Trakya Eğitim Dergisi
Cilt 13, Sayı 1
Ocak 2023, 242-259

Geliş Tarihi: 16.11.2021

ISSN: 2630-6301
Doi: 10.24315/tred.1024397
Araştırma Makalesi/
Research Article

Trakya Journal of Education
Volume 13, Issue 1
January 2023, 242-259

Yayına Kabul Tarihi:06.03.2022

MESLEKİ KIDEMİN KESME PUANI BELİRLEMEYE ETKİSİNİN GENELLENEBİLİRLİK KURAMI İLE İNCELENMESİ*

ANALYSING THE IMPACT OF PROFESSIONAL SENIORITY ON DETERMINING CUTSCORE WITH GENERALIZABILITY THEORY

Süleyman KESER[1], Nuri DOĞAN [2], Sümeyra SOYSAL[3]

ÖZ: Bu araştırmanın amacı, puanlayıcıların çalışma yılı olarak ele alınan mesleki kıdem farklılıklarının, farklı standart belirleme yöntemleri ile elde edilen kesme puanlarına etkisini incelemektir. Araştırmada Millî Eğitim Bakanlığı tarafından 2014-15 ve 2015-2016 eğitim ve öğretim yıllarında sekizinci sınıf öğrencilerine ortak sınav şeklinde uygulanan Temel Eğitimden Orta Öğretime Geçiş sınavı 1. dönem matematik sorularının ayırt edicilik açısından en iyi 20 maddesinden oluşturulan çoktan seçmeli matematik başarı testi kullanılmıştır. Söz konusu test, Ankara ilindeki ortaokullardan seçkisiz örnekleme yöntemi ile belirlenen 907 sekizinci sınıf öğrencisine uygulanmıştır. Maddelerin Angoff, Nedelsky ve Ebel standart belirleme yöntemleri ile değerlendirilmesi amacıyla hazırlanan uzman değerlendirme formu ile aynı okullarda görevli ve hali hazırda sekizinci sınıflarda ders vermekte olan 40 matematik öğretmeninin değerlendirmeleri alınmıştır. Elde edilen kesme puanları Genellenebilirlik Kuramı ile analiz edilmiştir. Yapılan analizler sonucunda; Angoff, Nedelsky ve Ebel standart belirleme çalışmalarında, mesleki olarak daha kıdemli olan puanlayıcı grubunun kesme puanları belirlemede göreceli olarak daha tutarlı olduğu görülmüştür. Kıdem seviyesinin artmasının yanı sıra puanlayıcı gruplarının çalışma yılı açısından homojenleşmesinin de belirlenen standartların tutarlılığını artırabileceği görülmüştür.

Anahtar sözcükler: Standart belirleme, Angoff, Ebel, Nedelsky, genellenebilirlik kuramı

ABSTRACT: The aim of this research is to analyze the effects of raters' professional seniority differences approached as working years on cutting scores obtained with different standard setting methods. In this research, the math test which is composed of the best 20 questions, in terms of separation, prepared by Ministry of Education and applied to 8th grade students as a common exam named "TEOG (Passing on Secondary Education from Primary Education)" in 2014-15 and 2015-16 education years was used. This test was applied to 907 8th grade students which were determined by random sampling method in Ankara. An evaluation form was prepared to assess the questions with Angoff, Nedelsky and Ebel standard setting methods. With the help of this form, evaluations carried out by 40 teachers who work at the same school and attend 8th grade classes were taken. Cutting scores, acquired at the end of this evaluation, were analyzed with Generalizability Theory. At the end of these analysis; in the standard setting studies of Angoff, Nedelsky and Ebel, it was seen that the raters which was more senior as working years was more consistent in specifying cutting scores. It was seen that, apart from increase of the level of seniority, more homogeneous of raters' experience can increase the consistency of standards

Keywords: Standard setting, Angoff, Ebel, Nedelsky, generalizability theory

Bu makaleye atıf vermek için:
Keser, S., Doğan, N., ve Soysal, S. (2023). Mesleki kıdemin kesme puanı belirlemeye etkisinin genellenebilirlik kuramı ile incelenmesi, *Trakya Eğitim Dergisi, 13*(1), 242-259

Cite this article as:
Keser, S., Doğan, N., & Soysal, S. (2023). Analysing the impact of professional seniority on determining cutscore wıth generalizability theory. *Trakya Journal of Education, 13*(1), 242-259

* Bu çalışma ikinci yazarın danışmanlığında yürütülen birinci yazarın yüksek lisans tez çalışmasından türetilmiştir.

[1] Ankara, Türkiye, s.kesser@hotmail.com, Orcid:0000-0002-6317-9568
[2] Prof. Dr., Hacettepe Üniversitesi, Eğitim Fakültesi, Ankara, Türkiye, nurid@hacettepe.edu.tr, Orcid: 0000-0001-6274-2016
[3] Dr. Öğr. Üyesi, Necmettin Erbakan Üniversitesi, Eğitim Fakültesi, Konya, Türkiye, sumeyrasoysal@hotmail.com, Orcid: 0000-0002-7304-1722

EXTENDED ABSTRACT

Introduction

Standard setting can be defined as the process of determining one or more cutoff points in tests. Cut-off scores have functions such as dividing a test score scale into one or more regions, creating categories, or classifying students according to their achievements. In most standard-setting methods, raters, who will make judgments about test items and students, play the most important role in the process. The qualifications and experiences of the raters in standard setting studies affect the quality of the cut-off scores. For this reason, it has been emphasized in the literature that the experience criterion is an important issue to be considered in the selection of raters to take part in standard setting studies.
The aim of this research is to examine the effect of the professional seniority differences, which are considered as the working years of the raters, on the cut-off scores. In the study, the differences between both standard-setting groups and standard-setting methods were tried to be revealed by comparing the parameters estimated by the Generalizability theory from the data obtained with separate rater groups at different levels in terms of professional seniority and different standard-setting methods. In the majority of the studies on standard setting in the literature, it has been mentioned a lot that the experts involved in the standard setting process should be experienced and the importance of their level of experience. However, no study has been found that provides a clear and quantitative explanation of how long this experience covers and how it should be defined. In this respect, with this study, it has been tried to contribute to the subject of what the qualifications of the raters to be selected in the standard setting processes should be. "How are the parameters estimated with the Generalizability Theory of the cut-off scores obtained by the Angoff, Nedelsky and Ebel standard setting methods by different rater groups?" searched for an answer to the question.

Method

The students in the study group of the research are 907 eighth grade students from 18 secondary schools selected by random sampling method in Ankara. The rater group consists of 40 mathematics teachers who work in the same schools and teach in the eighth grade. In the literature, there are studies stating that the level of knowledge of the raters in the subject area and their command of the subjects increase the reliability of the standard-setting processes. For this reason, instead of all mathematics teachers in schools, only secondary school mathematics teachers who teach in the eighth grade were evaluated because they knew the course content, were familiar with the TEOG questions and knew the students.
In order to examine the effect of the seniority differences on the standard setting process, the group consisting of 40 people was first divided into 2 groups according to the differences in seniority years. The seniority level is based on the years that the raters worked as an assigned at the schools under the Ministry of National Education. Raters with 1-10 years of professional work are "junior"; those who are 11 years, or more are called “Senior”. In order to increase the difference between seniority levels and to examine the effect of this change on the results of the study, the upper and lower groups 27% of the entire rater group consisting of 10 raters were formed. These groups are also named as “lower” and “upper” groups. Angoff, Ebel and Nedelsky standard-setting methods were used in the study. The cut-off scores obtained from these methods were analyzed separately for the senior-junior and upper-lower groups with Generalizability Theory.

Findings

In the three standard-setting methods, the judgments of both the senior rater group and the junior rater group made inconsistent assessments within themselves but reached more congruent judgments than the junior rater group of the senior rater group. The increase in the homogeneity of the rater groups in terms of seniority increased the rater consistency and the variance rate of the items in the model. In addition, it had been observed that the effect of unknown sources of variability on the cut-off scores obtained with the junior group in the Nedelsky method was quite high, and this effect decreased when the seniority of

the raters increased. In the Nedelsky method, it was more evident that the senior raters performed much better than the junior raters. It is possible to say that this situation may be due to the fact that the conceptualization process of the Nedelsky method is perceived as more difficult by the junior rater group and that the raters in this group do not sufficiently know the students and their abilities. In the Nedelsky method compared to other methods, it was found that the status of being senior or junior is more effective than homogeneity of seniority in determining the cut-off score.

Discussion and Conclusion

As a conclusion, raters with higher seniority and working years as close to each other as possible in the standard-setting studies in which Angoff, Nedelsky and Ebel methods would increase the reliability of the study and the consistency of the cut-off points. It was determined that Angoff was the most effective method in order to set a standard with sufficient reliability in the decision study with the variance components for item x rater design according to comparing senior and junior rater groups. Especially if junior raters will be involved in Nedelsky method, these raters should be given sufficient and satisfactory training on standard setting studies, otherwise it is considered that it would be appropriate not to include these raters in the study. It is recommended that the Generalizability Theory (GT), which enables both reliability and validity to be determined with a single analysis, and the effects of different surfaces such as raters, methods, and items on the process, to be used more in standard setting studies. Again, GT decision study can be used to determine the most appropriate number of raters or items required to apply a standard-setting study at the intended reliability and generalizability level.

GİRİŞ

Eğitimde alınacak kararların isabet oranı daha özelde, öğrencilere uygulanan çeşitli tür, amaç ve sayılardaki testlerin sonuçları ve alınan kararlarla ilgilidir. Öğrencilerin test puanları, gerçek yetenek seviyelerini isabetli bir şekilde yansıtabilmelidir. Ölçme uygulamaları neticesinde yüksek yeterlik seviyesindeki öğrenciler yüksek puanlar, düşük yeterlik seviyesindeki öğrenciler de düşük puanlar almalıdırlar (Wu & Tan, 2015). Eğitimde öğrencilerin başarılı veya başarısız olarak sınıflanmaları, girdikleri sınavlardan alacakları puanların türü, sertifika programları sonucunda sertifikaya hak kazanıp kazanmadıkları, karnelerinde "5", "3" gibi notlar almaları veya öğretmenlerin kadrolara atanıp atanmamaları gibi çok çeşitli kararlar, uygulanan sınavlar neticesinde ve sınavlardan alınan puanların önceden belirlenen ölçütlerle karşılaştırılması ile yani değerlendirme süreci uygulanarak alınmaktadır. Bahsi geçen ölçütlerin, belirli birtakım kurallara göre elde edilmesi işlemine standart belirleme adı verilmektedir. Daha öz bir anlatımla standart belirleme, testlerde bir veya daha fazla kesme puanı belirleme süreci olarak tanımlanabilir (Cizek & Bunch, 2007; Crocker & Algina, 2008; Tseng, Chiou & Sung, 2015). Kesme puanları bir test puan ölçeğini bir veya daha fazla bölgeye ayırmak, kategorileri oluşturmak veya öğrencileri başarılarına göre sınıflandırmak gibi işlevlere sahiptir (Cizek & Bunch, 2007). Bu tanımlarda bahsi geçen kesme puanlarının alanyazında genellikle ölçüt, geçme puanı ve performans standardı kavramları ile de eş anlamlı olarak kullanıldığı görülmektedir. Standart belirlemenin diğer bir tanımı da performans standartlarının test puan ölçeğinde rakamsal olarak yerleştirilmesi şeklinde yapılabilir (Hambleton, 2001). Bu tanımda bahsedilen husus eğitim süreci sonunda öğrenciler arasındaki öğrenme ve kazanımları edinme farklılıklarını ortaya koyabilmek amacıyla test puanları ölçeği üzerinde ayrım noktaları belirlenmesidir. Böylelikle öğrenciler, bu ayrım noktalarından faydalanılarak performans düzeylerine ayrılabilirler.

Standart belirleme çalışmaları bireylere, ailelere, kurumlara ve bir bütün olarak topluma çok çeşitli kazanımlar sunmaktadır. Lisans ve sertifika sınavları veya sınıflama amacıyla yapılan ölçme uygulamaları için açık standartlar belirlenmesi, yapılan ölçme uygulamalarına katılanların arasındaki rekabeti de artıran bir faktör olarak ortaya çıkar ve rekabetin artması olumlu bir durum olarak görülür. Ayrıca açık standartlar, ölçme uygulamalarına karşı kamuoyunda güven oluşmasını da sağlar. Her ne kadar hemen fark edilemese de standart belirlemenin eğitim, sağlık ve mesleki alanlarda kamuoyu güvenini ve itimadını artırmak gibi soyut faydaları yadsınamaz bir gerçektir (Cizek & Bunch, 2007). Belirtilen faydaları ve katkılarının yanı sıra elbette hiçbir ölçme ve değerlendirme uygulaması mükemmel değildir ve seçilen kesme puanından bağımsız olarak, ölçme uygulaması neticesinde kimi zaman yetersiz öğrenciler başarılı olurken, kimi zaman da yeterli öğrenciler başarısız olmaktadır. Bu

yanlış negatif ve yanlış pozitif hataların göreceli ihtimalleri seçilen kesme puanına göre değişiklik gösterebilmektedir (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014). Bu yüzden yanlış-negatif ve yanlış-pozitif sınıflandırma hatalarının ve yanlış kararların oranını uygun kesme puanları belirlemek suretiyle azaltmak, standart belirleyenlerin sorumluluğudur (Tseng vd., 2015). Bu gibi hataların azaltılabilmesi için istatistik modeller ve deneysel çalışmalar büyük önem taşımaktadır.

Standart belirleme yöntemlerinin büyük bir çoğunluğunda test maddeleri ve öğrenciler hakkında yargıda bulunacak olan puanlayıcılar, süreç içinde en önemli rolü oynamaktadırlar. Puanlayıcıların süreçte önemli bir rol oynaması da belirlenen standartların öznel ve puanlayıcıların özellikleri ile değer yargılarına bağlı olmasına, sonuç olarak da süreçte hataların olmasına neden olmaktadır. Glass'a (1978) göre basit bir işlem gibi gözükmesine rağmen, puanlama için seçilen grubun değer yargıları ve inançlarına göre karar vermeleri nedeni ile standart belirleme işlemleri birtakım belirsizliklere sebep olmaktadır. Ayrıca puanlayıcıların madde güçlüğünü ve öğrencilerin gerçek yetenek seviyelerini tahmin etmesi, uzun eğitim süreçlerine rağmen kolay bir iş değildir (Bejar, 1983). Elde edilen standartların çoğunlukla uzman kanısına dayandığı yöntemlerde bu durum güvenirlik ve dış geçerlik kanıtlarının yetersizliğine neden olduğu için eleştirilmektedir (Tseng vd., 2015). Bu yüzden standart belirlemek için seçilen uzmanların özelliklerinin neler olması gerektiği konusu büyük önem taşımaktadır. Uzmanların niteliğinin artırılması elde edilen kesme puanlarının da daha geçerli ve güvenilir olmasını sağlayacak ve alınacak kararların isabet oranını artırarak hataların da azaltılmasına yardımcı olacaktır.

Standart belirleme sürecine katılan uzmanlar, testin uygulandığı konuya ve bu konunun içeriğine hâkim oldukları kadar, öğrenci grubu hakkında da bilgi sahibi olmalıdırlar. Ancak bu sayede yargılarına anlam ve ağırlık katabilirler (Lim, Geranpayeh, Khalifa & Buckendahl, 2013). Ayrıca puanlayıcıları içeren bir standart belirleme süreci sonucunda güvenilir sonuçlar elde etmek ve puanlamalarının tutarlığını artırmak için seçilecek uzman grubunun yeterli nitelikte ve büyüklükte olması gerekmektedir (AERA vd., 2014). Tutarlılık kavramı ise standart belirleme uygulamalarının en önemli gerekliliğidir ve puanlayıcıların öğrencileri başarılı/başarısız olarak gruplandırmada ulaştıkları fikir birliği olarak düşünülebilir (McCann & Stanley, 2006).

Bahsedilenler ışığında standart belirlemenin eğitimde ve test geliştirme süreçlerinde büyük bir öneme sahip olduğu söylenebilir. Standart belirleme süreçleri için belirlenen koşulların değişkenliği arttıkça belirlenen standartların da değişkenliği artmakta ve tutarlılığı azalmaktadır (Kane, 2001). Standart belirlemedeki değişkenlik, tutarlılık ve isabet seviyesi üzerinde, süreçte rol oynayan puanlayıcıların özellikleri, davranışları ve katılık/cömertlik seviyeleri büyük etkiye sahiptir. Atılgan'a (2004) göre puanlayıcıların ölçme sürecine katılması durumunda daha çok değişkenlik kaynağının potansiyel hata kaynağı olarak dikkate alınması gerekmektedir. Buradan hareketle, ölçme ve standart belirleme süreçlerindeki hata kaynaklarını doğru tespit edebilmek ve hataları azaltabilmek amacıyla gerekli olan puanlayıcı niteliklerini tespit edebilmek önemlidir.

Standart belirleme çalışmalarında puanlayıcıların nitelikleri ve tecrübeleri (experience) süreç sonunda elde edilen kesme puanlarının kalitesini etkilemektedir. Bu nedenle alanyazında, standart belirleme çalışmalarında yer alacak puanlayıcıların seçiminde tecrübe kriterinin de göz önünde bulundurulması gereken önemli bir husus olduğuna vurgu yapılmıştır (AERA vd., 2014; Cizek & Bunch, 2007; Hambleton & Pitoniak, 2006). Fakat bahsedilen tecrübe kavramının net ve alanyazında ortak bir tanımı ortaya konulmamıştır. Tecrübe kavramının çalışmalarda kullanılmasından önce net olarak neyi ifade ettiğini ortaya koyabilmek önemlidir. En azından çalışmalarda kullanım amacına yönelik olarak işevuruk bir tanım yapılması gerekliliği kendini hissettirmektedir. Bu açıdan tecrübe konusunda yapılan tanımlamalara bakmak kavramın hangi açıdan ele alınması gerektiği hususunda fayda sağlayacaktır. Türk Dil Kurumu (TDK) tecrübeyi "deneyim, deney ve görgü" kelimeleri ile tanımlamıştır. Bu tanımların standart belirleme açısından ortak bir anlam ifade edecek şekilde yorumlanması mümkün görülmemektedir. Dewey (1997) tecrübenin sayısal olarak ifade edilmesinden çok niteliğinin ve kalitesinin önemli olduğundan bahsetmiştir. Ancak yapılan standart belirleme çalışmalarının önemli bir kısmında puanlayıcıların tecrübe seviyesine vurgu yapılmış ve bu tecrübe seviyesi çoğunlukla meslekteki çalışma yılı olarak ele alınmıştır (Hsieh, 2013; Kozaki, 2004; Lim, Geranpayeh, Khalifa & Buckendahl, 2013; Shulruf, Wilkinson, Weller, Jones, & Poole, 2016; Wu & Tan, 2015; Yousuf, Violato & Zuberi, 2015). Bu çalışmaların hiçbirinde tecrübenin veya mesleki kıdem olarak da adlandırılan meslekteki çalışma yılının belirlenen kesme puanlarını nasıl etkilediğini gösteren

bulgular ortaya konulmamıştır. Bu durum standart belirleme konusunda bir belirsizlik olarak ortaya çıkmaktadır.

Tecrübenin veya alanyazında tecrübe göstergesi olarak çokça ele alınan mesleki kıdem yılının, belirlenen kesme puanlarına etkisini incelemek bahsedilen belirsizliğin giderilebilmesi ve daha geçerli ve güvenilir standartlar belirlenebilmesi adına önem taşımaktadır. Yapılacak incelemelerde standart belirleme çalışmaları sonucunda elde edilecek kesme puanlarına etki edebilecek olan değişkenlik kaynaklarının sürece etkilerinin büyüklüğü, birbirine göre oranı, değişkenliği, tutarlılığı gibi birçok açıdan analiz edilmelidir. Bu analizlerin gerçekleştirilebilmesi amacıyla Klasik Test Kuramı (KTK), Genellenebilirlik Kuramı (GK), Madde Tepki Kuramına (MTK) ve MTK'nın bir uzantısı olan Çok Yüzeyli Rasch Ölçme Modeli'ne dayalı olarak çok çeşitli uygulamalar gerçekleştirilebilmektedir. Bu yöntemlerden GK ile yapılan analizlerde değişkenlik kaynağı veya yüzey olarak ele alınabilecek puanlayıcılar, yöntemler ve maddeler gibi farklı standart belirleme unsurlarının sürece olan etkileri, oluşturdukları değişkenlik miktarları ve toplam değişkenlik içinde yüzeylerin oranları, güvenirlik ve genellenebilirlik katsayıları gibi bilgilerin tek bir işlemle elde edilmesi mümkündür. Bu durum puanlayıcıların yüzey olarak ele alınabileceği ve puanlayıcı niteliklerinin belirlenen kesme puanlarına etkilerinin ortaya konulmasında GK analizlerinin kullanılmasının faydalı olabileceğini göstermektedir.

Bu araştırmanın amacı, standart belirleme süreçlerinde önemli bir yere sahip olan ve sonucu doğrudan etkileyen puanlayıcıların, çalışma yılı olarak ele alınan mesleki kıdem farklılıklarının, süreç sonunda belirlenen kesme puanlarına etkisinin Genellenebilirlik Kuramı ile incelenmesidir. Çalışmada mesleki kıdem açısından farklı seviyelerde olan ayrı puanlayıcı grupları ve farklı standart belirleme yöntemleri ile elde edilen veriler kullanılarak GK ile kestirilen parametrelerin karşılaştırılması sonucu; standart belirleyen gruplar ve standart belirleme yöntemleri arasındaki farklılıklar ortaya konulmaya çalışılmıştır. Alanyazında standart belirleme konusunda yapılan çalışmaların büyük bir çoğunluğunda, süreçte yer alan uzmanların tecrübeli olması gerektiğinden ve tecrübe seviyelerinin öneminden çokça bahsedilmiş olmasına rağmen, bu tecrübenin ne kadarlık bir süreyi, içeriği kapsadığı veya nasıl tanımlanması gerektiği hususlarında net ve niceliksel bir açıklama getiren herhangi bir çalışmaya rastlanılmamıştır. Bu açıdan, yapılan çalışma ile standart belirleme süreçlerinde seçilecek puanlayıcıların niteliklerinin neler olması gerektiği konusuna katkı sağlanmaya çalışılmıştır.

Problem ve Alt Problemler

Farklı puanlayıcı grupları tarafından Angoff, Nedelsky ve Ebel standart belirleme yöntemleri ile elde edilen kesme puanlarının GK ile kestirilen parametreleri nasıldır?

1. Farklı Puanlayıcı Grupları ve Angoff standart belirleme yöntemine göre oluşturulmuş maddeler x puanlayıcılar (m x p) deseninin varyans bileşenleri, G ve Φ katsayıları ve karar (K) çalışması sonuçları nasıldır?
2. Farklı Puanlayıcı Grupları ve Nedelsky standart belirleme yöntemine göre oluşturulmuş maddeler x puanlayıcılar (m x p) deseninin varyans bileşenleri, G ve Φ katsayıları ve karar (K) çalışması sonuçları nasıldır?
3. Farklı Puanlayıcı Grupları ve Ebel standart belirleme yöntemine göre oluşturulmuş maddeler x puanlayıcılar (m x p) deseninin varyans bileşenleri, G ve Φ katsayıları ve karar (K) çalışması sonuçları nasıldır?
4. Farklı Puanlayıcı Grupları için Maddeler x Yöntemler x Puanlayıcılar (m x y x p) Çaprazlanmış Desenine İlişkin Varyans Bileşenleri, G ve Φ Katsayıları Nasıldır?

YÖNTEM

Araştırmanın Modeli

Bu çalışmada mesleki kıdem yılı olarak farklı seviyelerindeki puanlayıcı gruplar tarafından Angoff, Nedelsky ve Ebel yöntemleri ile belirlenen standartların gruplar arasında farklılık gösterip göstermediği Genellenebilirlik Kuramı ile incelenmiştir. Bu nedenle araştırma esas itibari ile betimsel bir araştırmadır (Büyüköztürk, Kılıç Çakmak, Akgün, Karadeniz ve Demirel, 2016).

Örneklem

Araştırmanın evrenini Ankara ilindeki ortaokulların 8. sınıf öğrencileri ve öğretmenleri oluşturmaktadır. Kümeleme örnekleme yöntemi aracılığıyla seçkisiz olarak 18 ortaokul seçilmiştir. Araştırmada yer alan öğrenci grubunu seçilen bu 18 ortaokuldan seçkisiz olarak seçilen toplam 907 sekizinci sınıf öğrencisi oluştururken puanlayıcı grubunu, aynı okullarda görev yapan ve sekizinci sınıflarda ders veren 40 matematik öğretmeni oluşturmaktadır. Alanyazında puanlayıcıların konu alanındaki bilgi seviyesinin ve konulara hâkim olmasının standart belirleme süreçlerinin güvenirliğini artırdığını ifade eden çalışmalar mevcuttur. (Chang, 1999; Chang, Dziuban, Hynes & Olson, 1996). Bu nedenle ders içeriğini bilmeleri, TEOG sorularına aşina olmaları ve öğrencileri tanımaları nedeni ile okullardaki bütün matematik öğretmenleri yerine, yalnızca sekizinci sınıflarda ders veren ortaokul matematik öğretmenlerinin değerlendirmelerine başvurulmuştur. Mesleki kıdem yılları ve mesleki kıdem yıllarına göre üst ve alt gruba seçilen puanlayıcılar Tablo 1'de gösterilmiştir.

Tablo 1.
Araştırmaya katılan puanlayıcıların mesleki kıdem yılları

Kıdemli Grup			Kıdemsiz Grup		
Puanlayıcı	Kıdem Yılı	%27'lik Üst Gruba Seçilen Puanlayıcılar	Puanlayıcı	Kıdem Yılı	%27'lik Alt Gruba Seçilen Puanlayıcılar
1	13		21	10	
2	17		22	6	
3	15		23	9	
4	36	•	24	2	•
5	24	•	25	4	•
6	20	•	26	6	
7	34	•	27	8	
8	15		28	5	
9	11		29	2	•
10	13		30	4	•
11	20	•	31	4	•
12	35	•	32	4	
13	20	•	33	3	•
14	17		34	4	•
15	13		35	3	•
16	12		36	10	
17	17	•	37	6	
18	24	•	38	3	•
19	22	•	39	9	
20	14		40	1	•
Ortalama	19,6	25,5	*Ortalama*	4,95	3

Puanlayıcıların çalışma yılı olarak kıdem farklılıklarının standart belirleme sürecine olan etkisini inceleyebilmek maksadıyla, 40 kişiden oluşan grup kıdem yılı farklılıklarına göre öncelikle 2 gruba ayrılmıştır. Kıdem seviyesi olarak puanlayıcıların Millî Eğitim Bakanlığı bünyesindeki okullarda atamalı olarak çalıştıkları yıllar esas alınmıştır. Puanlayıcı grupları mesleki çalışma yılı olarak 1-10 yıl arasında olanlar ve 11 yıl ve üzerinde olanlar olarak belirlenmiştir. Bu gruplar "Kıdemsiz" ve "Kıdemli" gruplar olarak adlandırılmışlardır. Kıdemli ve kıdemsiz puanlayıcı gruplarının çalışma yıllarının ortalamaları arasında anlamlı farklılık olup olmadığını incelemek amacıyla öncelikle grupların dağılımlarının normalliklerine bakılmıştır. Grup büyüklüklerinin 50'den küçük olması nedeni ile normalliğe uygunluğun kontrolü Shapiro-Wilks testi ile yapılmıştır (Büyüköztürk vd., 2016). Normallik testi sonucunda kıdemli grup kıdem yılları normal dağılımdan anlamlı şekilde farklılık gösterirken ($p<0,05$), kıdemsiz grup kıdem yılları normal dağılım göstermiştir ($p>0,05$). Bu sonuçlara göre ortalama

kıdem yıllarının gruplara göre farklılık gösterip göstermediğinin incelenmesi parametrik olmayan bir yöntem olan Mann Whitney U-Testi ile yapılmıştır. U-Testi sonucuna göre kıdemli ve kıdemsiz puanlayıcı gruplarının kıdem yılları ortancaları arasında anlamlı bir farklılık olduğu bulunmuştur (U=0,00, p<0,05). Yıllar bazında kıdem seviyeleri arasındaki farkın artırılması ve bu değişikliğin çalışma sonuçlarına etkisinin incelenebilmesi amacıyla tüm puanlayıcı grubun %27'lik ve 10'ar puanlayıcıdan oluşan alt ve üst grupları oluşturulmuştur. Bu gruplar da "Alt" ve "Üst" gruplar olarak adlandırılmışlardır.

Veri Toplama Araçları

Bu araştırmada veri toplamak için iki araç kullanılmıştır. Öğrencilerden verilerin toplanması için 20 maddelik Çoktan Seçmeli Matematik Testi ve öğretmenlerden verilerin toplanması içinse Uzman Değerlendirme Formu kullanılmıştır.

Çoktan Seçmeli Matematik Testi: Öğrencilere uygulanan çoktan seçmeli matematik başarı testi 2014 ve 2015 yıllarında ortaokul sekizinci sınıf öğrencilerine uygulanan 1. dönem TEOG sınavı matematik sorularından ayırt ediciliği en yüksek ve güçlük düzeyleri orta seviyede olan 20 tanesi seçilerek oluşturulmuştur (MEB, 2015, 2016). Çalışmada çoktan seçmeli matematik testinin kullanılmasının nedeni, matematik testinin bilgi yapısının diğer içerik alanlarına oranla daha somut olmasından dolayı kesme puanı belirleme süreçlerini kolaylaştıracak, doğruluğu ve kararlılığı yükseltecek ve hatayı azaltarak diğer hata kaynaklarına yoğunlaşmayı kolaylaştırılabileceğinin düşünülmesidir (Tseng vd., 2015). Çalışmanın temel amacının puanlayıcıların kıdem seviyelerinden kaynaklanan farklılığın standart belirleme sürecine olan etkisinin ortaya çıkarılması olması nedeni ile testten ve diğer değişkenlerden kaynaklı hataların en aza indirilmesi amaçlanmıştır.

Uzman Değerlendirme Formu: Uzman Değerlendirme Formu öğrencilere uygulanan matematik testi sorularının öğretmenler tarafından Angoff, Nedelsky ve Ebel yöntemlerine göre değerlendirilebilmesini sağlayacak şekilde tasarlanmıştır. Değerlendirme formunda öğretmenlerin uygulamayı daha iyi anlayabilmelerini sağlamak amacıyla açıklamalar ve örnek puanlamalara yer verilmiştir. Bu açıklamalar ve örnek puanlamalardan, uygulama öncesindeki ve uygulama esnasında yararlanılmıştır.

Veri Toplama Araçlarının Uygulanışı

Çoktan Seçmeli Matematik Testinin Uygulanışı: 20 çoktan seçmeli maddeden oluşan matematik başarı testi toplamda 907 öğrenciye her sınıfta 40 dk. süreyle uygulanmıştır. Öğrencilere uygulanan matematik testine ait istatistikler Tablo 2'de yer almaktadır.

Tablo 2.
Matematik testi istatistikleri

K	*N*	S_j	$\overline{X}$	$\overline{P}$	*Ortalama* ($\overline{r}_{jx}$)	*Çarpıklık Katsayısı*	*Basıklık Katsayısı*	*KR-20*
20	907	4,89	14,04	0,70	0,56	- 0,57	0,71	0,88

K: Madde Sayısı, S_j: Standart Sapma, $\overline{X}$: Test Ortalaması, $\overline{P}$: Ortalama Güçlük, KR-20: Güvenirlik Katsayısı

Uzman Değerlendirme Formunun Uygulanışı: Standart belirleme uygulamalarından önce puanlayıcılara çalışmanın amacı, çalışmada yer alan standart belirleme yöntemleri, formların nasıl doldurulacağı, testin içeriği ve minimum yeterlik düzeyindeki öğrenci hakkında gerekli ve yeterli açıklama yapılmıştır. Çalışmada kullanılan TEOG soruları için belirlenmiş bir "minimum yeterlik seviyesi" tanımı yoktur. Bu nedenle puanlayıcılardan böyle bir tanım olmaksızın minimum yeterlik seviyesindeki öğrenciyi tahayyül ederek puanlama yapmaları istenmiştir. Bu sınırlılık, puanlayıcı örneklem grubunun öğrencilerle aynı okullardan ve seçkisiz olarak belirlenmesi ile giderilmeye çalışılmıştır. Puanlayıcılara yapılan açıklamaların amacı, puanlayıcı yargıları arasında oluşacak değişkenliği tamamen sıfırlamak değil, katılık/cömertlik açısından aşırı uç sayılabilecek değerlerin ortaya çıkmasının önüne geçmektir (Lumley & McNamara, 1995). Puanlayıcılara gerekli açıklamanın

yapılmasından sonra birer madde üzerinden deneme uygulaması yaptırılarak varsa hatalar ve yanlış anlaşılmalar düzeltilmesi sağlanmıştır.

Standart Belirleme Yöntemleri

Bu konuda birçok yöntem geliştirilmiştir. Bu yöntemler genellikle test merkezli (test-centered) ve testi alan merkezli (examinee-centered) yöntemler olarak gruplandırılmaktadır (Cizek, 1996; Kane, 1994). Test merkezli yöntemler testte yer alan her maddeyi inceleyen ve minimum yeterlik seviyesindeki öğrencinin bu maddeleri cevaplayabilme ihtimalleri hakkında yargılarda bulunan uzman görüşlerine dayalı yaklaşımları kapsamaktadır. Testi alan merkezli yöntemlerde ise süreçte yer alan uzmanlardan testte yer alan maddelerden çok testi alanlar hakkında yargılarda bulunmaları istenir. Test merkezli yöntemlere Angoff ve türleri, Nedelsky, Ebel ve Bookmark yöntemleri, testi alan merkezli yöntemlere ise Sınır Grup ve Karşıt Grup yöntemleri örnek olarak gösterilebilir (Cizek, 1996; Cizek & Bunch, 2007;Kane, 1994). Bu araştırmada test merkezli yöntemler tercih edilmiştir ve ele alınan yöntemler aşağıda özetlenmiştir.

Angoff Standart Belirleme Yöntemi: Angoff ve versiyonları, alanyazında en çok kullanılan test merkezli standart belirleme yöntemleridir (Tseng vd., 2015). Angoff tarafından 1971 yılında geliştirilmiştir. O tarihten itibaren 100'den fazla geliştirilmiş türünün alanyazında kullanılmış olduğu bilinmektedir (Hambleton & Pitoniak, 2006). Yöntemde puanlayıcılardan, test maddelerini inceleyerek, her madde için minimum yeterlik seviyesindeki öğrencinin maddeyi doğru cevaplayabilme ihtimalini tahmin etmeleri istenir. Puanlayıcılardan bütün maddeler için elde edilen ihtimal değerlendirmeleri toplanır ve puanlayıcı sayısına bölünmek suretiyle ortalaması alınarak puanlayıcı grubu ve test için nihai kesme puanı belirlenmiş olur. Angoff yöntemi Türkiye'de ve yurtdışında çokça kullanılmasına rağmen (Behuniak vd., 1982; Brennan & Lockwood, 1980; Çetin, 2011; Demir, 2014; Gündeğer ve Doğan, 2014; Kane, 1994; Yousuf, Violato & Zuberi, 2015) puanlayıcılar tarafından minimum yeterlik düzeyindeki öğrencinin maddeyi doğru cevaplayabilme olasılığının isabetli bir şekilde tahmin edilebilmesinin güç bir görev olduğundan dolayı eleştirilmektedir (Shepard, 1993).

Nedelsky Standart Belirleme Yöntemi: Nedelsky yöntemi sadece çoktan seçmeli test maddeleri için standart belirlemekte kullanılan bir yöntemdir. Bu yöntemde puanlayıcılardan, minimum yeterlik seviyesindeki öğrencilerin madde seçeneklerinden yanlış olanların kaç tanesini bilinçli olarak eleyebileceklerini tahmin etmeleri istenir. Bu tahmin neticesinde, her bir madde için puanlayıcıya göre, bilinçli olarak elenmeden kalacak olan seçenek sayısının çarpma işlemine göre tersi alınarak o madde için kesme puanı belirlenmiş olur (Kara ve Kelecioğlu, 2015; Violato, Marini & Lee, 2003). Maddeler için kesme puanları belirlendikten sonra bütün puanlayıcılardan elde edilen kesme puanlarının ortalaması ise testin kesme puanını verir (Cizek & Bunch, 2007; Hambleton & Pitoniak, 2006). Nedelsky yönteminin zayıf ve eleştirilen yönlerinden biri puanlayıcıların belirledikleri kesme puanlarının, seçenek sayısına göre değişiklik gösteren sınırlı sayıda olasılık değerine bağlı olmasıdır. Örneğin bu çalışmada yer alan dört seçenekli maddelerden oluşan bir testte, puanlayıcıların maddeler için belirleyebilecekleri kesme puanları 0, 0,33, 0,5 ve 1 arasında değişmektedir. Görüldüğü üzere, bu değerler sınırlı sayıda olmakla birlikte eşit aralıklı da değildir ve bu nedenle belirlenebilecek kesme puanlarındaki hata veya gerçek duruma yaklaşma ihtimali göreceli olarak azalmaktadır (Cizek & Bunch, 2007).

Ebel Standart Belirleme Yöntemi: 1972 yılında Robert Ebel tarafından ortaya konulan yöntemde puanlayıcılar iki aşamalı bir yargılama süreci sonucunda kesme puanı belirlemektedirler. İlk aşamada puanlayıcılar tarafından tek tek incelenen maddeler "güçlük" ve "uygunluk" ölçütlerine göre değerlendirilirler. Puanlayıcılar maddeleri "güçlük" açısından "kolay", "orta" ve "zor" seviyelerinden biri ile "uygunluk" açısından da "kabul edilebilir", "tartışılabilir", "önemli" ve "gerekli" uygunluk boyutlarından biri ile eşleştirmek durumundadırlar. Bu eşleştirme genelde, çalışma için hazırlanmış, güçlük ve uygunluk başlıklarına sahip satır ve sütunlardan oluşan 12 (3 x 4) hücreli tablolar içine madde numaralarının yazılması şeklinde gerçekleştirilir. İkinci aşamada ise hücrelere yerleştirilen maddeler için minimum yeterlik düzeyindeki öğrencinin her bir hücredeki maddeleri cevaplama ihtimalleri değerlendirilir. Yapılan değerlendirme sonucu tespit edilen ihtimallerin de hücreye yazılmasından sonra, her bir hücrede yer alan madde sayısı ve cevaplama ihtimalleri çarpılarak, çarpım sonuçlarının toplanması ile değerlendirmeyi yapan puanlayıcının teste ait kesme puanı belirlenmiş olur. Bütün puanlayıcıların kesme puanlarının ortalamasının alınması işlemi ise bize teste ait kesme puanını verir.

Bu yöntemde yüksek bir şekilde ilişkili olan güçlük ve uygunluk boyutlarını birbirinden ayırmanın puanlayıcılar için güç bir görev olduğu sorgulanmaktadır (Shepard, 1984).

Verilerin İşlenmesi ve Çözümlenmesi

Öncelikle farklı puanlayıcı gruplarının Angoff, Nedelsky ve Ebel yöntemleri ile belirledikleri kesme puanları hesaplanmıştır. GK ile cevaplanmaya çalışılan alt problemlerin analizleri EduG-6 programı ile gerçekleştirilmiştir (Cardinet, Johnson & Pini, 2009; Institut de Recherche et de Documantation Pedagogique [IRDP], 2010). Her bir alt problemin çözümünde, her bir puanlayıcı grubu için ayrı ayrı olmak üzere maddeler x puanlayıcılar (m x p) çaprazlanmış deseni ile maddeler x yöntemler x puanlayıcılar (m x y x p) çaprazlanmış deseni oluşturularak bu desenler için G ve katsayıları ile varyans bileşenleri bulunmuştur. Karar (K) çalışmasında m x p deseni için puanlayıcı sayıları beşer beşer artırılıp azaltılmış ve puanlayıcı sayılarının G ve katsayıları üzerinde etkisi incelenmiştir. K çalışmalarında hesaplanan genellenebilirlik ve güvenirliğin kabul edilebilir seviyelerde olması için en az 0,80 değerinin elde edildiği puanlayıcı sayıları tespit edilmeye çalışılmıştır (Brennan, 2001; Shavelson & Webb, 1991).

BULGULAR

Birinci Araştırma Problemi ile İlgili Bulgular

Birinci alt problem kapsamında "Farklı puanlayıcı grupları ve Angoff standart belirleme yöntemine göre oluşturulmuş maddeler x puanlayıcılar (m x p) deseninin varyans bileşenleri, G ve Φ Katsayıları, K Çalışması Sonuçları Nasıldır?" sorusuna yanıt aranmıştır. Öncelikle, Angoff yöntemi için kıdemsiz ve kıdemli gruplarının yargılarına dayalı olarak elde edilen kesme puanları ile oluşturulmuş m x p çaprazlanmış desenlerine ait varyans bileşenleri ve bu bileşenlerin varyans açıklama yüzdelerini gösteren ANOVA sonuçları Tablo 3'te özetlenmiştir.

Tablo 3.
Angoff yöntemi *ile kıdemli ve kıdemsiz gruplara ilişkin ANOVA sonuçları*

Puanlayıcı Grubu	Varyans Kaynağı	Kıdemsiz Grup		Kıdemli Grup	
		Sd	Varyans (%)	Sd	Varyans (%)
Tüm Grup (n=20)	Maddeler (m)	19	14,0	19	16,6
	Puanlayıcılar (p)	19	27,7	19	23,4
	Maddeler x Puanlayıcılar (mxp)	361	58,3	361	60
%27'lik Grup (n=10)	Maddeler (m)	19	34,5	19	27,6
	Puanlayıcılar (p)	9	4,2	9	4,0
	Maddeler x Puanlayıcılar (mxp)	171	61,3	171	68,4

Tablo 3 incelendiğinde tüm grupta kıdemsiz ve kıdemli puanlayıcılar ana etkisinin sırasıyla %27,7 ve %23,4; maddeler ana etkisinin sırasıyla %14 ve %16,6 ve maddeler x puanlayıcılar etkileşiminin sırasıyla %58 ve %60 varyans oranlarına sahip oldukları görülmektedir. Standart belirleme çalışmalarında maddelerden kaynaklı varyansın en büyük etkiye sahip olması amaçlanmaktadır (Taşdelen, Kelecioğlu ve Güler, 2010). Ancak çalışma sonuçlarına bakıldığı zaman her iki grupta da en düşük orana sahip varyans kaynağının maddeler olduğu görülmektedir. Bu durum, diğer varyans kaynaklarının süreç üzerinde maddelerden daha büyük etkiye sahip olduklarını ortaya koymaktadır ve istenen bir durum değildir. Öte yandan her iki grupta da ikinci en yüksek varyansın puanlayıcılara ait olması, puanlayıcıların kendi aralarında tutarlı puanlamalar yapmadığının bir göstergesi olarak değerlendirilebilir. Yine her iki grupta, puanlayıcılar ve maddelerin etkileşiminden ortaya çıkan varyans en büyüktür. Bu değerin büyük olması puanlayıcıların belirledikleri kesme puanlarının maddelere göre değişiklik gösterdiğini ifade etmekle birlikte modelde ele alınan yüzeyler dışında farklı değişkenlik ve hata kaynaklarının da standart belirleme süreçlerinde etkili olduğunu göstermektedir.

Tablo 3'te özetlenen %27'lik alt ve üst grup sonuçları incelendiğinde, madde ana etkisine ait varyansın alt grupta (kıdemsiz) %34,5 iken üst grupta (kıdemli) %27,6 olduğu gözlenmektedir. Puanlayıcı ana etkisine ait varyans alt grupta %4,2 iken üst grupta %4'tür. Bu bulgular her iki grupta, maddenin diğer varyans kaynaklarından daha fazla etkiye sahip olduğu ve puanlayıcıların aralarında tutarlı puanlamalar yaptığını göstermektedir. %27'lik grup ile tüm grup sonuçlarının karşılaşmaları, grupların mesleki kıdem açısından homojenliğinin artmasının madde kaynaklı varyansın artmasına ve puanlayıcı kaynaklı varyansın azalmasına neden olarak modelde istendik yönde ve önemli iyileşmeler sağladığını göstermektedir.

Angoff yöntemi ile yapılan standart belirleme çalışmasının sonuçlarını daha genel olarak karşılaştırılabilmek için G ve katsayıları hesaplanmış ve Tablo 4'te sunulmuştur.Tablo, şekil vb. kullanılması durumunda verilen örneğe uygun bir şekilde düzenlenmelidir. Tablo ve şekiller 1'den başlayarak numaralandırılmalı ve adlandırılmalıdır. Varsa kaynak altta belirtilmelidir.

Tablo 4.
Angoff yöntemi için puanlayıcı gruplarına ait G ve Φ katsayıları

Puanlayıcı Grubu		G	Φ
Kıdemsiz	Tüm grup	0,83	0,76
	%27'lik grup	0,85	0,84
Kıdemli	Tüm grup	0,85	0,80
	%27'lik grup	0,80	0,79

Angoff yöntemiyle elde edilen kesme puanları üzerinden puanlayıcı grupları için yürütülen G ve Φ çalışmaları sonuçlarının sunulduğu Tablo 4 incelendiğinde, kıdemsiz tüm gruba ait G değerinin kıdemli tüm gruba ait G değerinden 0,02 ve Φ değerinin ise 0,04 puan daha düşük olduğu görülmektedir. Bu durum kıdemli puanlayıcı grubunun kıdemsiz gruba oranla hem bağıl hem de mutlak değerlendirme bakımından daha tutarlı sonuçlar verdiğini ortaya çıkarmaktadır.

Puanlayıcı gruplarının %27'lik alt ve üst grupları ile yinelenen G çalışmaları sonucunda kıdemli grup için G ve Φ katsayılarında düşüş gözlemlenmiştir. Bu düşüşün puanlayıcı sayısının azalmasından kaynaklanmış olabileceği söylenebilir. Kıdemsiz grupta ise mutlak ve bağıl anlamda güvenirlik yükselmiştir. Puanlayıcı sayısının azalmasına rağmen güvenirliğin yükselmesi puanlayıcı yargılarındaki tutarlılığın artmasına delil olarak gösterilebilir. Bu durum da kıdem yıllarının yakınlaşması ve grubun kıdem yılı açısından daha homojen bir yapıda olmasından kaynaklanmış olabilir.

G çalışmalarından sonra Angoff yöntemi ile gerçekleştirilen standart belirleme çalışmalarında kıdemsiz ve kıdemli grupların optimum güvenirlik seviyesinde değerler verebilmesi için çalışmaların kaç puanlayıcıyla yapılması gerektiğini ortaya çıkaran K çalışmaları yürütülmüştür. Bu çalışmalar sonucunda ortaya çıkan değerler Tablo 5'te özetlenmiştir.

Tablo 5.
Angoff yöntemi için K çalışması değerleri

Puanlayıcı Grubu	Puanlayıcı Sayıları									
	15		20*		25		30		35	
	G	Φ	G	Φ	G	Φ	G	Φ	G	Φ
Kıdemsiz	0,78	0,71	**0,83**	0,76	0,86	**0,80**	0,88	0,83	0,89	0,85
Kıdemli	**0,81**	0,75	0,85	**0,80**	0,87	0,83	0,89	0,86	0,91	0,87

*Çalışmanın yürütüldüğü puanlayıcı sayısını göstermektedir.

Tablo 5'te puanlayıcı sayısının 5'er artırılıp ve azaltılarak yürütülen K çalışmasından elde edilen G ve Φ katsayıları görülmektedir. G ve Φ katsayıları için 0,80 değeri ölçüt alındığında Angoff yöntemi ile standart belirleme çalışmalarında uygun G değerine ulaşmak için kıdemsiz grupta 20, kıdemli grupta 15 puanlayıcıı; uygun Φ değerine ulaşmak kıdemsiz grupta 25, kıdemli grupta 20 puanlayıcı gereklidir. Bu sonuçlar, kıdemli gruptaki puanlayıcıların kıdemsiz gruptaki puanlayıcılardan göreceli olarak daha iyi performans gösterdiğini ortaya koymaktadır.

İkinci Araştırma Problemi ile İlgili Bulgular

İkinci alt problem kapsamında "Farklı puanlayıcı grupları ve Nedelsky standart belirleme yöntemine göre oluşturulmuş maddeler x puanlayıcılar (m x p) deseninin varyans bileşenleri, G ve Katsayıları, K Çalışması Sonuçları Nasıldır?" sorusuna yanıt aranmıştır. Öncelikle, Nedelsky yöntemi için kıdemsiz ve kıdemli gruplarının yargılarına dayalı olarak elde edilen kesme puanları ile oluşturulmuş m x p çaprazlanmış desenlerine ait varyans bileşenleri ve bu bileşenlerin varyans açıklama yüzdelerini gösteren ANOVA sonuçları Tablo 6'da özetlenmiştir.

Tablo 6.
Nedelsky yöntemi ile kıdemli ve kıdemsiz gruplara ilişkin ANOVA sonuçları

Puanlayıcı Grubu	Varyans Kaynağı	Kıdemsiz Grup		Kıdemli Grup	
		Sd	Varyans (%)	Sd	Varyans (%)
Tüm Grup (n=20)	Maddeler (m)	19	9,0	19	14,5
	Puanlayıcılar (p)	19	2,8	19	20,2
	Maddeler x Puanlayıcılar (mxp)	361	88,2	361	65,2
%27'lik Grup (n=10)	Maddeler (m)	19	0,0	19	30,9
	Puanlayıcılar (p)	9	6,4	9	6,3
	Maddeler x Puanlayıcılar (mxp)	171	93,6	171	62,8

Tablo 6 incelendiğinde, kıdemsiz grup için varyans kaynakları olarak puanlayıcılar ana etkisi tüm grupta %2,9 iken alt grupta %6,4'tür. Ayrıca maddeler ana etkisi tüm grupta %9 iken alt grupta varyans gözlenmemiştir. Maddeler x puanlayıcılar etkisinin tüm grupta %88,2 ve alt grupta %93,6 varyans oranlarına sahip olduğu görülmektedir. Bu sonuçlara göre maddelere ve puanlayıcılara ait her iki ana etkinin de varyans yüzdesinin oldukça düşük olduğu görülmektedir. Her ne kadar puanlayıcılara ait varyans açıklama yüzdesi düşük olsa da mxp varyans yüzdesinin büyük olması puanlayıcıların belirledikleri kesme puanlarının maddelere göre farklılaştığını ifade etmektedir. Ayrıca modelle açıklanamayan yüzeylerin süreçte çok etkili olduğunu söylemek mümkündür. Bu durumda kıdemsiz puanlayıcıların Nedelsky yöntemi ile belirledikleri standartların kalitelerinin düşük olduğu yorumunu yapmak yanlış olmayacaktır. Ayrıca, kıdemsiz grubun homojenliği arttıkça yani kıdem seviyesi oldukça azaldığında puanlayıcıların tutarsızlık oranının da arttığını söyleyebiliriz. Bu sonuçların, Nedelsky yönteminde öğrencilerin seçenekleri eleyebilme ihtimalinin tahmin edilmesi gibi karmaşık bir kavramsallaştırma sürecinin yer alıyor olmasından ve kıdemsiz puanlayıcıların bu sürece yaklaşımlarının farklılaşmasından kaynaklandığı düşünülebilir.

Tablo 6'da kıdemli grup için madde ana etkisi tüm grupta %14,5 varyans oranına sahip iken üst grupta bu oran %30,9'a yükselmiştir. Puanlayıcılar ana etkisinin varyansı tüm grupta %20,2 iken bu oran üst grupta %6,3'e düşmüştür. Bu değerler arasındaki fark bize kıdemli gruba oranla üst gruptaki puanlayıcıların daha tutarlı yargılamalar yaptığını ifade etmektedir. Grubun kıdem yılları birbirine yaklaştıkça puanlayıcı sayısındaki önemli azalmaya rağmen yargılarındaki tutarlılık artmıştır. Ayrıca Bu değerler arasındaki fark bize kıdemli gruba oranla üst gruptaki puanlayıcıların daha tutarlı yargılamalar yaptığını ifade etmektedir. Grubun kıdem yılları birbirine yaklaştıkça puanlayıcı sayısındaki önemli azalmaya rağmen yargılarındaki tutarlılık artmıştır. mxp etkileşimine ait varyans yüzdesi ise %65,2'den %62,8'e düşmüştür. Puanlayıcı sayısındaki azalmaya rağmen çalışmada bilinmeyen değişkenlik kaynaklarının etkisinin azalmasına, puanlayıcıların meslekteki çalışma yıllarının ve böylelikle de öğrencileri tanıma seviyelerinin artmasının da katkı sağladığı düşünülebilir.

Nedelsky yöntemi ile yapılan standart belirleme çalışmasının sonuçlarını daha genel olarak karşılaştırılabilmek için G ve Φ katsayıları hesaplanmış ve Tablo 7'de sunulmuştur.

Tablo 7.
Nedelsky yöntemi için puanlayıcı gruplarına ait G ve Φ katsayıları

Puanlayıcı Grubu		G	Φ
Kıdemsiz	Tüm grup	0,67	0,66
	%27'lik grup	0,00	0,00
Kıdemli	Tüm grup	0,82	0,77
	%27'lik grup	0,83	0,82

Tablo 7 incelendiğinde, kıdemli gruba ait değerlerin kıdemsiz gruba ait değerlerden daha yüksek olduğu görülmektedir. Bu durum Nedelsky yöntemi ile standart belirlemede kıdemli grubun kıdemsiz gruba oranla hem bağıl hem de mutlak değerlendirme bakımından daha güvenilir sonuçlar verdiğini ortaya koymaktadır. Kıdemli puanlayıcı grubuna ait G katsayısı üst puanlayıcı grubundan 0,01 puan; Φ katsayısı 0,05 puan daha düşüktür. Kıdemsiz puanlayıcı grubuna ait G katsayısı 0,67 ve Φ katsayısı 0,66 iken alt grupla yapılan çalışmada bu değerler sıfır olarak elde edilmiştir. Bu değerlere göre kıdem yılının azalmasının Nedelsky yöntemi ile belirlenen kesme puanlarının hem bağıl hem de mutlak anlamda güvenirliğini önemli derecede olumsuz yönde etkilediğini söylemek mümkündür.

G çalışmalarından sonra Nedelsky yöntemi ile gerçekleştirilen standart belirleme çalışmalarında kıdemsiz ve kıdemli grupların optimum güvenirlik seviyesinde değerler verebilmesi için çalışmaların kaç puanlayıcıyla yapılması gerektiğini ortaya çıkaran K çalışmaları yürütülmüştür. Bu çalışmalar sonucunda ortaya çıkan değerler Tablo 8'de özetlenmiştir.

Tablo 8.
Nedelsky yöntemi için K çalışması değerleri

Puanlayıcı Grubu	Puanlayıcı Sayıları									
	15		20*		25		30		35	
	G	Φ	G	Φ	G	Φ	G	Φ	G	Φ
Kıdemsiz	0,67	0,66	0,72	0,71	0,75	0,75	0,78	0,77	**0,80**	**0,80**
Kıdemli	**0,82**	0,77	0,85	**0,81**	0,87	0,84	0,89	0,86		

*Çalışmanın yürütüldüğü puanlayıcı sayısını göstermektedir.

Tablo 8'de puanlayıcı sayısının 5'er artırılıp ve azaltılarak yürütülen K çalışmasından elde edilen G ve Φ katsayıları görülmektedir. Katsayılara göre kıdemli gruptaki puanlayıcılar kıdemsiz gruptaki puanlayıcılara göre daha iyi performans göstermişlerdir. 0,80 değeri ölçüt alındığında Nedelsky yöntemi ile uygun G katsayısına ulaşmak için kıdemli grupta 20 kıdemsiz grupta 40, Φ katsayısına ulaşmak için kıdemli grupta 25, kıdemsiz grupta 40 puanlayıcı ile standart belirleme çalışmalarının gerçekleştirilmesinin uygun olacağı değerlendirilmektedir.

Üçüncü Araştırma Problemi ile İlgili Bulgular

Üçüncü alt problem kapsamında "Farklı puanlayıcı grupları ve Ebel standart belirleme yöntemine göre oluşturulmuş maddeler x puanlayıcılar (m x p) deseninin varyans bileşenleri, G ve Φ Katsayıları, K Çalışması Sonuçları Nasıldır?" sorusuna yanıt aranmıştır. Öncelikle, Ebel yöntemi için kıdemsiz ve kıdemli gruplarının yargılarına dayalı olarak elde edilen kesme puanları ile oluşturulmuş m x p çaprazlanmış desenlerine ait varyans bileşenleri ve bu bileşenlerin varyans açıklama yüzdelerini gösteren ANOVA sonuçları Tablo 9'da özetlenmiştir.

Tablo 9.
Ebel yöntemi ile kıdemli ve kıdemsiz gruplara ilişkin ANOVA sonuçları

Puanlayıcı Grubu	Varyans Kaynağı	Kıdemsiz Grup		Kıdemli Grup	
		Sd	Varyans (%)	Sd	Varyans (%)
Tüm Grup (n=20)	Maddeler (m)	19	10,1	19	12,4
	Puanlayıcılar (p)	19	41,7	19	41,4
	Maddeler x Puanlayıcılar (mxp)	361	48,2	361	46,2
%27'lik Grup (n=10)	Maddeler (m)	19	47,9	19	30,9
	Puanlayıcılar (p)	9	4,0	9	6,3
	Maddeler x Puanlayıcılar (mxp)	171	48,1	171	62,8

Tablo 9 incelendiğinde, kıdemsiz grup için varyans kaynakları olarak puanlayıcılar ana etkisi tüm grupta %41,7 iken alt grupta %4'e düşmektedir. Ayrıca maddeler ana etkisinin varyansı tüm grupta %10,1 iken alt grupta %47,9'e yükselmiştir. Benzer şekilde kıdemli grup için puanlayıcılar ana etkisi tüm grupta %41,4 iken alt grupta %6,3'e düştüğü; madde ana etkisinin varyasının tüm grupta 12,4 iken üst grupta 30,9'a yükseldiği görülmektedir. Tüm grup sonuçları, kıdem düzeyindeki heterojenliğin artması ile puanlayıcıların maddeleri güçlük ve uygunluk açısından değerlendirirken oldukça farklı kararlar verdiğini göstermektedir. Ancak modeller arasındaki varyans oranlarındaki değişmeler alt ve üst gruplardaki puanlayıcıların tüm gruplara göre daha tutarlı değerlendirmeler yaptığını ve madde güçlükleri arasındaki farkların daha fazla ortaya çıktığını göstermektedir. Bu durum standart belirleme süreçlerinde istenilen bir durumdur. Angoff ve Nedelsky yöntemlerinden elde edilen bulgular ile tutarlı olarak Ebel yönteminde de kıdem yılları birbirine daha yakın olan puanlayıcıların maddeler hakkındaki yargılarının da birbirine daha fazla yakınsadığı gözlenmiştir. Ek olarak, belirsiz kaynaklardan gelen etki olarak da değerlendirebileceğimiz m x p etkisinin ise Angoff ve Nedelsky yöntemi ile yapılan çalışmalara oranla daha az olduğunu görmekteyiz.

Ebel yöntemi ile yapılan standart belirleme çalışmasının sonuçlarını daha genel olarak karşılaştırılabilmek için G ve Φ katsayıları hesaplanmış ve Tablo 10'da sunulmuştur.

Tablo 10.
Ebel yöntemi için puanlayıcı gruplarına ait G ve Φ katsayıları

Puanlayıcı Grubu		G	Φ
Kıdemsiz	Tüm grup	0,81	0,69
	%27'lik grup	0,91	0,90
Kıdemli	Tüm grup	0,84	0,74
	%27'lik grup	0,84	0,82

Tablo 10 incelendiğinde, kıdemli gruba ait değerlerin kıdemsiz gruba ait değerlerden genel olarak daha yüksek olduğu görülmektedir. Ancak grupların kıdem yılları arasındaki farkın artırılması ve grupların kıdem yılı olarak daha homojen bir yapıya kavuşturulmasının standart belirleme üzerine etkisini incelemek için grupların %27'lik alt ve üst kısımlarında yer alan puanlayıcılarla yapılan analizler neticesinde G ve Φ katsayılarında artış olduğu görülmektedir. Bu durum Ebel yöntemi ile standart belirleme çalışmalarında kıdemli ve kıdem yılı olarak daha homojen yapıdaki puanlayıcı gruplarının yer almasının bağıl ve mutlak anlamda güvenirliği artıracağı şeklinde yargıya varmamız için yeterli delili sunmaktadır.

G çalışmalarından sonra Ebel yöntemi ile gerçekleştirilen standart belirleme çalışmalarında kıdemsiz ve kıdemli grupların optimum güvenirlik seviyesinde değerler verebilmesi için çalışmaların kaç puanlayıcıyla yapılması gerektiğini ortaya çıkaran K çalışmaları yürütülmüştür. Bu çalışmalar sonucunda ortaya çıkan değerler Tablo 11'da özetlenmiştir.

Tablo 11.
Ebel yöntemi için K çalışması değerleri

Puanlayıcı Grubu	Puanlayıcı Sayıları									
	15		20*		25		30		35	
	G	Φ	G	Φ	G	Φ	G	Φ	G	Φ
Kıdemsiz	0,76	0,63	**0,81**	0,69	0,84	0,74	0,86	0,77	0,88	**0,80**
Kıdemli	**0,80**	0,60	0,84	0,74	0,87	0,78	0,89	**0,81**	0,90	0,83

*Çalışmanın yürütüldüğü puanlayıcı sayısını göstermektedir.

Tablo 11'de puanlayıcı sayısının 5'er artırılıp ve azaltılarak yürütülen K çalışmasından elde edilen G ve Φ katsayıları görülmektedir. Katsayılara göre kıdemli gruptaki puanlayıcılar kıdemsiz gruptaki puanlayıcılara göre daha iyi performans göstermişlerdir. G ve Φ katsayıları için 0,80 değeri ölçüt alındığında Ebel yöntemi ve kıdemsiz grupla yapılacak standart belirleme çalışmalarında uygun G değerine ulaşmak için 20, uygun Φ değerine ulaşmak için 35 puanlayıcının, kıdemli grupla yapılacak standart belirleme çalışmalarında ise uygun G değerine ulaşmak için 15, uygun Φ değerine ulaşmak için 30 puanlayıcının gerekli olduğu söylenebilir. Ancak puanlayıcı grubunun yeterli kıdem seviyesinde ve kıdem yılları olarak homojen bir yapıda seçilmesi durumunda optimum güvenirlik değerine sahip bir çalışma yapabilmek için gerekli olacak puanlayıcı sayısının düşebileceğini de söylemek mümkündür.

Dördüncü Alt Problemi ile İlgili Bulgular

Dördüncü alt problem kapsamında "Farklı puanlayıcı grupları için maddeler x yöntemler x puanlayıcılar (m x y x p) çaprazlanmış desenine ilişkin varyans bileşenleri, G ve Φ Katsayıları Nasıldır?" sorusuna yanıt aranmıştır. Bu problemin çözülebilmesi amacıyla puanlayıcı grupları için ayrı ayrı olmak üzere m x y x p çaprazlanmış deseni için G çalışması yürütülmüş ve varyans bileşenleri ile varyans yüzdeleri Tablo 12'de sunulmuştur.

Tablo 12.
Kıdemli ve kıdemsiz gruplar ile mxpxy çapraz desenine ilişkin ANOVA sonuçları

Puanlayıcı Grubu		Kıdemsiz Grup		Kıdemli Grup	
	Varyans Kaynağı	Sd	Varyans (%)	Sd	Varyans (%)
Tüm Grup (n=20)	Maddeler (m)	19	10,5	19	12,2
	Yöntem (y)	2	0,0	2	0,0
	Puanlayıcılar (p)	19	0,0	19	4,6
	Puanlayıcılar x Yöntemler (pxy)	38	16,8	38	27,1
	Maddeler x puanlayıcılar (mxp)	361	3,5	361	0,4
	Yöntemler x Maddeler (yxm)	38	0,0	38	0,0
	Maddeler x Puanlayıcılar x Yöntemler (mxpxy)	722	69,2	722	55,7
%27'lik Grup (n=10)	Maddeler (m)	19	6,4	19	26,7
	Yöntem (y)	2	0,0	2	0,0
	Puanlayıcılar (p)	9	0.7	9	0,3
	Puanlayıcılar x Yöntemler (pxy)	18	4,4	18	3,7
	Maddeler x puanlayıcılar (mxp)	171	11,1	171	18,1
	Yöntemler x Maddeler (yxm)	38	0,0	38	3,8
	Maddeler x Puanlayıcılar x Yöntemler (mxpxy)	342	77,4	342	47,4

Tablo 12'de özetlenen kıdemsiz grupla elde edilen kesme puanları ile yürütülen çalışmaya ait varyans bileşenleri incelendiğinde en büyük varyans yüzdesine sahip olan varyans bileşeninin tüm grupta %69,2 ve alt grupta %77,4 ile m x y x p etkileşimine ait olduğu görülmektedir. Bu değer bize

puanlayıcıların farklı yöntem ve maddeler için yargılarında farklılaşmalar olduğunu belirtmektedir. Ayrıca belirsiz hata kaynaklarının süreç üzerindeki etkisinin oldukça yüksek olduğu söylenebilir. p x y etkileşimi varyans açıklama yüzdesi tüm grupta %16,8 iken alt grupta %4,4'e düşmüştür. Bu değişimden dolayı puanlayıcı grubu kıdem anlamında homojenleştikçe standart belirleme yöntemlerdeki farklı tutumları da azalmaktadır diyebiliriz. Tecrübesiz puanlayıcı grubu özellikle Nedelsky yönteminde diğer yöntemlere oranla daha tutarsız yargılarda bulunmuştur. p x m etkileşiminden ortaya çıkan varyans bileşeninin ise toplam varyans içindeki oranı tüm grupta %3,5 iken alt grupta %11'e çıkmıştır. Grup kıdem anlamında homojenleştikçe madde yüzeyinin süreç içindeki etkisinin arttığını söylemek mümkündür.

Tablo 12 incelendiğinde kıdemli grup için modelde en büyük varyans açıklama yüzdesine sahip olan varyans bileşeninin tüm grupta %55,7 ve üst grupta %47,4 ile m x y x p etkileşimine ait olduğu görülmektedir. Ancak bu oranlar kıdemsiz grup sonuçlarına göre oldukça düşüktür. Bu değerler karşılaştırıldığında farklı yöntemler ve farklı maddeler için kıdemli gruptaki puanlayıcıların kıdemsiz gruptaki puanlayıcılara oranla daha tutarlı yargılarda bulundukları söylenebilir. Puanlayıcı varyans bileşenine ait varyans açıklama yüzdesi kıdemli tüm grupta %4,6 iken kıdemli üst grupta %0,3'e düşerken p x y etkileşimi varyans açıklama yüzdesi tüm grupta %27,1 iken üst grupta %3,7'e düşmüştür. Madde yüzeyine ait varyans açıklama yüzdesi kıdemli tüm grupta %12,2'den kıdemli üst grupta %26,7'ye yükselmiş ve p x m etkileşiminin varyans oranı tüm grupta %3,5 iken alt grupta %11'e yükselmiştir. Bu veriler birlikte değerlendirildiğinde üç ayrı standart belirleme yaklaşımıyla yapılan standart belirleme çalışmalarında puanlayıcıların kıdem yıllarının artmasının ve birbirine yaklaşmasının puanlayıcıların yargılarının tutarlılığının artmasına, çalışmadaki hata varyansının azalmasına ve maddeler arasındaki farklılıkların ortaya çıkarılmasına yardımcı olduğu yorumunu yapmak mümkündür.

Puanlayıcı gruplarının m x y x p çaprazlanmış deseni için yürütülen genellenebilirlik çalışmaları sonuçlarının daha genel bir anlamda karşılaştırılabilmesi için yapılan analiz sonucunda elde edilen G ve Φ katsayıları Tablo 13'te sunulmuştur.

Tablo 13.
mxpxy çapraz desenine ilişkin G ve Φ katsayıları

Puanlayıcı Grubu		G	Φ
Kıdemsiz	Tüm grup	0,89	0,87
	%27'lik grup	0,64	0,62
Kıdemli	Tüm grup	0,93	0,88
	%27'lik grup	0,85	0,85

Tablo 13 incelendiğinde, her koşulda kıdemli grup için elde edilen katsayılar kıdemsiz gruba göre daha yüksek çıkmıştır. Bu durum kıdemli grubun kıdemsiz gruba oranla standart belirleme çalışmasında daha güvenilir sonuçlar verdiğinin bir göstergesi olarak yorumlanabilir. Kıdemli ve kıdemsiz puanlayıcı grupları için ayrı ayrı tüm gruptan elde edilen bağıl ve mutlak güvenirlik değerleri alt ve üst gruplarla elde edilen katsayılardan daha yüksektir. Özellikle kıdemsiz puanlayıcı grubunda güvenirlikte dikkate değer bir azalma olduğu görülmektedir. Bu durumun Nedelsky yöntemi ile kıdemsiz puanlayıcıların belirledikleri kesme puanlarının tutarsız olmasından kaynaklandığını söylemek mümkündür. Kıdemli grupta yapılan analizlerde ise G katsayısında 0,08, Φ katsayısında ise 0,03'lük bir azalma meydana gelmiştir. Bu azalmaya puanlayıcı sayısındaki 10 kişilik farklılığın neden olduğu düşünülebilir. Kıdemli ve üst puanlayıcı grupları genel olarak her üç yöntemle de genellenebilirliği ve güvenirliği yeterli seviyede yargılarda bulunmuşlardır.

TARTIŞMA, SONUÇ ve ÖNERİLER

Angoff, Ebel ve Nedelsky standart belirleme yöntemleri ile puanlayıcı grupları için ayrı ayrı oluşturulmuş tek değişkenlik kaynaklı maddeler x puanlayıcılar (m x p) çaprazlanmış deseni için varyans bileşenleri ve G ve Φ katsayıları incelenmiştir. Bu inceleme sonucunda Angoff, Nedelsky ve

Ebel yöntemleri ile matematik başarısını ölçmeye yönelik hazırlanmış 20 maddelik test için kesme puanı belirleme çalışmalarında 20'şer öğretmenden oluşan hem kıdemli puanlayıcı grubun hem de kıdemsiz puanlayıcı grubun yargılarının kendi içlerinde tutarsız sonuçlar verdiği, maddelere ait varyans açıklama yüzdesinin düşük olması nedeni ile maddeler arasındaki farklılıkların yeterince ortaya çıkarılamadığı görülmüştür. Ancak kıdemli puanlayıcı grubun kıdemsiz puanlayıcı grubuna oranla daha uyumlu yargılara vardığı ortaya çıkmıştır. Puanlayıcı gruplarının kıdem yılları olarak homojenleşmesi puanlayıcı tutarlılığının ve maddelerin sürece etkisinin önemli derecede artmasına sebep olduğu sonucuna ulaşılmıştır. Mutlak ve bağıl anlamda güvenirlik değerlerine göre kıdemli puanlayıcıların kıdemsiz puanlayıcılara oranla daha iyi performans gösterdiğini söylemek mümkündür. Bu bulgulara ek olarak Nedelsky yönteminde kıdemsiz grupla elde edilen kesme puanlarına bilinmeyen değişkenlik kaynaklarının etkisinin oldukça fazla olduğu, puanlayıcı kıdemlerinin artması durumunda bu etkinin azaldığı görülmüştür. Nedelsky yöntemi ile standart belirlemede kıdemli puanlayıcıların kıdemsiz puanlayıcılara oranla çok daha iyi performans gösterdiği daha belirgin olarak ortaya çıkmıştır. Bu duruma Nedelsky yönteminin kavramsallaştırılma sürecinin kıdemsiz puanlayıcı grup tarafından daha güç olarak algılanmasının ve bu grupta yer alan puanlayıcıların öğrencileri ve öğrencilerin yeteneklerini yeterince tanımamalarından kaynaklanmış olabileceğini söylemek mümkündür. Nedelsky yönteminde diğer yöntemlere oranla kıdem yılı olarak homojenliğin sürece etkisinin kıdemin sürece etkisinden daha düşük olduğunu söylemek mümkündür.

Angoff, Ebel ve Nedelsky standart belirleme yöntemleri, kıdemli ve kıdemsiz puanlayıcı grupları için tek değişkenlik kaynaklı maddeler x puanlayıcılar (m x p) çaprazlanmış deseninde puanlayıcı sayılarının artırılması ve azaltılması yoluyla yürütülen karar çalışmasında G ve Φ katsayılarına göre çalışmalar için en uygun puanlayıcı sayısına ulaşılmaya çalışılmıştır. Buna göre yeterli güvenirlikte bir standart belirleme çalışması yürütmek için en verimli yöntemin Angoff olduğunu söylemek mümkündür. Ebel yönteminin performansı da Angoff yöntemine oldukça yakındır. Nedelsky yönteminin ise özellikle göreceli olarak daha kıdemsiz puanlayıcı grupları ile yapılan standart belirleme çalışmalarında oldukça verimsiz ve çok puanlayıcı ile çalışma gerekliliğini ortaya çıkaran bir yöntem olduğu ifade edilebilir. Ayrıca, kıdem yılı açısından heterojen puanlayıcı grubuyla elde edilebilecek tutarlı ve genellenebilir değerlendirmeler daha az sayıdaki daha homojen ve kıdemli puanlayıcı gruplarıyla elde edilebilir. İkinci durumun tercih edilmesi, standart belirleme süreçlerinde birçok noktada zaman ve kaynakların daha verimli kullanılmasını sağlayacaktır.

Bu araştırmanın sonuçlarından yola çıkarak uygulamaya dönük olarak şu öneriler söylenebilir: Angoff, Nedelsky ve Ebel yöntemleri kullanılan ve öğretmenlerin puanlayıcı olarak yer aldığı standart belirleme çalışmalarında görev yılı olarak daha yüksek kıdeme sahip ve mümkün olduğunca birbirine yakın çalışma yıllarında puanlayıcılar seçilmesinin çalışmanın güvenirliğini ve çalışma sonucunda belirlenen kesme puanlarının tutarlığını artıracağı düşünülmektedir. Özellikle Nedelsky yöntemi ile standart belirleme çalışmalarında yer alan puanlayıcıların çalışma yılı olarak kıdemsiz olması durumunda bu puanlayıcılara standart belirleme çalışmaları hakkında yeterli ve tatmin edici sürelerde eğitim verilmesinin, aksi taktirde bu puanlayıcıların çalışmada yer almamasının uygun olacağı değerlendirilmektedir. Tek bir analiz ile hem güvenirlik hem de geçerliğin belirlenebilmesini ve puanlayıcılar, yöntemler, maddeler gibi farklı yüzeylerin sürece etkilerinin değerlendirilebilmesini sağlayan Genellenebilirlik Kuramının standart belirleme çalışmalarında daha fazla kullanılması önerilir. Yine, amaçlanan güvenirlik ve genellenebilirlik seviyesinde bir standart belirleme çalışması yürütmek için gerekli olan en uygun puanlayıcı veya madde sayısının tespit edilmesinde GK karar çalışmasına başvurulabilir.

Gelecek araştırmalara yönelik olarak benzer bir araştırmanın Matematik dersi dışında farklı derslere ait testlerin kullanıldığı standart belirleme süreçleri için de yürütülmesi önerilebilir. Yine bu araştırmada puanlayıcıların yalnızca yıl bazında kıdemleri dikkate alınmıştır. Fakat, standart belirleme süreçlerinde en önemli unsur olan puanlayıcıların niteliklerinin sonuçlara etkisinin daha detaylı olarak incelenebilmesi amacıyla farklı niteliklerdeki (akademik başarı durumu, yaş vb.) ve farklı alt yapılardan gelen puanlayıcılar (akademisyen, uzman, yönetici, mezun olunan okul türü, daha önce çalışılan öğrenci profili ve okul türü vb.) ile de benzer çalışmalar yürütülebilir. Standart belirleme uygulamalarının puanlayıcıların bir araya geldikleri ve görüş alışverişinde bulunabilecekleri bir panel şeklinde yapıldığında sonuçların nasıl etkileneceğini inceleyen bir çalışma yararlı olabilir.

KAYNAKÇA

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME) (2014). *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.

Atılgan, H. (2004). *Genellenebilirlik Kuramı Ve Çok Değişkenlik Kaynaklı Rasch Modelinin Karşılaştırılmasına İlişkin Bir Araştırma.* (Yayımlanmamış Doktora Tezi). Hacettepe Üniversitesi, Ankara.

Behuniak, P., Archambault, X. F., & Gable, R. K. (1982). Angoff and Nedelsky Standard Setting Procedures: Implications for the Validity of Proficiency Test Score Interpretation. *Educational and Psychological Measurement*, 42, 247-255.

Bejar, I. I. (1983). Subject Matter Expert's Assesment of Item Statistics. *Applied Psychological Measurement, 7*(3), 303-310.

Brennan, R. L. (2001). *Generalizability Theory - Statistics for Social Sciences and Public Policy*. New York: Springer.

Brennan, R. L., & Lockwood, R. E. (1980). A Comparison of the Nedelsky and Angoff Cutting Score Procedures Using Generalizability Theory. *Applied Psychological Measurement, 4*(2), 219-240.

Büyüköztürk, Ş., Kılıç Çakmak, E., Akgün, Ö. E., Karadeniz, Ş. ve Demirel, F. (2016). *Bilimsel Araştırma Yöntemleri*. Ankara: Pegem Akademi.

Cardinet, J., Johnson, S., & Pini, G. (2009). *Applying Generalizability Theory using EduG (Quantitative Methodology Series)*. New York, NY: Routledge.

Chang, L. (1999). Judgmental Item Analysis of the Nedelsky and Angoff Standard-Setting Methods. *Applied Measurement in Education, 12*(2), 151-165.

Chang, L., Dziuban, C. D., Hynes, M. C., & Olson, A. H. (1996). Does a Standard Reflect Minimal Competency of Examinees or Judge Competency? *Applied Measurement in Education, 9*(2), 161-173.

Cizek, G. J. (1996). An NCME instructional module on: Setting passing scores. *Educational Measurement: Issues and Practice, 15*(2), 20-31.

Cizek, G. J., & Bunch, M. B. (2007). *Standard Setting A Guide to Establishing and Evaluating Performance Standards on Tests.* Thousand Oaks, California: Sage Publications.

Crocker, L., & Algina, J. (2008). *Introduction to Classical and Modern Test Theory*. Mason, Ohio: Cengage Learning.

Çetin, S. (2011). *İşaretleme ve Angoff Standart Belirleme Yöntemlerinin Karşılaştırılması.* (Yayımlanmamış Doktora Tezi). Hacettepe Üniversitesi, Ankara.

Demir, O. (2014). *Angoff, Nedelsky ve Ebel standart belirleme yöntemleri ile belirlenen kesme puanlarının karşılaştırılması*. (Yayımlanmamış Yüksek Lisans Tezi). Abant İzzet Baysal Üniversitesi, Bolu.

Dewey, J. (1997). *Experience and Education*. New York, NY: Touchstone.

Glass, G. V. (1978). Standards and Criteria. *Journal of Educational Measurement, 15*(4), 237-261.

Gündeğer, C. ve Doğan, N. (2014). Angoff, Yes/No ve Ebel Standart Belirleme Yöntemlerinin Karşılaştırılması. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi, 5*(1), 53-60.

Hambleton, R. K. (2001). Setting performance standards on educational assessments and criteria for evaluating the process. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives*. Mahwah, NJ, Lawrence Erlbaum Associates**:** 89-116.

Hambleton, R. K., & Pitoniak, M. J. (2006). Setting performance standards. *In R. L. Brennan (Ed.), Educational Measurement (4 ed.).* Westport, CT: American Council on Education & Praeger.

Hsieh, M. (2013). An application of Multifaceted Rasch measurement in the Yes/No Angoff standard setting procedure. *Language Testing, 30*(04), 491-512.

Institut de Recherche et de Documantation Pedagogique [IRDP]. (2010). *EDUG User Guide*. Neuchatel, Switzerland: Swiss Society for Research in Education Working Group.

Kane, M. (1994). Validating the Performance Standards Associated With Passing Scores. *Review of Educational Research, 64*(3), 425-461.

Kane, M. T. (2001). So much remains the same: Conception and status of validation in setting standards. *In G. J. Cizek (Ed.), Setting performance standards: concepts, methods, and perspectives (pp. 53-88).* Mahwah, NJ: Lawrence Erlbaum.

Kara, Y. ve Kelecioğlu, H. (2015). Puanlayıcı Niteliklerinin Kesme Puanlarının Belirlenmesine Etkisinin Genellenebilirlik Kuramı'yla İncelenmesi. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme, 6*(1), 58-71.

Kozaki, Y. (2004). Using GENOVA and FACETS to set multiple standards on performance assessment for certification in medical translation from Japanese into English. *Language Testing, 21*(1), 1-27.

Lim, G. S., Geranpayeh, A., Khalifa, H., & Buckendahl, C. W. (2013). Standard Setting to an International Reference Framework: Implications for Theory and Practice. *International Journal of Testing, 13*(1), 32-49.

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

McCann, R. G., & Stanley, G. (2006). The Use of Rasch Modeling to Improve Standard Setting. *Practical Assesment, Research & Evaluation, 11*(2).

MEB. (2015b). 2014-2015 Eğitim Öğretim Yılı I. Dönem Ortak Sınavı Test ve Madde İstatistikleri. Ankara. (http://odsgm.meb.gov.tr/test/analizler/docs/2014-2015-1-Donem-Ortak-Sinavlar-Genel-Bilgiler.pdf, Erişim tarihi: 17/10/2016)

MEB. (2016). 2015-2016 Eğitim Öğretim Yılı I. Dönem Ortak Sınavı Test ve Madde İstatistikleri. Ankara. (http://odsgm.meb.gov.tr/test/analizler/docs/2015-2016-ortak-sinav-1-donem-madde-istatistikleri.pdf, Erişim tarihi: 17/10/2016)

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer (Vol. 1)*. Newbury Park, CA: Sage Publications.

Shulruf, B., Wilkinson, T., Weller, J., Jones, P., & Poole, P. (2016). Insights into the Angoff method: results from a simulation study. *BMC Med Educ, 16*, 134.

Taşdelen, G., Kelecioğlu, H. & Güler, N. (2010). Nedelsky ve Angoff Standart Belirleme Yöntemleri ile Elde Edilen Kesme Puanlarının Genellenebilirlik Kuramı ile Karşılaştırılması. *Mersin Üniversitesi Eğitimde Ölçme ve Psikolojide Ölçme ve Değerlendirme II. Ulusal Kongresi, Mersin.*

Tseng, F.-L., Chiou, J.-M., & Sung, Y.-T. (2015). *A Validity Study For Yes/No Angoff Standard Setting Method Using Cluster Analysis*. Paper presented at the 12th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), Taipei, Chinese.

Violato, C., Marini, A., & Lee, C. (2003). A validity study of expert judgment procedures for setting cutoff scores on high-stakes credentialing examinations using cluster analysis. *Eval Health Prof, 26*(1), 59-72.

Wu, S. M., & Tan, S. (2015). Managing rater effects through the use of FACETS analysis: the case of a university placement test. *Higher Education Research & Development, 35*(2), 380-394.

Yousuf, N., Violato, C., & Zuberi, R. W. (2015). Standard Setting Methods for Pass/Fail Decisions on High-Stakes Objective Structured Clinical Examinations: A Validity Study. *Teach Learn Med, 27*(3), 280-291.