek ve iki değişkenlik kaynaklı tümüyle çaprazlanmış desenler ile analiz edilmiştir. Elde edilen bulgularda ,her iki standart belirleme yönteminde uzman grubunun öğretmen grubuna göre belirgin olmamakla beraber daha tutarlı yargılar verdiği görülmüş; bu bulgular hesaplanan G ve Phi katsayılarıyla da desteklenmiştir.

Anahtar Kelimeler: standart belirleme, 1-0 yöntemi, Nedelsky yöntemi, genellenebilirlik kuramı, puanlayıcı tutarlılığı

Abstract
Rater group, who give judgments about cutting scores, is the most important factor of the standard setting procedures. Reliability of the cutoff score(s), will be directly affected by the consistency of the raters' judgments. In this study, performance of school teachers and domain experts were compared within the scope of the Generalizability Theory during the standard setting procedure. With this comparison, it was aimed to demonstrate the effect of rater qualifications on cutoff scores. 1-0 and Nedelsky methods which are two of the student centered standard setting methods were adopted in this study. The data that had been collected from raters was analyzed with single and multiple facet crossed designs. According to results of the analyses that were conducted under different designs, it was seen that domain experts gave more consistent judgments compared to school teachers; but this distinction was not so notable. More detailed findings such as G and Phi coefficients of decision studies were also covered in the subsequent sections.

Key Words: standard setting, 1-0 method, Nedelsky method, generalizability theory, rater consistency

GİRİŞ

Eğitim-öğretim süreçlerinde öncelikle ölçme işlemi gerçekleşmekte, geçerli ve güvenilir ölçme sonuçları göz önünde bulundurularak belirlenen ölçüt ya da ölçütler doğrultusunda değerlendirme yapılmaktadır. Ölçütler, değerlendirme sürecinde bireyler hakkında değer yargısına varıp bireyleri ölçülen yapının düzeylerine göre başarılı/başarısız, temel/yeterli/üst düzey gibi birtakım kategorilere ayırmak için kullanılmaktadır. Bu ölçütler, anılan kategorileri birbirinden ayıran puanlar olarak ifade edilebilir. Bu puanlar, alanyazında kesme puanı ya da geçme puanı şeklinde ifade edilmektedir. Değerlendirme sürecinde, bireyleri ölçülen yapıdaki durumlarına göre sınıflandırmada kullanılacak kesme puanlarını belirleme

* Bu çalışmanın bir bölümü, IV. Eğitimde ve Psikolojide Ölçme ve Değerlendirme Kongresi'nde sözlü bildiri olarak sunulmuştur.
** Araş. Gör., Anadolu Üniversitesi, Eğitim Fakültesi, Eskişehir-TÜRKİYE, yusufkara@anadolu.edu.tr
*** Prof. Dr., Hacettepe Üniversitesi, Ankara-TÜRKİYE, hulyaebb@hacettepe.edu.tr

süreci ise standart belirleme olarak ifade edilmektedir (Cizek, 2001). Standart belirleme süreçleri için geliştirilen birçok yöntem bulunmaktadır.

Standart belirleme yöntemleri, birçok araştırmacı tarafından değişik şekilde sınıflandırılmıştır. Bu sınıflandırmalardan en yaygın kabul göreni test ve öğrenci merkezli yöntemler şeklindedir (Jeager, 1989). Test merkezli yöntemlerde, puanlayıcılar doğrudan test maddeleri hakkında yargılarda bulunarak testin içeriğine göre kesme puanı belirlemektedirler. Öğrenci merkezli yöntemlerde ise, testin ölçtüğü özellik bakımından bireylerin yeterlik düzeyleri hakkında yargılarda bulunmaktadırlar. Özet olarak, test merkezli yöntemlerde odak test maddeleri iken; öğrenci merkezli yöntemlerde ise odak öğrencilerin yetenek düzeyleri olmaktadır.

Bu çalışmada, test merkezli yöntemlerden 1-0 (yes/no olarak da adlandırılmaktadır) ve Nedelsky yöntemleri ele alınmıştır. Aşağıda bu iki yöntem hakkında özet bilgilere yer verilmiştir.

1-0 Yöntemi

1-0 yöntemi, Angoff yöntemine alternatif olarak Impara ve Plake (1998) tarafından geliştirilmiştir. Bu yöntemde, puanlayıcıların minimum yeterlik düzeyindeki bir öğrenciyi göz önüne alarak, bu düzeydeki bir öğrencinin testin her bir maddesini yanıtlama durumu hakkında bir yargıya varmaları söz konusudur. Puanlayıcılar her bir maddeyi, minimum yeterlik düzeyindeki bir öğrenci tarafından doğru yanıtlanabileceğini düşündüklerinde 1; maddeye yanıt verilemeyeceğini ya da yanlış yanıtlandırılacağını düşündüklerinde ise 0 şeklinde puanlamaktadırlar. Puanlamanın tamamlanmasının ardından her bir puanlayıcı için toplam yargı puanı hesaplanmaktadır. Son aşamada tüm puanlayıcılara ait yargı puanlarının ortalaması alınarak test için kesme puanı elde edilmektedir.

Nedelsky Yöntemi

Bu yöntem Nedelsky (1954) tarafından geliştirilmiştir. Nedelsky yöntemi sadece çoktan seçmeli testler için uygulanabilmektedir. Yöntemin uygulanmasında ilk olarak "F-D grubu" olarak da adlandırılan, yeterli yetenek düzeyine sahip olanlar ile olmayanların ayrıldığı sınırda yer alıp iki düzeyden birine dâhil edilemeyen bireylerin oluşturduğu grup belirlenmektedir. Bu çalışma için F-D grubu, 2011 SBS matematik testi için geçme-kalma sınırında yer alacak öğrenci grubunu ifade etmektedir. Test için geçme-kalma sınırının ve dolayısıyla F-D grubunun sahip olduğu yeterlik düzeyinin işevuruk tanımı puanlayıcıların kendi yargılarına dayanarak ortaya çıkmaktadır. F-D grubu belirlendikten sonraki aşamada, puanlayıcılar testte yer alan her madde için F-D grubundaki bir öğrencinin her maddeye ait seçeneklerden yanlış olduğunu düşünüp eleyeceği seçenek sayısını tahmin etmeye çalışmaktadırlar. Sonraki adımda, her soru için tüm seçenek sayısından elenen seçenek sayısı çıkarılmakta ve bu farkın çarpma işlemine göre tersi alınarak her maddeye ait 'Nedelsky değeri' hesaplanmaktadır (Cizek ve Bunch, 2007; Tanrıverdi, 2006). Her madde için hesaplanacak Nedelsky değerleri toplanarak teste ait kesme puanı elde edilmektedir. Yapılan işleme ait formül aşağıdaki şekilde ifade edilebilir:

$$\boldsymbol{\tau} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} (\mathbf{q}\boldsymbol{i} - \mathbf{t}\boldsymbol{i})^{-\mathbf{1}}$$

τ: Puanlayıcının belirlediği kesme puanı
qi: i maddesindeki seçenek sayısı
ti: Minimum yeterlikteki bir öğrencinin eleyeceğini düşündüğü seçenek sayısına ilişkin puanlayıcının tahmini
n: Testteki madde sayısı

Kesme puanlarının belirlenmesi, standart belirleme sürecinde görev alacak puanlayıcıların test maddeleri veya öğrenciler hakkında verecekleri yargılara dayanmaktadır. Standart belirleme süreçlerinin doğrudan puanlayıcı yargılarına dayanması, sürecin niteliğinin puanlayıcıların verecekleri kararlar doğrultusunda şekilleneceğini göstermektedir (Koffler, 1980). Dolayısıyla, standart belirleme sürecinin başarısını etkileyen en kritik unsur puanlayıcı grubu olmaktadır. Cizek ve Bunch (2007) standart belirleme süreçlerinde puanlayıcı grubunun önemini vurgulamış ve bu süreçlerdeki en büyük değişkenlik kaynağının, kullanılan değişik standart belirleme yöntemlerinden ziyade süreçlerde görev alan puanlayıcı grubu olduğunu belirtmişlerdir. Bu çalışmada, farklı puanlayıcı gruplarının kesme puanı belirleme sürecindeki tutarlılıklarının incelenmesinde Genellenebilirlik Kuramı'ndan (GK) faydalanılmıştır. Dolayısıyla, aşağıdaki bölümlerde GK ile ilgili temel bilgilere yer verilmiştir.

Genellenebilirlik Kuramı

GK, Klasik Test Kuramı'nın (KTK) bir uzantısı olup, temelleri varyans analizine dayanan ve güvenirliğin belirlenmesinde değişik hata kaynaklarını da dikkate alabilen bir kuram olarak tanımlanmaktadır (Brennan, 2001). GK'nın KTK'dan farklılaştığı en temel nokta, ölçme hatası kavramına bakış açısıdır. KTK'da gerçek puan varyansının dışındaki tüm artık varyanslar tesadüfi hata varyansı olarak nitelendirilmektedir. Ayrıca KTK'da, güvenirlik belirlenirken kullanılan değişik yöntemler olduğunda hata varyansının kaynağı da değişik şekillerde ele alınmaktadır. Örneğin test-tekrar test yöntemiyle güvenirlik kestiriminde zaman faktörü hata kaynağı olabilmekteyken, iç tutarlığın incelendiği yöntemlerde ise maddelere ait değişimler hata kaynağı olmaktadır. GK'da hata varyansına sebep olan faktörler sadece tesadüfî olarak nitelendirilmekle kalmayıp yapılacak analizler ile ortaya koyulabilmekte; ayrıca değişik hata kaynaklarının aynı anda incelenebilmesinin verdiği olanakla tek bir güvenirlik değerine ulaşılabilmektedir. Ayrıca, varyans analizi ile hata varyansına sebep olan değişik faktörler aynı anda belirlenebilmekte; buna ek olarak iki faktörün etkileşiminin varyansı da kestirilebilmektedir. Anılan tüm bu özellikler, GK'yı KTK'ya göre daha avantajlı kılmaktadır.

GK'da ölçme hatalarının olası kaynakları "değişkenlik kaynağı" olarak adlandırılmaktadır (Brennan, 2001; Shavelson ve Webb, 1991). Örneğin bir testin farklı puanlayıcılar tarafından değerlendirildiği durumlarda "farklı puanlayıcılar" bir değişkenlik kaynağı olmaktadır. Yine aynı örnek üzerinden devam edildiğinde, farklı puanlayıcı grubunda her bir puanlayıcı ise değişkenlik kaynağının bir "koşulu" olarak nitelendirilir. Ölçme süreçlerinde, eldeki değişkenlik kaynaklarının sayısına göre bir veya birden fazla sayıda değişkenlik kaynaklı evrenler belirlenebilmektedir. Değişkenlik kaynaklarının sayısı artırıldığında, ölçme sürecine karışan hata kaynaklarının daha iyi irdelenmesi mümkün olabilmektedir. GK'da değişkenlik kaynakları çalışmanın amacına göre sabit veya tesadüfi olarak ele alınabilmektedir (Shavelson ve Webb, 1991). Değişkenlik kaynağının sabit olarak alınması, sadece mevcut puanlayıcı grubu için bir güvenirlik analizi yürütüleceğini ifade eder. Öte yandan değişkenlik kaynağının tesadüfi olarak alınması ise, tüm puanlayıcıların evrenine bir genelleme yapılacağını ifade eder.

GK'da, kurulacak olan desenler çaprazlanmış veya yuvalanmış olabilir. Yukarıdaki örnek üzerinden açıklanmaya çalışılırsa, her puanlayıcı her bir bireyi değerlendirirse "çaprazlanmış" bir desen söz konusu olur. Bunun aksine, eğer her bir bireyi sadece belli bir puanlayıcı değerlendirirse, bu durumda desen "yuvalanmış" olarak nitelendirilir. Puanlayıcılar p, bireyler ise b ile gösterildiğinde çaprazlanmış desen (pxb); yuvalanmış desen ise (p:b) biçiminde sembolize edilir. Ayrıca bazı değişkenlik kaynaklarının çaprazlanmış, bazılarının da yuvalanmış olduğu karışık desenlerin de kurulabilmesi

mümkündür. Kurulan desenler üzerinde, "Genellenebilirlik (G)" ve "Karar (K)" çalışmaları olmak üzere iki farklı analiz yürütülmektedir. G çalışmalarında temel odak, değişkenlik kaynaklarına ait varyans bileşenlerinin belirlenip gözlenmesi iken; K çalışmalarındaki odak ise kurulan desenlerin ve değişkenlik kaynaklarındaki koşul sayısının manipüle edilerek istenen güvenirlik düzeyini sağlayan desenlerin belirlenmesidir. G ve K çalışmaları sonucunda bağıl değerlendirmelere yönelik olarak G katsayısı; mutlak değerlendirmeler için ise Phi (Φ) katsayısı kestirilmektedir. G katsayısı, KTK'da güvenirlik katsayısı ile özdeşleştirilebilir. Φ katsayısı ise kesme puanlarının söz konusu olduğu ölçme durumlarında kullanılmaktadır (Brennan, 2001; Shavelson ve Webb, 1991).

Araştırmanın Amacı ve Önemi

Bu araştırma ile değişik puanlayıcı gruplarının, 1-0 ve Nedelsky standart belirleme süreçlerinde ne derece tutarlı yargılarda bulunduklarının belirlenmesi; dolayısıyla değişik niteliklere sahip puanlayıcıların, belirlenen kesme puanlarına olan etkisinin ortaya konulması amaçlanmaktadır.

Bireyler hakkında önemli kararların alınmasına yön veren standart belirleme süreçlerinde en önemli unsurun puanlayıcı grubu olması, puanlayıcıların ne gibi nitelikler taşıması gerektiği sorusunu akıllara getirmektedir. Impara ve Plake (1998) standart belirleme süreçlerinde yer alacak puanlayıcıların genel anlamda "alanlarında uzman" olan kişiler olarak nitelendirilebileceğini belirtmişlerdir. Öte yandan, Jeager (1991) tarafından ifade edilen özelliklerden en dikkat çekici olanları; uzmanlık alanında gerekli yetkinliğe sahip olma, güçlü bir özdenetim becerisine sahip olma, alanlarındaki problemleri niteliksel bir biçimde analiz edebilme ve gelişmiş bir anlamsal (semantic) hafızaya sahip olma şeklinde ifade edilmektedir. Eğitimde ve Psikolojide Ölçme Standartları'nda, (AERA/APA/NCME, 1999) standart belirleme süreçlerinde yer alacak puanlayıcıların yeterli sayıda olup puanlayıcı evrenini temsil edecek nitelikte olması gerektiği belirtilmiştir. Ayrıca puanlayıcı grubunun, standart belirleme süreçlerinin tekrarlanması varsayımı altında, sonuçların büyük bir değişkenlik göstermeyecek nitelikte olması gerektiği vurgulanmaktadır.

Yukarıda anıldığı gibi, standart belirleme süreçlerinde puanlayıcıların verecekleri yargıların tutarlılık göstermesi, bu süreçlerin geçerliği açsından şüphesiz büyük önem taşımaktadır. Reid, (1991) yapılan birçok araştırmanın bulgularına dayanarak puanlayıcıların bu süreçlerde maalesef belirgin bir tutarlılık gösteremediğini belirtmiştir. Alanyazında yer alan standart belirleme ile ilgili araştırmalar (ör: Demir, 2014; Gündeğer, 2012; Impara ve Plake, 1998; Hein ve Skaggs, 2010; Tanrıverdi, 2006; Taşdelen, 2009) incelendiğinde bu çalışmalarda puanlayıcı olarak genellikle okul öğretmenlerinin yer aldığı görülmektedir. Öte yandan üniversitelerdeki alan uzmanlarının (ör: Ömür ve Selvi, 2010; Taşdemir, 2013) hatta okul yöneticilerinin (ör: David, Buckendahl ve Plake, 2008) yer aldığı çalışmaların da olduğu gözlenmiştir. Bu bağlamda, farklı niteliklere sahip puanlayıcıların yargı tutarlılıklarının karşılaştırılmasının; dolaysıyla puanlayıcı niteliklerinin kesme puanının belirlenmesi sürecine olan etkisinin incelenmesinin önemli olduğu düşünülmektedir.

YÖNTEM

Bu araştırmada, 1-0 ve Nedelsky standart belirleme süreçlerinde rol alan farklı özellikteki puanlayıcı gruplarının, kesme puanının belirlenmesindeki tutarlılıkları GK kullanılarak karşılaştırılmaya çalışılmıştır. Dolayısıyla araştırmanın betimsel türde olduğu söylenebilir.

Çalışma Grubu

Araştırmanın amacı doğrultusunda, farklı nitelikteki iki grup olarak ortaokul matematik öğretmenleri ve ilköğretim matematik eğitimi alanında en az yüksek lisans derecesine sahip öğretim elemanları (alan uzmanları) ile çalışılmıştır. Çalışma grubunun belirlenmesinde, seçkisiz örnekleme tekniklerinden ölçüt örnekleme yöntemi kullanılmıştır. Bu örnekleme yönteminde, araştırmanın amacı doğrultusunda belirli niteliklere sahip bireylerin seçilmesi mümkün olmaktadır (Büyüköztürk, Çakmak, Akgün, Karadeniz ve Demirel, 2010). Belirlenen öğretmen grubu, 2013-2014 öğretim yılında Ankara, Bursa ve Hatay illerinin değişik ortaokullarında görev yapan 15 öğretmenden oluşmaktadır. Uzman grubu ise Hacettepe, Orta Doğu Teknik ve Anadolu Üniversitelerinin İlköğretim Matematik Eğitimi bölümlerinde görev yapmakta olup en az yüksek lisans (bilim uzmanlığı) derecesine sahip 15 öğretim elemanından oluşmaktadır. Jeager (1989) standart belirleme süreçlerinde 15 puanlayıcının yeterli olacağını belirtmiştir. Öte yandan, Taşdelen (2009) standart belirleme süreçlerini GK çerçevesinde ele aldığı çalışmasında, .90 civarı güvenirlik için 15 puanlayıcının yeterli olacağını rapor etmiştir. Bu bilgiler ışığında, puanlayıcı sayısı her iki grup için 15 olarak belirlenmiştir.

Veri Toplama Yöntemleri

Veri toplama sürecinde, öğretmen ve öğretim elemanlarına, 18 maddeden oluşan 2011 yılı Seviye Belirleme Sınavı (SBS) matematik sorularının yer aldığı formlar verilmiştir. Puanlayıcılardan, her bir standart belirleme yöntemi çerçevesinde her bir maddeye ilişkin yargılarını formlarda belirtilen yerlere yazmaları istenmiştir. Bu süreçte öncelikle "minimum yeterlik düzeyindeki öğrenci" ve "F-D grubunda yer alan öğrenci" şeklinde tanımlanan bir öğrencinin, SBS göz önünde bulundurulduğunda ne gibi yeterliklere sahip olması gerektiği konusunda bir kanı oluşturmaları için puanlayıcılara gerekli açıklamalar yapılmıştır. Bu çalışma için F-D grubunun, "2011 SBS matematik testinden geçmek için gereken minimum yeterlik düzeyine sahip olan öğrenci grubu" olduğu vurgulanmıştır. Puanlayıcılardan geri bildirimler alınarak, her bir puanlayıcının anılan kavramları açık bir biçimde anladığından emin olunmuştur. Ayrıca gerçek uygulama öncesinde, denemelik maddeler üzerinden puanlamalar yaptırılarak puanlayıcıların sürece ilişkin soru işaretleri giderilmiştir.

Verilerin Analizi

Puanlayıcılardan toplanan veriler GK çerçevesinde analiz edilmiştir. Öncelikle öğretmen ve uzman grupları için 1-0 ve Nedelsky yöntemlerine göre ayrı ayrı maddeler x puanlayıcılar tümüyle çaprazlanmış deseni (mxp) oluşturulmuş; tek değişkenlik kaynaklı bu desenler için varyans bileşenleri ile G ve Φ katsayıları hesaplanmıştır. Tek değişkenlik kaynaklı desenlerin ardından, her iki yöntemin birer koşul olarak ele alındığı maddeler x yöntemler x puanlayıcılar çaprazlanmış deseni (mxyxp) yine öğretmen ve uzman grupları için ayrı olarak oluşturulmuştur. İki değişkenlik kaynaklı bu desenlerin çözümlenmesiyle yine varyans bileşenleri, G ve Φ katsayıları elde edilmiştir. Öğretmen ve uzman grupları, öncelikle her iki yöntem için ayrı olarak kurulan (mxp) deseninden kestirilen varyans bileşenleri ve genellenebilirlik katsayıları bağlamında karşılaştırılmıştır. Daha sonra yine iki grup için yapılan (mxyxp) deseni analiz sonuçları da aynı bağlamda karşılaştırılarak elde edilen bulgular sunulmuştur. Analiz aşamasının son bölümünde ise yine iki ayrı grup için kurulan (mxp) desenlerine ilişkin karar çalışmaları yürütülmüştür. Bu karar çalışmalarıyla, ilgili grup ve desende daha güvenilir sonuçlar elde edilebilmesi için gerekli olan en uygun puanlayıcı sayısı belirlenmeye çalışılmıştır.

BULGULAR

1-0 Yöntemi mxp Desenlerine İlişkin Bulgular

Aşağıda yer alan Tablo 1'de 1-0 yöntemi için öğretmen grubu ile yürütülen G çalışmasına ait varyans bileşenleri ve bu bileşenlerin toplam varyansı açıklama yüzdeleri görülmektedir.

Tablo 1. *Öğretmen Grubu İçin 1-0 Yöntemi İle mxp Desenine İlişkin Varyans Bilgileri*

Varyans Kaynağı	Sd	Kareler Toplamı	Kareler Ortalaması	Varyans Miktarı	Varyans Yüzdesi
Maddeler (m)	17	12.652	0.744	0.037	16.7
Puanlayıcılar (p)	14	2.430	0.174	0.001	0.0
Maddeler x Puanlayıcılar (mxp)	238	44.237	0.186	0.186	83.3
Toplam	269	59.319			%100

Tablo 1'deki bulgular incelendiğinde 1-0 yöntemi için öğretmen grubu ile yapılan standart belirleme çalışmasında, madde ana etkisinin varyans oranının %16.7; puanlayıcılardan kaynaklı varyans oranının %0 ve madde-puanlayıcı etkileşiminden kaynaklı varyans oranının %83.3 olduğu görülmektedir. Standart belirleme süreçlerinde maddeler GK analizinin amacıdırlar ve dolayısıyla istenen durum maddelerden kaynaklı varyansın mümkün olduğunca yüksek olmasıdır (Taşdelen, 2009). Fakat tablodaki bulgular incelendiğinde, madde ana etkisinin açıkladığı varyans % sinin birer hata kaynağı olan diğer iki değişkenlik kaynağının açıkladığı miktarlara göre düşük olduğu söylenebilir. Öte yandan puanlayıcı ana etkisinin açıkladığı varyans yüzdesi ise 0 olarak bulunmuştur. Bu durum, öğretmen yargılarının 1-0 yönteminde değişkenlik göstermediğini; daha açık bir ifadeyle öğretmenlerin tutarlı yargılar verdiğini belirtmektedir. Bu desen için toplam varyansta %83.3 lük dikkat çekici en büyük varyansı açıklayan hata kaynağı ise madde ve puanlayıcı etkileşimi olmuştur. Puanlayıcı ana etkisi için varyans miktarının sıfır çıkıp; etkileşim için kayda değer bir varyans miktarının olması, puanlayıcıların maddelere ilişkin yargılarının her madde için aynı kararlılık düzeyinde olmamasından kaynaklanmaktadır (Brennan, 2001; Shavelson ve Webb, 1991; Taşdelen, 2009). Bu durum ayrıca maddelerin güçlük düzeylerinin puanlayıcılar tarafından farklı algılandığına işaret etmektedir (Gündeğer, 2012).

Aşağıda yer alan Tablo 2'de ise 1-0 yöntemi için uzman grubu ile yürütülen G çalışmasına ait varyans bileşenleri ve bu bileşenlerin toplam varyansı açıklama yüzdeleri görülmektedir.

Tablo 2. *Uzman Grubu İçin 1-0 Yöntemi ile mxp Desenine İlişkin Varyans Bilgileri*

Varyans Kaynağı	Sd	Kareler Toplamı	Kareler Ortalaması	Varyans Miktarı	Varyans Yüzdesi
Maddeler (m)	17	15.052	0.885	0.046	18.8
Puanlayıcılar (p)	14	4.074	0.291	0.005	2.2
Maddeler x Puanlayıcılar (mxp)	238	46.059	0.194	0.194	79.0
Toplam	269	65.185			%100

Tablo 2'deki bulgular incelendiğinde 1-0 yöntemi için uzman grubu ile yapılan standart belirleme çalışmasında, madde ana etkisinin varyans oranının %18.8; puanlayıcılardan kaynaklı varyans oranının %2.2 ve madde-puanlayıcı etkileşiminden kaynaklı varyans oranının %79 olduğu görülmektedir. Madde ana etkisinin açıkladığı varyans, öğretmen grubunda olduğu gibi diğer değişkenlik kaynaklarına göre nispeten düşük kalmıştır. Öte yandan puanlayıcı ana etkisi ise %2.2'lik bir varyans açıklama oranına sahiptir. Bu durum, az da olsa alan uzmanlarının kendi içerisinde öğretmen grubuna göre

daha tutarsız yargılar verdiğini göstermektedir. Madde-puanlayıcı etkileşimi ise yine en yüksek varyans miktarını açıklamış olup, alan uzmanlarının her maddeye ilişkin aynı kararlılıkta yargı vermedikleri söylenebilir.

Her iki grup için 1-0 yöntemi bulguları beraber değerlendirildiğinde, tüm test maddeleri için uzman grubunun öğretmen grubuna göre çok az miktarda tutarsızlık gösterdiği; ancak mxp etkileri incelendiğinde, öğretmen grubunun belirli maddelerdeki yargılarının uzman grubu yargılarına göre daha değişken olduğu söylenebilir. Öte yandan madde ana etkisinin varyans oranının uzman grubu için biraz daha fazla olması, yine uzman grubu lehine bir sonuç olarak değerlendirilebilir. Varyans oranlarının ayrı ayrı incelenmesinin ardından, daha genel bir karşılaştırma yapılabilmesi için G ve Φ katsayıları aşağıda yer alan Tablo 3'de sunulmuştur.

Tablo 3. *1-0 Yöntemi İçin İki Ayrı Grupla Yapılan G Çalışmalarına Ait Katsayılar*

Grup	G katsayısı	Φ katsayısı
Öğretmenler	0.75	0.75
Uzmanlar	0.78	0.78

Tablo 3'deki bulgular incelendiğinde, her iki grupla yürütülen 1-0 yöntemi standart belirleme çalışmalarının orta düzeyde güvenirliğe sahip olduğu görülmektedir. G katsayısı Cronbach α katsayısı ile aynı anlama sahip olup, genellenebilirlik kuramı çerçevesinde analiz edilen ölçme modelinin güvenirliğini belirtmektedir. Öte yandan Φ katsayısı, yürütülen ölçme sürecinde ölçmenin amacı konumunda olan koşulların mutlak olarak (diğer koşulların ölçek üzerindeki konumlarından bağımsız olarak) bir ölçek üzerine ne kadar iyi yerleştirildiğini belirtmektedir (Cardinet, Johnson ve Pini, 2010; Güler, 2009). Dolaysıyla, uzman grubunun 1-0 yönteminde öğretmen grubuna göre hem bağıl hem de mutlak değerlendirme bağlamında .03 farkla daha güvenilir sonuçlar verdiği söylenebilir.

Genellenebilirlik çalışmalarının ardından, her iki puanlayıcı grubu için yürütülen karar çalışmalarına ilişkin bulgular, aşağıda yer alan Tablo 4'de sunulmuştur.

Tablo 4. *1-0 Yöntemine İlişkin Karar Çalışmalarına Ait Bulgular*

	Puanlayıcı Sayıları									
	15*		20		30		40		50	
Grup	G	Φ	G	Φ	G	Φ	G	Φ	G	Φ
Öğretmenler	.750	.750	.800	.800	.850	.850	.889	.889	.909	.909
Uzmanlar	.781	.777	.827	.823	.877	.874	.905	.903	.923	.921

**Desenler için yürütülen G çalışmalarında 15 puanlayıcı mevcuttur.*

Tablo 4'deki bulgular incelendiğinde, 1-0 yönteminde öğretmenlerin sayısının kademeli olarak 15'ten 50'ye çıkarılmasının, G katsayısını .75'den .909'a yükselteceği görülmektedir. Öğretmen grubu için Φ katsayıları da aynı şekilde değişkenlik göstermektedir. Öte yandan, uzman sayısının kademeli olarak 50'ye çıkarılması ise G katsayısını .781'den .923'e; Φ katsayısını ise .777'den .921'e yükseltmektedir. Burada vurgulanması gereken nokta, hem G hem de Φ katsayıları açısından uzman grubunun öğretmen grubuna göre az da olsa daha iyi performans göstermiş olmasıdır.

Nedelsky Yöntemi mxp Desenlerine İlişkin Bulgular

Aşağıda yer alan Tablo 5'de Nedelsky yöntemi için öğretmen grubu ile yürütülen G çalışmasına ait varyans bileşenleri ve bu bileşenlerin toplam varyansı açıklama yüzdeleri görülmektedir.

Tablo 5. *Öğretmen Grubu İçin Nedelsky Yöntemi İle mxp Desenine İlişkin Varyans Bilgileri*

Varyans Kaynağı	Sd	Kareler Toplamı	Kareler Ortalaması	Varyans Miktarı	Varyans Yüzdesi
Maddeler (m)	17	51.052	3.003	0.147	11.0
Puanlayıcılar (p)	14	108.163	7.726	0.385	28.8
Maddeler x Puanlayıcılar (mxp)	238	191.170	0.803	0.803	60.2
Toplam	269	350.385			%100

Tablo 5'deki bulgular incelendiğinde Nedelsky yöntemi için öğretmen grubu ile yapılan standart belirleme çalışmasında, madde ana etkisinin varyans oranının %11; puanlayıcılardan kaynaklı varyans oranının %28.8 ve madde-puanlayıcı etkileşiminden kaynaklı varyans oranının %60.2 olduğu görülmektedir. Madde etkisinin varyans oranı, diğer hata kaynaklarının varyans oranından daha düşük çıkmıştır. Ayrıca, puanlayıcıların tutarsızlığının göstergesi olan puanlayıcı etkisinin varyans oranın yüksek olması dikkat çekmektedir.

Öğretmen grubu için Nedelsky yöntemi bulguları ile 1-0 yönteminin bulguları karşılaştırıldığında, özellikle puanlayıcı tutarlılığının 1-0 yöntemine göre belirgin bir biçimde düştüğü görülmektedir. Bu durumun, Nedelsky yönteminde verilen yargıların 1-0 yöntemindeki yargılara göre daha karmaşık olmasından kaynaklandığı ve Taşdelen (2009) tarafından elde edilen bulgularla örtüştüğü söylenebilir.

Tablo 6. *Uzman Grubu İçin Nedelsky Yöntemi ile mxp Desenine İlişkin Varyans Bilgileri*

Varyans Kaynağı	Sd	Kareler Toplamı	Kareler Ortalaması	Varyans Miktarı	Varyans Yüzdesi
Maddeler (m)	17	51.885	3.052	0.159	14.6
Puanlayıcılar (p)	14	75.652	5.404	0.263	24.1
Maddeler x Puanlayıcılar (mxp)	238	159.281	0.669	0.669	61.3
Toplam	269	286.819			%100

Tablo 6'daki bulgular incelendiğinde Nedelsky yöntemi için uzman grubu ile yapılan standart belirleme çalışmasında, madde ana etkisinin varyans oranının %14.6; puanlayıcılardan kaynaklı varyans oranının %24.1 ve madde-puanlayıcı etkileşiminden kaynaklı varyans oranının %61.3 olduğu görülmektedir. Madde ana etkisi varyansının diğer tüm hata varyanslarından düşük çıkması dikkat çekmektedir. Uzman grubu için Nedelsky yöntemi bulguları ile 1-0 yönteminin bulguları karşılaştırıldığında, öğretmen grubundakine benzer biçimde puanlayıcı tutarlılığının düştüğü görülmektedir.

Her iki grup için Nedelsky yöntemi bulguları, açıklanan varyans oranları bağlamında değerlendirildiğinde, uzmanların 1-0 yönteminde olduğu gibi öğretmenlere göre biraz daha tutarlı yargılar verdiği söylenebilir. Genel bir değerlendirme için, iki ayrı grupla yürütülen G çalışmalarından elde edilen G ve Φ katsayıları aşağıda yer alan Tablo 7'de sunulmuştur.

Tablo 7. *Nedelsky Yöntemi İçin İki Ayrı Grupla Yapılan G Çalışmalarına Ait Katsayılar*

Grup	G katsayısı	Φ katsayısı
Öğretmenler	0.73	0.65
Uzmanlar	0.78	0.72

Tablo 7'deki bulgular incelendiğinde, her iki grupla yürütülen Nedelsky yöntemi standart belirleme çalışmalarının orta düzeyde katsayılara sahip olduğu görülmektedir. Uzman grubuna ait G ve Φ katsayıları, öğretmen grubuna göre daha yüksek çıkmıştır. Bu durum, varyans bileşenleri bulguları ile paralel olarak, uzman grubunun çok belirgin

olmamakla beraber öğretmen grubuna göre daha tutarlı yargılar verdiğini göstermektedir. Tabloda dikkati çeken bir nokta da, öğretmen grubu için Φ katsayısının .70'in altına düşmüş olmasıdır. Son olarak, tablodaki değerler 1-0 yöntemi katsayıları ile karşılaştırıldığında, öğretmenlerin Nedelsky yöntemine ilişkin G ve Φ katsayılarının düştüğü görülmektedir. Uzman grubunda ise Φ katsayısı düşmüş; fakat G katsayısı ise aynı kalmıştır. Nedelsky yönteminde G katsayısının uzman grubu için düşüş göstermemesi, yargı tutarlılığının daha zor olduğu bu yönteme karşı uzman grubunun daha dayanıklı olduğuna işaret etmiştir.

Genellenebilirlik çalışmalarının ardından, her iki puanlayıcı grubu için yürütülen karar çalışmalarına ilişkin bulgular, aşağıda yer alan Tablo 8'de sunulmuştur.

Tablo 8. *Nedelsky Yöntemine İlişkin Karar Çalışmalarına Ait Bulgular*

	Puanlayıcı Sayıları									
	15*		20		30		40		50	
Grup	G	Φ	G	Φ	G	Φ	G	Φ	G	Φ
Öğretmenler	0.733	0.649	0.785	0.712	0.846	0.787	0.880	0.832	0.901	0.861
Uzmanlar	0.781	0.719	0.826	0.773	0.877	0.836	0.905	0.872	0.922	0.895

**Desenler için yürütülen G çalışmalarında 15 puanlayıcı mevcuttur.*

Karar çalışmalarına ilişkin bulgular incelendiğinde, Nedelsky yöntemi için elde edilen bulguların 1-0 yöntemi bulgularına benzer olduğu görülmektedir. Uzman grubu için her durumda katsayı değerlerinin öğretmen grubundan belirgin olmayan biçimde yüksek olduğu görülmektedir. Öte yandan Nedelsky yöntemi için elde edilen katsayılar, her iki grup için de 1-0 yönteminden daha düşük bulunmuştur.

mxyxp Desenlerine İlişkin Bulgular

Bu bölümde, standart belirleme yöntemlerinin de bir değişkenlik kaynağı olarak alındığı mxyxp desenlerine ilişkin iki puanlayıcı grubuna ait bulgulara yer verilmiştir. İlk olarak öğretmen grubu için mxyxp desenine ait varyans bileşenleri bulguları aşağıda yer alan Tablo 9'da sunulmuştur.

Tablo 9. *Öğretmen Grubu İçin mxyxp Deseni Varyans Bileşenleri Bulguları*

Varyans Kaynağı	Sd	Kareler Toplamı	Kareler Ortalaması	Varyans Miktarı	Varyans Yüzdesi
Maddeler (m)	17	48.437	2.849	0.074	6.5
Yöntemler (y)	1	38.400	38.400	0.063	5.6
Puanlayıcılar (p)	14	57.937	4.138	0.098	8.6
Maddeler x Yöntemler (mxy)	17	15.267	0.898	0.036	3.1
Maddeler x Puanlayıcılar (mxp)	238	148.730	0.625	0.312	27.5
Yöntemler x Puanlayıcılar (yxp)	14	52.656	3.761	0.189	16.6
Maddeler x Yöntemler x Puanlayıcılar (mxyxp)	238	86.678	0.364	0.364	32.1
Toplam	539	448.104			%100

Tablo 9'daki bulgular incelendiğinde, mxyxp etkisinin en büyük varyans açıklama yüzdesine sahip olduğu görülmektedir. Bu etki, öğretmenlerin farklı yöntem ve farklı maddeler için yargılarında değişkenlik olduğunu göstermektedir. Ayrıca, bu etki her türlü tesadüfî hatayı da içinde barındırmaktadır (Shavelson ve Webb, 1991). Elbette güvenilir bir ölçme süreci için istenen durum, bu etkiye ait varyansın mümkün olduğunca düşük olmasıdır. mxyp etkisinin ardından, mxp etkisi %27.5; yxp etkisi ise %16.6'lık varyans açıklama yüzdelerine sahip olmuştur. Bu bulgular ışığında, öğretmenlerin farklı madde ve yöntem koşullarında çok tutarlı yargılar vermedikleri söylenebilir. Öte yandan p ana etkisi

%8.6, m ana etkisi %6.5 ve y ana etkisi ise %5.6'lık varyans açıklama yüzdelerine sahip olmuştur.

Aşağıda yer alan Tablo 10'daki bulgular incelendiğinde, öğretmen grubunda olduğu gibi uzman grubu için de en büyük varyans açıklama yüzdelerine sahip olan değişkenlik kaynakları sırasıyla mxyxp, mxp ve yxp olmuştur. Yukarıda da anıldığı gibi, istenen durum m ana etkisinin diğer tüm etkilerden daha fazla varyans açıklama yüzdesine sahip olmasıdır.

Tablo 10. *Uzman Grubu İçin mxyxp Deseni Varyans Bileşenleri Bulguları*

Varyans Kaynağı	Sd	Kareler Toplamı	Kareler Ortalaması	Varyans Miktarı	Varyans Yüzdesi
Maddeler (m)	17	53.750	3.161	0.086	9.0
Yöntemler (y)	1	45.646	45.646	0.079	8.2
Puanlayıcılar (p)	14	48.844	3.488	0.081	8.4
Maddeler x Yöntemler (mxy)	17	13.187	0.775	0.031	3.3
Maddeler x Puanlayıcılar (mxp)	238	133.555	0.561	0.280	29.0
Yöntemler x Puanlayıcılar (yxp)	14	30.881	2.205	0.105	10.9
Maddeler x Yöntemler x Puanlayıcılar (mxyxp)	238	71.785	0.301	0.301	31.2
Toplam	539	397.650			%100

M ve p etkileri iki grup için beraber değerlendirildiğinde, uzman grubunun yine çok az farkla daha tutarlı yargılar verdiği söylenebilir. İkili çapraz etkiler karşılaştırıldığında, uzman grubu için mxp etkisinin açıkladığı varyans oranı %1.5 daha yüksek çıkmıştır. Bu bulgu, ikili çapraz desenlerdeki ile aynıdır. Bu desende özellikle incelenmek istenen yxp etkisi, öğretmenlerde %5.7 daha fazla varyans açıklama oranına sahip olmuştur. İki grup için aradaki bu fark belirgin olmamakla beraber, yxp etkisi bakımından öğretmenlerin uzmanlara göre az da olsa değişken yargılar verdikleri söylenebilir. Daha açık bir ifadeyle, uzmanlar öğretmenlere göre değişen yöntemlerde tutarlılıklarını az da olsa korumuşlardır. Bu durum kendini iki yöntem için oluşturulan mxp desenlerinde hesaplanan G katsayılarının öğretmenler için farklı çıkarken, uzmanlar için değişmemesinde de göstermiştir.

SONUÇLAR ve TARTIŞMA

Bu araştırmada 1-0 ve Nedelsky standart belirleme yöntemlerinde, öğretmen ve uzmanlardan oluşan iki farklı grubun kesme puanı belirleme sürecinde verdikleri yargıların karşılaştırılması amaçlanmıştır. Bu amaç doğrultusunda ilk olarak iki puanlayıcı grubu için mxp desenleri her iki yöntem için GK ile analiz edilmiştir. Tek değişkenlik kaynaklı desenlerin ardından, yine her iki grup için ayrı ayrı yöntemlerin de bir değişkenlik kaynağı olarak modellendiği mxyxp desenleri oluşturulmuş ve aynı biçimde GK ile analiz edilmiştir. Oluşturulan tüm desenler için varyans açıklama miktarları, yüzdeleri ve desenler için hesaplanan G ve Φ katsayıları rapor edilmiştir.

Elde edilen bulgular doğrultusunda, 1-0 yöntemi için uzman grubunun, öğretmen grubuna göre önemsenmeyecek düzeyde daha iyi sonuçlar verdiği görülmüştür. Gerek varyans açıklama oranları; gerekse G ve Φ katsayıları arasındaki farklar çok küçük olduğundan, iki grubun da kesme puanı belirleme sürecindeki performanslarında çok belirgin bir fark olmadığı söylenebilir. Nedelsky yönteminde de aynı durum söz konusu olmuştur. Uzman grubunun az farkla daha iyi değerler elde ettiği görülmüş; fakat bu farkın belirgin olmaması, yine her iki grubun da yakın performans gösterdiği sonucuna işaret etmiştir. Bu bağlamda, alan uzmanlarının standart belirleme sürecinde öğretmenlere göre belirgin bir avantaj sağlamadığı söylenebilir. Buna benzer bir durum Noricini, Shea ve Kanya (1988) tarafından yapılan araştırmada da rapor edilmiştir. Anılan araştırma, Angoff yöntemi ile gerçekleştirilen ve tıp eğitimi ile ilgili bir standart belirleme çalışması ile ilgili

olup; araştırma sonucunda, özel bir tıp alanı ile ilgili uzmanlaşmanın sınır grubun performansının belirlenmesine önemli bir etki etmediği rapor edilmiştir.

Nedelsky yöntemi için elde edilen bulgularda vurgulanması gereken bir nokta öğretmen grubu için Φ katsayısının .65 olarak hesaplanmış olmasıdır. Dolayısıyla öğretmen grubunun maddelere ilişkin verdikleri yargılara ait mutlak güvenirlik katsayısının diğer tüm katsayılara göre .70'in altında kaldığı sonucuna varılmıştır. Elde edilen bu durumun, başka öğretmen örneklemleri ile yapılacak çalışmalarla irdelenmesi önerilebilir. Öte yandan uzman grubunun, Nedelsky yöntemindeki yargıda bulunma zorluğuna karşı öğretmenlerden daha dayanıklı olduğu görülmüştür. Bu durum, Nedelsky yönteminde puanlayıcılardan istenen yargıları uzmanların daha iyi anladığını göstermektedir.

Her iki yöntem için öğretmen ve uzman gruplarına yönelik yürütülen karar çalışmaları sonucunda hesaplanan katsayılarda, puanlayıcı sayısının kademeli olarak artırılması halinde uzman grubunun öğretmen grubuna göre her durumda yaklaşık olarak 0.02'lik avantaj sağladığı görülmüştür. Bu fark çok büyük olmamakla beraber, değişik senaryolarda aynı güvenirliğe ulaşabilmek için öğretmen sayısına göre uzman sayısının daha az olacağı söylenebilir. Öte yandan, karar çalışmaları sonucunda, .90 civarı bir güvenirlik elde edebilmek için her iki yöntemde de 40 civarı puanlayıcıya ihtiyaç duyulacağı sonucuna ulaşılmıştır. Bu sayı, her iki yöntem için 10 puanlayıcının yeterli olacağını belirten Taşdelen (2009) ve Angoff yöntemi için 10-15 arası puanlayıcının yeterli olacağını belirten Hurtz ve Hertz'in (1999) bulgularına göre daha yüksek kalmaktadır. Bu durumun, başka örneklemlerde yürütülecek çalışmalar ile daha detaylı bir biçimde araştırılması önerilmektedir.

Mxyp desenlerinde iki grup için tüm etkilere ait varyans açıklama yüzdeleri, genellikle birbirlerine yakın bulunmuştur. Dolayısıyla, mxp desenlerinden elde edilen sonuçlara benzer biçimde yine iki grubun standart belirleme sürecinde birbirlerine yakın performans gösterdiği sonucuna ulaşılmıştır. Ancak mxyxp desenlerinde iki grup için en belirgin farklılık, yxp (yöntem x puanlayıcı etkileşimi) bileşeninde görülmüştür. Uzman grubu için bu bileşen, öğretmen grubuna göre %5.7 daha az varyans açıklamıştır. Elde edilen bu bulgu, yukarıda da anıldığı gibi farklı standart belirleme yöntemlerine karşı uzman grubunun, öğretmen grubuna karşı biraz daha dayanıklı olduğunu göstermektedir.

Mxyxp desenlerine ilişkin her iki grup için yürütülen analizlerden elde edilen sonuçlarda, maddeler için yüksek olması istenen varyans miktarı düşük kalmış; aksine hata kaynaklarına ait varyans miktarları görece daha yüksek olmuştur. Clauser ve diğerleri (2009) tarafından yapılan araştırmada, puanlayıcıların ilk turda yürüttüğü sürecin ardından bir tartışma panelinin yapılmasının daha olumlu sonuçlar doğuracağı belirtilmiştir. Panel sonrasında ikinci tur için yapılan GK analizinde maddelere ait varyans miktarının arttığı; puanlayıcı ve puanlayıcı-madde etkileşimine ait varyansın ise düştüğü rapor edilmiştir. Dolayısıyla, yapılması planlanan standart belirleme çalışmalarında puanlayıcıların etkileşim içinde olması ve tartışmalarının da sağlanabileceği bir panelin planlaması ile istenen iyileşmelerin sağlanabileceği önerilmektedir.

Son söz olarak, standart belirleme süreçlerinde, yalnızca okul öğretmenlerinin değil; ilgili alan uzmanlarının da görev almasının süreci daha iyileştireceği söylenebilir. Öğrenciler hakkında önemli kararların alınacağı sınavlar için kesme puanlarının belirlenmesinde, çoğulcu bir yaklaşım benimsenerek uzmanların da sürece katılmasının verimli olacağı düşünülmektedir. Ayrıca, yukarıda önerildiği gibi sürecin puanlayıcıların etkileşim kurup tartışabileceği bir panel şeklinde gerçekleştirilmesi, hem tutarlılık anlamında hem de belirlenecek kesme puanlarının nitel kalitesi anlamında olumlu sonuçlar doğuracaktır.

KAYNAKLAR

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). Standards for educational and psychological testing. Washington, DC: American Educational Research Association.

Brennan, R. L. (2001). Generalizability theory. New York: Springer Verlag.

Büyüköztürk, Ş., Çakmak, E., Akgün,Ö. E., Karadeniz, Ş. ve Demirel, F. (2010). Bilimsel Araştırma Yöntemleri. 7. Baskı. Ankara: Pegem A Yayıncılık.

Cardinet, J., Johnson, S., & Pini, G. (2010). Applying generalizability theory using EduG. In G. Marcoulides (Series Ed.), Quantitative methodology series. New York: Routledge.

Cizek, G. J. (Ed.). (2001). Setting performance standards: Concepts, methods, and perspectives. Mahwah, NJ: Lawrence Erlbaum Associates.

Cizek, G. J. & Bunch, M. B. (2007) Standard Setting: A guide to establishing and evaluating performance standards on tests. Thousand Oaks, CA: Sage.

Clauser, B. E., Harik, P., Margolis, M. J., McManus, M. C., Mollon, J., Chis, L. & Williams, S. (2009). An empirical examination of the impact of group discussion and examinee performance information on judgments made in the Angoff standard-setting procedure. Applied Measurement in Education, 22, 1-21.

Davis, S. L., Buckendahl, C. W. & Plake, B. S. (2008). When adaptation is not an option: an application of multilingual standard setting. Journal of Educational Measurement, 45(3), 287-304.

Demir, O. (2014). Angoff, Nedelsky ve Ebel standart belirleme yöntemleri ile belirlenen kesme puanlarının karşılaştırılması. Yayımlanmamış yüksek lisans tezi. Abant İzzet Baysal Üniversitesi, Eğitim Bilimleri Enstitüsü, Bolu.

Güler, N. (2009). Genellenebilirlik Kuramı ve SPSS ile GENOVA programlarıyla hesaplanan G ve K çalışmalarına ilişkin sonuçların karşılaştırılması. Eğitim ve Bilim, 34(154), 93-103.

Gündeğer, C. (2012). Angoff, Yes/No ve Ebel standart belirleme yöntemlerinin karşılaştırılması. Yayımlanmamış yüksek lisans tezi. Hacettepe Üniversitesi, Sosyal Bilimler Enstitüsü, Ankara.

Hein, S. F. & Skaggs, G. (2010). Conceptualizing the classroom of target students: a qualitative ınvestigation of panelists' experiences during standard setting. Educational Measurement: Issues and Practice, 29(2), 36-44.

Hurtz, G. M. & Hertz, N. R. (1999). How many raters should be used for establishing cutoff scores with the Angoff method? A Generalizability Theory study. Educational and Psychological Measurement, 59(6), 885-897.

Impara, J.C. & Plake, B.S. (1998). Teacher's ability to estimate item difficulty: a test of assumptions in the Angoff standard setting method. Journal of Educational Measurement, 35, 69-81.

Jaeger, R. M. (1989). Certification of student competence. In R. L. Linn (Ed.), Educational measurement (3rd ed., pp. 485–514). New York: Macmillan Publishing Co.

Jaeger, R. M. (1991). Selection of judges for standard setting. Educational Measurement: Issues and Practice, 10(2), 3–14.

Koffler, S. L. (1980). A comparison of approached for setting proficiency standards. Journal of Educational Measurement, 17, 167–178.

Nedelsky, L. (1954). Absolute grading standards for objective tests. Educational and Psychological Measurement, 14, 3–19.

Norcini, J. J., Shea, J. A. & Kanya, D. T. (1988). The effect of various factors on standard setting. Journal of Educational Measurement, 25(1), 57-65.

Ömür, S. ve Selvi, H. (2010). Angoff, Ebel ve Nedelsky yöntemleriyle belirlenen kesme puanlarının sınıflama tutarlılıklarının karşılaştırılması. Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi, 1(2), 102-103.

Reid, J. (1991). Training judges to generate standard setting data. Educational Measurement: Issues and Practice, 10(2), 11–14

Shavelson, R. J. & Webb, N. M. (1991). Generalizability Theory: A Primer. Sage Publications, USA.

Tanrıverdi, S. (2006). Standart belirleme yöntemlerinin standart belirleme üzerine etkisi. Yayımlanmamış yüksek lisans tezi. Hacettepe Üniversitesi, Sosyal Bilimler Enstitüsü, Ankara.

Taşdelen, G. (2009). Nedelsky ve Angoff standart belirleme yöntemlerinin genellenebilirlik kuramı ile karşılaştırılmasına ilişkin bir araştırma. Yayımlanmamış yüksek lisans tezi. Hacettepe Üniversitesi, Sosyal Bilimler Enstitüsü, Ankara.

Taşdemir, F. (2013). Angoff (1-0), Nedelsky ve Sınır Değerleri Saptama yöntemleri ile bir testin sınıflama doğruluklarının incelenmesi. Yayımlanmamış doktora tezi. Ankara Üniversitesi, Eğitim Bilimleri Enstitüsü, Ankara.

EXTENDED ABSTRACT

Introduction

It is well known that measurement is the first stage of the educational evaluation. After conducting a reliable measurement study, one will get some useful information which are usually called as "scores" for making decisions about students. These decisions usually will be as "failed/passed" or as "beginner/intermediate/advanced" to categorize students according to their proficiency levels of the measured construct(s). In order to make these decisions properly, students' scores are compared with predefined scores which are called "cutoff scores" in the measurement literature. Cutoff scores are defined at the end of the standard setting procedures, depending on the raters' judgments about test items and examinees. Because these cutoff scores will be completely shaped by "judgments" of raters, it can definitely be said that the rater group is the most important factor of the standard setting procedures. This importance is underlined by many researchers who are qualified with standard setting procedures and also by NCME standards about setting performance standards of students. In various standard setting procedures, raters who have different qualifications like school teachers, domain experts, even school principals can take part. For this reason, it is sensible to ask the questions: "How these raters who have different qualifications will perform during the standard setting procedures?" or more apparently "How the different qualifications of the raters will affect the accuracy of the cutoff scores?". In this study it was aimed to find answers to these questions with Generalizability Theory.

Method

According to this aim, two separate rater groups who are comprised of school teachers and domain experts were asked to make judgments about 2011 SBS (Level Determination Exam for High School Applications) mathematics test items. There were 18 multiple choice items within this test. School teachers who involved in this study were working at the different public middle schools in different provinces. Domain experts were comprised of research assistants, and faculty members who have at least master's degree in elementary mathematics education. 15 raters were involved in each group based on the recommendations of former researchers.

Only 1-0 and Nedelsky standard setting methods which are two commonly used test centered methods were adopted for this study. After the data had been collected, generalizability analyses were conducted with EduG software for one and two facet completely crossed designs. Analyses were conducted for the two groups separately and both the variance components and G/Phi coefficients were reported for all methods and rater groups. After these generalizability studies, decision studies were also done as final analyses in order to define the optimum rater number in two groups for a reliable standard setting procedure.

Results and Discussion

According to explained variance values of the first two generalizability studies with one facet designs, for both of the standard setting methods, domain experts performed slightly better than the school teachers. This difference was well observed with Nedelsky method which requires more complicated judgments and definition of the F-D groups' features, compared to 1-0 method. Based on these findings, it can be said that the expert group performed slightly better than the teacher group, especially with more demanding Nedelsky method. After the inspection of the variance components, it was also seen that G and Phi

coefficients were slightly higher for domain experts group. This result is also favoring domain experts in terms of consistency; but again it should be noted that the difference was not very distinctive.

Depending on the findings of the decision studies that were conducted separately for the two groups, it can be said that in order to get a reliability level around .90 with either teacher or expert group, there will be a need of approximately 40 raters for both methods. This number seems to be high when compared to another researcher's' results and it can be suggested to be studied with different samples of raters who will have same qualifications. When two facet crossed design's findings inspected, it was seen that domain experts' judgments were more robust to method changes. In other words, *method x rater* interaction effects' variance rates were found to be %5.7 lower for domain experts. This result can be considered as a sign of the domain experts' consistency compared to teacher group.

In the light of the all these results, it is reasonable to say that the interaction between raters which can be enabled with a panel study can reduce the inconsistent judgments of different raters. This will led to more reliable and valid standard setting procedures and consequently to correct decisions that will be made about students.