Necatibey Eğitim Fakültesi Elektronik Fen ve Matematik Eğitimi Dergisi (EFMED)
Cilt 13, Sayı 1, Haziran 2019, sayfa 105-124. ISSN: 1307-6086

Necatibey Faculty of Education Electronic Journal of Science and Mathematics Education
Vol. 13, Issue 1, June 2019, pp. 105-124. ISSN: 1307-6086

Araştırma Makalesi / Research Article

# INVESTIGATION OF CLASSIFICATION INDICES ON TIMSS-2015 MATHEMATIC-SUBTEST THROUGH BAYESIAN AND NONBAYESIAN ESTIMATION METHODS*

**Serpil ÇELİKTEN [1], Mehtap ÇAKAN [2]**

[1] Dicle Üniversitesi, Diyarbakır, sserpilcelikten@gmail.com,

[2] Gazi Üniversitesi, Ankara, cakanmehtap@hotmail.com,

Received : 16.05.2019 Accepted : 19.06.2019

Doi: 10.17522/balikesirnef.566446

*Abstract* – Purpose of this study is to compare the classification accuracy and consistency indices at different sample sizes in terms of Bayesian estimation methods with MAP, EAP and Nonbayesian estimation methods with MLE, WLE in the framework of IRT. In this direction, ability estimations based on MLE, WLE, MAP and EAP were obtained for each sample size. Then, for each condition of sample size, classification accuracy and consistency indices were calculated by using the Rudner' s appoach. According to the findings of study, it is seen that classification indices based on Nonbayesian methods are more accurate and consistent than the indices obtained based on Bayesian methods. Among Nonbayesian methods, it is concluded that MLE leads the more accurate and consistent classification indices than WLE. However, when the post hoc tests and effect sizes are investigated, it is seen that all pairs that results in significant difference have small effect in practice.

*Key words*: Bayesian-Nonbayesian estimation methods, classification accuracy and consistency indexes, item response theory

## Summary

### Introduction

Decision is a crucial part of the educational measurement processes for both the individual and the society because development of the many fields are related with a well-designed testing processes. At this point the term of classification arises with the terms of classification accuracy

* This paper is presented at the 6th International Congress on Measurement and Evaluation in Education and Psychology, as oral presentation.

and classification consistency which are two important index to evaluate the quality of the testing processes (Cizek ve Bunch, 2007). Classification accuracy refers to the validity of the classification (Lee, Hanson ve Brennan, 2000) whereas classification consistency refers to the reliability of the classifications (Barnett ve Macmann, 1992). Since these indexes provides validity and reliability evidences, it is important to use these indexes to evaluate the classification decisions resulting from the testing processes.

There are many methods to calculate classification accuracy and consistency indices. IRT which has been getting attention due to its usability at many fields is an effective way to obtain these indexes through single administration. Under IRT, there are many approaches such as Rudner's approach or Lee's approach available to calculate classification indexes. Selection of the related approach depends on the aim of the study by the way the background of the data and the measurement processes. Moreover, selection of the estimation methods is an important aspect that could influence the classification results. At this point, Bayesian and Nonbayesian ability estimation methods could be considered to trace the effect of estimation methods on classification accuracy and consistency. Since Bayesian and Nonbayesian methods are two different approaches with different mathematical background, it is important to evaluate the results at different conditions. When studies were investigated, it was seen that different sample sizes could be a factor that influence the measurement results in terms of Bayesian and Nonbayesian methods (Kadane, 2015; Lee & Song, 2004; Mislevy, 1986). Therefore, sample size could be a factor to be manipulated so as to provide suggestions for the appropriate method selection at many changing conditions.

In this way, purpose of this study is to compare the classification accuracy and consistency indexes at different sample sizes in terms of Bayesian estimation methods with MAP, EAP and Nonbayesian estimation methods with MLE, WLE in the framework of IRT by using Rudner's approach on the data obtained through the Mathematics subtest of TIMSS 2015.

**Methodology**

In this study, classification accuracy and consistency indices were were obtained based on differerent ability estimation methods and these were obtained at the different sample sizes with two condition small (N=461) and large (N=965). Since this study investigated the classification indexes at different conditions in a comparative way, this study could be called as a causal-comparative study (Fraenkal, Wallen ve Hyun, 2008).

The study was conducted on the data obtained from the Mathematic subtest of TIMSS 2015, in the framework of IRT. For the analyses of the IRT, required assumptions were checked. Then the Bayesian and Nonbayesian ability estimations were obtained. These analyses were

conducted through R studio. For the significance tests and post hoc test, friedman and wilcoxon's signed rank teste were used, respectively. For these anaylses SPSS was used.

**Results**

At the condition of small sample size, classification accuracy values are higher with Nonbayesian methods than Bayesian methods with statistically significant difference. To compare in detail, MLE (0,78) is higher than the WLE (0,71) whereas classification accuracy values are same for MAP (0,70) and EAP (0,70) at this condition. When post hoc tests were conducted, it was seen that for the pairs MLE-MAP, MLE-EAP, MLE-WLE and WLE-EAP, there are statistically significant differences with low effect sizes. Moreover, results of classification consistency are similar to the results of classification accuracy.

Similarly, at the condition of large sample size, classification accuracy values are higher with Nonbayesian methods than Bayesian methods with statistically significant difference. To compare in detail, different from the results of small sampe size classification accuracy values are same for MLE (0,76) and WLE (0,76) whereas the classification accuracy obtained depending on the method of EAP (0,75) ise higher than the MAP (0,71). When post hoc tests were conducted, it was seen that for the pairs MLE-MAP, MLE-EAP, WLE-EAP, WLE-MAP and MAP-EAP, there are statistically significant difference with low effect sizes. Moreover, results of classification consistency are similar to the results of classification accuracy.

**Conclusion and Discussion**

In this study, classification indexes are higher with Nonbayesian ability estimation methods than Bayesian methods. Moreover, classification indexes with MLE is the highest. This result is consistent with the results obtained in Wyse ve Hao (2012)' s study. They investigated the classification indexes based on Bayesian and Nonbayesian methods both on the real data sets and simulated data sets. They found that on the real data set Nonbayesian methods are better whereas on the simulated data set, classification indexes are higher with Bayesian methods. Since the study is based on TIMSS 2015 data, a real data set, it was concluded that Nonbayesian methods could be expected to lead higher classification values and it could be related to the shape of ability distribution which is more restriced with real data sets. Moreover, when the post-hoc tests were conducted to reveal the pairs that are statistically significant from each other with the effect sizes, it was seen that despite significant differences between many pairs, all of them present low effect sizes in practice. This result could be associated with the distributional features of the ability estimations because distribution for each method is aproximately normal.

When studies are investigated (Lathrop ve Cheng, 2014; Wyse ve Hao, 2012; Zhang, Du, Chen, Xin, ve Chen, 2017), it was understood that difference between the methods are growing when the distributions getting away from the normality.

# BAYESIAN VE NONBAYESIAN KESTİRİM YÖNTEMLERİNE DAYALI OLARAK SINIFLAMA İNDEKSLERİNİN TIMSS-2015 MATEMATİK TESTİ ÜZERİNDE İNCELENMESİ†

**Serpil ÇELİKTEN [1], Mehtap ÇAKAN [2]**

[1] Dicle Üniversitesi, Diyarbakır, sserpilcelikten@gmail.com,

[2] Gazi Üniversitesi, Ankara, cakanmehtap@hotmail.com,

Gönderme Tarihi: 16.05.2019 Kabul Tarihi: 19.06.2019

Doi: 10.17522/balikesirnef.566446

*Özet* – Bu araştırmanın amacı modern test kuramı olan MTK çerçevesinde, Nonbayesian kestirim yöntemlerinden MLE, WLE ve Bayesian kestirim yöntemlerinden MAP ve EAP ile elde edilen yetenek kestirimlerine göre bireylerin sınıflandırılması sonucunda elde edilen sınıflama doğruluğu ve sınıflama tutarlılığı indekslerini farklı örneklem koşullarında karşılaştırmaktır. Bu doğrultuda MTK çerçevesinde her bir örneklem koşulu için MLE, WLE, MAP ve EAP kestirim yöntemlerine dayalı olarak yetenek kestirimleri elde edilmiştir. Sonrasında her bir koşul için, MTK'ya dayalı sınıflama yaklaşımlarından biri olan Rudner'in yaklaşımı kullanılarak sınıflama doğruluğu ve tutarlılığı indeksleri elde edilmiştir. Çalışmanın bulgularına göre Nonbayesian yetenek kestirimlerine dayalı olarak elde edilen sınıflama indekslerinin, Bayesian yöntemlerinden daha doğru ve tutarlı olduğu gözlenmiştir. Nonbayesian yöntemler arasında ise en doğru ve tutarlı sınıflama indekslerinin MLE ile kestirilen yeteneklere dayalı olarak elde edildiği sonucuna ulaşılmıştır. Ancak yapılan ikili karşılaştırma testleri ve pratik anlamlılık değerlerinin incelenmesi sonucunda anlamlı çıkan tüm etkilerin pratikteki etkisinin küçük olduğu gözlenmiştir.

*Anahtar kelimeler*: Bayesian-Nonbayesian kestirim yöntemleri, sınıflama doğruluğu ve tutarlılığı indeksleri, madde tepki kuramı

## Giriş

Eğitim ile ilgili ölçmelerde nihai amaç bireyler hakkında karar vermektir. Bireylere ilişkin ölçme sonuçları kullanılarak yapılan bu karar verme süreci hem bireyler hem de toplum için hayati önem taşımaktadır. Çünkü verilen bu kararlar bireylerin geleceğini şekillendirdiği gibi, sınırlı kaynakların en iyi şekilde kullanılmasını sağlayarak toplumsal gelişmelere de ışık tutmaktadır. Bu nedenle, bireylerin potansiyellerinin uygun bir şekilde kullanılabilmesi için karar verme sürecinin etkili bir şekilde gerçekleştirilmesi gerekmektedir (Cizek ve Bunch,

† Bu çalışma 6. Uluslararası Eğitimde ve Psikolojide Ölçme ve Değerlendirme Kongresi'nde sözlü bildiri olarak sunulmuştur.

2007). Bu doğrultuda sınıflama geçerliğine işaret eden sınıflama doğruluğu (Lee, Hanson ve Brennan, 2000) ile sınıflama güvenirliğine işaret eden sınıflama tutarlılığı (Barnett ve Macmann, 1992) kavramları ele alınmaktadır. Sınıflama doğruluğu tek bir uygulama üzerinde kesme puanının bilindiği varsayılarak gözlenen sınıflama ile gerçek sınıflamanın örtüşme derecesidir. Farklı bir ifadeyle, sınıflama doğruluğu, bireyin gerçek puanına dayalı olarak yapılan sınıflama ile bireyin gözlenen puanına dayalı yapılan sınıflamanın örtüşmesine ilişkin derecenin ifade edilmesinde kullanılmaktadır (Lee, Hanson ve Brennan, 2000). Sınıflama tutarlılığı ise aynı uygulamanın paralel tekrarları üzerinde bireyin aynı kategoride sınıflandırılma derecesini ifade etmektedir. Yani sınıflama tutarlılığı, aynı uygulamaya ilişkin elde edilen iki gözlenen veri setine dayalı olarak elde edilen bir indekstir (Lee, 2010). Bu indeksler kullanılarak sınıflama kararları incelenmektedir. Bireylere ilişkin sağlıklı kararların verilmesi için bireylerin belli düzeylere atandığı durumlarda söz konusu sınıflamaların doğruluk ve tutarlılıklarının incelenmesi bu sürecin geçerlik ve güvenirliklerine ilişkin kanıt sağlaması açısından önem taşımaktadır.

Eğitim alanında bireylerin puanlarının karşılaştırılabilirliğine izin vermesi ve daha adil kararların verilebilmesi adına objektif testler sıklıkla tercih edilmekte ve buna dayalı olarak da bireylerin ilgili alandaki yetenek düzeyleri belirlenerek çeşitli kararlar verilmektedir. Bu doğrultuda bireyler merkezi ya da kurum sınavlarıyla bir takım ölçme ve değerlendirme süreçlerine tabi tutularak yönlendirilmektedir. Ancak bu ulusal sınavların yanı sıra eğitim politikalarının iyileştirilmesine katkıda bulunmayı amaçlayan ve belli aralıklarla uygulanan bir takım uluslararası sınavlar da mevcuttur. Bu amaca dayalı olarak yürütülen projelerden biri de Uluslararası Matematik ve Fen Eğilimleri Araştırması (Trends in International Mathematics and Science Study-TIMSS)' dır. TIMSS çeşitli ülkelerden katılan öğrencilerin performanslarını ölçerek ve bununla birlikte öğrencilere, öğretmenlere, idarecilere ve velilere bir takım anketler uygulayarak veri toplamaktadır. Bu verilere dayalı olarak da ülkelerin eğitim politikalarını düzeltmelerine yönelik kanıt temelli kararlar verilmesine katkıda bulunmaktadır (TIMSS International Study Center). Bu uygulamalarda 4. ve 8. sınıf düzeyindeki öğrenciler ele alınmakta olup bu öğrencilere fen ve matematik alanlarında testler uygulanmaktadır.

TIMSS ülkelerin geneline ilişkin başarı düzeylerini belirlemeyi ve karşılaştırmayı amaçlamaktadır. Bu doğrultuda düşük (low), orta (intermediate), yüksek (high) ve ileri (advanced) olmak üzere dört tane uluslararası kıyaslama noktası (benchmark) belirlenmiştir. Her bir alan için söz konusu kıyas noktalarındaki birey yüzdeleri incelenerek ülkelere dair karşılaştırmalar yapılmaktadır (TIMSS International Study Center). Bu doğrultuda bireylerin

sınıflandırılabileceği 5 düzeyin mevcudiyeti söz konusudur. Bu düzeyler düşükten yükseğe doğru sırasıyla alt düzey altı, alt düzey, orta düzey, üst düzey ve ileri düzey şeklinde adlandırılmıştır (Büyüköztürk, Çakan, Tan ve Atar, 2011).

Bireyler hakkında verilecek kararların doğru bir şekilde yapılabilmesi için bireylerin ilgili yeterlik düzeylerinde sınıflanmasına ilişkin geçerlik ve güvenirliklerin incelenmesi bireyler hakkında verilecek kararlar açısından birçok alanda önem taşımaktadır. Çünkü öğrencilerin belli yeterlik düzeylerine ulaşmış olması onların ilgili alanda sahip olabileceği bilişsel beceriler hakkında bilgi sağlamaktadır. Bu doğrultuda öğrencilerin yeterlik düzeylerinde sınıflandırılmasına ilişkin doğruluk ve tutarlılık incelemelerinin yapılması birçok alan için ihtiyaçtır. Matematik de bu alanlardan biri olarak ele alınabilir. Çünkü Matematik problem çözme, matematiksel okuryazarlık, muhakeme etme, kavram bilgisi ve işlem becerileri gibi önemli bilişsel becerilerin kullanılmasını gerektirmektedir (Altun, 2010). Bu becerilerin ölçülebilmesi noktasında ise gerekli yeterlik düzeylerinin uygun yöntemlerle ele alınabilmesi verilecek kararların uygunluğu açısından önemli aşamalardan biridir. Ölçme ve değerlendirmede ele alınacak yöntemsel süreçlerin doğru bir şekilde yönetilmesi hem öğrencilere sağlıklı geri bildirimlerin verilmesi hem de öğrencilere doğru yönlendirmelerin yapılabilmesi adına öğrenme-öğretme sürecine önemli ölçüde katkıda bulunmaktadır. Çünkü etkin bir ölçme ve değerlendirme sürecinin altındaki temel hususlardan biri öğrencilerden elde edilen söz konusu ölçümlerin incelenmesinde kullanılacak teknik süreçlerin sağlıklı bir şekilde yürütülmesidir.

Ölçme ve değerlendirme sürecinde kullanılmak üzere her bir farklı süreç için çeşitli yöntemler mevcuttur ve söz konusu yöntemler arasından uygun yönteme doğru bir şekilde karar verebilmek test uygulamalarının kritik basamaklarından biridir. Çünkü pratikteki karar verme durumlarına ilişkin sonuçların geçerlik ve güvenirliği, ölçme ve değerlendirme sürecindeki yöntemlerin seçimi ve uygulanmasıyla doğrudan bağlantılıdır. Bu noktada sonuçların doğru bir şekilde yorumlanması ve bireylerin doğru yönlendirilmesi açısından uygun yöntem ile sınıflama güvenirlik ve geçerliğinin incelenmesi önem taşımaktadır.

Sınıflama geçerlik ve güvenirliğini incelemeye yönelik geçmişten bugüne klasik test kuramı (KTK) (Cohen, 1960; Huynh, 1976; Subkoviak, 1976; Livingston ve Lewis, 1995) ve madde tepki kuramı (MTK) (Rudner, 2001; 2005; Lee, 2010) altında birtakım çalışmalar yürütülmüştür. Sınıflama kararlarına ilişkin güvenirlik incelemeleri 1960' lı yıllardan itibaren yapılmaya başlanılmış ve bu doğrultuda çeşitli indeksler önerilmiştir (Cohen, 1960; Hambleton ve Novick, 1973; Swaminathan, Hambleton ve Algina, 1974). Sınıflama tutarlılığını incelemek

üzere önerilen ilk zamanlardaki bu indeksler bireylerin birbirine alternatif olacak iki farklı uygulamadan elde edilmiş puan çiftlerine dayalı olarak hesaplanmaktadır. Bu doğrultuda bireylerin iki farklı uygulama sonucunda geçti/kaldı kararlarına ilişkin uyuşma yüzdesi temel alınarak çeşitli hesaplamalar sunulmuştur. Ancak pratikte alternatif iki uygulama her zaman mevcut olmamaktadır. Bu durum ise verilen kararların tutarlılığının incelenmesinde engel oluşturmaktadır. Dolayısıyla buna çözüm üretmek için tek uygulamaya dayalı sınıflama tutarlılığı indeksleri önerilmiştir (Livingston ve Lewis, 1975; Subkoviak, 1976). Geliştirilen bu yöntemler ile tek uygulamadan elde edilen puanlara dayalı olarak bireylerin geçti/kaldı kararlarının tutarlılığı incelenmektedir. Elde edilen birey puanlarının dağılımsal özellikleri ve maddelerin puanlama şekline göre (iki kategorili, çok kategorili, karma) seçilecek yöntemler farklılık göstermektedir. Bireylerin toplam puanı kullanılarak belli bir kesme puanına dayalı olarak sınıflama tutarlılığı hesaplanmaktadır. Dolayısıyla bu indeksler KTK altında ele alınabilmektedir. Ancak son yıllarda MTK' nın ölçme ve değerlendirme süreçlerinde önemli bir yer tuttuğu ve çeşitli alanlarda etkin bir şekilde kullanıldığı görülmektedir. Çünkü MTK' da KTK' dan farklı olarak bireylerin madde yanıt örüntüleri dikkate alınmakta ve daha hassas kestirimler yapılabilmektedir. MTK altında elde edilen birey puanları bireylerin yanıt örüntülerine dayalı olarak elde edildiği için doğası gereği bu puanlar KTK altında elde edilen puanlardan farklılaşmaktadır. Söz konusu farklılaşmadan dolayı KTK' ya dayalı geliştirilen indekslerin MTK altında elde edilen birey puanlarına dayalı sınıflama çalışmaları için kullanılması sonuçların doğruluğunu düşürebilmektedir. Bu durum MTK' ya dayalı sınıflama indekslerinin geliştirilmesinin önünü açmıştır. Bu doğrultuda başta Lee' nin ve Rudner' in indeksi olmak üzere sınıflama doğruluğu ve sınıflama tutarlılığını değerlendirmek amacıyla çeşitli indeksler geliştirilmiştir (Bourque, Goodman, Hambleton ve Han, 2004; Guo, 2006; Lee, 2010; Rudner, 2001; Rudner, 2005).

Sınıflama geçerliği ve güvenirliğine ilişkin çalışmalar incelendiğinde MTK' ya dayalı sınıflama yaklaşımlarının KTK' ya dayalı sınıflama yaklaşımlarından farklılaştığı ve dağılımsal özelliklerin farklılaştığı durumlarda MTK' ya dayalı yöntemlerin daha iyi sonuç verdiği gözlenmiştir (Lathrop ve Cheng, 2014). Farklı bir ifadeyle belirtilecek olursa bireylerden elde edilen test puanlarının dağılımının normale yaklaşması ve normalden uzaklaşmasına göre yöntemlerden elde edilen sınıflama sonuçları farklılaşabilmektedir. Dolayısıyla kuramlara bağlı olarak yöntem seçiminin sınıflama kararları üzerinde etkisi olduğu sonucuna ulaşılmaktadır. Bunun yanı sıra sınıflama çalışmalarını etkileyen çeşitli faktörler mevcuttur. Bu faktörlerden biri ölçme sonuçlarının doğruluğudur (Yang, Poggio ve Glasnapp, 2006). MTK altında elde edilen sınıflama indekslerinden Rudner' in indeksi yetenek kestirimlerinin standart hata

değerlerine ilişkin vektörü de dikkate alarak sınıflama doğruluğu ve tutarlılığı indekslerine ilişkin hesaplamaları yapmaktadır. Dolayısıyla bu durum farklı kestirim yöntemlerinden elde edilen yetenek kestirimlerine dayalı olarak sınıflama indekslerinin farklılaşabileceğine işaret etmektedir. Ancak ele alınacak yetenek kestirim yöntemlerinin de bir takım faktörlere karşı duyarlılığı değişebilmektedir. Dolayısıyla bu yöntemlerin etkilendiği faktörlerin de ele alınması mevcut koşullardaki uygun yöntem seçiminde önem taşımaktadır. Örneğin Bayesian ve Nonbayesian yaklaşımlar altında yer alan kestirim yöntemlerinin etkililiğinin farklı örneklem koşullarına göre değişebileceği, küçük örneklemin söz konusu olduğu koşullarda Bayesian yöntemlerin daha iyi sonuçlar verebileceği belirtilmiştir (Kadane, 2015; Lee & Song, 2004; Mislevy, 1986). Pratikteki test etme durumları ele alınacak olursa eğitim alanında büyük örneklem koşullarının her zaman mümkün olamayacağı gerçeğinden dolayı küçük örneklem koşullarında en iyi sonuç verecek yöntemin önerilmesi önem taşımaktadır. Dolayısıyla farklı kestirim yöntemlerinin sınıflama çalışmalarındaki etkisinin küçük ve büyük olmak üzere farklı örneklem koşularında incelenmesinin uygun yöntem seçimine katkı sağlayacağı düşünülmektedir.

Bu doğrultuda, araştırmanın amacı MTK çerçevesinde, Nonbayesian kestirim yöntemlerinden Maksimum Olabilirlik Kestirim (MLE) yöntemi, Ağırlıklandırılmış Olabilirlik Kestirim (WLE) yöntemi; Bayesian kestirim yöntemlerinden ise Maksimum a Posteriori (MAP) ve Expected a Posteriori (EAP) kestirim yöntemlerinden elde edilen yetenek kestirimlerine dayalı olarak sınıflama doğruluğu ve tutarlılığı indekslerini karşılaştırmaktır. Bu teknik karşılaştırma TIMSS 2015 Matematik alt testinden elde edilen veri üzerinde TIMSS' in ele aldığı 5 düzey sayısı göz önünde bulundurularak ve Rudner'in yaklaşımı kullanılarak küçük ve büyük olmak üzere iki farklı örneklem koşulu dikkate alınarak yapılmıştır. Bu amaç doğrultusunda aşağıdaki araştırma sorularına yanıt aranmıştır:

*Araştırma soruları*

1) Küçük (461 kişilik) örneklem koşulunda, MLE, WLE, EAP ve MAP kestirim yöntemlerinden elde edilen yetenek kestirimlerine göre bireylerin sınıflandırılması sonucunda elde edilen,

   a. Sınıflama doğruluğu indeksleri birbirinden istatistiksel açıdan anlamlı olarak farklılaşmakta mıdır?

b. Sınıflama tutarlılığı indeksleri birbirinden istatistiksel açıdan anlamlı olarak farklılaşmakta mıdır?

2) Büyük (965 kişilik) örneklem koşulunda, MLE, WLE, EAP ve MAP kestirim yöntemlerinden elde edilen yetenek kestirimlerine göre bireylerin sınıflandırılması sonucunda elde edilen,

a. Sınıflama doğruluğu indeksleri birbirinden istatistiksel açıdan anlamlı olarak farklılaşmakta mıdır?

b. Sınıflama tutarlılığı indeksleri birbirinden istatistiksel açıdan anlamlı olarak farklılaşmakta mıdır?

## Yöntem

### *Araştırmanın Modeli*

Bu çalışmada çeşitli kestirim yöntemlerinden elde edilen yetenek kestirimlerine dayalı olarak bireylerin sınıflandırılması sonucunda elde edilen sınıflama doğruluğu ve tutarlılığı indeksleri farklı örneklem koşullarında karşılaştırmalı olarak incelenmektedir. Bu doğrultuda çalışma nedensel karşılaştırma araştırması olarak ele alınmıştır (Fraenkal, Wallen ve Hyun, 2008).

### *Veri Toplama Aracı*

İncelemeler TIMSS-2015 matematik alt testinin 12. kitapçığından elde edilen veri üzerinde gerçekleştirilmiştir. TIMSS-2015 Matematik alt testi dört alt içerik alanından oluşmaktadır; sayılar (number), cebir (algebra), geometry (geometry) ve veri & olasılık (data and chance) alanlarıdır. Toplam 30 maddeden oluşan bu kitapçıktaki maddeler 0-1 şeklinde ikili kodlanarak ilgili analizler gerçekleştirilmiştir.

### *Çalışma Grubu*

Çalışmanın analizlerini gerçekleştirmek için Türkiye'nin yanı sıra 5 ülkenin (Hong Kong, Kazakistan, Avustralya, Ürdün) verisi dikkate alınmıştır. Çalışma için ülkeler seçilirken TIMSS' in ölçek puanı dikkate alınmıştır. Bu doğrultuda 1921 kişilik bir veri seti oluşturulmuştur. Analizler iki farklı örneklem koşulu üzerinde gerçekleştirileceğinden 1921 kişilik gruptan 1000 ve 500 olmak üzere iki tane seçkisiz örneklem oluşturulmuştur. Oluşturulan örneklemler üzerinde yetenek kestirimleri yapıldıktan sonra bazı bireylere ilişkin kestirimlerin MLE yöntemi altında yakınsamadığı görülmüştür. Yöntemlerin

karşılaştırılabilirliği için bu bireyler çalışma gruplarından çıkarılmıştır. Bu doğrultuda 1000 kişilik gruptan 965; 500 kişilik gruptan ise 461 kişi geriye kalmıştır. Sınıflama doğruluğu ve tutarlılığı analizleri nihai olarak oluşturulan bu iki grup üzerinde gerçekleştirilmiştir.

*Veri Analizi*

Verilerin analizi için öncelikle MTK varsayımları (tekboyutluluk ve yerel bağımsızlık) kontrol edilmiştir. Tekboyutluluk için R studio programında psycho R paketi kullanılarak tetrakorik korelasyon matrisine dayalı açımlayıcı faktör analizi yapılmıştır. Yerel bağımsızlık varsayımı için de mirt R paketi ile Yen' in Q3 istatistikleri incelenmiştir. Yapılan incelemeler sonucunda ilgili varsayımların sağlandığı görülmüştür. Varsayımların kontrollerinden sonra model-veri uyumuna bakılmış ve modelin veriye uyum sağladığı sonucuna ulaşılmıştır. Yetenek kestirimleri mirt R paketi ile üç parametreli lojistik model altında MLE, WLE, MAP, ve EAP olmak üzere dört farklı kestirim yöntemine dayalı olarak elde edilmiştir. Sınıflama çalışmasına geçmeden önce kümeleme analizi ile standart belirleme çalışması yapılmıştır. Standart belirleme çalışması için TIMSS' in belirlediği 5 düzey dikkate alınarak K-Means yaklaşımı ile kümeleme analizi yapılmıştır. Kesme puanlarının oluşturulması için Sireci, Robin ve Patelis (1999)' in çalışmasında bahsedilen sınır grup yöntemi kullanılmıştır. Bu doğrultuda her bir koşul altında her bir kestirim yöntemi için ayrı ayrı standart belirleme çalışması yapılarak kesme puanları belirlenmiştir.

Bu analizlerin ardından yetenek kestirimleri vektörü, yetenek kestirimlerinin standart hatalarının vektörü ve kesme puanları vektörü oluşturulmuş ve bu 3 vektöre dayalı olarak her bir koşul için sınıflama doğruluğu ve sınıflama tutarlılığı indeksleri hesaplanmıştır. Bu indeksler cacIRT R paketi ile Rudner (2001, 2005)' in yaklaşımına dayalı olarak elde edilmiştir. Sınıflama doğruluğu ve sınıflama tutarlılığı indeksleri birey bazında elde edilmiştir. Farklı bir ifadeyle her bir koşul altında her bir birey için sınıflama doğruluğu ve sınıflama tutarlılığı olmak üzere iki farklı değer elde edilmiştir. Her bir örneklem koşulunda bireylerin sınıflama doğruluklarının ve sınıflama tutarlılıklarının istatistiksel açıdan anlamlı olarak farklılık gösterip göstermediğini incelemek amacıyla tekrarlı ölçümler için ANOVA yapılmasına karar verilmiş ancak her bir koşul altında her bir birey için hesaplanan sınıflama doğruluğu ve sınıflama tutarlılığı indekslerine ilişkin ilgili parametrik testin varsayımları sağlanmadığından bu testin nonparametrik karşılığı olan Friedman testi kullanılmıştır. Bu doğrultuda her bir örneklem koşulunda biri sınıflama doğruluklarının biri de sınıflama tutarlılıklarının incelenmesi olmak üzere iki, toplamda ise dört farklı Friedman testi yürütülerek incelemeler yapılmıştır.

Karşılaştırma testi olarak Wilcoxon' un işaretçi sıralar testi kullanılmıştır. Anlamlı çıkan ikili karşılaştırmalara yönelik etki büyüklükleri ise $r = {}^{z}/_{\sqrt{N}}$ formülü (Pallant, 2007) ile hesaplanmış ve sonuçlar Cohen (1988)' in kriterine göre değerlendirilmiştir. Her bir Friedman testinden sonra gerçekleştirilen karşılaştırma testlerinin sayısı altı olduğu için, I tip hatayı kontrol etmek amacıyla alfa düzeyi .05/6=0.008 olarak dikkate alınmıştır.

## Bulgular ve Yorumlar

*Birinci Araştırma Sorusuna Yönelik Bulgular*

*Küçük Örneklem Koşulunda Elde Edilen Sınıflama Doğruluklarının Karşılaştırılmasına İlişkin Bulgular*

Küçük örneklem koşulunda elde edilen sınıflama doğruluklarına ilişkin farklılığın istatistiksel açıdan anlamlılığını incelemek amacıyla Friedman testi yürütülmüş olup bu teste ilişkin sonuçlar Tablo 1' de sunulmuştur:

**Tablo 1** Küçük Örneklem Koşulunda Elde Edilen Sınıflama Doğruluklarının Karşılaştırılmasına İlişkin Friedman Testi Sonuçları

| Yöntemler | N | Medyan | Sıra ortalamaları | Ki kare | sd | p değeri |
|---|---|---|---|---|---|---|
| MLE | 461 | 0,78 | 2,79 | | | |
| WLE | 461 | 0,71 | 2,51 | 37,88 | 3 | 0 ,00* |
| EAP | 461 | 0,70 | 2,39 | | | |
| MAP | 461 | 0,70 | 2,30 | | | |

*p ≤ .01

Friedman testinin sonuçlarına göre küçük örneklem koşulunda farklı yetenek kestirim yöntemleri altında elde edilen sınıflama doğrulukları arasında anlamlı farklılık vardır ($\chi$2:3, 461 = 37.88 , p < .05). Medyan değerleri incelendiğinde ise bu değerlerin sırasıyla MLE (Md=0,78), WLE (Md=0,71), EAP (Md=0,70) ve MAP (Md=0,70) şeklinde MLE' den EAP' a doğru azaldığı ancak MAP ve EAP' a dayalı elde edilen medyan değerlerinin aynı olduğu görülmektedir. Hangi yöntem çiftleri arasında istatistiksel olarak anlamlı farklılık olduğunu ortaya çıkarmak için Wilcoxon' un işaretçi sıralar testi kullanılmıştır.

Wilcoxon' un işaretçi sıralar testine göre MLE ile diğer bütün yöntemler arasındaki farklılığın istatistiksel olarak anlamlı çıktığı ve bu farklılığın her bir karşılaştırma çifti için MLE' nin lehine olduğu gözlenmiştir. Bunun yanı sıra WLE-EAP arasındaki farklılık da manidar çıkmıştır. Anlamlı çıkan bu farklılıklara ilişkin etki büyüklükleri hesaplandığında

MLE-EAP ve MLE-MAP yöntem çiftleri için 0,18 (küçük etki); MLE-WLE yöntem çiftleri için 0,13 (küçük etki) ve WLE-EAP yöntem çiftleri için ise 0,10 (küçük etki) olduğu görülmektedir.

*Küçük Örneklem Koşulunda Elde Edilen Sınıflama Tutarlılıklarının Karşılaştırılmasına İlişkin Bulgular*

Küçük örneklem koşulunda elde edilen sınıflama tutarlılıklarına ilişkin farklılığın istatistiksel açıdan anlamlılığını incelemek amacıyla Friedman testi yürütülmüş olup bu teste ilişkin sonuçlar Tablo 2' de sunulmuştur:

**Tablo 1** Küçük Örneklem Koşulunda Elde Edilen Sınıflama Tutarlılıklarının Karşılaştırılmasına İlişkin Friedman Testi Sonuçları

| Yöntemler | N | Medyan | Sıra ortalamaları | Ki kare | sd | p değeri |
|---|---|---|---|---|---|---|
| MLE | 461 | 0,65 | 2,87 | | | |
| WLE | 461 | 0,58 | 2,53 | 62,39 | 3 | 0 ,00* |
| EAP | 461 | 0,57 | 2,35 | | | |
| MAP | 461 | 0,57 | 2,25 | | | |

*p ≤ .01

Friedman testinin sonuçlarına göre küçük örneklem koşulunda farklı yetenek kestirim yöntemleri altında elde edilen sınıflama tutarlılıkları arasında anlamlı farklılık vardır ($\chi$2:3, 461 = 62.39 , p < .05) Medyan değerleri incelendiğinde ise bu değerlerin sırasıyla MLE (Md=0,65), WLE (Md=0,58), EAP (Md=0,57) ve MAP (Md=0,57) şeklinde MLE' den EAP' a doğru azaldığı ancak MAP ve EAP a dayalı elde edilen medyan değerlerinin aynı olduğu görülmektedir. Hangi yöntem çiftleri arasında istatistiksel olarak anlamlı farklılık olduğunu ortaya çıkarmak için Wilcoxon' un işaretçi sıralar testi kullanılmıştır.

Wilcoxon' un işaretçi sıralar testine göre MAP-EAP yöntem çiftleri hariç bütün yöntem çiftleri arasındaki farklılık istatistiksel olarak anlamlı bulunmuştur. Anlamlı çıkan bu farklılıklara ilişkin etki büyüklükleri hesaplandığında MLE-MAP için 0.25 (küçük etki); MLE-EAP için 0.21 (küçük etki); MLE-WLE için 0.16 (küçük etki); WLE-MAP için 0.09 (küçük etki) ve WLE-EAP için 0.10 (küçük etki) olduğu görülmektedir.

*İkinci Araştırma Sorusuna Yönelik Bulgular*

*Büyük Örneklem Koşulunda Elde Edilen Sınıflama Doğruluklarının Karşılaştırılmasına İlişkin Bulgular*

Büyük örneklem koşulunda elde edilen sınıflama doğruluklarına ilişkin farklılığın istatistiksel açıdan anlamlılığını incelemek amacıyla Friedman testi yürütülmüş olup bu teste ilişkin sonuçlar Tablo 1' de sunulmuştur:

**Tablo 1** Büyük Örneklem Koşulunda Elde Edilen Sınıflama Doğruluklarının Karşılaştırılmasına İlişkin Friedman Testi Sonuçları

| Yöntemler | N | Medyan | Sıra ortalamaları | Ki kare | sd | p değeri |
|---|---|---|---|---|---|---|
| MLE | 965 | 0,76 | 2,59 | | | |
| WLE | 965 | 0,76 | 2,71 | 74.24 | 3 | 0 ,00* |
| EAP | 965 | 0,75 | 2,46 | | | |
| MAP | 965 | 0,71 | 2,23 | | | |

*p ≤ .01

Friedman testinin sonuçlarına göre büyük örneklem koşulunda farklı yetenek kestirim yöntemleri altında elde edilen sınıflama doğrulukları arasında anlamlı farklılık vardır ($\chi$2:3, 461 = 74.24 , p < .05). Medyan değerleri incelendiğinde ise MLE (Md=0,76), ve WLE (Md=0,76)' nin aynı olduğu Bayesian yöntemlerin ise daha düşük medyan değerlerine sahip olduğu görülmüştür. Hangi yöntem çiftleri arasında istatistiksel olarak anlamlı farklılık olduğunu ortaya çıkarmak için Wilcoxon' un işaretçi sıralar testi kullanılmıştır.

Wilcoxon' un işaretçi sıralar testine göre MLE-WLE yöntem çiftleri hariç bütün yöntem çiftleri arasındaki farklılık istatistiksel olarak anlamlı bulunmuştur. Anlamlı çıkan bu farklılıklara ilişkin etki büyüklükleri hesaplandığında MLE-MAP için 0.16 (küçük etki); MLE-EAP için 0.06 (küçük etki); WLE-EAP için 0.10 (küçük etki); WLE-MAP için 0.16 (küçük etki) ve MAP-EAP için 0.08 (küçük etki) olduğu görülmektedir.

*Büyük Örneklem Koşulunda Elde Edilen Sınıflama Tutarlılıklarının Karşılaştırılmasına İlişkin Bulgular*

Büyük örneklem koşulunda elde edilen sınıflama tutarlılıklarına ilişkin farklılığın istatistiksel açıdan anlamlılığını incelemek amacıyla Friedman testi yürütülmüş olup bu teste ilişkin sonuçlar Tablo 2' de sunulmuştur:

**Tablo 1** Büyük Örneklem Koşulunda Elde Edilen Sınıflama Tutarlılıklarının Karşılaştırılmasına İlişkin Friedman Testi Sonuçları

| Yöntemler | N | Medyan | Sıra ortalamaları | Ki kare | sd | p değeri |
|---|---|---|---|---|---|---|
| MLE | 965 | 0,62 | 2,61 | | | |
| WLE | 965 | 0,62 | 2,77 | 123,35 | 3 | 0 ,00* |
| EAP | 965 | 0,61 | 2,48 | | | |
| MAP | 965 | 0,58 | 2,14 | | | |

*p ≤ .01

Friedman testinin sonuçlarına göre büyük örneklem koşulunda farklı yetenek kestirim yöntemleri altında elde edilen sınıflama doğrulukları arasında anlamlı farklılık vardır ($\chi$2:3, 461 = 123.35 , p < .05). Medyan değerleri incelendiğinde ise MLE (Md=0,62), ve WLE (Md=0,62)' nin aynı olduğu Bayesian yöntemlerin ise daha düşük medyan değerlerine sahip olduğu görülmüştür. Hangi yöntem çiftleri arasında istatistiksel olarak anlamlı farklılık olduğunu ortaya çıkarmak için Wilcoxon' un işaretçi sıralar testi kullanılmıştır.

Wilcoxon' un işaretçi sıralar testine göre MLE-WLE yöntem çiftleri hariç bütün yöntem çiftleri arasındaki farklılık istatistiksel olarak anlamlı bulunmuştur. Anlamlı çıkan bu farklılıklara ilişkin etki büyüklükleri hesaplandığında MLE-MAP için 0.22 (küçük etki); MLE-EAP için 0.07 (küçük etki); WLE-EAP için 0.13 (küçük etki); WLE-MAP için 0.21 (küçük etki) ve MAP-EAP için 0.13 (küçük etki) olduğu görülmektedir.

## Sonuç ve Tartışma

Bu çalışmada büyük ve küçük olmak üzere iki farklı örneklem koşulu altında dört farklı yetenek kestirim yöntemine dayalı olarak elde edilen sınıflama doğruluğu ve sınıflama tutarlılığı indeksleri karşılaştırmalı olarak incelenmiştir. Bu doğrultuda TIMSS 2015 Matematik alt testinden elde edilen veri üzerinde Bayesian yöntemlerden MAP, EAP; Nonbayesian yöntemlerden de MLE ve WLE' ye dayalı olarak MTK altında yetenek kestirimleri elde edilmiş ve bu yetenek kestirimlerine dayalı olarak Rudner' in yaklaşımı kullanılarak sınıflama doğruluğu ve sınıflama tutarlılığı indeksleri hesaplanmıştır. Çalışmadan elde edilen bulgulara göre hem büyük hem de küçük örneklem koşulunda, Nonbayesian kestirim yöntemlerinden olan MLE yöntemi ile elde edilen yetenek kestirimlerine dayalı olarak yapılan sınıflamanın diğer bütün yöntemlerden istatistiksel olarak daha doğru ve tutarlı olduğu gözlenmiştir. Wyse ve Hao (2012) çalışmalarında Rudner-temelli indekse dayalı olarak gerçek

ve simülasyon veri üzerinde hesaplamalar yapmıştır. Çalışmadan elde ettikleri bulgulara göre simülasyon veri üzerinde yapılan incelemede EAP ve MAP kestiricilerine dayalı olarak hesaplanan indekslerin daha iyi sonuç verdiği gözlenirken gerçek veri üzerinde yapılan incelemede EAP ve MAP kestiricilerine dayalı elde edilen sınıflama doğruluğu ve tutarlılıklarının MLE' den daha düşük olduğu gözlenmiştir. Bu çalışmada kullanılan TIMSS verisinin gerçek bir veri seti olması ve bireylerin yetenek kestirimlerine ilişkin dağılımın normale yakın olmasından dolayı elde edilen sonuçların Wyse ve Hao (2012)' nun çalışmasından elde edilen sonuçlarla örtüştüğü söylenebilir. Ayrıca Wyse ve Hao (2012)' nun çalışmalarında farklı dağılım durumlarına göre söz konusu incelemeler yapılmıştır. Dağılımın normale yakın olduğu 30 maddelik test koşullarında MLE yetenek kestiricisi ile Bayesian yetenek kestiricilerine dayalı elde edilen sınıflama indeksleri arasındaki farklılığın MLE' nin lehine artış gösterdiği gözlenmiştir. Bu husus dikkate alınarak bu çalışmada ele alınan yetenek kestirimlerinin dağılımları incelendiğinde söz konusu dağılımların genel olarak normale yakın olduğu gözlenmiştir. Dolayısıyla yetenek dağılımlarının normale yakın olduğu bu çalışmada da sınıflama indekslerinden elde edilen sonuçların MLE kestirim yönteminin lehine olması beklenen bir durum olarak ele alınabilir.

Yöntemler arasında istatistiksel farklılıkların ortaya çıktığı durumlarda yapılan ikili karşılaştırma ve etki büyüklüğü incelemeleri, söz konusu istatistiksel farklılıkların pratikteki etkisinin küçük olduğunu göstermiştir. Pratikteki istatistiksel etkinin küçük olması dağılımsal özelliklerin bir sonucu olarak ele alınabilir. Farklı bir ifadeyle belirtilecek olursa yetenek kestirimlerine ilişkin dağılımların genel olarak normale yakın olmasından dolayı yöntemler arasındaki farklılaşmanın pratikteki etkisinin küçük olması olasıdır. Çünkü normal dağılımdan bariz sapmaların olduğu durumlarda farklı yöntemlere ilişkin sonuçlar arasında farklılıkların arttığı görülmektedir (Lathrop ve Cheng, 2014; Wyse ve Hao, 2012; Zhang, Du, Chen, Xin, ve Chen, 2017). Lathrop ve Cheng (2014) çalışmalarında parametrik yaklaşımlardan Lee (2010) ve Livingston ve Lewis (1995)' in yaklaşımlarını kullanılarak elde edilen sınıflama indeksleri ile nonparametrik yaklaşıma dayalı olarak sınıflama doğruluğu ve tutarlılığı indekslerini çeşitli koşullarda test etmişlerdir. Koşulların oluşturulmasında yetenek dağılımları manipüle edilerek normal, çarpık ve karma (bimodal-iki modlu dağılım) olmak üzere üç farklı dağılımsal özellik ele alınmıştır. Bu çalışma sonuçlarına dayalı olarak Lathrop ve Cheng (2014) dağılımsal özelliklerin sınıflama indekslerini etkilediğini göstermişlerdir. Normal dağılım koşulunda farklı sınıflama yöntemlerinden elde edilen sonuçların genel olarak benzer olduğu ancak çarpık dağılım koşulunda kesme puanının dağılımdaki yerine bağlı olarak değerlerin farklılaştığı görülmüştür. Lathrop ve Cheng (2014)' in çalışmasından elde edilen sonuçlar bu çalışmadan

elde edilen kestirim yöntemleri arasındaki istatistiksel farklılığın pratikteki etkisinin küçük olmasına ilişkin sonucu desteklemektedir. Çünkü her bir kestirim yöntemi ile elde edilen yetenek dağılımları normale yakındır ve bu pratikteki farklılığın küçük olmasının muhtemel bir sebebi olarak ele alınabilir.

Çalışmada küçük ve büyük olmak üzere iki farklı örneklem koşulunda da Nonbayesian yöntemlerin Bayesian yöntemlerden daha doğru ve tutarlı olduğu sonucuna ulaşılmıştır. Ancak çalışmalar (Lee & Song, 2004; Kadane, 2015) küçük örneklem koşulunda Bayesian yöntemlerin Nonbayesian yöntemlere göre daha avantajlı olabileceğini göstermektedir. Bu çalışmanın yetenek kestirimlerinden elde edilen standart hata değerleri incelendiğinde Bayesian ve Nonbayesian yöntemlere bağlı olarak standart hatanın miktarsal değişiminin bu çalışmalarla örtüştüğü gözlenmiştir. Yani küçük örneklem koşullarında elde edilen yetenek kestirimlerinin standart hata değerleri daha küçüktür. Ancak sınıflama doğruluğu ve tutarlılığının hesaplanmasını önemli ölçüde etkileyen faktörlerden biri de bireylerin yetenek kestirimlerine dayalı olarak atanacağı düzey sayısıdır (Ercikan ve Julian, 2002; Lathrop ve Cheng, 2014). Bireylerin atanacağı sınıf sayısının artış ve azalışına koşullu olarak sınıflama indeks değerleri birbirinden farklılaşmaktadır. Kesme puan noktalarında biriken birey sayılarının artışı bireyler hakkında yanlış karar verme olasılıklarını arttırdığından düzey sayısının yüksek olduğu durumlarda sınıflama doğruluğu ve tutarlılıkları azalmaktadır. Farklı bir ifadeyle sınıflama kararları düzey sayısının artışından doğrudan etkilenmektedir çünkü düzey sayısı arttıkça ele alınacak kesme puanlarının da sayısı artmakta ve bu nedenle daha çok birey kesme puan noktalarında yığılmaktadır. Bu çalışmada TIMSS' in belirttiği 5 düzey ele alınarak incelemeler yapıldığından çalışmanın sonuçlarının Nonbayesian yöntemlerin lehine olmasının altında yatan bir faktörün de düzey sayısı olabileceği düşünülmektedir. Çünkü yetenek kestirimleri elde edilirken yöntemlerin matematiksel alt yapısının doğası gereği bayesian yöntemlerden MAP dağılımın ortancasına, EAP ise dağılımın ortalamasına çekilme eğilimi göstermektedir. Bu durum Bayesian yöntemlerden elde edilen yetenek kestirimlerinin Nonbayesian yöntemlere göre daha dar bir ranjda yer almasıyla sonuçlanmaktadır. Dolayısıyla hem ranjın dar olması hem düzey sayısının ve buna koşullu olarak da kesme puan sayısının yüksek olması sebebiyle Bayesian yöntemlerin kullanıldığı durumlarda daha fazla birey kesme puanına yakın yerlerde birikmiştir. Bu durum ise sınıflama sonuçları üzerinde baskın bir faktör olarak hem küçük hem de büyük örneklem koşullarında sonuçların Nonbayesian yöntemlerin lehine sonuçlanmasının bir nedeni olarak ele alınabilir.

## Öneriler

Bu çalışma tek bir yetenek dağılımına dayalı olarak gerçekleştirilmiştir. Ancak yetenek dağılımının normale yakın olması da sonuçları etkileyen önemli bir faktör olarak ele alınabilir. Dolayısıyla farklı dağılım koşullarının dikkate alındığı çalışmaların yapılması önerilmektedir.

Matematiksel alt yapılarından kaynaklı olarak Bayesian ve Nonbayesian yöntemlere dayalı elde edilen yetenek kestirimleri farklı dağılımsal özelliklerle sonuçlanabilmektedir. Dolayısıyla, yetenek dağılımının ranjının da bir koşul olarak ele alındığı çalışmaların yapılması, farklı sayıdaki yeterlik düzeylerinin ele alınması, ve dağılımsal özelliklere koşullu olarak düzey sayısının Bayesian ve Nonbayesian yöntemlerden elde edilen yeteneklere göre sınıflama sonuçlarını nasıl etkilediğinin incelenmesi önerilmektedir.

Bu çalışma maddelerin iki kategorili olarak ele alındığı başarı testi üzerinde yapılmıştır. Dolayısıyla, çok kategorili ve karma testlerin de ele alındığı çalışmaların yapılması önerilmektedir. Ayrıca düzey belirleme çalışmalarının önem kazandığı psikolojik testler üzerinde de söz konusu teknik incelemelerin yapılması önerilmektedir.

## Kaynakça

Altun, M. (2010). Matematik Öğretimi. Bursa: Pegem Akademi.

Büyüköztürk , S. Çakan, M., Tan, S., & Atar, H. Y. (2014). TIMSS 2011 ulusal matematik ve fen raporu 8. sınıflarRetrieved fromhttp://timss.meb.gov.tr/wp-content/uploads/TIMSS-2011-8-Sinif.pdf

Barnett, D. W., & Macmann, G. M. (1992). Decision reliability and validity: contribution and limitations of alternative assessment systems. The Journal of Special Education. 25(4), 431-452.

Bourque, M. L., Goodman, D., Hambleton, R. K., & Han, N. (2004). Reliability estimates for the ABTE tests in elementary education, professional teaching knowledge, secondary mathematics and English/language arts (Final Report). Leesburg, VA: Mid-Atlantic Psychometric Services.

Cizek, G.J. ve Bunch, M. B. (2007). Standard Setting: A Guide to Establishing and Evaluating Performance Standards on Tests. London: Sage.

Cohen J (1960). A Coefficient of Agreement for Nominal Scales. Educational and Psychological Measurement, 20(1), 37–46.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Hillsdale, NJ: Lawrence Earlbaum Associates.

Ercikan, K., & Julian, M. (2002). Classification accuracy of assigning student performance to proficiency levels: Guidelines for assessment design. Applied Measurement in Education, 15, 269-294.

Fraenkal, J. R., Wallen, N. E., & Hyun, H. H. (2008). How to design and evaluate research in education (7th ed.). New York: M Graw Hill.

Guo, F. (2006). Expected classification accuracy using the latent distribution. Practical assessment. Research & Evaluation, 11(6), 1-9.

Hambleton, R. K, & Novick, M. (1973). Toward an integration of theory and method for criterion-referenced tests. Journal of Educational Measurement, 10(3), 159-170.

Huynh, H. (1976). On the reliability of decisions in domain-referenced testing. Journal of Educational Measurement, 13, 253-264.

Kadane, J. B. (2015). Bayesian methods for prevention research. Prevention Science, 16, 1017–1025. 10.1007/s11121-014-0531-x

Lathrop, Q. N., & Cheng, Y. (2014). A nonparametric approach to estimate classification accuracy and consistency. Journal of Educational Measurement, 51, 318-334.

Lee, W., Hanson, B. A., & Brennan, R. L. (2000). Procedures for computing classification consistency and accuracy ındices with multiple categories. ACT : Inc, Research Report.

Lee, S. Y., & Song, X. Y. (2004). Evaluation of the Bayesian and maximum likelihood approaches in analyzing structural equation models with small sample sizes. Multivariate Behavioral Research, 39, 653–686. DOI: 10.1207/s15327906mbr3904_4

Lee, W. (2010). Classification consistency and accuracy for complex assessments using item response theory. Journal of Educational Measurement, 47, 1-17.

Livingston SA, Lewis C (1995). "Estimating the Consistency and Accuracy of Classifications Based on Test Scores." Journal of Educational Measurement, 32(2), 179–197.

Mislevy, R. (1986). Bayes modal estimation in item response models. Psychometrika, 51, 177–195 Seong, T. J., Kim, S. H. & Cohen, A. S. (1997, March). A comparison of procedures for ability estimation under the graded response model. Paper presented at the Annual Meeting of the American Educational Research Association. Chicago

Pallant, J. (2007). SPSS Survival Manual: A Step by Step Guide to Data Analysis using SPSS for Windows. New York: Open University Press.

Rudner, L. M. (2001). Computing the expected proportions of misclassified examinees. Practical Assessment Research & Evaluation, 7(14). Available online: http://pareonline.net/getvn.asp?v=7&n=14

Rudner, L. M. (2005). Expected classification accuracy. Practical Assessment Research & Evaluation, 10(13). Available online: http://pareonline.net/getvn.asp?v=10&n=13.

Sireci, S. G., Robin, F., & Patelis, T. (1999). Using cluster analysis to facilitate standard setting. Applied Measurement in Education, 12, 301-325.

Subkoviak, M. J. (1976). Estimating reliability from a single administration of a criterion-referenced test. Journal of Educational Measurement, 13, 265-276.

Swaminathan H, Hambleton RK, Algina J (1974). "Reliability of Criterion-Referenced Tests: A Decision-Theoretic Formulation." Journal of Educational Measurement, 11(4), 263–267.

Yang, X., Poggio, J. C., &Glasnapp, D. R. (2006). Effects of estimation bias on multiple-category classificationwith an IRT-baesd adaptive classification procedure. Educational and Psychological Measurement,31, 275-291.

Wyse, A. E., & Hao, S. (2012). An evaluation of item response theory classification accuracy and consistency indices. Applied Psychological Measurement, 36, 602-624.

Zhang, S., Du, J., Chen, P., Xin, T., & Chen, F. (2017). Using Procedure Based on Item Response Theory to Evaluate Classification Consistency Indices in the Practice of Large-Scale Assessment. Frontiers in Psychology, 8, 1676. http://doi.org/10.3389/fpsyg.2017.01676