



Açıköğretim Sistemi'nde uygulanan çoktan seçmeli bir teste ait madde güçlüklerinin ardışık aralıklar yöntemiyle ölçeklenmesi

Dr. Murat Doğan ŞAHİN^a

^a Anadolu Üniversitesi Açıköğretim Fakültesi

Özet

Bu çalışmanın amacı, Açıköğretim Sistemi'nde uygulanan bir çoktan seçmeli sınava ait maddelerin alan uzmanlarının görüşleri doğrultusunda ardışık aralıklar yöntemi ile ölçeklenmesi ve bu ölçek değerleri ile sınavın uygulanması sonucu elde edilen madde güçlük değerleri arasındaki ilişkinin hesaplanmasıdır. Bu amaçla öncelikle, Anadolu Üniversitesi'nde görev yapan 72 İngiliz Dili eğitimcısından toplanan verilerle İngilizce-I sınavına ait 20 maddeye ilişkin güçlük değerleri ölçeklenmiştir. Hesaplanan ölçek değerleri ile sınavın uygulanması sonucu elde edilen güçlük değerleri arasındaki korelasyonun ise düşük düzeyde ve istatistiksel olarak anlamlı olmadığı sonucuna varılmıştır. Buna göre aynı amaçla her sınav döneminde farklı madde setleri kullanılarak yapılan sınavların güçlük düzeylerinin birbirine denk olmasının sadece uzmanların görüşleri doğrultusunda sağlanmasının oldukça güç olduğu sonucuna varılmıştır. Sonraki çalışmalarda farklı test kuramları ile farklı madde ve farklı uzman gruplarından elde edilecek ölçek değerlerinin karşılaştırılması önerilmiştir.

Anahtar Sözcükler: Ardışık Aralıklar Yöntemi, Ölçekleme, Madde Güçlüğü.

Abstract

The purpose of this study is scaling the item difficulties of an exam applied in Open Education System by applying successive interval method based on the classification of the judges and calculating the correlation coefficient between these scale values and the item difficulty values obtained after the test application. For this purpose, the difficulty values of 20 items belonging to the English-I test were scaled in line with the opinions of 72 English language educators working at Anadolu University. The correlation coefficient calculated between the scale values and item difficulties obtained after the test implementation was low and statistically non-significant. Accordingly, it is considered that it is very difficult to ensure that the difficulty levels of exams made by using different sets of items at each examination period are equal to each other only in accordance with the opinions of experts. It has been proposed that different scale values gathered from different item and expert groups using different test theories should be used in the next studies.

Keywords: Successive Interval Method, Scaling, Item Difficulty.

Kaynak Gösterme

Şahin, M. D. (2018). Açıköğretim Sistemi'nde uygulanan çoktan seçmeli bir teste ait madde güçlüklerinin ardışık aralıklar yöntemiyle ölçeklenmesi. *AUAd*, 4(1), 8-22.

Giriş

Türkiye’de, ulusal ölçekteki merkezi sınavların neredeyse tamamı çoktan seçmeli formattadır. Objektif bir değerlendirme imkanı sunmasının yanı sıra ölçme araçlarında aranan temel özelliklerden biri olan kullanılabilirlik açısından da öğrenci başarısının ve öğretim sürecinin değerlendirilmesi amacıyla kullanılabilir diğer yöntemlere göre daha avantajlı olması, çoktan seçmeli testlerin yüksek önem arz eden (high stakes) sınavlarda tercih edilmesine neden olmaktadır. Diğer ulusal çaptaki sınavlarda olduğu gibi ülkemizde açık ve uzaktan öğrenme kapsamında öğretim faaliyeti gerçekleştiren üniversiteler de öğretimin etkililiği ile ilgili dönüt almak ve öğrenci başarısını değerlendirmek amacıyla çoğunlukla çoktan seçmeli testlerden yararlanmaktadırlar. Çeşitli gerekçelerle sınav sorularının kamuoyuyla paylaşılmak durumunda olması, her bir sınav uygulamasında yeni soruların kullanılması gerekliliğini beraberinde getirmektedir. Bu durum da aynı amaçla farklı zamanlarda uygulanan sınavların ortalama güçlüklerinin ve test maddelerinin güçlük düzeylerine göre dağılımlarının farklılaşması sorununu beraberinde getirmektedir. Çoğu zaman bu testler bir ya da birkaç alan uzmanının maddelerin güçlük düzeyine ilişkin görüşlerine dayanarak oluşturulmaktadır. Ancak uygulama öncesinde yapılacak bu öngörülerin tutarlılığı bir muammadır.

Alanyazında uzmanların madde istatistiklerine ilişkin öngörülerini/tahminleriyle uygulama sonrası elde edilen değerlerin uyumuna ilişkin herhangi bir çalışma olmadığı görülmektedir. Alan uzmanlarının görüşleri, uygulanmalarına gerek olmaksızın maddelerin işlevliliği hakkında doğru bilgi edinmemizi sağlar mı? Bu bilgilere dayalı olarak madde istatistikleri uygulama öncesinde öngörülen bir madde havuzu oluşturmak mümkün müdür? Bu ve bu kapsamdaki sorulara bu araştırmanın sonucunda yanıt bulunacağı değerlendirilmektedir.

Araştırma Sorunsalı

Bu çalışmanın amacı, Açıköğretim Sistemi’nde yer alan bir çoktan seçmeli sınavda kullanılan maddelerin güçlük düzeylerinin alan uzmanlarının görüşleri doğrultusunda ölçeklenmesi sonucu elde edilen ölçek değerleri ile testin uygulanması sonucu elde edilen madde güçlüklerinin (p) arasında bir ilişki olup olmadığının belirlenmesidir. Bu amaçla 2016-2017 güz dönemi ara sınavında kullanılan 20 soruluk İngilizce-I testinin 1789 kişiden elde edilen madde güçlük parametreleri ile alan uzmanlarının görüşlerine dayalı olarak testi oluşturan maddelerin güçlüklerine ilişkin sınıflama yargıları kanununa dayalı ardışık aralıklar yöntemiyle elde edilen ölçek değerleri arasındaki ilişki araştırılmıştır. Elde edilen sonuçlar doğrultusunda testin

oluşturulması aşamasında parametreleri bilinmeyen maddelerin uzman görüşleri doğrultusunda güçlük düzeylerinin belirlenip belirlenemeyeceği tartışılmıştır.

Belirtilen amaç doğrultusunda bu araştırma, aşağıda yer alan sorulara yanıt aramaktadır:

- Uzmanlar tarafından yapılan sınıflama sonucu maddelerin güçlük düzeylerine ilişkin elde edilen ölçek değerleri nasıldır?
- Sınavın uygulanması sonrasında elde edilen güçlük değerleri ile maddelerin güçlük düzeylerinin uzmanlar tarafından sınıflanması sonucu elde edilen ölçek değerleri arasında anlamlı bir ilişki var mıdır?

İlgili Alanyazın

Çalışmanın Kuramsal Temelleri

Çoktan seçmeli bir madde setinin belirli bir gruba uygulanmasıyla elde edilen veriler kullanılarak madde istatistikleri elde edilir. Bu istatistikler, maddenin işlerliğine yönelik yorum yapmamızı sağlayan değerlerdir. Madde analizi sonucu elde edilen bu istatistikler, madde güçlüğü ve madde ayırt ediciliğidir. Madde güçlüğü, bir maddeyi doğru yanıtlayanların tüm yanıtlayıcılara oranıdır. Ayırt edicilik değeri ise maddenin testin bütünüyle olan korelasyonuna ya da testi alan bireyler toplam puanlarına göre sıralandıklarında, bu sıralamada %27'lik üst ve alt gruplar arasındaki maddeyi doğru yanıtlama oranı farkına eşittir (Turgut ve Baykul, 2010). Bu araştırma kapsamında, maddeye dair bu iki temel istatistikten madde güçlüğü üzerinde durulacaktır. “p” ile sembolize edilen ve 0-1 arasında değer alan madde güçlüğü sıfıra yaklaştıkça maddenin zorlaştığı, 1'e yaklaştıkça ise maddenin yüksek bir yüzde tarafından doğru yanıtladığı yani kolaylaştığı sonucuna varılır. Bir başka ifadeyle bu değer 0'a yaklaştı madde güçlüğü artarken, 1'e yaklaştıkça madde güçlüğü azalır.

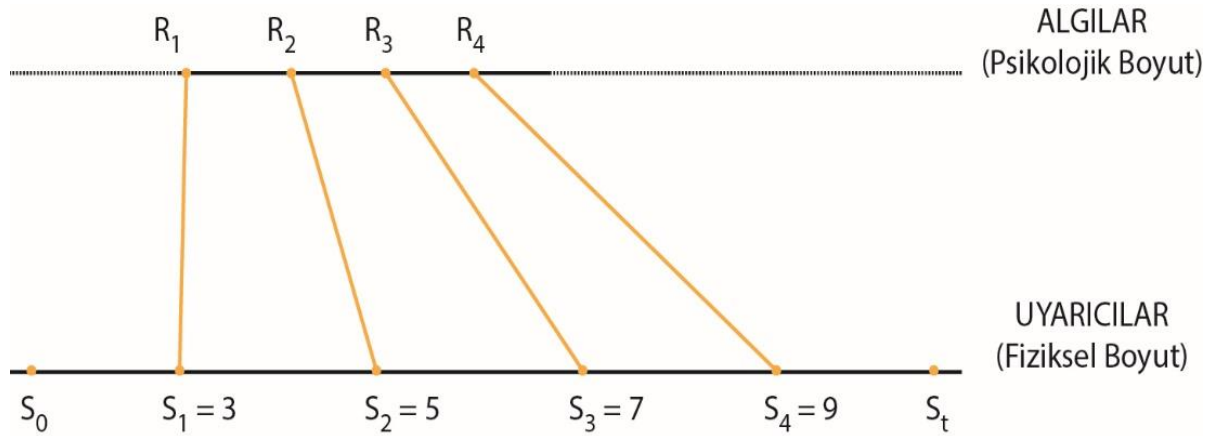
Bir test geliştirme sürecinde, yüksek ayırt edicilik ve orta güçlükteki maddeler seçilmek istenir (Gelbal, 2013). Çok büyük ve heterojen gruplara uygulanan testlerde ise testle ölçülen yetenek açısından birbirinden çok farklı bireylere uygulama yapıldığından, testte farklı güçlük düzeylerinden maddelerin yer alması beklenir. Testin uygulanma amacıyla doğru orantılı olarak güçlük düzeylerine göre madde sayıları değişkenlik gösterebilmektedir. Testler bu duruma göre hazırlanmaya çalışılmakla beraber madde istatistikleri uygulama sonrasında elde edildiğinden öngörülen güçlük düzeyinde maddelerden oluşan, ortalama güçlüğü tahmin edilebilen bir test elde etmek oldukça zor olmaktadır. Özellikle yüksek önem arz eden ulusal çaptaki sınavlardaki en büyük problemlerden biri, test güvenliği açısından parametreleri bilinmeyen maddelerin uygulanmak durumunda olunmasıdır. Bu durum da çoğu zaman güçlük

düzeyi beklenenin üzerinde ya da altında olan sınavların ortaya çıkmasına neden olmaktadır. Maddelerin sadece yazarı tarafından güçlük düzeyinin doğru tahmin edilmesi ise neredeyse imkansızdır.

Uzaktan öğretim kapsamında gerçekleştirilen ölçme-değerlendirme faaliyetleri, yüz yüze yapılan uygulamalar göz önüne alındığında bir takım sınırlılıklar içermektedir (Puspitasari, 2010). Öğrenci sayısının fazla olduğu uzaktan öğretim uygulamalarında, yüz yüze öğretim uygulamalarında öğrenciyle girilen etkileşim kullanılarak öğrencileri değerlendirme imkanı bulunmamaktadır (Koçdar, Karadağ & Şahin, 2016). Bu durum, öğrencilerin ders başarıları ile ilgili dönüt almak isteyen kurumları çoktan seçmeli testlerle sınavlarını yürütme zorunluluğuna itmektedir. Ülkemizde yapılan tüm ulusal çaptaki sınavlarda olduğu gibi açık ve uzaktan öğretim faaliyetleri gösteren kurumlar da uyguladıkları çoktan seçmeli sınavları kamuoyuyla paylaşmak durumundadırlar. Bu nedenle bir sınavda kullanılan bir maddenin, başka bir sınavda tekrar kullanılması, kamuoyuyla paylaşılan soruları çalışmak suretiyle sınavlarına hazırlanan öğrencilerin başarı düzeyi hakkında doğru kestirim yapılamamasına neden olabilmektedir. Buna engel olmak adına akla gelen ilk çözüm, her sınavda yeni madde setlerinin kullanılmasıdır. Her sınavda yeni madde setlerini kullanmak sınavın amacı ve güvenliği açısından önemli olmakla birlikte bu durum farklı problemleri beraberinde getirmektedir. Bunlardan ilk akla gelenler; sınırlı bir kaynağa bağlı olarak hazırlanan testlerin dersin amacının dışında soru sorularak geçerliği düşük testler oluşturulmasına neden olması ve yıllara göre ortalama güçlüğü değişkenlik gösteren testlerin ortaya çıkmasına neden olunmasıdır.

Her ne kadar sınav hazırlama sürecinde görev alan (soru yazarı, sınav uzmanı vb.) uzmanların görüşlerine göre farklı dönemlerde aynı amaçla uygulanan başarı testlerinin zorlukları benzer düzeyde tutulmaya çalışılsa da sadece bir ya da birkaç uzman görüşüne dayanarak bunu sağlamak oldukça zordur. Buna bağlı olarak , uzman sayısı artışıyla beraber maddelerin güçlük düzeylerinin ne derece doğru tahmin edilebileceği sorusunu akla gelmektedir.

Uyarıcıların gerçek büyüklükleri algılanan büyüklükleri farklılık gösterirler. Psikofizik, bu durumu aşağıdaki biçimde şematize eder:



Şekil 1. Uyarıcılara ait fiziksel değerler ile algılanan psikolojik değerlere ilişkin gösterim (Turgut ve Baykul (1992, s.10)'dan alınmıştır.)

Yukarıdaki örnekte görüldüğü gibi uyarıcıların gerçekteki değeri ile algılanan değeri farklılık gösterebilir. Psikofizik, uyarıcılara ait fiziksel ve algılanan büyüklükler arasında bir ilişki bulmayı hedefler. Bununla beraber, bu büyüklüklerin hatalı olarak ölçülmesi istatistiksel olarak doğru bir ilişkinin elde edilmesine engel olur. Bu nedenle, psikofizik boyutlardaki değişkenlerin birim ve sayılarla ifade edilmesine girilmiş ve böylece ölçekleme yöntemleri ortaya çıkmıştır. Denek yargıcı kararlarına ve denek tepkilerine dayalı olmak üzere temel olarak ölçekleme yöntemleri ikiye ayrılır (Turgut ve Baykul, 1992).

Thurstone, psikofizik ölçme alanındaki çalışmaları psikolojik ölçmeye taşımak için önemli çalışmalar yürüten bilim insanları arasında yer alır. Bireye ve nesnelere dair bazı özellikler fiziksel büyüklüklere ilişkin olduğundan, Thurstone tarafından geliştirilen yargı yaklaşımının bu tür özelliklerin ölçeklenmesinde kullanılması kaçınılmazdır (Erkuş, 2014). Yargıcı kararlarına dayalı ölçekleme yöntemleri; ikili karşılaştırma, sınıflama, mutlak yargılar, sıralama yargıları gibi bir çok farklı yöntemi içermektedir (Turgut ve Baykul, 1992). Bu çalışma kapsamında sınıflama yargılarına dayalı ardışık aralıklar yöntemi kullanılmıştır. Ölçeklenecek uyarıcı sayısının ikili karşılaştırma yapmak için uygun olmadığı durumlar için alternatif olarak ortaya çıkan bu yöntem, ilk kez Thurstone tarafından ele alınmıştır (Turgut ve Baykul, 1992).

Sınıflama yargıları kanunu, uyarıcıların ardışık aralıklarla sınıflandığı durumlarda kullanılır. Bu yöntemin temel amacı, aralıkların sınırlarıyla uyarıcıların ölçek değerleri arasındaki ilişkiyi ortaya koymaktır. Bu yöntem kapsamında gözlemciler K tane uyarıcının tamamı verilir. Gözlemcilerden beklenen, bu uyarıcıların her birinin önceden belirlenen sıralı

sınıflardan birine yerleştirmeleridir (Turgut ve Baykul, 1992). Buna dayalı olarak yapılan istatistiksel işlemler sonucunda uyarıcılar ölçeklenmiş, bir başka ifadeyle bir sayı doğrusunun üzerine ölçek değeri en küçük olan uyarıcı 0'da olacak biçimde yerleştirilmiş olurlar. Sınıflama işlemi yapılırken her bir sınıfın, sınır değerlerinin belirlenmesi gerekir. Bir sınıfın sınır değeri ile uyarıcının ölçek değeri arasındaki fark şu şekilde belirlenir (Torgerson, 1958; akt, Turgut ve Baykul, 1992):

$$(1) \quad t_g - S_j = z_{jg} \sqrt{a_j^2 + a_g^2 - 2 \cdot r_{jg} \cdot a_j \cdot a_g}; \quad (j = 1, 2, \dots, K)$$

$$(g = 1, 2, \dots, n)$$

K: Uyarıcı sayısı

n-1: Sınıf sayısı

t_g: g sınır noktasının ortalama değeri

σ_g: g sınırına ait gözlemci yargılarının standart sapması

S_j: U_j uyarıcısının ölçek değeri

σ_j: U_j sınırına ait gözlemci yargılarının standart sapması

r_{jg}: U_j uyarıcısı ve g sınırının algılanan değerleri arasındaki korelasyon

z_{jg}: U_j uyarıcısına ait g sınırının hemen altındaki sınıfa konulması sayısına ait oranla ilgili birim normal sapma

Bu eşitlik, sınıflama yargıları kanununun genel halidir. Ancak bu genel hal uygulanabilirlik açısından pratik olmadığından, ek sayılılarla bazı basitleştirmelere gidilmiş ve A, B, C ve D hali olmak üzere dört hal denklemi elde edilmiştir. Bu denklemlerden çoğunlukla D hali kullanılmakta olup, ikili karşılaştırmalardaki V. hal eşitliğine benzemektedir (Turgut ve Baykul, 1992). D hali için kullanılan nihai denklem aşağıda görülmektedir:

$$(2) \quad S'_j = \frac{1}{K \cdot n} \cdot \sum_{j=1}^K \cdot \sum_{g=1}^n z_{jg} - \frac{1}{n} \cdot \sum_{g=1}^n z_{jg}$$

ya da daha sade haliyle

$$(3) \quad S'_j = \bar{z}_{..} - z_j. \quad (j = 1, 2, \dots, K)$$

formülleri kullanılabilir.

Son aşamada elde edilen veri setinin kurulan model ile uyumunu araştırmak amacıyla iç tutarlılık hesaplanır. Z' matrisini elde etmek için kullanılan formül şöyledir:

$$(4) \quad z'_{jg} = t_g - S_j \quad (j=1, 2, \dots, K)$$

$$(g=1, 2, \dots, n)$$

Sonraki adımda ise tablo yardımıyla Z' matrisine karşılık gelen teorik yığılmalı oranlar matrisi (P') oluşturulur ve ardından P' ile P matrisi arasındaki farklara dayalı elde edilen matris kullanılarak aşağıdaki eşitlikte görülen işlemler yapılır. Bu yolla ardışık aralıklar yönteminde D hali için iç tutarlılık katsayısı (A.D.) elde edilmiş olur:

$$(5) \quad AD = \frac{1}{K.n} \sum_{j=1}^K \sum_{g=1}^n |p'_{jg} - p_{jg}|$$

Yöntem

Araştırma Modeli

Bu araştırmanın amacı, alan uzmanlarının sınıflama yargılarına dayalı olarak çoktan seçmeli bir testteki maddelerin güçlük düzeylerinin ölçeklenmesi ve bu ölçek değerleri ile testin uygulaması sonrasında elde edilen madde güçlük değerleri arasındaki ilişkinin incelenmesidir. Buna dayanarak bu araştırmanın tarama modelinde bir araştırma olduğu değerlendirilmektedir. Karasar (1991)'a göre tarama modelindeki araştırmalar günümüzde ya da geçmişte var olan bir durumun olduğu biçimiyle betimlenmeye çalışıldığı araştırma modelleridir.

Çalışma Grubu

Bu çalışmada Açıköğretim Sistemi ara sınavında uygulanan İngilizce sınavı maddeleri uzman görüşleri doğrultusunda ölçeklenmiştir. Söz konusu uzman grup, Anadolu Üniversitesi Yabancı Diller Yüksek Okulu (YDYO)'nda görev yapmakta olan İngiliz Dili eğitimcilerinden oluşmaktadır. Oluşturulan veri toplama aracı YDYO'da görev yapan tüm öğretim elemanlarına gönderilmiş, analizler gönüllü olarak katılım gösteren 72 alan uzmanından elde edilen veriler kullanılarak yapılmıştır. Araştırma kapsamında sadece Anadolu Üniversitesi YDYO öğretim elemanlarından veri toplanmasının sebebi, bu uzman grubunun Açıköğretim Sistemi ve bu sistemdeki öğrenci profili hakkında bilgi sahibi olmasıdır.

Veri Toplama Araçları

Araştırma kapsamında kullanılan veri setinden ilki, 2016-2017 öğretim yılı Açıköğretim Sistemi güz ara sınavı İngilizce-I dersindeki 20 maddenin güçlük değerleridir. 1789 öğrencinin verdiği yanıtlardan elde edilen madde güçlük değerleri, İstatistik Bilgi Edinme ve Değerlendirme biriminin izniyle kullanılmıştır.

Araştırma kapsamındaki söz konusu 20 çoktan seçmeli test maddesinin ölçeklenmesi işlemi için söz konusu madde seti, Anadolu Üniversitesi YDYO’da görev yapmakta olan öğretim elemanlarına sunularak maddeleri güçlük düzeylerine göre “çok kolay – kolay – orta – zor – çok zor” şeklinde derecelendirilmiş 5 kategoride sınıflamaları, bir başka ifadeyle her maddeyi güçlük düzeyleri için belirlenen bu sıralı sınıflardan birine yerleştirmeleri istenmiştir.

Veri Toplama Süreci

Araştırma kapsamında uzmanlardan ölçekleme işlemi amacıyla hazırlanan veri toplama aracı, çevrim içi olarak Anadolu Üniversitesi YDYO’da görev yapan alan uzmanlarına 2017 yılı Ekim ayında sunulmuştur. Ortalama yanıtlama süresi 5-7 dakika olarak belirlenmiştir. Veri toplama aracı bir hafta süreyle aktif olmuş, süre sonunda 72 gönüllü katılımcıdan veri toplanmıştır.

Veri Analizi

Araştırma kapsamındaki uzmanların tüm maddeleri sınıflaması zorunlu kılındığından, analizlerde D hali ile tam veri matrisi üzerinden işlemler yürütülmüştür. Ölçekleme işleminin ardından iç tutarlılık katsayısı elde edilmiştir.

Analizler sonucunda Z yığılmalı birim sapmalar matrisini elde etmek amacıyla; 72 uzmanın 20 çoktan seçmeli maddenin güçlük düzeylerine yönelik sınıflamaları frekans matrisine dönüştürülmüş, sorasında ise yığılmalı frekans matrisi elde edilmiştir. Bu matrislerde satırlar uyarıcıları (20 adet çoktan seçmeli madde), sütunlar ise dereceleri (çok kolaydan çok zora 5 kategori) ifade eder. Yığılmalı frekans matrisinde yer alan değerlerin toplam gözlemci sayısına (maddeleri güçlüklerine göre kategorilendiren uzman sayısına) bölünmesiyle P yığılmalı oranlar matrisi elde edilir. Bu matrisin sütun sayısı, frekans ve yığılma frekans matrislerinin sütun sayısından bir eksiktir. Bu durumun sebebi, yığılmalı frekans matrisinin son sütunundaki değer kişi sayısına bölünmesiyle elde edilen 1,00 oranları için Z değerlerinin bulunamayacak olmasıdır. Sonraki adımda, P matrisindeki oranlardan, “p oranlarını z birim normal sapmalarına dönüştüren tablo” kullanılarak birim normal sapmalar matrisi (Z matrisi) elde edilmiştir.

D haliyle tam veri matrisinden ölçekleme işlemi yapmak amacıyla, Z matrisinin sütun ortalamaları alınarak sınır değeri (t_g) elde edilmiştir. Matrisin genel ortalaması bulunduğundan

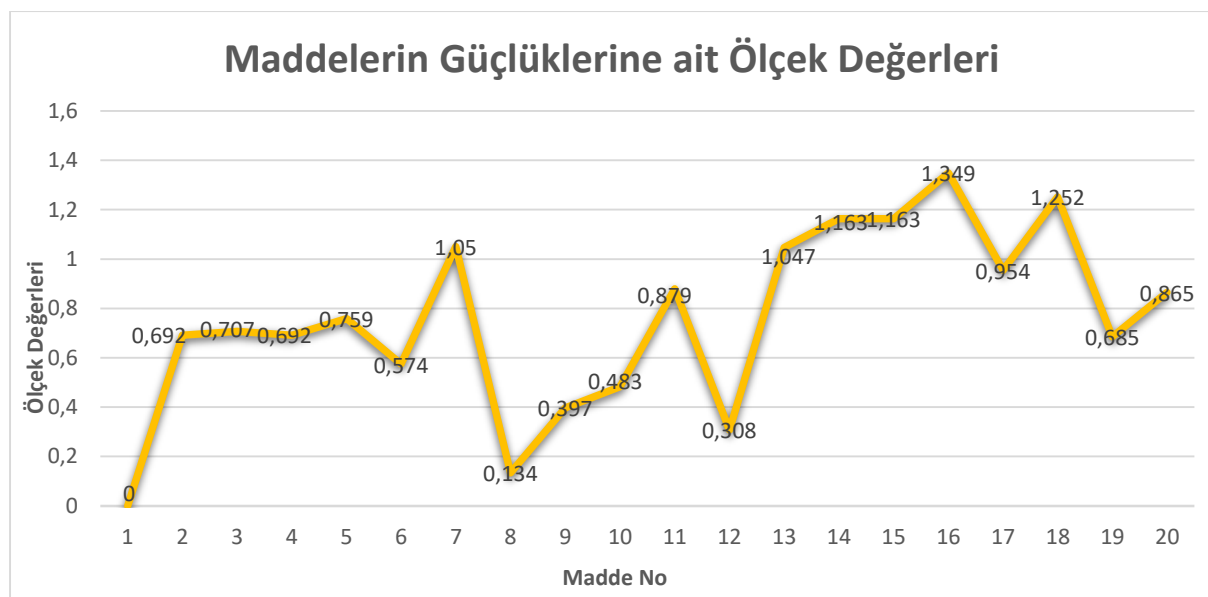
sonra, genel ortalamadan satır ortalamaları çıkarılarak eşitlik (2) ile uyarıcıların ölçek değerleri (S_j) kestirilmiştir. Son adımda ise, en küçük ölçek değerini 0,00'a eşitleyecek değer tüm ölçek kestirimlerine eklenerek ölçekleme işlemi sonlandırılmıştır.

İç tutarlılığı hesaplamak amacıyla öncelikle eşitlik (4) kullanılarak, Z matrisi üzerindeki ölçekleme işlemleri sırasında elde edilen her bir sınıfın sınır değerinden (t_g) her bir uyarıcının ölçek değeri (S_j) çıkarılarak teorik yığılmalı birim normal sapmalar matrisi (Z') elde edilir. Tablodan yararlanılarak Z' değerlerine karşılık gelen P' değerleriyle bir matris oluşturulduktan sonra bu matristeki değerlerden, daha önce ölçekleme işlemi için hazırlanmış olan P matrisindeki değerler her bir hücre birbirine örtüşecek biçimde çıkarılarak P'-P matrisi elde edilir. Son aşamada bu matristen elde edilen değerler (5) nolu eşitlikteki yerlerine konarak iç tutarlılık katsayısı elde edilmiş olur.

Son olarak, maddelerin güçlüğüne ilişkin ardışık aralıklar yöntemiyle elde edilen ölçek değerleri ile testin uygulanması sonucu elde edilen madde güçlükleri arasındaki korelasyon hesaplanmıştır. burada dikkat edilmesi gereken husus, ölçekleme işleminde ölçek değeri arttıkça maddenin zorlaştığı düşünülürken, p değerinin yükselmesinin maddenin kolaylaştığını göstermesidir. Bu nedenle korelasyon ölçek değerleri ile güçlük değerinin 1'den çıkarılmasıyla elde edilen "1-p" değeri arasında hesaplanmıştır.

Bulgular ve Yorumlar

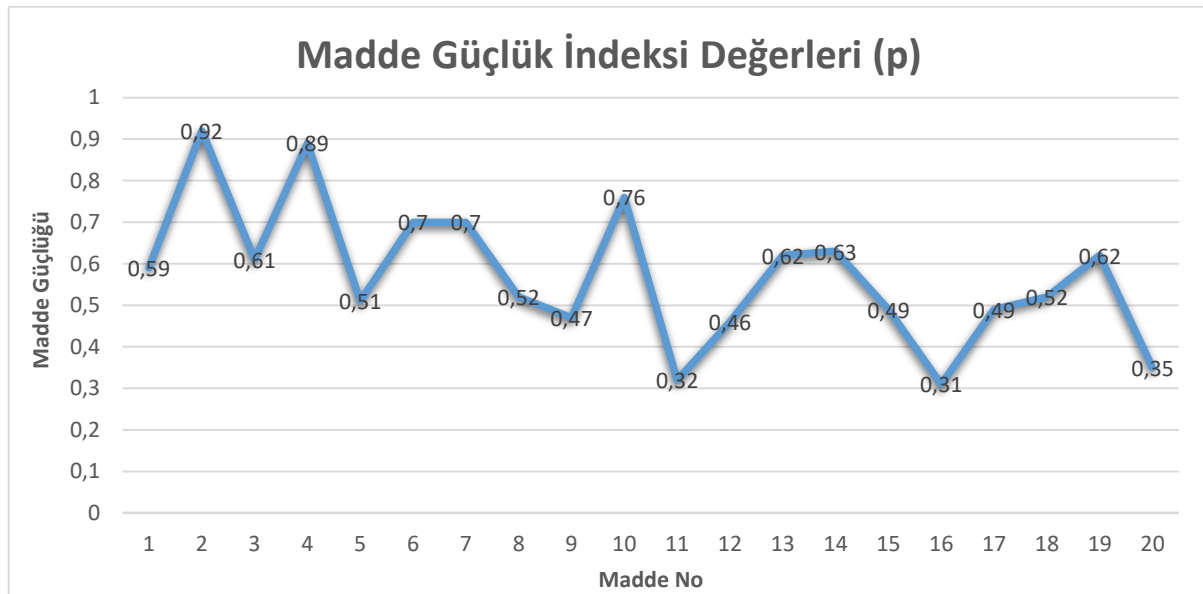
Maddelerin güçlüklerine ait ölçek değerleri aşağıdaki grafikte görülmektedir.



Şekil 2. Ardışık aralıklar yöntemiyle ölçekleme işlemi sonrasında elde edilen değerler

Elde edilen sonuçlar incelendiğinde, uzmanların görüşleri doğrultusunda en kolay maddenin 1 numaralı madde olduğu görülmektedir. En zor madde ise 1.349'luk ölçek değeri ile 16. maddedir. 2 – 5 numaralı maddelerin güçlük düzeyleri orta düzeyde ve birbirine oldukça yakındır. Elde edilen sonuçlar, uzman görüşlerine göre testin 13 nolu sorudan itibaren daha önceki kısma göre daha zor olduğunu göstermektedir. Özellikle 13 – 18 numaralı soruların ölçek değerlerinin dar bir ranjda olması, oldukça yakın bir güçlükte değerlendirildikleri şeklinde yorumlanabilir. Uzman kararlarına dayanarak yapılan ölçekleme işlemi sonrasında en zor maddenin 16. madde olduğu görülmektedir.

2016-2017 öğretim yılı İngilizce-I dersi güz ara sınavına ait madde güçlük indeksi değerleri aşağıdaki grafikte görülmektedir.



Şekil 3. Testin uygulanması sonrasında elde edilen madde güçlük değerleri

Şekil 3'te görüldüğü gibi maddelere ait güçlük değerleri .31 ile .92 arasında değişmektedir. Bir sorunun yanıtlanma oranı arttıkça p değerinin 1'e yaklaştığı dolayısıyla maddenin kolaylaştığı göz önüne alındığında, en kolay maddenin 2 numaralı madde olduğu görülmektedir. .89'luk güçlük değeri ile ikinci en kolay madde ise 4 numaralı maddedir. Testte yer alan maddelerin on beşinin .49 – .76 aralığında güçlük değerlerine sahip olması dikkat çeken bir başka husustur. Ayrıca testin ortalama güçlüğü de .57 olarak hesaplanmıştır. Bu istatistikler, testin güçlük seviyesinin kolaya yakın olduğu biçiminde yorumlanabilir. Bununla beraber testteki en zor maddenin de .31'lik bir güçlük indeksine sahip olduğu görülmektedir.

Dikkat çeken bir husus ise ölçekleme sonucu en büyük ölçek değeri elde ederek uzmanların görüşleri doğrultusunda testin en zor maddesi olduğu öngörülen 16. maddenin, uygulama sonucu en küçük p değerine sahip olması, yani uygulama sonucunda da testin en zor maddesi olmasıdır.

Yukarıda görülen ölçek değerleri için elde edilen iç tutarlılık katsayısı .011 olarak bulunmuştur. Buna bağlı olarak ölçekleme işlemine çok düşük miktarda bir hata karıştığı sonucuna varılmıştır.

Uzman görüşlerine dayalı olarak test maddelerinin ölçek değerlerinin elde edilmesinin ardından araştırma kapsamındaki diğer soruya yanıt aranmıştır. Buna göre, maddelerin güçlük düzeylerine yönelik sınıflama yargıları kanununa dayalı ardışık aralıklar yöntemiyle elde edilen ölçek değerleriyle, maddelerin gerçek güçlük değerleri arasındaki korelasyon katsayısı hesaplanmıştır. Ölçekleme işlemi sonucu elde edilen ölçek değeri arttıkça sorunun tahmin edilen güçlük düzeyi artarken, madde güçlük indeksi olan p değerinde tersine bir durum söz konusudur. Buna göre p değeri arttıkça madde güçlüğü azalmaktadır. Bu nedenle maddelere ait p değerleri 1'den çıkarılmış, bu işlem sonrası elde edilen değer ile madde güçlüğüne dair ölçek değerleri arasındaki korelasyon hesaplanmıştır. Buna göre elde edilen pearson momentler çarpımı korelasyon katsayısı .20, spearman sıra farkları koelasyon katsayısı ise .19 olarak bulunmuştur. Her iki korelasyon katsayısının da .05 düzeyinde istatistiksel olarak anlamlı olmadığı görülmüştür. Bir başka ifadeyle maddelerin, uzmanların sınıflama yargılarına dayalı ölçek değerleri ile testin uygulanması sonucu elde edilen p değerleri arasında anlamlı bir ilişki olmadığı sonucuna varılmıştır.

Sonuçlar

Türkiye'de, ulusal düzeyde yüksek önem arz eden sınavları organize eden ve uygulayan kurumların bilgi edinme kanunu gereği sınav sonrasında soruları kamuoyuyla paylaşmak durumunda olmaları, aynı amaçla yeniden uyguladıkları sınavlarda daha önce uygulanmamış orjinal sorular kullanma zorunluluğunu beraberinde getirmektedir. Bu durum, özellikle sınav yapılan derse ait kaynağın içerğinin sınırlı olduğu açık ve uzaktan öğrenme faaliyeti yürüten kurumların gerçekleştirildiği sınavlarda geçerlik problemini ve sınav güçlüğüünün yıllara göre farklılaşması sonucunu beraberinde getirmektedir.

Bu çalışmada, Açıköğretim Sistemi'nde yer alan öğrenci profili hakkında bilgi sahibi olan 72 kişilik bir uzman grubunun görüşleri doğrultusunda sınıflama yargıları kanununa dayalı ardışık aralıklar yöntemi kullanılarak bir testteki maddeler güçlüklerine göre

ölçeklenmiştir. Sonrasında ise elde edilen ölçek değerleri ile testin uygulanması sonucu elde edilen madde güçlük değerleri arasındaki korelasyon hesaplanmıştır. Buna göre alan uzmanlarının görüşleri doğrultusunda madde güçlük düzeyine yönelik elde edilen ölçek değerleri ile uygulama sonucu elde edilen gerçek güçlük değerleri arasında düşük düzeyde ve istatistiksel olarak anlamlı olmayan bir ilişki olduğu görülmüştür.

Bu sonuçtan yola çıkarak uygulama öncesinde maddenin güçlük düzeyine yönelik öngörünün gerçek sonuçlarla tutarlılık göstermediği sonucuna varılmıştır. Geniş bir uzman grubunun katkılarıyla elde edilen güçlük tahminlerinin gerçek verilerle uyuşmadığı düşünüldüğünde, çok daha küçük grupların görüşleri doğrultusunda hazırlanan sınavların güçlük düzeyine yönelik tahminler konusunda oldukça itinalı davranılması gerektiği söylenebilir. Bununla beraber göz önünde bulundurulması gereken bir başka husus ise Klasik Test Kuramı'nda elde edilen madde ve yetenek parametrelerinin grup bağımlı olmasıdır. Alan uzmanlarının sınıflamalarını Açıköğretim Sistemi öğrenci profilinin göz önüne alarak yapmaları istenirken, madde analizinin elde edildiği grubun genel öğrenci profilini ne derece temsil ettiği konusunda bir fikir sahibi olunmadığı göz ardı edilmemelidir.

Öneriler

Bu çalışma sonucunda alan uzmanlarının maddelerin güçlük düzeylerine ilişkin sınıflama yargıları sonucu elde edilen ölçek değerleri ile maddelerin uygulama sonucu elde edilen güçlük değerleri arasında anlamlı bir ilişki olmadığı görülmüştür. Çalışma kapsamında alan uzmanlarının maddeleri güçlük düzeylerine göre sınıflamaları istenirken kendilerine, madde parametrelerinin hesaplanması ve yorumlanması konusunda ne derece bilgi sahibi oldukları ve Açıköğretim Sistemi'nde yer alan öğrenci profillerini ne derece tanıdıklarına yönelik iki soru daha yöneltilmiştir. Burada; katılımcıların sınıflama yargılarını bu iki hususa göre ayrı ayrı ölçeklemek, buna bağlı olarak elde edilen ölçek değerlerinin farklılaşma durumunu ve hangi alt gruplardan elde edilen ölçek değerlerinin gerçek p değerleri ile daha tutarlı sonuçlar verdiği saptanmak amaçlanmıştır. Ancak söz konusu iki hususta bazı kategorilerdeki uzman sayısının çok az olması ölçeklemenin sağlıklı yapılmasına engel olacağından tek bir ölçekleme değeri elde edilmiştir. Sonraki araştırmalarda bu iki hususun yanına literatürden yararlanılarak başka önemli değişkenler de eklenerek ölçekleme işleminin farklı gruplar için ayrı ayrı yapılabileceği düşünülmektedir.

Klasik Test Kuramı'nda elde edilen madde parametreleri grup bağımlıdır. Bu nedenle elde edilen parametrelerin madde analizinin yapıldığı gruptan başka bir gruba uygulanması

sonucu nasıl bir sonuç elde edileceği bilinmemektedir. Sonraki çalışmalarda; Çüm, Gelbal ve Tsai (2016)'nın çalışmalarına göre farklı örneklemlerde birbiriyle tutarlı madde parametrelerinin elde edildiği sonucuna ulaşılan Sato Test Kuramı kullanılarak elde edilen madde parametreleri ile uzman görüşlerine dayalı ölçekleme işlemi sonunda elde edilen sınıflamaların karşılaştırılması söz konusu olabilir. Ayrıca geniş katılımlı bir sınav uygulaması sonrasında farklı örneklemlerden elde edilen madde güçlük değerlerinin, farklı uzman grupların madde güçlük düzeylerini sınıflamaları ile elde edilen ölçekleme sonuçlarıyla karşılaştırılmasının, hangi örneklemden elde edilen gerçek p değerleri ile hangi grubun yargıları sonucu elde edilen ölçek değerleri arasında daha uyumlu bir sonuç olduğu hususunda daha ayrıntılı bilgi vereceği düşünülmektedir.

Bu sonuçlara göre değerlendirilmesi gereken önemli bir diğer konu, ulusal çaptaki sınavların uygulanması sonrası soruların kamuoyuyla paylaşılması ile ilgilidir. Şeffaflık amacıyla sınav sonrasında soruların açıklanması, sonraki dönemlerde aynı içeriğe yönelik orjinal soru ortaya koyamama, güçlük düzeyinin bilinmediği soruların sorulması gibi öğrenci başarı kestiriminde hataya neden olan durumları ortaya çıkarmaktadır. Ülkemizde özellikle son yıllarda ortaya çıkan sınav objektifliği ve güvenliğine yönelik haklı bir hassasiyet sonucu soruların sınav sonrası kamuoyuyla paylaşımı kaçınılmaz hale gelmektedir. Oysa ki sınava girenlerin yetenek düzeylerini daha iyi kestirebilmek adına madde parametreleri bilinen uygulanmış soruların kamuoyuna açıklanmadan bir soru havuzunda tutulması, bilimsel olarak bilgisayarda bireyselleştirilmiş testler (BBT), paralel testler vb. üst düzey uygulamaların ülkemizde de gelişmesine katkıda bulunacaktır.

Kaynakça

- Çüm, S., Gelbal, S. & Tsai, C.P. (2016). Sato Test Kuramı yöntemleriyle farklı örneklemlerden elde edilen madde parametrelerinin tutarlılığının incelenmesi. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, 7(1), 170-181.
- Erkuş, A. (2014). *Psikolojide ölçme ve ölçek geliştirme*. Ankara: Pegem Akademi.
- Gelbal, S. (2013). *Ölçme ve değerlendirme*. Eskişehir: Anadolu Üniversitesi Açıköğretim Fakültesi Yayınları.
- Karasar, N. (1991). *Bilimsel araştırma yöntemi: kavramlar, ilkeler, teknikler*. Bahçelievler/Ankara.
- Koçdar, S., Karadağ, N. & Şahin, M. D. (2016). Analysis of difficulty and discrimination indices of multiple-choice questions according to cognitive levels in an open and distance learning context. *The Turkish Online Journal of Educational Technology*, 15(4), 16-24.
- Puspitasari, K.A. (2010). Student assessment. *Policy and practice in Asian distance education* (Ed: T. Belawati ve J. Baggaley). New Delhi: SAGE, pp.60-65.
- Turgut, F. & Baykul, Y. (1992). *Ölçekleme teknikleri*. Ankara: ÖSYM Yayınları.
- Turgut, F. & Baykul, Y. (2010). *Eğitimde ölçme ve değerlendirme*. Ankara: Pegem Akademi.

Yazar Hakkında

Dr. Murat Dođan ŞAHİN



Lisans eğitimini Hacettepe Üniversitesi Fen Bilgisi Öğretmenliği'nde, yüksek lisans ve doktora eğitimini ise Hacettepe Üniversitesi Eğitimde Ölçme ve Değerlendirme Anabilim Dalı'nda tamamladı. Temel ilgi alanları arasında çok aşamalı ve bireyselleştirilmiş bilgisayarlı testler, yapısal eşitlik modellemesi, istatistiksel analiz ve psikometri yer almaktadır.

Tel (İş): +90 2223350580/5874

E-posta: mdsahin@anadolu.edu.tr