Elementary Education Online, 2018; 17(2): pp. 500-509
İlköğretim Online, 2018; 17(2): s.500-509 . [Online]: http://ilkogretim-online.org.tr
doi 10.17051/ilkonline.2018.418890

# Yazılı Yoklamalarda Puanlama Yönteminin Test ve Madde İstatistiklerine Etkisi

# The Impact of Scoring Method in Written Examinations on Test and Item Statistics

**Tülin ACAR,** Parantez Eğitim Araştırma Danışmanlık ve Yayıncılık Hizmetleri, *totbicer@gmail.com*

**Öz.** Bu araştırmanın amacı, ayrıntılı ve genel puanlama olmak üzere farklı iki yöntemle puanlanan yazılı yoklamalardan elde edilen test ve madde istatistiklerini karşılaştırmaktır. Araştırmada, yazılı yoklama soruları ayrıntılı ve genel puanlama yapılarak puanlanmıştır. Farklı iki yönteme göre elde edilen sınav puanlarının güvenirliği arasında anlamlı bir fark bulunamamıştır. Ancak, farklı iki yönteme göre elde edilen sınav puanlarının ortalaması arasında anlamlı bir fark bulunmuştur. Ayrıntılı puanlama yapılması durumunda genel puanlamaya göre test ve madde istatistiklerinin daha yüksek olduğu gözlenmiştir. Ayrıntılı ve genel puanlama yöntemine göre puanlanan yazılı yoklama sorularının madde güçlükleri karşılaştırıldığında 2 ve 5. sorularda anlamlı bir fark bulunmuştur. Ancak 1, 3, ve 4. soruların güçlük değerlerinde anlamlı bir fark bulunamamıştır. Farklı iki puanlama yöntemine göre soruların madde ayırıcılıkları arasında da anlamlı bir fark bulunamamıştır. Yazılı yoklamalarda puanlama yönteminin niteliği ile birlikte test ve madde istatistiklerinin yorumlanması gerektiği sonucu elde edilmiştir.

**Anahtar Sözcükler:** Yazılı yoklamalar, puanlama yöntemleri, test istatistikleri, madde istatistikleri

**Abstract.** The purpose of this study is to compare the test and item statistics obtained from written examinations scored by two different methods – detailed and general scoring. The questions of written examinations are scored by detailed and general scoring in the study. No significant difference was found between the reliability of exam grades given by the two methods. However, a significant difference was found between the average grades given by the two different methods. It was observed that detailed scoring yielded higher test and item statistics than general scoring. A comparison of the difficulties of items in written exam questions scored by detailed and general scoring yielded significant differences in questions 2 and 5. However, difficulty values of the questions 1, 3, and 4 did not yield any significant difference. No significant difference was found between the item discrimination of the questions scored by the two scoring methods, either. The conclusion that test and item statistics should be interpreted in addition to the scoring method of written examinations was made.

**Keywords:** Written examinations, Scoring methods, Test Statistics, Item Statistics

Geliş tarihi: 15.05.2017 Kabul tarihi: 19.10.2017 Yayımlanma tarihi: 26.04.2018

## SUMMARY

### Introduction

The purpose of this study is to show empirically how scoring a written exam by detailed and general scoring methods affects the reliability, arithmetic average, item difficulty and item discrimination of exam grades. Therefore, it is important for providing teachers and researchers with empirical data for choosing a scoring method in written exams. In this study, written exam papers of a course that requires mathematical operations were scored by two different methods: general scoring (GS) and detailed scoring (DS). The impact of scoring by these two methods on the test and item properties was examined.

General Scoring (GS) stands for the scoring for complete and correct answers. It involves not giving a full or partial point even if an answer is partially correct. Detailed Scoring (DS), on the other hand, is defined as giving a partial point for the correct operations for the solution even if an incorrect result is found.

### Method

Five open-ended questions were prepared and implemented by a professor for the midterm exam of 79 students who take an undergraduate course of Statistics. The data collection instrument of this study was the papers of the written examination used for implementation of the study. Midterm exam papers were re-scored by the researcher according to the "scoring key" that had been prepared beforehand for DS. On the other hand, GS was done by the lecturer of the course.

### Results

The internal consistency reliability of the exam was found 0.49 for DS, and 0.34 for GS. Whether there is a statistical difference between these two reliability factors was tested by Fisher's Z-test after turning the reliability factors into Z-statistics. No significant difference was found between the reliability factors of written examinations scored by DS and GS ($p>0.05$).

A significant difference with a significance level of 0.05 ($p<0.05$) was found between the arithmetic average of exam grades given by DS and the arithmetic average of exam grades given GS. This difference is in favor of DS.

Whether there is a statistical difference between the difficulty values of the exam questions calculated by the two scoring methods was tested by testing the difference between the two ratios. A significant difference was found between the item difficulties of the questions 2 and 5 scored by DS and GS ($p<0.05$). However, even if a detailed and general scoring was made, no significant difference was found between the item difficulty values of the questions 1, 3, and 4 ($p>0.05$).

Item discrimination values of the questions scored by the two different methods were first converted into z-statistics and significance of the difference was tested by Fisher's Z-statistic. It was observed that item discrimination of exam questions did not differentiate by the two different scoring methods ($p>0.05$).

### Discussion and Conclusion

It is known that adopting a detailed scoring method minimizes errors in measurement results and increases the precision of measurement results. Therefore test and article statistics of written examinations depend are affected by whether the detailed scoring method is used or not. How the scoring is done should be taken into consideration in interpreting such statistics. If the steps of calculation or reasoning are inquired rather than the result of the operation in math-oriented courses such as Statistics, detailed scoring should be considered.

Selection of detailed or general scoring method affects the reliability of grades. Partial or detailed scoring for written examinations that are used frequently in testing the knowledge and skills of students affects reliability and internal consistency of exam grades positively. It should be considered important to increase the knowledge of teachers as exam graders about partial scoring. In other words, despite the fact that detailed or general scoring in statistics and other math-oriented courses affects the level of reliability, no statistical difference was observed between the reliability factors yielded from the two different methods of scoring.

The fact that the average grades of written examinations scored by the two scoring methods are different draws attention to the scoring methods of written examinations that are used frequently in teaching. Especially if cognitive behaviors of students are to be decided based on exam grades, partial or detailed scoring should be preferred.

It was found in the study that whether the scoring method is detailed or not causes differentiation between item difficulties of written examinations. Therefore, difficulty values of the questions of written examinations should be interpreted on the basis of the scoring method.

Item discrimination of questions scored by DS was found to be higher than the item discrimination of questions scored by GS. Again, this shows that the method of scoring is an important factor in evaluating the item discrimination of questions.

## GİRİŞ

Eğitim sisteminin geliştirilmesi, var olan eksikliklerin saptanması ve yeterli bir geri bildirimin verilebilmesi için eğitimde ölçme ve değerlendirme yöntem ve tekniklerinin doğru kullanımı oldukça önemlidir. Bugün eğitimde kullanılan sınav türleri, yazılı yoklamalar, doğru-yanlış testleri, kısa cevaplı testler, sözlü sınavlar, çoktan seçmeli testler başta olmak üzere çeşitlilik göstermektedir. Eğitimde kullanılan sınavların ölçülen hedef-davranışa uygunluğuna, testin/sınavın kullanışlılığına, puanlama nesnelliğine, puanların güvenirliğine gibi pek çok niteliğe göre avantajları ve dezavantajları söz konusudur (Atılgan ve diğerleri, 2006; Roid&Haladyna,1982, Turgut, 1992). Özellikle öğrencilerin üst düzey bilgi ve becerilerinin ölçülmesi söz konusu olduğunda "yazılı yoklamalar" tercih sebebidir.

Yazılı yoklamalarda daha çok öğrencinin bir mukayese yapması, bir durumu tanımlaması, değerlendirmesi veya açıklaması istenmektedir (Wiliam & Karnes, 1968:246). Bu tip yazılı yoklamalarda, öğrenci ilgili konu ve hedef-davranış düzeyinde kıyaslama yapmada, değerlendirmede veya yargılamada ifade serbestliğine sahiptir. Yazılı yoklamalar, dikkatli olarak hazırlanır ve puanlanırsa elde edilen puanlar, öğrencilerin başarıları hakkında fazlaca bilgi verici olmaktadır. Yazılı yoklamaların puanlama süreci, eğitimde bilgi ve becerileri ölçmede kullanılan diğer sınav türlerine göre daha meşakkatlidir. Puanlamayı yapan öğretmenin ya da puanlayıcının ön yargıları, öğrenci hakkındaki düşünceleri puanlama sürecine bir tür hata olarak karışabildiği için yazılı yoklamaların puanlamasında yansızlık, nesnellik temel bir hassasiyet oluşturmaktadır (Hopkins, 1998). Yazılı yoklamaların değerlendirilmesinde birbirlerinden bağımsız farklı puanlayıcıların aynı puanlama anahtarını kullanmalarına rağmen yapmış oldukları puanlamalar, farklılıklar gösterebilmektedir (Iramaneerat & Yudkowsky). Puanlamalardaki farklılıklar da sınavın test ve madde istatistiklerini önemli ölçüde etkilemektedir (Kan, 2005). Dolayısıyla sınav türlerinin nitelikleri ve madde-test özelliklerine ilişkin sonuçları iyi bilinip işe koşulması gerekmektedir.

Bilişsel alan davranışlarını ölçmede kullanılan yazılı yoklamaların puanlama sürecindeki sıkıntılar, diğer test türlerine göre daha fazladır. Puanlama sürecindeki sıkıntıları azaltmak için alan yazında farklı puanlama tekniklerinin kullanıldığı gözlenmektedir. Yazılı yoklamaların puanlanmasında kullanılan teknikler, genel izlenimle puanlama, sınıflama yoluyla puanlama, sıralama yoluyla puanlama veya puanlama anahtarına göre çeşitlenmektedir (Atılgan ve diğerleri, 2006:214). Bir yazılı yoklamanın puanlanmasında birden fazla teknikle puanlama yapmak mümkündür. Ancak puanlayıcı açısından zaman ve bireysel dikkat faktörü yazılı yoklamaların puanlama sürecinde birden fazla tekniğin aynı anda kullanılmasını azaltmaktadır.

Yazılı yoklamaların puanlamasında özellikle kısmi puanlamanın yapılabiliyor olması puanlama açısından öğrenciye bir avantaj sağlamaktadır. Ancak puanlayıcı tutumları, kısmi puanlamanın miktarında farklılıklar oluşturmaktadır. Alan yazında puanlayıcıların puanlamada cimri ya da esnek tutumlarının olduğu bilinmektedir. Yine puanlayıcının puanlama tutumu, ölçülen hedef-davranışın sayısal ya da sözel alanı temsil eden bir bilgi ve beceri olup olmamasına göre farklılaşmakta olduğu elbette göz ardı edilmemelidir.

Alanyazında yazılı yoklamaların puanlama yöntemlerinin karşılaştırıldığı araştırmalardan ziyade daha çok çoktan seçmeli test maddelerinin farklı puanlanmasının test ve madde istatistiklerine etkisinin incelendiği çalışmalara (Akkuş ve Baykul, 2001; Özdemir, 2003; Yurdugül, 2010) rastlamak mümkündür. Yazılı yoklamaların şans başarısından arınık olması avantajı ile birlikte puanlama yönteminin seçimi ve puanlayıcıların tutumu ve tutarlılığı daha çok ön plana çıkmaktadır. Alan yazında açık uçlu sorularda puanlayıcıların daha çok genellenebilirlik kuramına dayalı olarak tutarlılığının karşılaştırıldığı veya farklı kuramlara göre yapılan puan kestirimlerinin karşılaştırıldığı çalışmalar (Güler ve Gelbal, 2010; Güler ve Taşdelen Teker,2015; İlhan, 2016) gözlenmektedir.

### Araştırmanın Amacı

Bu araştırmanın amacı, yazılı bir yoklamanın ayrıntılı ve genel puanlama olmak üzere farklı iki yöntemle puanlanmasının, sınav puanlarının güvenirliğini, aritmetik ortalamasını, madde güçlüklerini ve madde ayırıcılıklarını nasıl değiştirdiğini ampirik bir yoldan göstermektir.

Dolayısıyla, yazılı yoklamalarda puanlama yönteminin seçimi konusunda öğretmenlere, araştırmacılara deneysel bir bilgi sağlaması bakımından da önemlidir. Bu bağlamda araştırmada şu sorulara cevaplar aranmıştır:

1- Ayrıntılı ve genel puanlama yöntemine göre puanlanan yazılı yoklamanın sınav puanlarının güvenilirlikleri farklılaşmakta mıdır?
2- Ayrıntılı ve genel puanlama yöntemine göre puanlanan yazılı yoklamanın sınav puanlarının ortalamaları arasında bir fark var mıdır?
3- Ayrıntılı ve genel puanlama yöntemine göre puanlanan sınav sorularının madde güçlükleri arasında bir fark var mıdır?
4- Ayrıntılı ve genel puanlama yöntemine göre puanlanan sınav sorularının madde ayırıcılıkları arasında bir fark var mıdır?

### Tanımlar ve Kısaltmalar

Genel Puanlama (GP): Cevap anahtarına göre tam, eksiksiz, doğru yanıtlara verilen puanlamadır. Bir başka deyişle, verilen cevaplar kısmi olarak doğru olsa bile *tam* ya da *kısmi puan* verilmeden yapılmış puanlamadır.
Ayrıntılı Puanlama (AP): İşlem sonucu yanlış bulunmuş olsa dahi çözüm yolunda yapılan doğru işlemlerin her biri için yapılan kısmi puanlamadır.

### Araştırmanın Sınırlılığı

Araştırma, yazılı yoklama tipindeki sorularla; ayrıntılı ve genel puanlama yöntemine göre puanlamayla; yazılı yoklamaya esas olan ara sınavla; lisans düzeyinde sayısal alan temelli bir ders olan İstatistik dersi ile sınırlıdır.

## YÖNTEM

### Araştırmanın Türü

Bu araştırmada matematiksel işlem gerektiren sayısal içerikli bir dersin yazılı yoklama kâğıtları, genel puanlama (GP) ve ayrıntılı puanlama (AP) olmak üzere iki farklı yöntemle puanlanmıştır. Farklı iki yöntemle yapılan puanlamanın, test ve madde özelliklerine etkisi incelenmiştir. Dolayısıyla bu araştırma, kuram geliştirmeye veya kuramı test etmeye yönelik bir araştırma olduğundan araştırmanın türü temel araştırma niteliğindedir (Kaptan, 1998).

### Çalışma Grubu

Eğitim Bilimleri Bölümünde lisans düzeyinde İstatistik dersi alan 79 üniversite öğrencisi, araştırmanın çalışma grubunu oluşturmaktadır.

### Veri Toplama Aracı

Lisans düzeyindeki öğrencilerin İstatistik dersi ara sınavları için açık uçlu 5 soru, öğretim elemanı tarafından hazırlanmış ve uygulamada kullanılan yazılı yoklama sınav kâğıtları, bu çalışmada veri toplama aracı olarak kullanılmıştır.

### Verilerin Çözümlenmesi

Ara sınav kâğıtları, AP için önceden hazırlanmış olan "puanlama anahtarına" göre araştırmacı tarafından puanlanmıştır. GP ise dersin öğretim elemanı tarafından yapılmıştır. Farklı yöntemleri kullanan iki puanlayıcının yapmış olduğu puanlamalar arasındaki tutarlılık Spearman'ın sıra farkları korelasyon katsayısı ile hesaplanmıştır. 0.87 olarak hesaplanan korelasyon katsayısı, pozitif yönlü, yüksek düzeyde bir ilişkiyi göstermek üzere anlamlı bulunmuştur($p<0.05$). Araştırma sorularının çözümü için kullanılan istatistikler hakkında bilgiler aşağıda belirtilmiştir:

*Madde güçlük indeksi ($p_j$):* Maddenin doğru cevaplanma yüzdesidir. Ancak açık uçlu sorularda soruyu cevaplayanların elde ettikleri puanların aritmetik ortalamasının, o madde için belirlenen en yüksek puana oranıdır.

*Madde ayırıcılık gücü indeksi($r_{jx}$):* Bir sorunun ayırıcılığı, yoklanan davranışa sahip olan öğrencileri o davranışa sahip olmayanlardan ayırabilme gücüdür ve korelasyon istatistikleri ile hesaplanmaktadır (Özçelik,1997:123). Bu araştırmada madde ve test puanlarının sürekli değişken olması sebebiyle madde ayırıcılık indeksleri pearson momentler çarpımı korelasyon katsayısı ile hesaplanmıştır(Baykul,1997:215). $r_{jx}$ bir korelasyon katsayısı olduğundan (-1.00) ile (+1.00) arasında değerler alır. Pozitif değerler maddenin testin bütünüyle aynı yönlü ilişki içinde olduğunu; 0.00 (sıfır) değeri maddenin ve testin ölçülen değişken arasında herhangi bir ilişki bulunmadığını; negatif değerler ise testin ve maddenin ölçtükleri değişkenler arasında ters yönlü bir ilişki bulunduğunu gösterir (Büyüköztürk, 2002; Özdamar, 2002: 549). $r_{jx}$ korelâsyonunun 1.00'e yakın değerler alması istendiktir. Çünkü 1.00'e yakın bir değer alan bir madde, testin tümü ile ölçülen özelliğe sahip olan öğrenci ile olmayanı iyi ayırıyor olduğunun bir göstergesidir.

*Test puanlarının güvenirliği:* Güvenirlilik, tesadüfî hatalardan arınıklık derecesi olarak tanımlanır (Turgut,1992). Bu tanımın yanı sıra güvenirlilik aynı zamanda test maddelerinin testin tümüyle olan tutarlılığıdır. Güvenirlilik 0.00 (sıfır) ile +1.00 arasında değişmektedir. Testin güvenirliliği 1.00'e yakın değerler alması istenen bir durumdur. Bu araştırmada test tipi yazılı yoklama olduğundan ve madde ile test puanlarının sürekli değişkenler olması gerekçesi ile sınav puanlarının güvenirliği Cronbach $\alpha$ katsayısı ile hesaplanmıştır (Nunnally,1970:126).

*Fisher'ın Z testi:* İki farklı puanlama yöntemi ile hesaplanan madde ayırıcılık gücü indeksleri birer korelasyon katsayısıdır. Bu nedenle korelasyon katsayıları ve güvenirlik katsayıları z istatistiğine dönüştürülmüştür. Z istatistiğine dönüştürülen iki korelasyon katsayısı arasındaki fark (0'dan farklı olup olmadığı ) Fisher Z testi ile aşağıdaki formül ile test edilmiştir (Akhun, 1982:33).

$$Z = \frac{Z_{r1} - Z_{r2}}{\sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}}$$

*Madde güçlüklerinin istatistiksel testi:* Farklı iki yöntemle puanlanan sınav sorularının güçlüklerinin anlamlılık testi için iki oran farkının testi kullanılmıştır. Madde güçlükleri, hesaplama yöntemi gereği bir oran bilgisi taşımaktadır. İki oran farkının testi aşağıda belirtilen formülle hesaplanmaktadır (Baykul,1997: 356).

$$P = \frac{(p_1 - p_2)}{\sqrt{\frac{p_{1*}q_1}{n_1} + \frac{p_{2*}q_2}{n_2}}}$$

*Bağımlı örneklemler için t testi:* Farklı iki yöntemle puanlanan sınav puanlarının ortalamaları arasında istatistiksel olarak bir farkın olup olmadığı bağımlı örneklemler için t testi ile test edilmiştir (Büyüköztürk, 2002). 79 öğrencinin sınav kâğıdı ilkin dersin öğretim elemanı tarafından genel puanlama ile puanlanmıştır. Ardından aynı sınav kâğıtları araştırmacı tarafından ayrıntılı olarak puanlanmıştır. Aynı sınav kâğıtlarının farklı yöntemlerle iki kez puanlanmış olması tekrarlı bir durumu ortaya çıkarmaktadır. Dolayısıyla bağımlı örneklemler için t testi kullanılmıştır.

Araştırmada anlamlılık düzeyi 0.05 olarak alınmıştır.

## BULGULAR

### Ayrıntılı ve genel puanlama yöntemine göre puanlanan yazılı yoklamanın sınav puanlarının güvenilirlikleri farklılaşmakta mıdır?

Sınav kâğıtları, AP ve GP yapıldıktan sonra elde edilen sınav puanlarının güvenirlik katsayıları Cronbach $\alpha$ istatistiği ile hesaplanmıştır. Bu iki güvenirlik katsayısı arasında istatistiksel bir farkın olup olmadığı güvenirlik katsayıları Z istatistiğine dönüştürülüp Fisher'ın Z testi ile test edilmiştir ve bulgular Tablo 1 de sunulmuştur.

**Tablo 1.** *AP ve GP Yapılan Sınav Puanlarının Güvenirliklerinin Karşılaştırılması*

| Puanlama Yöntemi | Öğrenci Sayısı | Madde Sayısı | $r_x$ | Zr. | Fisher'ın Z değeri | p |
|---|---|---|---|---|---|---|
| AP | 79 | 5 | 0,49 | 0,536 | 0,182 | 0,857 |
| GP | 79 | 5 | 0,34 | 0,354 | | |

Buna göre AP ile puanlanan sınavın iç tutarlılık anlamında güvenirliği 0.49; GP ile puanlandığında güvenirliği 0.34 olarak bulunmuştur. Her iki puanlama için hesaplanan iç tutarlılık anlamındaki güvenirlik katsayıları düşük bulunmuştur. Araştırmada güvenirlik katsayılarının farklı puanlama yöntemlerine göre nasıl değiştiği incelenmiş olduğundan güvenirliğin neden düşük çıktığı sorusu araştırma kapsamı dışında tutulmuştur. Ancak soru sayısı ile güvenirlik ilişkisi hatırlandığında bir neden olarak soru sayısının az olması, güvenirliğin düşük bulunmasının bir sebebi olarak değerlendirilebilir. AP ile hesaplanan güvenirlik katsayısının, GP ile hesaplanan güvenirlik katsayısından daha yüksek olduğu gözlenmiştir. Yazılı yoklamaların ayrıntılı puanlanması ölçme sonuçlarının duyarlılığını arttırmaktadır (Roid&Haladyna,1982: 63). Dolayısıyla bu bulgu, puanlama duyarlılığının artması, ölçme sonuçlarının güvenirliğini arttır çıkarımını desteklemiştir. Ancak, AP ve GP ile puanlanan yazılı yoklamanın güvenirlik katsayıları arasında anlamlı bir fark bulunamamıştır ($p>0.05$).

**Ayrıntılı ve genel puanlama yöntemine göre puanlanan yazılı yoklamanın sınav puanlarının ortalamaları arasında bir fark var mıdır?**

AP ve GP ile puanlanan elde edilen puanların ortalamaları arasında anlamlı bir farkın olup olmadığı bağımlı örneklemler için t testi istatistiği ile hesaplanmış ve bulgular Tablo 2'de sunulmuştur.

**Tablo 2.** *AP ve GP ile puanlanan Sınav Puanlarının Karşılaştırılması*

| Puanlama Yöntemi | Öğrenci Sayısı | Ortalama | S. sapma | Sd | t | p |
|---|---|---|---|---|---|---|
| AP | 79 | 25,70 | 6.73 | 78 | 12,73 | 0.000* |
| GP | 79 | 20,92 | 5.58 | | | |

*P<0,05

AP yapılan yazılı yoklamanın sınav puanlarının aritmetik ortalaması ile GP yapılan yazılı yoklama puanlarının aritmetik ortalaması arasında 0.05 manidarlık düzeyinde anlamlı bir fark bulunmuştur($p<0.05$). Bulunan bu fark AP lehinedir. Bir başka deyişle, eğitim sisteminde yazılı yoklamaların puanlama yönteminin seçimi öğrenciler için özellikle geçme kalma gibi kararlarının verilmesinde çok önemli bir rolü olduğunu göstermektedir.

**Ayrıntılı ve genel puanlama yöntemine göre puanlanan sınav sorularının madde güçlükleri arasında bir fark var mıdır?**

GP ve AP yapılan her bir sorunun madde güçlükleri hesaplanmıştır. İki farklı puanlama yöntemine göre hesaplanan sınav sorularının güçlük değerleri arasında istatistiksel olarak bir farkın olup olmadığı iki oran farkının testi ile sınanmıştır. Sonuçlar Tablo 3'de özetlenmiştir.

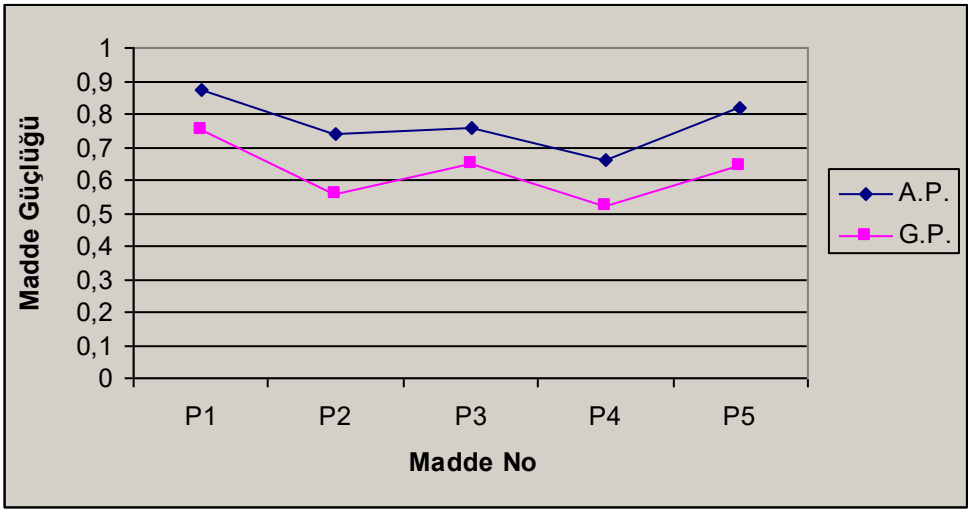
**Tablo 3.** *AP ve GP Yapılan Maddelerin Güçlüklerine Ait İstatistikler*

| Puanlama yöntemi | $P_1$ | $P_2$ | $P_3$ | $P_4$ | $P_5$ | $P_{ort.}$ |
|---|---|---|---|---|---|---|

| | | | | | | |
|---|---|---|---|---|---|---|
| AP | 0,87 | 0,74 | 0,76 | 0,66 | 0,82 | 0,77 |
| GP | 0,75 | 0,56 | 0,65 | 0,52 | 0,64 | 0,63 |
| Z değeri | 1,92 | 2,37* | 1,52 | 1,81 | 2,54* | 1,92 |

*P<0.05

2. ve 5. sorularının AP ile GP ile hesaplanması durumunda elde edilen madde güçlükleri arasında anlamlı bir fark bulunmuştur (p<0.05). Ancak ayrıntılı ve genel puanlama yapılmış olmasına rağmen 1., 3., ve 4. soruların güçlük değerleri arasında anlamlı bir fark bulunamamıştır(p>0.05). AP ve GP ile puanlanmış soruların madde güçlük değerlerinin grafikleri şekil 2'de gösterilmiştir.




**Şekil 1.** *AP ve GP yapılan soruların madde güçlükleri*

AP yapılan soruların madde güçlük indekslerinin 0,66 ile 0,87 arasında değişmekte olduğu; GP yapılan soruların madde güçlüklerinin ise 0,52 ile 0,75 arasında değiştiği görülmektedir. AP yapılan yazılı yoklama sorularının madde güçlüklerinin, GP yapılan sınav sorularının madde güçlüklerinden daha yüksek değerlere sahip olduğu gözlenmiştir.

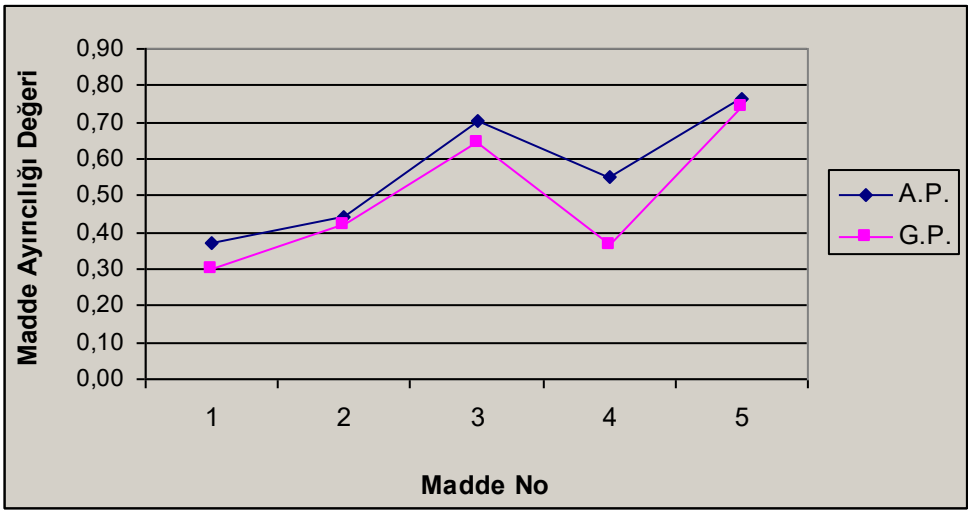
**Ayrıntılı ve genel puanlama yöntemine göre puanlanan sınav sorularının madde ayırıcılıkları arasında bir fark var mıdır?**

GP ve AP yapılan yazılı yoklama sorularının madde ayırıcılıkları hesaplanmış ve sonuçlar Tablo 4'de özetlenmiştir.

**Tablo 4.** *AP ve GP Yapılan Maddelerin Ayırt Etme Gücü İndekslerine Ait Fisher'ın Z İstatistikler*

| Puanlama Türü | $r_{1x}$ | $r_{2x}$ | $r_{3x}$ | $r_{4x}$ | $r_{5x}$ | $r_{jx(ort)}$ |
|---|---|---|---|---|---|---|
| AP | 0,37 | 0,44 | 0,71 | 0,55 | 0,76 | 0,55 |
| GP | 0,30 | 0,42 | 0,65 | 0,37 | 0,74 | 0,51 |
| Fisher'ın Z değeri | 0,48 | 0,15 | 0,69 | 1,41 | 0,28 | 0,34 |

İki farklı yöntem ile puanlanan soruların madde ayırıcılık değerleri ilkin z istatistiğine dönüştürmüş ve farkın manidarlığı Fisher'ın Z testi ile sınanmıştır. Sınav sorularının madde ayırıcılıkları, iki farklı puanlama yöntemine göre farklılaşmadığı gözlenmiştir (p>0.05). AP ve GP ile puanlanmış soruların madde ayırıcılıklarının grafikleri Şekil 2'de gösterilmiştir.




**Şekil 2.** *AP ve GP yapılan soruların madde ayırıcılıkları*

AP yapılan sınav sorularının madde ayırıcılıkları 0,37 ile 0,76 arasında değişmektedir. GP yapılan sınav sorularının madde ayırıcılıklarını 0,30 ile 0,74 arasında değişmekte olduğu gözlenmiştir. AP yapılan sınav sorularının madde ayırıcılıkları, GP yapılan sınav sorularının madde ayırıcılıklarından daha yüksek değerlere sahip olduğu bulunmuştur.

## TARTIŞMA ve SONUÇ

Yazılı yoklamaların puanlanmasında ayrıntılı bir puanlama yönteminin seçilmesi, ölçme sonuçlarına karışan hatayı miniminize edip ölçme sonuçlarının duyarlılığını arttırdığı bilinmektedir. Dolayısıyla ayrıntılı puanlama yönteminin, yazılı yoklamalarda kullanılıp kullanılmaması durumu yazılı yoklamaların test ve madde istatistiklerini etkilemektedir. Bu istatistiklerin yorumlanmasında puanlamanın nasıl yapıldığının dikkate alınması gerekmektedir. Özellikle İstatistik gibi sayısal içerikli derslerin yazılı yoklamalarında işlemin sonucunun bulunmasından ziyade hesaplama ya da düşünme basamaklarının niteliği yoklanıyorsa puanlama yönteminin ayrıntılı yapılması önemsenmelidir.

Yazılı yoklamalarda ayrıntılı ya da genel puanlama yönteminin seçimi, sınav puanlarının güvenirliğine etki etmektedir. Öğrencilerin bilgi ve becerilerinin yoklanmasında sıkça kullanılan yazılı yoklamaların puanlanmasında yapılacak olan kısmi veya ayrıntılı bir puanlama, sınav puanlarının güvenirliğini, iç tutarlılığını olumlu etkilemektedir. Sınav puanlayıcısı olarak öğretmenlerin kısmi puanlamaya ilişkin bilgilerinin artırılması önemsenmelidir. Bir başka deyişle, sayısal içerikli bir ders olan istatistik ve benzeri derslerde ayrıntılı puanlama ya da genel puanlama yapılması elde edilen güvenirlik miktarını değiştirmesine rağmen farklı iki puanlamadan elde edilen güvenirlik katsayıları arasında istatistiksel bir fark gözlenmemiştir.

İki farklı puanlama yöntemine göre puanlanan yazılı yoklamaların ortalamalarının farklı olduğu gözlenmiş olması öğretim sürecinde sıkça kullanılan yazılı yoklamaların puanlama yöntemine bir kez daha dikkati çekmektedir. Özellikle sınav sonuçlarına göre öğrencilerin bilişsel alandaki davranışları konusunda karar verilecekse kısmi veya ayrıntılı puanlama tercih edilmelidir.

Araştırmada puanlama yönteminin ayrıntılı olup olmaması yazılı yoklama sorularının madde güçlüklerinde farklılaşmaya yol açtığı bulunmuştur. Dolayısıyla yazılı yoklama sorularının güçlük değerleri, puanlamanın yapılış yöntemine göre yorumlanmalıdır.

AP yapılan soruların madde ayırıcılıkları, GP yapılan madde ayırıcılıklarından daha yüksek bulunmuştur. Dolayısıyla soruların madde ayırıcılıklarının değerlendirilmesinde puanlamanın nasıl yapıldığı yine önemli bir faktör olmaktadır.

## KAYNAKÇA


Akhun, İ.(1982). *Hipotez testi ile ilgili bir araştırma*. Ankara: Ankara Üniversitesi Eğitim Fakültesi Yayınları No: 110.

Akkuş, O ve Baykul, Y. (2001). Çoktan seçmeli test maddelerini puanlamada, seçenekleri farklı biçimlerde ağırlıklandırmanın madde ve test istatistiklerine olan etkisinin incelenmesi. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi*. 20, 9 – 15.

Baykul, Y. (1997). *İstatistik: Metodlar ve uygulamalar*. Ankara: Anı yayıncılık

Büyüköztürk, Ş. (2002). *Sosyal Bilimler için veri analizi el kitabı*. Ankara: Pagem yayıncılık

Hopkins, K. (1998). *Educational and psychigical measurement and evaluation*. USA: Allyn&Bacon

Doğan, N.(2006) Yazılı Yoklamalar (ed. Hakan Atılgan) *Eğitimde ölçme ve değerlendirme*. Ankara: Anı Yayıncılık

Güler ve Gelbal, Açık uçlu matematik sorularının güvenirliğinin klasik test kuramı ve genellenebilirlik kuramına göre incelenmesi. *Kuram ve Uygulamada Eğitim Bilimleri / Educational Sciences: Theory & Practice*.10 (2), 989-1019

Güler, N. ve Teker Taşdelen, G. (2015). Açık uçlu maddelerde farklı yaklaşımlarla elde edilen puanlayıcılar arası güvenirliğin değerlendirilmesi. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi.* 6 (1), 12-24.

İlhan, M. (2016). Açık uçlu sorularla yapılan ölçmelerde klasik test kuramı ve çok yüzeyli rasch modeline göre hesaplanan yetenek kestirimlerinin karşılaştırılması. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi (H. U. Journal of Education)* 31(2), 346-368.

Iramaneerat, C. & Yudkowsky, R. (2007). Rater errors in a clinical skills assessment of medical students. *Evaluation & the Health Professions. 30 (3), 266-283.*

Kan, A. (2005). Yazılı yoklamaların puanlanmasında puanlama cetveli ve yanıt anahtarı kullanımının (farklı) puanlayıcı güvenirliğine etkisi. *Eğitim Araştırmaları*.19, 207-219

Kaptan, S. (1998). *Bilimsel araştırma ve istatistik teknikleri*. Ankara: Tekışık Web Ofset

Nunnally, J.C. (1970). *Introduction to psychological measurement.* New York : McGraw-Hill

Özçelik, A. D. (1997). *Test hazırlama kılavuzu.* Ankara: ÖSYM yayınları

Özdamar, K. (2002). *Paket programlar ile istatistiksel veri analizi 1*. Eskişehir: ETAM A.Ş. Matbaa Tesisleri, Kaan Kitapevi

Özdemir, D. (2003). Çoktan seçmeli testleri puanlama yöntemlerine bir bakış. *Eğitim Araştırmaları Dergisi.* 4(12),121-122.

Roid H.G & Haladyna, T. M. (1982). *A technology for test-item writing*. Newyork: Academic Press

Turgut, M. F. (1992). *Eğitimde ölçme ve değerlendirme*, Ankara: Saydam Matbaacılık

Yurdugül, H. (2010). Farklı madde puanlama yöntemlerinin ve farklı test puanlama yöntemlerinin karşılaştırılması. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*. 1(1), 1-8.

Wiliam J. M. & Karnes, M. R. Çev.: İbrahim Yurt (1968) , *Eğitimde başarının ölçülmesi*. Ankara: Ajans Türk Matbaası