Journal of Computer and Education Research
December 2021 Volume 9 Issue 18 (ISSN:2148-2896)
http://dergipark.org.tr/jcer

Research Article/Araştırma Makalesi

# A Comparison of Difficulty Indices Predicted by Experts and Calculated Empirically in Multiple Choice Items

Neşe GÜLER *[1] Mustafa İLHAN [2] Gülşen TAŞDELEN TEKER [3]

[1] İzmir Demokrasi University, İzmir, Turkey, gnguler@gmail.com
[2] Dicle University, Diyarbakir, Turkey, mustafailhan21@gmail.com
[3] Hacettepe University, Ankara, Turkey, gulsentasdelen@gmail.com
*Corresponding Author: mustafailhan21@gmail.com

**Article Info**

***Received:*** *26 September 2021*
***Accepted:*** *17 November 2021*

**Keywords:** Item difficulty, expert opinion, estimation of item difficulty index, item analysis

**10.18009/jcer.1000934**

**Publication Language:** Turkish

**Abstract**

In this study, we aimed to compare the difficulty indices predicted by experts and calculated empirically for multiple choice test items. The participants of the research consisted of 10 experts from the field of measurement and evaluation in education, and 222 teacher candidates who study at the education faculty of a state university in Turkey. We collected research data via a measurement-evaluation achievement test developed by the ourselves and containing 25 multiple choice items. The research results revealed that there were positive correlations range through .25 to .71 between item difficulties estimated by experts and the difficulty indices calculated empirically. Nevertheless, we did not observe a clear pattern among the correlation coefficients that can be attributed to the titles of the experts or to whether they had taught the related course before.

**To cite this article:** Güler, N., İlhan, M., & Taşdelen-Teker, G. (2021). Çoktan seçmeli maddelerde uzmanlarca öngörülen ve ampirik olarak hesaplanan güçlük indekslerinin karşılaştırılması. *Journal of Computer and Education Research*, *9*(18), 1022-1036. DOI: 10.18009/jcer.1000934

# Çoktan Seçmeli Maddelerde Uzmanlarca Öngörülen ve Ampirik Olarak Hesaplanan Güçlük İndekslerinin Karşılaştırılması

**Makale Bilgisi**

***Geliş:*** *26 Eylül 2021*
***Kabul:*** *17 Kasım 2021*

**Anahtar kelimeler:** Madde güçlüğü, uzman kanısı, madde güçlük indeksinin kestirimi, madde analizi

**10.18009/jcer.1000934**

**Yayım Dili:** Türkçe

**Öz**

Bu araştırmada, çoktan seçmeli maddeler için uzmanlarca öngörülen ve ampirik olarak hesaplanan güçlük indekslerinin karşılaştırılması amaçlanmıştır. Araştırmanın katılımcıları; ölçme ve değerlendirme alanından 10 uzman ile Türkiye'de bir devlet üniversitesinin eğitim fakültesinde öğrenim gören 222 öğretmen adayından oluşmuştur. Çalışmanın verileri araştırmacılar tarafından geliştirilen ve çoktan seçmeli 25 madde içeren ölçme değerlendirme başarı testi ile toplanmıştır. Araştırma sonuçları, uzman kanılarına dayalı madde güçlükleri ile ampirik olarak hesaplanan güçlük indeksleri arasında .25 ile .71 arasında değişen pozitif yönlü korelasyonlar bulunduğunu ortaya koymuştur. Fakat elde edilen korelasyon katsayıları arasında uzmanların unvanlarına ya da daha önce ilgili dersi yürütüp yürütmemelerine bağlanabilecek net bir örüntü gözlenmemiştir.

# Summary
# A Comparison of Difficulty Indices Predicted by Experts and Calculated Empirically in Multiple Choice Items

Neşe GÜLER *[1] Mustafa İLHAN [2] Gülşen TAŞDELEN TEKER [3]

[1] İzmir Demokrasi University, İzmir, Turkey, gnguler@gmail.com
[2] Dicle University, Diyarbakir, Turkey, mustafailhan21@gmail.com
[3] Hacettepe University, Ankara, Turkey, gulsentasdelen@gmail.com
*Corresponding Author: mustafailhan21@gmail.com

## Introduction

In classical test theory, item difficulty refers to the ratio of the number of individuals who answered the item correctly to the total number of people who answered the item. In item response theory, on the other hand, it is defined as the ability level where the probability of answering the item correctly is .50. The common point of these two definitions is that item difficulty is conceptualized as a statistic that is calculated after the test is administered to individuals. However, in order to develop a test with high reliability and to decide on the cut-off score to be taken into account in the assessment using test results, there is a need to estimate item difficulties based on expert opinions as well as empirically calculating.

Firstly, the items that contributed the most to the test reliability are the items with moderate difficulty. Therefore, most of the items in a test are expected to be of moderate difficulty, and the number of easy and difficult items is expected to be less. Yet, what difficulty level items should be included in the test differs also according to the purpose of the assessment and the group to which the test is applied. For example, a test to identify students with learning disabilities should generally consist of easy items. On the other hand, in a test where only a small number of people will be selected from among a large number of examinees, most of the items must be difficult. In order to provide these, the person(s) who prepared the test and the experts whose opinions were consulted about the test should make a prediction about the difficulty level of the items.

Secondly, there is not always a chance to make a trial application for the developed test. For instance, in high-stake tests, piloting is not possible due to security issues. Similarly, teachers often do not have the opportunity to conduct pilot trials in classroom exams.

Finally, in some standard-setting studies (e.g., Angoff), cut-off scores are obtained by reference to experts' predictions of the probability of an individual at the minimum proficiency level answering the item correctly. In this context, item difficulty is not calculated purely empirically; sometimes it is predicted based on expert opinions. Considering all these issues listed, it is important to investigate the consistency between the actual item difficulties and the experts-predicted difficulty indices. From the point of this view, we aimed to compare the difficulty indexes predicted by experts and calculated empirically in the study.

## Method

The participants of the research consisted of 10 experts from the field of measurement and evaluation in education, and 222 teacher candidates who study at the education faculty of a state university in Turkey. We collected research data via a measurement evaluation achievement test developed by the ourselves and containing 25 multiple choice items. After the we created the test items, we sent them to the 10 experts and these experts independently from each other predicted the difficulty index of each item. They made their predictions not as easy, moderate or difficult; but as numeric values (such as .45). We estimated, on the other hand, the item difficulty indices empirically on the data we obtained by administering the test to teacher candidates. For this, we employed test analysis program (TAP). Then, we calculated the correlation coefficients between the empirically estimated difficulty indices and the item difficulties predicted by the experts. In addition, we analyzed the predictions by 10 experts according to the two-facet Rasch model including items and experts by means of FACETS package program. Thus, based on the predictions of 10 experts, we reached a pooled prediction of the item difficulty indices. We tested the consistency between these pooled predictions and the empirically calculated item difficulties via correlation analysis. In the study, we performed the correlation analyzes in the IBM SPSS Statistics 22 package program.

## Results

According to the research findings, the correlation coefficients between the difficulty indices estimated by the experts and calculated empirically vary between .25 and .71. While the calculated correlation coefficients were significant in four of the 10 experts, they were not statistically significant in the other six experts. After we calculated the correlations between expert predictions and empirical item difficulties separately for each expert, we performed

Rasch analysis to reach a pooled prediction of the item difficulties based on the predictions by 10 experts. We found the correlation coefficient between actual item difficulties, and pooled predictions of experts as .61. Furthermore, when we looked at the positions of the experts on the variable map obtained from the Rasch analysis, we did not observe a clear pattern in the differences between their predictions for item difficulties, which could be attributed to their titles/degrees or to whether they had taught the related course before.

## Discussion and Conclusion

We found that the correlation between the predictions of item difficulties and the actual difficulty indices varied from one expert to another. This finding is in line with the results reported in the study of Tinkelman (1947), Lorge and Diamon (1954) and, Quereshi and Fisher (1977). We did not observe a pattern in the difference between the predictions of experts which can be attributed to their titles, conducting an measurement and evaluation course before, and recognizing the group in which the test was applied. In fact, the percentage of people who answered the item correctly is a sampling dependent parameter. Therefore, it is expected that the expert who teaches the course in the group in which the test is applied will make more accurate predictions for item difficulty compared to other experts. However, research results did not confirm this expectation. In addition, although it is thought that experts who conducted assessment and evaluation courses before, will predict item difficulties more accurately than experts who have no teaching experience, the results obtained did not coincide with this idea. Another important result reached in the research was that there was a strong positive relationship between the pooled predictions of experts and actual item difficulties. This finding is in line with the results of existing studies in the literature (Tinkelman, 1947; Quereshi & Fisher, 1977).

## Giriş

Eğitimde ve psikolojide ölçülmesi amaçlanan tutum, ilgi, başarı, yetenek, zekâ vb. özellikler doğrudan gözlenemez; ancak dolaylı olarak ölçülebilir. Dolayısıyla bu özellikleri ölçmek için tutum ölçeği, başarı testi, ilgi envanteri gibi ölçme araçları kullanılır. Bu ölçme araçları madde ya da soru adını alan uyarıcılardan oluşur ve bireyin ölçme aracındaki maddelere verdiği cevapların ölçülmesi hedeflenen özelliğe sahip oluş düzeyinin bir yansıması olduğu kabul edilir. Bu anlamda ölçme işlemi sonucunda doğru ölçümlere ulaşılabilmesi, ölçme aracındaki maddelerin ölçülmek istenen özelliği başka değişkenlerle karıştırmadan ve olabildiğince hatasız şekilde ölçmesine bağlıdır. Diğer bir deyişle, doğru ölçümler elde etmenin yolu geçerli ve güvenilir maddelerden geçmektedir.

Maddenin geçerlik ve güvenirliği incelenirken niteliksel ve niceliksel değerlendirmeler yapılır. Niteliksel değerlendirmelerde madde; bilimsel açıdan doğruluğu, içeriğinin testin amacına uygunluğu ve biçimsel özellikleri (ifadelerin ve anlatımın açıklığı, yazım ve dil bilgisi kurallarına uygunluğu vb.) bakımından gözden geçirilir. Nicel değerlendirme ise madde analizi sürecini kapsar ve temel olarak madde ayırt edicilik ile güçlük indekslerinin hesaplanmasını gerektirir (Urbina, 2014). Madde ayırt ediciliği, maddenin ölçülen özelliğe yüksek ve düşük düzeyde sahip olan bireyleri birbirinden ne kadar iyi ayırt edebildiğinin bir ölçüsüdür (Domino & Domino, 2006) ve ölçme aracında yer alacak maddelere karar verirken ilk bakılması gereken madde istatistiği olarak ifade edilir. Bununla birlikte, maddenin testin uygulandığı grup için ne düzeyde ayırt edici olduğunu ve ayırt edicilik indeksinin en fazla hangi değeri alabileceğini belirleyen unsur, madde güçlük indeksidir (Salkind, 2018; Uyar, 2019). Bundan dolayı, madde yazarken ve ölçme aracının nihai formunda yer alacak maddelere karar verirken madde güçlüğü açısından da bir değerlendirme yapmak gerekir.

Madde güçlüğü farklı ölçme kuramlarında değişik biçimlerde tanımlanır. Klasik test kuramında (KTK) madde güçlüğü, maddeyi doğru cevaplayan birey sayısının, maddenin uygulandığı kişi sayısına oranına karşılık gelir (Cohen & Swerdlik, 2018). Bu hesaplamaya göre madde güçlük indeksi 0 ile 1 aralığında değişen değerler alır ve 0'a yaklaştıkça madde zorlaşırken 1'e yaklaştıkça madde kolaylaşır (Frey, 2015). Madde tepki kuramında (MTK) ise madde güçlüğü maddeyi doğru yanıtlama olasılığının .50 olduğu yetenek düzeyini ifade eder (Crocker & Algina, 1986). MTK'ya göre madde güçlüğü kuramsal olarak $-\infty$ ile $+\infty$ aralığında uzansa da pratikte –3 ile +3 aralığında değişen değerler alır (Baker, 2001). Madde

güçlüğü için bu iki kurama göre yapılan tanımların ortak noktası; güçlük indeksinin testin bireylere uygulanmasından sonra hesaplanan bir istatistik olarak kavramsallaştırılmasıdır. Ancak, güvenirliği yüksek bir test geliştirebilmek ve test sonucunda yapılacak değerlendirmede esas alınacak kesme puanına karar verebilmek için madde güçlüğünü ampirik olarak hesaplamanın yanı sıra uzman görüşlerine dayalı olarak da tahmin etmeye ihtiyaç vardır.

Öncelikle, testin ayırt ediciliğine ve beraberinde ölçümlerin güvenirliğine en çok katkı getiren maddeler orta güçlükteki maddelerdir (Haladyna & Rodriguez, 2013). Bu sebeple bir testin çoğunlukla orta güçlükte maddelerden oluşması, kolay ve zor madde sayısının ise daha az olması istenir (Özçelik, 2010). Bununla beraber teste hangi güçlükte maddelerin dâhil edilmesi gerektiği, değerlendirmenin amacına ve testin uygulandığı gruba göre de farklılaşır (Whiston, 2017). Mesela öğrenme güçlüğü yaşayan öğrencilerin tespitine yönelik bir test, genel olarak kolay maddelerden oluşmalıdır (Kilmen, 2012). Öte yandan çok sayıda başvuran aday arasından az sayıda kişinin seçileceği bir sınavda maddelerin büyük kısmı zor olmalıdır. Test sonuçlarının ölçüt dayanıklı değerlendirme amacıyla kullanılacağı durumlarda ise güçlük indeksi kesme puanına yakın maddelerin tercih edilmesi önerilir. Söz gelimi, geçme notu 60 ise test genel itibariyle güçlük indeksi .60'a yakın maddeler içermelidir (Mohan, 2016). Bunların sağlanabilmesi, testi hazırlayan kişinin/kişilerin ve test hakkında görüşüne başvurulan uzmanların maddelerin güçlük düzeyine ilişkin bir öngörüde bulunmasıyla mümkün olabilir.

İkinci olarak, literatürde ölçüt dayanaklı ya da mutlak değerlendirme şeklinde ifade edilen değerlendirmelerde kesme puanı, geçme puanı veya ölçüt puan olarak tanımlanan puanın belirlenmesi oldukça önemlidir. Çünkü bu puanlar kullanılarak bireyler hakkında geçti/kaldı ya da başarılı/başarısız gibi önemli kararlar alınır. Kesme puanının belirlendiği bu süreç standart belirleme olarak isimlendirilir ve standart belirleme çalışmalarının bir kısmında uzman görüşlerine başvurulup uzmanlardan maddelerin güçlük düzeylerine ilişkin öngörüde bulunması beklenir. Bu amaçla özellikle son yıllarda yaygın şekilde kullanılmaya başlanan test/sınav yazılımlarında, uzmanlardan yazdıkları maddeleri test havuzuna kaydederken tahmini bir güçlük değerini de sisteme girmeleri istenir. Bu bilgi genelde "Kaydettiğiniz/Sisteme girdiğiniz madde öğrencilerin yüzde kaçı tarafından doğru cevaplanır?" tarzında bir soruya verilen cevapla alınır. Böylece literatürde Angoff olarak adlandırılan standart belirleme yöntemi ile geçme puanı belirlenebilir. Bunun yanı sıra

madde yazarlarından, sisteme girdikleri çoktan seçmeli test maddesinin hangi seçeneklerinin minimum yeterlik düzeyindeki öğrenciler tarafından elenebileceğine dair bir öngörüde bulunmalarının istendiği yazılımlar da mevcuttur. Madde seçenekleri için uzmanlardan bu tür bir bilgi alınması halinde kesme puanı belirlenirken literatürde Nedelsky olarak adlandırılan yöntem işe koşulabilir. Yani hem Angoff hem de Nedelsky standart belirleme yönteminde kesme puanına esas oluşturan uzman görüşlerdir. Buna bağlı olarak uzmanların maddenin sınır yeterlikteki bireyler tarafından elenebilecek seçeneklerine ve güçlük düzeyine dair doğru tahminlerde bulunması önem kazanmaktadır.

Son olarak, geliştirilen test için her zaman bir deneme uygulaması yapma şansı bulunmaz. Örneğin, lise ya da üniversiteye giriş sınavları gibi ulusal çapta yürütülen yüksek riskli sınavlarda güvenlik sorunları sebebiyle pilot uygulama yapmak mümkün olmaz. Benzer şekilde, öğretmenlerin sınıf içinde yaptıkları sınavlarda çoğu zaman deneme uygulaması gerçekleştirme olanakları bulunmaz. Bu nedenle sözü edilen durumlarda uzmanların ve test geliştiricilerin madde güçlüklerine ilişkin tahminde bulunması gerekir.

Yukarıda sıralanan hususların tümü göz önüne alındığında ampirik olarak kestirilen ve uzmanlarca tahmin edilen güçlük indeksleri arasındaki uyumun incelenmesi araştırmaya değer bir konu haline gelmektedir. Bu kapsamda çalışmada, uzmanlarca öngörülen ve ampirik olarak hesaplanan güçlük indekslerinin karşılaştırılması amaçlanmaktadır. Alanyazın incelendiğinde konuyla ilgili benzer çalışmaların (Baykul & Sezer, 1993; Bazvand vd., 2019; Enright & Bejar, 1989; Impara & Plake, 1998; Quereshi & Fisher, 1977; Lorge & Diamon, 1954, Taube & Newman, 1996; Tinkelman, 1947) olduğu görülmektedir. Bu araştırmada bahsi geçen çalışmalardan farklı olarak uzmanlarca öngörülen güçlük indekslerinin; uzmanın unvanına, derse girme tecrübesinin bulunup bulunmamasına ve ilgili dersi testin uygulandığı grupta yürütme durumuna göre farklılık gösterip göstermediğine de bakılacaktır. Bu farklılık, çalışmanın özgün tarafını oluşturmaktadır. Dolayısıyla araştırmanın alanyazına katkı sağlayacağı düşünülmektedir.

## Yöntem

*Katılımcılar*

Araştırmanın katılımcıları; eğitimde ölçme ve değerlendirme alanından 10 uzman ile 2018–2019 öğretim yılı bahar döneminde Türkiye'de bir devlet üniversitesinin eğitim fakültesinde öğrenim gören 222 öğretmen adayından oluşmaktadır. Öğretmen adaylarının 42'si (%18.92) Almanca öğretmenliği, 38'i (%17.12) ilköğretim matematik öğretmenliği, 29'u

(%13.06) İngilizce öğretmenliği, 30'u (%13.51) resim öğretmenliği, 71'i (%31.98) sosyal bilgiler öğretmenliği ve 12'si (%5.41) tarih öğretmenliği programlarına kayıtlıdır. Çalışma kapsamında madde güçlük indeksleri için tahminde bulunan uzmanlar hakkındaki bilgiler ise Tablo 1'de sunulmuştur.

**Tablo 1.** Çalışma grubundaki ölçme ve değerlendirme uzmanlarına ilişkin bilgiler*

| Uzman Sayısı | Açıklama |
|---|---|
| 1 | - Eğitimde ölçme ve değerlendirme alanında doktor unvanına sahiptir.<br>- Ders verme tecrübesi bulunmaktadır.<br>- Testin uygulandığı öğretmen adaylarının ölçme değerlendirme dersini yürüten öğretim üyesidir. |
| 3 | - Eğitimde ölçme ve değerlendirme alanında doktor unvanına sahiptir.<br>- Ders verme tecrübesi bulunmaktadır.<br>- Lisans düzeyinde farklı devlet üniversitelerinde ölçme değerlendirme dersi yürütmüştür. |
| 2 | - Eğitimde ölçme ve değerlendirme alanında doktor unvanına sahiptir.<br>- Ders verme tecrübesi bulunmamaktadır. |
| 4 | - Eğitimde ölçme ve değerlendirme alanında doktora eğitimine devam etmektedir.<br>- Ders verme tecrübesi bulunmamaktadır. |

** Uzmanların, verilerin toplandığı dönemdeki unvanlarını içermektedir. Makalenin yazarları olan ve çalışmada kullanılan başarı testini geliştiren araştırmacılar da bu uzman grubu içerisinde yer almaktadır.*

*Veri Toplama Aracı*

Çalışmada veri toplama aracı olarak araştırmacılar tarafından geliştirilen ve çoktan seçmeli 25 madde içeren ölçme değerlendirme başarı testi kullanılmıştır. Testte; ölçme ve değerlendirmede temel kavramlar, ölçmede hata, korelasyon, ölçme araç ve sonuçlarında bulunması gereken özellikler, ölçme araçlarının sınıflandırılması ve eğitimde kullanılan geleneksel ölçme araçları konuları ile ilgili maddeler bulunmaktadır. Testte yer alan maddelerin ölçülmek istenen kapsama uygunluğuna ilişkin üç ölçme değerlendirme uzmanının, dilbilgisi açısından uygunluğuna ilişkin ise bir dilbilimcinin görüşü alınmıştır. Alınan görüşler doğrultusunda teste son hali verilmiş ve test öğretmen adaylarına uygulanmıştır. Uygulamadan elde edilen verilerin analizinde ulaşılan sonuçlara göre; maddelerin güçlük indeksleri .21 ile .93 arasında ($\bar{X}_{güçlük}$=.60) değişirken, madde ayırt ediciliğine ilişkin nokta çift serili korelasyon katsayıları .23 ile .75 arasında ($\bar{X}_{ayırt\ edicilik}$=.48) sıralanmaktadır. Ölçümlere ait KR–20 iç tutarlık katsayısı ise .71 olarak hesaplanmıştır.

*Verilerin Toplanması ve Analizi*

Test maddeleri oluşturulduktan sonra 10 ölçme değerlendirme uzmanı birbirinden bağımsız olarak her bir maddenin güçlük indeksine ilişkin tahminde bulunmuştur. Tahminler kolay, orta veya zor şeklinde değil; sayısal değerler olarak yapılmıştır (örneğin, .45 gibi). Testin öğretmen adaylarına uygulanmasıyla elde edilen veriler üzerinden ise madde güçlük indeksleri ampirik olarak kestirilmiştir. Bunun için test analiz programından (Test Analysis Program-TAP) yararlanılmıştır. Ardından ampirik olarak kestirilen güçlük indeksleri ile uzmanlarca öngörülen madde güçlükleri arasındaki korelasyon katsayıları hesaplanmıştır. Ayrıca, 10 uzman tarafından yapılan tahminler, FACETS paket programında *maddeler* ve *uzmanlar* şeklinde iki yüzeyli bir desen ile Rasch modeline göre analiz edilmiştir. Böylece 10 uzmanın tahmininden hareketle, madde güçlük indekslerine ilişkin ortak bir tahmine ulaşılmıştır. Madde güçlüklerine ilişkin tahminlerin Rasch modeline göre analiz edilmesiyle ulaşılan değerler ile ampirik olarak hesaplanan madde güçlükleri arasındaki tutarlılık, yine korelasyon analizi ile incelenmiştir. Çalışmada korelasyon analizleri, IBM SPSS Statistics 22 paket programında yapılmıştır.

## Bulgular

Testte yer alan 25 madde için uzmanlarca öngörülen ve ampirik olarak hesaplanan güçlük indeksleri arasındaki korelasyonlar Tablo 2'de sunulmuştur. Tabloda; doktor unvanına sahip olup testin uygulandığı grupta ölçme ve değerlendirme dersini yürüten öğretim üyesi DS; doktora unvanına sahip olup lisans düzeyinde ölçme ve değerlendirme dersi yürütme tecrübesi bulunan uzmanlar T+1, T+2 ve T+3; doktora unvanına sahip olup ders yürütme tecrübesi bulunmayan uzmanlar T–1 ve T–2; doktora öğrencileri ise DÖ1, DÖ2, DÖ3 ve DÖ4 şeklinde kodlanmıştır.

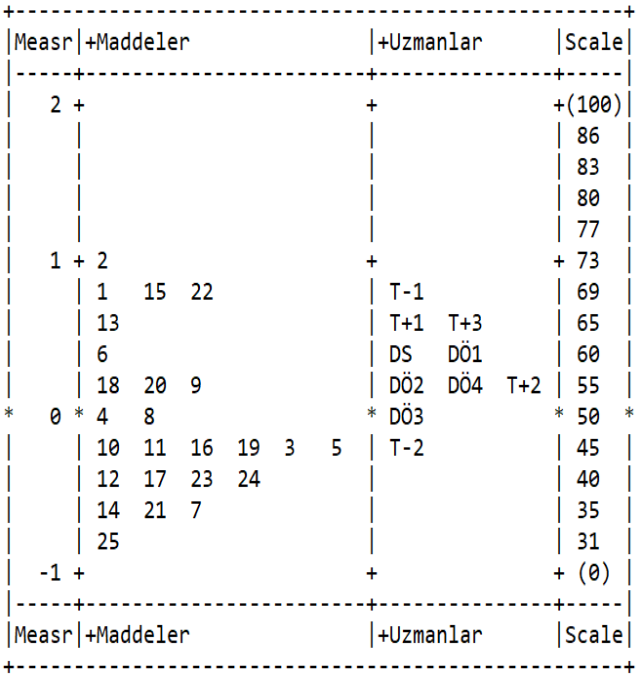
**Tablo 2.** Maddeler için uzmanlarca öngörülen ve ampirik olarak hesaplanan güçlük indeksleri arasındaki korelasyon katsayıları

| Uzmanlar | DS | T+1 | T+2 | T+3 | T–1 | T–2 | DÖ1 | DÖ2 | DÖ3 | DÖ4 |
|---|---|---|---|---|---|---|---|---|---|---|
| r | .47* | .71* | .39 | .49* | .38 | .26 | .25 | .45* | .34 | .39 |

*$p$<.05

Tablo 2'ye göre uzmanlarca tahmin edilen ve ampirik olarak hesaplanan güçlük indeksleri arasındaki korelasyon katsayıları .25 ile .71 arasında değişmektedir. Hesaplanan korelasyon katsayıları 10 uzmanın dördünde (DS, T+1, T+3 ve DÖ2) anlamlı bulunurken;

diğer altı uzmanda istatistiksel açıdan anlamlı çıkmamıştır. Uzman tahminleri ile ampirik madde güçlükleri arasındaki korelasyonlar her bir uzman için ayrı ayrı hesaplandıktan sonra 10 uzman tarafından yapılan tahminlerden yola çıkarak madde güçlük indekslerine ilişkin ortak bir tahmine ulaşmak için Rasch analizi uygulanmıştır. İki yüzeyli (uzmanlar ve maddeler) Rasch analizinde rapor edilen logit cetvel Şekil 1'de sunulmuş ve uzmanlar için Tablo 1'de yapılan kodlama logit cetvelde de benimsenmiştir.




**Şekil 1.** Rasch analizinden elde edilen logit cetvel

Şekil 1'e bakıldığında uzmanların logit cetvel üzerinde farklı noktalarda yer aldığı görülmektedir. Bunun nedeni uzmanların madde güçlükleri için yaptıkları tahminler arasındaki farklılıklardır. Nitekim Rasch analizi sonuçlarına göre uzman yüzeyinde anlamlı farklılık tespit edilmiş ($\chi^2$=526.7, *sd*=9, $p$<.01), diğer bir deyişle uzmanların madde güçlükleri için yaptığı tahminler arasındaki farkın istatistiki bakımdan anlamlı olduğu saptanmıştır. Yine Şekil 1'de görüldüğü gibi uzmanlar arasında unvanlarına veya daha önce ölçme-değerlendirme dersi yürütmüş olma durumlarına göre bir kümelenme söz konusu değildir. Madde güçlükleri için ampirik olarak hesaplanan değerler ve Rasch analizi sonucunda ulaşılan uzman tahminleri, aralarındaki korelasyon katsayısı ile birlikte Tablo 3'te verilmiştir.

**Tablo 3.** Madde güçlükleri için ampirik olarak hesaplanan değerler ve Rasch analizi sonucunda ulaşılan uzman tahminleri

| Madde No | Tahminler (Rasch Analizi-Logit Birimi) | Ampirik Güçlük İndeksleri | Madde No | Tahminler (Rasch Analizi-Logit Birimi) | Ampirik Güçlük İndeksleri |
|---|---|---|---|---|---|
| M1 | .77 | .88 | M14 | –.53 | .62 |
| M2 | 1.08 | .74 | M15 | .79 | .72 |
| M3 | –.19 | .62 | M15 | –.26 | .21 |
| M4 | .03 | .45 | M17 | –.32 | .55 |
| M5 | –.17 | .40 | M18 | .25 | .84 |
| M6 | .36 | .71 | M19 | –.22 | .29 |
| M7 | –.63 | .45 | M20 | .14 | .62 |
| M8 | –.05 | .67 | M21 | –.51 | .73 |
| M9 | .25 | .57 | M22 | .90 | .93 |
| M10 | –.23 | .45 | M23 | –.33 | .78 |
| M11 | –.23 | .62 | M24 | –.30 | .34 |
| M12 | –.40 | .58 | M25 | –.77 | .42 |
| M13 | .56 | .76 | | | |
| r = .61, p < .01 | | | | | |

Tablo 3'e göre, uzman tahminleri ile ampirik madde güçlükleri arasındaki korelasyon katsayısı .61 olarak bulunmuştur. Rowntree (1981) korelasyon katsayısı için mutlak değerce .00 ile .20 arasındaki değerlerin çok zayıf, .20 ile .40 arasındaki değerlerin zayıf, .40 ile .60 arasındaki değerlerin orta düzey, .60 ile .80 arasındaki değerlerin güçlü ve .80 ile 1.00 arasındaki değerlerin çok güçlü ilişkiyi ifade ettiğini belirtmektedir. Buna göre, uzmanların madde güçlüklerine ilişkin öngörülerinin Rasch modeline göre analiz edilmesiyle ulaşılan değerler ile ampirik olarak hesaplanan güçlük indeksleri arasında pozitif yönlü, anlamlı ve güçlü sayılabilecek bir ilişki olduğu anlaşılmaktadır.

## Tartışma, Sonuç ve Öneriler

Bu araştırmada çoktan seçmeli maddeler için uzmanlarca tahmin edilen ve ampirik olarak hesaplanan güçlük indeksleri karşılaştırılmıştır. Ulaşılan bulgular madde güçlüklerine ilişkin gerçek indeksler ile tahmin edilen değerler arasındaki korelasyonun bir uzmandan diğerine farklılaştığını ortaya koymuştur. Bu bulgu, Tinkelman (1947), Lorge ve Diamon (1954) ile Quereshi ve Fisher'in (1977) çalışmasında ulaşılan sonuçlarla paraleldir. Tinkelman (1947) yaptığı çalışmada gerçek test uygulamasından önce madde güçlük indeksinin ne kadar doğru tahmin edilebileceği sorusuna cevap aramıştır. Araştırmada, 30 uzman 100

maddelik çoktan seçmeli bir testte her bir maddeye doğru cevap vermesi muhtemel aday yüzdesi hakkında tahminde bulunmuştur. Tahminlerin geçerliği, öngörülen güçlük düzeyleri ile gerçek güçlük indeksleri karşılaştırılarak test edilmiştir. Sonuçta, tahminlerin geçerliğinin bir uzmandan diğerine farklılaştığı saptanmıştır. Lorge ve Diamon (1954), 14 uzmandan 45 maddelik bir aritmetik testindeki maddelerin güçlük indekslerini tahmin etmesini istemiştir. Yapılan tahminler ile maddelerin gerçek güçlük indeksleri arasındaki korelasyonlar her bir uzman için ayrı ayrı hesaplanmış ve hesaplanan katsayıların .40 ile .82 arasında değiştiği tespit edilmiştir. Aynı şekilde, Quereshi ve Fisher (1977) 44 maddenin güçlük indeksi için beş uzmanın yaptığı tahminleri, maddelere ait gerçek güçlük indeksleri ile karşılaştırmıştır. Araştırmada, tahmin edilen ve ampirik olarak hesaplanan güçlük indeksleri arasındaki göreli uyumun bir uzmandan diğerine farklılaştığı rapor edilmiştir.

Araştırmada uzmanların madde güçlük indeksleri için yaptıkları tahminlerdeki farklılıkta; unvanlarına, daha önce ölçme değerlendirme dersi yürütme ve testin uygulandığı grubu tanıma durumlarına bağlanabilecek net bir örüntü gözlenmemiştir. Aslında, maddeyi doğru cevaplayan kişi yüzdesi örnekleme bağımlı bir parametredir. Dolayısıyla testin uygulandığı grupta ders yürüten uzmanının diğer uzmanlara kıyasla madde güçlüğü için daha doğru tahminler yapması beklenir. Ancak araştırma sonuçları, bu beklentiyi doğrulamamıştır. Ayrıca, daha önce ölçme ve değerlendirme dersi yürütmüş uzmanların ders yürütme tecrübesi olmayan uzmanlara göre madde güçlüklerini daha doğru tahmin edeceği düşünülmesine rağmen ulaşılan sonuçlar bu düşünceyle de örtüşmemiştir. Bu durum örneklem özellikleri ile ilişki olabilir. Şöyle ki bu çalışma lisans düzeyinde ve ölçme değerlendirme dersi kapsamında uygulanan bir başarı testi üzerinden gerçekleştirilmiştir. Üniversitede ders yürüten öğretim elemanlarının öğrencileri tanıma olanakları ilköğretim ve lise kademelerine kıyasla daha azdır. Buna bağlı olarak benzer bir çalışmanın ilköğretim ve lise kademesinde yürütülmesi önemlidir. Daha açık bir anlatımla, lise ya da ortaokul öğrencilerine uygulanacak bir başarı testi için hesaplanan madde güçlük indeksleri ile bu testin uygulandığı sınıflarda derse giren ilgili branştaki öğretmenlerin madde güçlük tahminleri karşılaştırılıp bu araştırmada ulaşılan sonuçların genellenebilirliği sınanmalıdır.

Araştırmada ulaşılan bir diğer önemli sonuç, 10 uzmanın yaptığı tahminlerin Rasch modeline göre analiz edilmesiyle ulaşılan değerler ile gerçek madde güçlükleri arasında pozitif yönlü güçlü bir ilişki bulunmasıdır. Bu bulgu, Tinkelman (1947) ile Quereshi ve Fisher'in (1977) araştırmasının sonuçları ile aynı eksendedir. Sonuç itibariyle, araştırma

bulguları genel anlamda konuya ilişkin mevcut literatür ile desteklenmektedir. Ancak, beklenenin aksine uzmanların unvanları, ders yürütme tecrübeleri ve testin uygulandığı grubu tanıma durumları ile madde güçlüklerine ilişkin tahminleri arasında bir ilişki gözlenmemiştir. Bu sonucun bu araştırmaya özel mi olduğunu belirlemek ve genellenebilirliği hakkında bir yargıya varmak için konuya ilişkin ileri araştırmaların yapılması gereklidir.

Özetlemek gerekirse uzman kanılarına dayalı madde güçlüklerinin geçerliği ve hangi uzmanların daha doğru tahminler yapabildiği günümüzde üzerinde daha fazla durulması gereken bir konuya dönüşmüştür. Çünkü günümüzde çevrimiçi uygulanan birçok sınavda, madde havuzundan istenen niteliklerde maddelerin seçimine dayanan sınav yazılımlarından yararlanılmakta ve bu yazılımların arka planındaki algoritma, güçlük ve ayırt edicilik gibi madde istatistiklerinin sisteme işlenmiş olmasını gerektirmektedir. Madde havuzuna kaydedilen maddelerin gerçek güçlük değerleri bilinmiyorsa genelde ilgili maddelerin güçlük indeksleri için uzmanların yaptığı tahminler sisteme kaydedilmektedir. Dolayısıyla seçme, yerleştirme, durum ya da düzey belirleme gibi amaçlarla oluşturulan testlerin ve test puanları kullanılarak yapılan standart belirleme çalışmalarının geçerlik ve güvenirliği uzmanlarca öngörülen madde güçlüklerinin doğruluğu ile yakından ilişkilidir. Bu durum uzman kanılarına dayalı madde güçlüğüyle ilgili çalışmaların önemini yansıtmaktadır. Bu araştırma bahsedilen düşünceyle ortaya çıkmış olup konu hakkında farklı öğretim düzeylerinde ve gruplarda benzer çalışmaların yapılması alanyazına katkı sağlayacaktır.

*Bilgilendirme*

*Bu çalışmada kullanılan verilerin 2020 yılı öncesine ait olduğu araştırmacılar tarafından onaylanmıştır. Çalışma, 19–22 Haziran 2019 tarihleri arasında Ankara Üniversitesinde düzenlenen VI. Uluslararası Avrasya Eğitim Araştırmaları Kongresi'nde (VI. International Eurasian Educational Research Congress) sözlü bildiri olarak sunulmuştur*

*Yazar Katkı Beyanı*

***Neşe GÜLER:*** *Alanyazın taraması, kavramsallaştırma, metodoloji, veri toplama aracının geliştirilmesi ve verilerin toplanması, inceleme-yazma ve düzenleme.*

***Mustafa İLHAN:*** *Alanyazın taraması, metodoloji, veri toplama aracının geliştirilmesi, verilerin toplanması ve analizi, inceleme-yazma ve düzenleme.*

***Gülşen TAŞDELEN TEKER:*** *Alanyazın taraması, metodoloji, veri toplama aracının geliştirilmesi ve verilerin toplanması, inceleme-yazma ve düzenleme.*

## Kaynakça


Baker, F. (2001). *The basics of item response theory* (2nd ed.). ERIC Clearinghouse on Assessment and Evaluation.

Baykul, Y., & Sezer, S. (1993). Deneme yapılamayan durumlarda madde güçlük ve ayırıcılık gücü indekslerinin ve bunlara bağlı test istatiklerinin kestirilmesi [Özet]. *Eğitim ve Bilim, 17*(83).

Bazvand, A. D., Kheirzadeh, S., & Ahmadi, A. (2019). On the statistical and heuristic difficulty estimates of a high stakes test in Iran. *International Journal of Assessment Tools in Education*, *6*(3), 330–343. https://doi.org/10.21449/ijate.546709

Cohen, R. J., & Swerdlik, M. E. (2018). *Psychological testing and assessment: An introduction to tests and measurement* (9th ed.). NY: McGraw-Hill Education.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. NY: Holt, Rinehart and Winston.

Domino, G., & Domino, M. L. (2006). *Psychological testing: An introduction* (2nd ed.). NY: Cambridge University.

Enright, M. K., & Bejar, I. I. (1989). *An analysis of test writers' expertise: Modeling analogy item difficulty.* Alınan yer https://files.eric.ed.gov/fulltext/ED395014.pdf

Frey, B. B. (2015). *100 questions (and answers) about tests and measurement*. CA: Sage.

Haladyna, T. M., & Rodriguez, M. C. (2013). *Developing and validating test items*. NY: Routledge.

Impara, J. C., & Plake, B. S. (1998). Teachers' ability to estimate item difficulty: A test of the assumptions in the Angoff standard setting method. *Journal of Educational Measurement*, *35*(1), 69-81. https://doi.org/10.1111/j.1745-3984.1998.tb00528.x|

Kilmen, S. (2012). Madde analizi, madde seçimi ve yorumlanması. N. Çıkrıkçı Demirtaşlı, (Ed.), *Eğitimde ölçme ve değerlendirme* içinde (s. 363–385). Ankara: Öz Baran Ofset.

Lorge, I., & Diamon, L. K. (1954). The value of information to good and poor judges of item difficulty. *Educational and Psychological Measurement, 14*(1), 29–33. https://doi.org/10.1177/001316445401400103

Mohan, R. (2016). *Measurement, evaluation and assessment in education*. PHI Learning Pvt.

Özçelik, D. A. (2010). *Test hazırlama kılavuzu*. Ankara: PegemA.

Quereshi, M. Y., & Fisher, T. L. (1977). Logical versus empirical estimates of item difficulty. *Educational and Psychologıcal Measurement*, *37*(1), 91–100. https://doi.org/10.1177/001316447703700110

Rowntree, D. (1981). *Statistics without tears: A primer for non-mathematicians*. Ally & Bacon.

Salkind, N. J. (2018). *Tests & measurement for people who (think they) hate tests & measurement* (3rd ed.). CA: Sage.

Taube, K. T., & Newman, L. S. (1996, 8–12 April). *The accuracy and use of item difficulty calibrations estimated from judges' ratings of item difficulty* [Conference presentation]. Annual Meeting of the American Educational Research Association, New York.

Tinkelman, S. (1947). Difficulty prediction of test items. *Teachers College Contributions to Education*, *941*, 55.

Urbina, S. (2014). *Essentials of psychological testing* (2nd ed.). Wiley.

Uyar, Ş. (2019). Madde puanları üzerinde istatistiksel işlemler. N. Doğan, (Ed.), *Eğitimde ölçme ve değerlendirme* içinde (s. 377–399). Ankara: Pegem Akademi.

Whiston, S. C. (2017). *Principles and applications of assessment in counseling* (5th ed.). Cengage Learning.