Necatibey Eğitim Fakültesi Elektronik Fen ve Matematik Eğitimi Dergisi (EFMED)
Cilt 14, Sayı 1, Haziran 2020, sayfa 606-628. ISSN: 1307-6086

Necatibey Faculty of Education Electronic Journal of Science and Mathematics Education
Vol. 14, Issue 1, June 2020, pp. 606-628. ISSN: 1307-6086

Araştırma Makalesi / Research Article

Investigation of Solutions of Mathematical Problems Using Multiple Representations in Terms of İnter-Rater Reliability

Çiğdem AKIN ARIKAN [1], Feride ÖZYILDIRIM GÜMÜŞ [2]

[1] Ordu University, Faculty of Education, Ordu, akincgdm@gmail.com,
http://orcid.org/0000-0001-5255-8792

[2] Aksaray University, Faculty of Education, Aksaray, ferideozyildirimgumus@gmail.com,
http://orcid.org/0000-0002-1149-0039

Received : 11.02.2020 Accepted : 15.06.2020
Doi: 10.17522/balikesirnef.687639

Abstract The main purpose of this study is to examine the inter-rater reliability of the math problems, presented with four different representations (graphs, tables, equations, and verbal) on the basis of multiple representations in mathematics education. For this purpose, the generalizability theory is used in the scoring of the problem solutions and it is aimed to compare the G and Phi coefficients obtained in cases where crossed design is used. The open-ended mathematics problems are used in the research and students' solutions are graded by using a rubric. 54 students in the eighth grade of a public school participated in the study by solving the problems, and five mathematics teachers participated as a rater of students' solutions. As a result of the research, it was found that the problems in all representation types were effective on the inter-rater reliability, while the biggest difference was in the graphic representation.

Key words: mathematics education, multiple representation, inter-rater reliability, generalizability theory

---

Corresponding author: Çiğdem AKIN ARIKAN, Ordu University, akincgdm@gmail.com

Summary

Introduction

Expressing a problem or concept in different ways in mathematics can positively contribute to the learning processes of students with individual differences. At this point, the importance of using multiple representations in mathematics education becomes more significant. Hitt (1999) also mentioned that one of the aims of mathematics education is that

students can switch from one type of representation to another, supporting this view. Herbel-Eisenmann (2002) stated that multiple representations were collected in four categories: graphical, table, equation and verbal like problem states. According to Greeno and Hall (1997), using only one type of representation during courses negatively affects students' awareness about the advantages and disadvantages of different representations. For this reason, it is important for teachers to encourage their students to use multiple representations and to run those representations during the assessment and evaluation processes.

When considering mathematics classes, written exams are more effective for evaluating student performance. However, one of the biggest threats that can occur during the grading process of written exams is the subjectivity of raters. This is why a rubric was used to minimize rater subjectivity while scoring the problem solutions obtained in this research.

In this context, the purpose of this study was determined as the evaluation of inter-rater reliability for the grades obtained from students' solutions for math problems presented in four different representations according to the generalizability theory. In this process, how the student (o), rater (p) and representation (g) variables changed in the completely crossed pattern was examined.

Methodology

This study is described as descriptive research since the aim was to determine the inter-rater reliability of the grades obtained with a rubric from a test containing math problems using multiple representations.

The study group in the research consists of 54 students attending the eighth grade in a public school and five elementary mathematics teachers who graded the students' solutions.

Open-ended mathematical problems were used as one of the data collection tools. In determining the representations to be used, the four categories of graphical, table, equation and verbal mentioned by Herbel-Eisenmann (2002) were adopted. There were a total of 16 items in the related test. The items were not specifically directed to a learning area, but were prepared by the researchers in accordance with the eighth grade mathematics curriculum. In addition, a rubric developed by İlhan (2016) was used to grade students' solutions for the math problems. The rubric has four-point scoring as insufficient (1), needs to improve (2), good (3) and very good (4). In this context, a student can get minimum 0 points and maximum 4 points for a problem.

Moreover, generalizability theory was used to estimate the inter-rater reliability in the evaluation. Generalizability theory is a statistical model that deals with multiple sources of error

in the observed points and is based on classical test theory and variance analysis (Brennan, 2001; Shavelson & Webb, 1991). The data for grades obtained by the students were calculated by using the squares averages, variance components, percentages and G and Phi coefficients for the sources of variability with the main effect and interaction effect. For this process, the Edu G 6.1 program was used within the scope of the generalizability theory of o x p x g pattern which crossed student (o), representation (g) and rater (p) variables.

Results

In the framework of the generalizability theory, measurements made due to the high variance value predicted for the student in a fully cross-over pattern can determine the differences between the students. The variance value predicted for the rater main effect means that the raters' degree of rigidity or generosity differ slightly during the grading processes of students. The fact that the estimated variance component of the representations (G) is lower compared to the variance components of the other main effects indicates that the levels of the notations differ in terms of difficulty. As a result of the analysis made for each representation type separately, although there was a difference from the rater in all representations, the lowest rater variance was found for verbal representation and the highest for graphical representation.

Conclusion and Discussion

In the study, the variance component predicted for the main effect of the student is the largest proportion of the total variance. This shows that the test can distinguish the differences between students. In addition, it was determined that the effects arising from the rater difference were not high, but the raters differed slightly when evaluating the students. Swartz et al. (1999) also stated that the raters investigated within the framework of generalizability theory are not a serious source of variability if they are well trained. Another result is that the levels of the presented problems are not different in terms of difficulty. In addition, it was determined that the performance of students differs from each other in representation types; in other words, each student was more successful in one representation type than the other types.

One of the remarkable findings was the differentiation of stiffness/generosity during the grading the problems' solutions. The highest level of difference was experienced in grading the graphical and table representations. It is thought that this situation is due to the fact that these two display types are based more on visuality and that the viewers' perspective about solutions may differ.

Looking at this point, increasing this type of research will enable both valid and reliable tests to be created and problem solutions of students to be evaluated more accurately using multiple representations. In addition, using different types of rubrics, whether the situation

differs in multiple representations in mathematics can be examined. Another suggestion offered within the scope of the study is to encourage students to use different types of evaluation processes in classroom environments and to create learning environments that enable them to easily switch from one type of representation to another.

Çoklu Gösterimlerin Kullanıldığı Matematik Problemlerine Ait Çözümlerin Puanlayıcı Güvenirliği Açısından İncelenmesi

Çiğdem AKIN ARIKAN [1], Feride ÖZYILDIRIM GÜMÜŞ [2]

[1] Ordu Üniversitesi, Eğitim Fakültesi, Ordu, akincgdm@gmail.com,
http://orcid.org/0000-0001-5255-8792
[2] Aksaray Üniversitesi, Eğitim Fakültesi, Aksaray, frideozyildirimgumus@gmail.com,
http://orcid.org/0000-0002-1149-0039

Gönderme Tarihi: 11.02.2020 Kabul Tarihi: 15.06.2020
Doi: 10.17522/balikesirnef.687639

Özet –Bu çalışmanın temel amacı, matematik eğitiminde çoklu gösterimler temelinde ele alınan dört farklı gösterim (grafik, tablo, denklem ve sözel) ile sunulan matematik problemlerinin puanlayıcı güvenirliğini incelemektir. Bu amaçla, problem çözümlerinin puanlanması sürecinde genellenebilirlik kuramı işe koşulmuş ve tümüyle çaprazlanmış desenin (oxgxp) kullanıldığı durumlarda elde edilen G ve Phi katsayılarının karşılaştırılması amaçlanmıştır. Araştırmada kullanılan matematik problemleri, açık uçlu olarak sunulmuş ve öğrenci çözümleri dereceli puanlama anahtarı kullanılarak puanlanmıştır. Çalışmaya bir devlet okulunun sekizinci sınıfında öğrenim gören 54 öğrenci sunulan problemleri çözerek, beş matematik öğretmeni ise öğrenci çözümlerini puanlamak üzere puanlayıcı olarak katılmıştır. Araştırma sonucunda farklı gösterimdeki problemlerin puanlayıcı güvenirliği üzerinde etkili olduğu ve en büyük farklılığın grafik gösteriminde olduğu belirlenmiştir.

Anahtar kelimeler: matematik eğitimi, çoklu gösterim, puanlayıcılar arası güvenirlik, genellenebilirlik kuramı

Sorumlu yazar: Çiğdem AKIN ARIKAN, Ordu Üniversitesi, akincgdm@gmail.com

Giriş

Matematik eğitiminde bireysel farklılıkları göz önüne almanın, daha etkili bir öğrenme ortamı sunacağı düşünülmektedir. Bir problemi veya kavramı farklı şekillerde ifade etmek, bireysel farklılıklara sahip öğrencilerin öğrenme süreçlerine olumlu yönde katkı sağlayabilir. Bu noktadan yola çıkıldığında matematik eğitiminde çoklu gösterimleri kullanmanın önemi ortaya çıkmaktadır. Çünkü matematiksel fikirleri organize etme, matematiksel olguları modelleme ve problem çözme sürecinde çoklu temsilleri oluşturmak ve kullanmak gereklidir (NCTM, 2000). Hitt (1999) bu görüşü destekler nitelikte matematik eğitiminin asıl

amaçlarından birinin, öğrencilerin bir gösterim türünden diğerine tereddüt duymadan geçiş yapabilmeleri olduğundan söz etmiştir.

Goldin ve Shteingold (2001)'a göre çoklu gösterim, aynı bilgiyi birden fazla matematiksel sunumla göstermektir. Greeno ve Hall (1997) problem çözme sürecinde öğrencilerin çizimler yaptıklarından, notlar yazdıklarından, tablolar ve denklemler oluşturduklarından söz etmiştir. Bu şekilde de öğrenciler bir problemi farklı açılardan görebilmektedirler (Dufour-Janvier, Bednarz ve Belanger, 1987). Ayrıca çoklu gösterimleri kullanmak, daha derinlemesine ve esnek bir anlama fırsatı sunar (Keller ve Hirsch, 1998) ve öğrencilerin matematiksel fikirler arasında bağlantı kurmalarına, resmin tamamını görmelerine ve bir problemin çözümü için birden fazla yol olduğunu fark etmelerine imkan sağlar (Cleaves, 2008). Problem çözme becerisine önemli derecede fayda sağladığından (Schultz ve Waters, 2000), problem çözme sürecinde sıklıkla kullanılmasının gerekli olduğu düşünülmektedir.

Gösterimler temel olarak içsel ve dışsal olmak üzere iki şekilde gruplanmaktadır. İçsel gösterimler öğrencinin zihninde yer alan (Cobb, Yackel ve Wood, 1992; Erbilgin, 2003), dışsal gösterimler ise öğrencinin çevresinde yer alan gösterimlerdir (Cobb ve diğerleri,1992). Çoklu gösterimlere ait kategoriler incelendiğinde ise alan yazında farklı kategoriler olduğu görülmüştür. Örneğin Cleaves (2008) çoklu temsillerin sayısal (değerlerin tablo şeklinde sunulması), grafiksel, resimsel (resim veya diyagram şeklinde), sözel (hikaye ya da tanım şeklinde), sembolik (eşitlik) ve fiziksel (manipülatifler ya da somut materyaller) olmak üzere altı kategori altında toplandığını belirtmiştir. Öte yandan Herbel-Eisenmann (2002) ise temelde grafik, tablo, denklem ve problem durumları olmak üzere dört kategoride toplandığını belirtmiştir.

NCTM (2000)'e göre grafikler, semboller ve denklemler gibi farklı gösterimler okul matematiğinde birlikte kullanmak yerine ayrı ayrı kullanılmaktadır. Bu durum öğrencinin gösterimler arasındaki ilişkileri görmesini ve anlamlandırmasını zorlaştırabilir. Greeno ve Hall (1997) derslerde sadece bir tür gösterimin kullanılmasının, öğrencilerin farklı gösterimlerin iyi ve kötü yönlerini görmesini ve gösterimleri kullanmalarını olumsuz etkilediğini belirtmiştir. Bu nedenle de öncelikle öğretmenlerin derslerde çoklu gösterimlere yer vermeleri ve öğrencileri de farklı gösterimleri birlikte kullanmaları konusunda cesaretlendirmeleri gerekmektedir. Bu noktadan yola çıkıldığında, öğretmenlerin çoklu gösterimlerin kullanımını yaygınlaştırmada ve öğrencilerini teşvik etmede, ölçme değerlendirme süreçlerini de işe koşmasının önemli olduğu ifade edilebilir.

Birgin ve Gürbüz (2008)'e göre, ölçme ve değerlendirme sürecinin doğru şekilde yürütülmesi, öğrenciyi ve öğrenim sürecini daha yakından inceleme fırsatı sunar. Bir başka ifade ile doğru ölçme ve değerlendirme yöntemlerinin doğru şekilde kullanılması, öğrencinin kavram yanılgılarını ve öğrenme eksikliklerini belirlemenin yanı sıra, öğretim yöntem ve teknikleri ile öğrenme ortamlarının da etkinliğini görmeyi sağlar. Bu nedenle hazırlanan ölçme değerlendirme aracının ve bu ölçme aracı ile sergilenen performansın doğru değerlendirilmesinin son derece önemli olduğu söylenebilir.

Alan yazın incelendiğinde öğretmenlerin ölçme değerlendirme sürecinde ağırlıklı olarak geleneksel yöntemlerden olan yazılı yoklamalar, çoktan seçmeli, kısa cevaplı ve boşluk doldurmalı testleri tercih ettikleri (Güven ve Eskitürk; 2007), öğrencilerin ise uzun cevap gerektiren sınavlara göre çoktan seçmeli testleri daha fazla tercih ettikleri (Struyven, Dochy ve Janssens, 2005) belirlenmiştir. Diğer yandan matematik dersi özelinde düşünüldüğünde bu ölçme yöntemlerinden yazılı sınavların, çoktan seçmeli testlere göre öğrencinin performansını değerlendirmede daha etkili olduğu söylenebilir. Çünkü yazılı sınavlarda, öğretmen öğrencinin kavram yanılgılarını ve öğrenme eksikliklerini daha net görebilirken, şans başarısından uzak olması ve adım adım puanlama şansı sunmasıyla da daha etkin bir değerlendirme imkanı sunabilmektedir. Ancak yazılı sınavların puanlanması sırasında devreye girebilecek olan en büyük tehditlerden biri de puanlayıcının öznelliğidir.

Dereceli Puanlama Anahtarı ve Puanlayıcı Güvenirliği

Dereceli puanlama anahtarı, öğrencilerin performanslarını değerlendirmek için önceden belirlenmiş kriterlerden oluşan puanlama kılavuzları olarak tanımlanmaktadır (Mertler, 2000). Dereceli puanlama anahtarlarıyla, öğrencilerin ortaya çıkardığı ürün ve bu ürünün ortaya çıkması için yapması gereken davranışlar birlikte değerlendirilebilir. Bütünsel ve analitik olmak üzere iki dereceli puanlama anahtarı çeşidi bulunmaktadır (Nitko, 2001). Bütünsel dereceli puanlama anahtarı, öğrencilerin performans düzeyleri için tek puan verildiği durumda kullanılır. Analitik dereceli puanlama anahtarı ise, belirlenmiş bazı ölçütler çerçevesinde bir ürünü, süreci veya performansı meydana getiren parçaların ayrı ayrı puanlanması sürecinde kullanılır (Moskal, 2000).

Bir bireyi bir pozisyona yerleştirme ya da kabul etme gibi bir durum söz konusu olduğunda, puanlayıcıların bireylere verdikleri puanlar konusunda hem fikir olmaları yararlı bilgiler sağlayabilmektedir (Goodwin, 2001). Bu nedenle puanlama sırasında dereceli puanlama anahtarı kullanılması, puanlayıcılar arasında birlik sağlamada, bir başka ifade ile objektifliği sağlamada yardımcı olmaktadır. Ancak iyi hazırlanmış puanlama anahtarı bile, açık

uçlu maddelerde puanlamanın objektif olması konusunda belli bir noktaya kadar işe yaramaktadır (Tekindal, 2000). Bu noktada işe koşulabilecek bir diğer süreç ise puanlayıcı güvenirliğidir.

Aiken (2000)'a göre, farklı maddeler ve bireyler için ikiden fazla puanlayıcının yaptığı puanlamaların tutarlılık derecesi, puanlayıcı güvenirliği olarak tanımlanmaktadır. Puanlayıcılar arası güvenirlik bireylerin, olayların, özellik ya da davranışlarını derecelendirmek ve puanlamak için öznel görüşlere ihtiyaç duyulduğunda önemli olmaktadır (Goodwin, 2001). Ancak puanlayıcı güvenirliği sağlanmadığı takdirde, bir öğrencinin puanının puanlayıcıdan puanlayıcıya değişebilme durumu ortaya çıkabilmekte ve öğrenciler de aldıkları puanların genellikle puanlayıcının öznel yargısına dayandığını belirtmektedirler (Moskal ve Leydens, 2000). Bu nedenle puanlayıcıdan kaynaklanan öznel yargıları ve farklılıkları tamamen ortadan kaldırmasa da iyi tasarlanmış bir problemlerin çözümüne ilişkin cevap anahtarı ya da dereceli puanlama anahtarı gerekebilmektedir. Çünkü dereceli puanlama anahtarı kullanımı puanlamayı daha anlamlı hale getirmektedir (Bresciani, Zelna ve Anderson, 2004).

Araştırma kapsamında elde edilen problem çözümlerinin puanlanması sırasında, puanlama sınırlarının iyi tanımlanması için bir puanlama anahtarına ihtiyaç olduğuna karar verilmiştir. Kullanılacak puanlama anahtarının türünün belirlenebilmesi için alan yazın incelenmiş ve genellenebilirlik çalışmaları kapsamında kullanılan puanlama anahtarı türünün bir farklılık oluşturmadığı sonucuna ulaşılmıştır (Ömür ve Erkuş, 2013). Bu çalışma kapsamında bazı problem çözümlerini değerlendirirken adım adım parçalara ayırmanın mümkün olamayacağı kanısı ile bütüncül puanlama anahtarı kullanmanın uygun olacağı düşünülmüştür.

Puanlayıcı güvenirliği için klasik test kuramına, madde tepki kuramına ve genellenebilirlik kuramına dayalı yöntemler bulunmaktadır. Alan yazın incelendiğinde açık uçlu veya performansa dayalı sınavlar için puanlayıcı güvenirliği çalışmaların olduğu görülmektedir (Büyükkıdık ve Anıl 2015; Doğan ve Anadol, 2017; Güler ve Gelbal 2010, Güler ve Teker, 2015; Kan, 2005, Yılmaz ve Başusta, 2015). Güler ve Teker (2015) tarafından yapılan çalışmada puanlayıcılar arası güvenirliği belirlemek için korelasyon, ortalamaların karşılaştırılması, uyuşma yüzdesi ve genellenebilirlik kuramı kullanılmış ve bu yöntemler arasında genellenebilirlik kuramının en kullanışlı yöntem olduğunu belirtmişlerdir. Bu çalışma kapsamında da puanlayıcı güvenirliği belirlemek için genellenebilirlik kuramından yararlanılmıştır. Ayrıca alan yazın incelendiğinde, gösterimlerden elde edilen puanların puanlayıcı güvenirliğini ortaya koyan yeterli çalışmaya da rastlanmamıştır. Bu nedenle,

çözümlerinde farklı gösterimlerin bulunduğu matematik problemlerinin farklı puanlayıcılar tarafından puanlandığında güvenirliğin belirlenmesinin alana katkı sağlayacağı düşünülmektedir. Çünkü alan yazında puanlayıcı güvenirliğini inceleyen çalışmaların, matematik problemlerinde farklı gösterimleri yeterince ele almadıkları gözlenmiştir. Bu nedenle de elde edilen bulguların yeni çalışmalara da fikir sunması beklenmektedir.

Bu bağlamda, bu araştırmanın amacı dört gösterim türünün bulunduğu matematik problemleri için çözümlerin bütüncül bir puanlama anahtarı kullanarak puanlanmasıyla elde edilen puanların genellenebilirlik kuramına göre öğrenci (o), puanlayıcı (p) ve gösterim (g) değişkenlerinin tümüyle çaprazlanmış deseninde nasıl değiştiğini incelemektir. Bu amaç doğrultusunda aşağıdaki sorulara cevap aranmıştır.

a. Matematik beceri testinin dereceli puanlama anahtarıyla puanlanmasıyla öğrenci, gösterim ve puanlayıcı ana etkileri ve etkileşim etkilerine ait kestirilen varyans bileşeni nasıldır?

b. Matematik beceri testinin dereceli puanlama anahtarıyla puanlanmasıyla elde edilen puanlamaların genellenebilirlik katsayısı [G] ve güvenirlik katsayısı [Phi] katsayıları nasıldır?

Yöntem

Araştırmanın Türü

Bu araştırma, dereceli puanlama anahtarıyla puanlanan çoklu gösterimlerin kullanıldığı matematik problemlerini içeren bir testin puanlayıcı güvenirliğinin incelendiği betimsel bir araştırmadır. Betimsel araştırmalarda amaç var olan verilen durumu olabildiğince tam ve olduğu gibi belirlemektir (Büyüköztürk, Kılıç Çakmak, Akgün, Karadeniz ve Demirel, 2012). Betimsel araştırmalar, öğrenci gruplarının başarılarını belirlemek, öğretmenlerin, yöneticilerin davranışlarını tanımlamak bireylerin tutumlarını belirlemek için yapılır (Büyüköztürk ve diğerleri, 2012).

Araştırmanın etik kurul onay belgesi, Ordu Üniversitesi Sosyal ve Beşeri Bilimler Etik komisyonundan 23.06.2020 tarihinde alınmıştır.

Araştırma Grubu

Bu araştırmada, seçkisiz olmayan örnekleme yöntemlerinden uygun örnekleme yöntemi kullanılmıştır. Uygun örneklemede, araştırmacılar katılımcıları coğrafi olarak yakın, kolay erişilen, araştırma için uygun ve gönüllü bireylerden seçmektedir (Dörnyei, 2007; Gravetter ve Forzano, 2012).

Çalışma grubunu Ordu ili Altınordu ilçesinde bulunan bir devlet ortaokulunun 2019-2020 eğitim öğretim yılında sekizinci sınıfta öğrenim gören 29'u (%53,7) erkek ve 25'ü (% 46,3) kız olmak üzere toplam 54 öğrenci oluşturmaktadır. Ayrıca puanlayıcı olarak beş ilköğretim matematik öğretmeni araştırmada yer almıştır ve puanlayıcılara ilişkin bilgiler Tablo 1'de verilmiştir.

Tablo 1 Puanlayıcılara ait Bilgiler

Puanlayıcı	Cinsiyet	Kıdem	Eğitimi
P1	K	7	İlköğretim matematik öğretmeni-Bilgisayar ve öğretim teknolojileri eğitimi alanında Yüksek Lisans Mezunu
P2	E	7	İlköğretim matematik öğretmeni-İlköğretim matematik eğitimi alanında Yüksek Lisans Tez aşamasında
P3	K	11	İlköğretim matematik öğretmeni- İlköğretim matematik eğitimi alanında Yüksek Lisans Tez aşamasında
P4	K	6	İlköğretim matematik öğretmeni- İlköğretim matematik eğitimi alanında Yüksek Lisans Tez aşamasında
P5	K	7	İlköğretim matematik öğretmeni

Veri Toplama Araçları

Bu çalışmada kapsamında veri toplama aracı olarak farklı gösterimlerin kullanıldığı, açık uçlu matematiksel problemler ve bir dereceli puanlama anahtarı (İlhan, 2016) kullanılmıştır. Kullanılan gösterimlerin belirlenmesinde Herbel-Eisenmann (2002)'in ele aldığı grafik, tablo, denklem ve problem durumları olmak üzere dört kategori temel alınmıştır. Söz konusu açık uçlu matematiksel problemlerin yer aldığı ölçme aracında tablo, grafik, denklem ve sözel ifade olmak üzere dört gösterimin kullanıldığı toplam 16 madde bulunmaktadır. Maddeler özel olarak bir öğrenme alanına yönelik olmayıp, MEB (2018) ortaokul matematik öğretim programında sekizinci sınıf düzeyindeki öğrencilere uygun olarak araştırmacılar tarafından hazırlanmıştır. Ayrıca maddeler hazırlanırken, her bir madde aynı özelliği ölçen farklı bir gösterim şekliyle sunacak şekilde testteki başka bir maddeyle eşleşmiştir. Ölçme aracında yer alan maddelere ilişkin özellikler Tablo 2'de sunulmuştur.

Tablo 2 Maddelerin Özellikleri

Madde No / Gösterim Biçimi	Eşlendiği Madde No / Gösterim Biçimi	Madde No / Gösterim Biçimi	Eşlendiği Madde No / Gösterim Biçimi
1 /sözel	12 / grafik	9 / tablo	16 /grafik
2 / denklem	13 / denklem	10 / sözel	7 / grafik
3 / tablo	5 / grafik	11 / denklem	4 /grafik
4 / grafik	11 / denklem	12 / tablo	1 /sözel
5 / grafik	3 / tablo	13 / sözel	2 /denklem
6 /denklem	14 / tablo	14 / tablo	6 /denklem
7 / grafik	10 / sözel	15 / denklem	8 /sözel
8 / sözel	15 / denklem	16 / grafik	9 / tablo

Tablo 2'de görüldüğü üzere grafik, tablo, denklem ve sözel olmak üzere dört farklı gösterim türünün her birinden dörder tane olmak üzere toplam 16 açık uçlu madde ölçme aracında yer almaktadır. Söz konusu maddeler hazırlandıktan sonra belirtilen çoklu gösterim türüne ve sınıf seviyesine uygunluğu açısından alan uzmanlarına değerlendirmeleri için sunulmuştur. Söz konusu alan uzmanlarından iki tanesi matematik eğitimi alanında akademisyen olarak görev yapmakta iken, iki tanesi de matematik öğretmeni olarak görev yapmaktadır. Alınan görüşler doğrultusunda bazı maddelerin ifadelerinde değişikliğe gidilmiştir. Ayrıca elde edilen uzman görüşlerine ait puan ortalamalarının 4 puanın (5 puan üzerinden) üzerinde olduğu belirlenmiş ve maddelerin hem belirlenen gösterim türüne hem de sınıf seviyesine uygun olduğu sonucuna ulaşılmıştır.

Ayrıca, öğrencilerin açık uçlu matematik problemlerine verdikleri yanıtların puanlanmasında İlhan (2016) tarafından geliştirilen dereceli puanlama anahtarı kullanılmıştır. Bu dereceli puanlama anahtarının seçilmesinin nedeni geçerlik ve güvenirlik çalışmalarının yapılmış olması ve matematik problemlerinin çözümü için geliştirilmiş olmasıdır. Söz konusu dereceli puanlama anahtarında yetersiz (0), geliştirilmesi gerek (1), iyi (2) ve çok iyi (3) şeklinde dörtlü bir puanlama bulunmaktadır. Bu bağlamda bir öğrenci bir problem için en az 0 puan, en fazla 3 puan alabilmektedir. Dereceli puanlama anahtarı Tablo 3'de verilmiştir.

Tablo 3 Soruların puanlanmasında kullanılan dereceli puanlama anahtarı

3 (Çok iyi)	Uygun çözüm yolu kullanılmıştır. Çözüme yönelik olarak yapılan işlemlerde herhangi bir hata bulunmamaktadır. Doğru sonuca ulaşılmıştır. Problemi çözmek için yapılan işlemler açık, ayrıntılı ve örnek yanıt niteliğindedir.
2 (iyi)	Problem büyük ölçüde anlaşılmıştır. Uygun çözüm yolu kullanılmasına rağmen küçük işlem hatalarından ya da anlaşılmayan nedenlerden dolayı doğru sonuca ulaşılmamıştır. Doğru sonuca ulaşılmıştır. Ancak çözüme nasıl ulaşıldığına dair yeterli açıklama bulunmamaktadır.
1 (Geliştirilmesi Gerek)	Problem kısmen anlaşılmıştır. Uygun çözüm yolu ile başlangıç yapılmış, fakat devamı getirilememiştir. Kullanılan çözüm yolu doğru olmakla birlikte, yapılan işlemlerde önemli hatalar bulunmaktadır. Dolayısıyla doğru sonucuna ulaşılamamıştır.
0 (Yetersiz)	Problem anlaşılmamıştır. Problemi cevaplamak için kullanılan stratejiler tamamen yanlıştır ve çözüme yönelik herhangi bir yarar sağlamamaktadır. Herhangi bir işlem veya açıklama yapılmamıştır. "Bilmiyorum", "Çok zor bir soru" gibi ifadeler kullanılmış ya da problemde sunulan veriler tekrar edilmiştir

İşlem Yolu

Araştırmayla ilgili gerekli açıklamalar yapıldıktan sonra, katılımın zorunlu olmadığı ve test sonuçlarının not verme amacıyla kullanılmayacağı öğrencilere bildirilmiştir. Ayrıca matematik öğretmeni ve araştırmacı tarafından formda yer alan soruların hepsini

cevaplamalarının önemli olduğu vurgulanmıştır. Öğrenciler soruları cevapladıktan sonra, öğrenci kâğıtları numaralandırılmış ve beş puanlayıcı için çoğaltılmıştır. Puanlayıcılara öğrencilere ait yazılı kağıtlarıyla birlikte puanlama için dereceli puanlama anahtarı verilerek puanlama için gerekli materyaller sağlanmıştır. Puanlama öncesinde, puanlayıcılara değerlendirmede kullanacakları dereceli puanlama anahtarı tanıtılmış ve puanlarken dikkat etmeleri gerekenler noktasında bütün puanlayıcıların bir arada olduğu bir panel düzenlenmiştir. Beş puanlayıcı 54 öğrencinin 16 maddeye verdikleri yanıtları bağımsız şekilde puanlamış ve bu şekilde puanlayıcılara ait veri setleri oluşturulmuştur.

Verilerin Analizi

Performans ve açık uçlu sınavların değerlendirmelerinde puanlayıcı güvenirliğini kestirmek için sıklıkla Klasik Test Kuramı [KTK] kullanılmaktadır. KTK sadece tek bir hata kaynağını ele alır ve varyans kaynaklarının etkileşimlerini belirleyemez. Genellenebilirlik kuramı ise gözlenen puanlardaki çoklu hata kaynaklarını birlikte ele alan ve KTK ve varyans analizine dayanan istatistiksel bir modeldir (Brennan, 2001; Shavelson ve Webb, 1991). Bir başka ifadeyle, genellenebilirlik kuramı durumlar, test formları, puanlayıcılar ve maddeler gibi olası birden fazla hata kaynağını ve bu olası hata kaynaklarının etkileşimlerini kestirebilir (Shavelson ve Webb, 1991). Genellenebilirlik kuramı, genellenebilirlik [G] çalışması ve karar [K] çalışması olmak üzere iki aşama içerir. G çalışması yürüten araştırmacı, öncelikle ölçme yaptığı örneklemi ölçmenin evrenine genelleme derecesi ile ilgilenirken, K çalışması ise farklı durumlarda güvenilirliğin nasıl değişeceğini araştırır. G çalışmasının amacı, yeterli genellenebilirliğe sahip bir karar çalışmasının (K) planlanmasına yardımcı olmaktır (Crocker ve Algina, 2008). G Kuramında yüzey (değişkenlik kaynağı), zaman, madde ve puanlayıcı gibi benzerlik gösteren ölçme durumlarına denir (Brennan, 1992). Bir yüzeyin sabit veya rastgele olarak ele alınması araştırmacıya bağlıdır. Eğer araştırmacı elde ettiği sonuçlar evrene genellemek istiyorsa değişkenlik kaynağını rastgele ele almalı, ancak eğer elde ettiği sonuçlar evrene genellemek istemiyorsa ya da sadece o değişkenlik kaynağının ele aldığı durumları belirtmek istiyorsa sabit yüzey olarak ele alabilir. Dolayısıyla bu araştırmada belirlenen gösterimlerdeki durum incelenmek istendiğinden sabit yüzey olarak ele alınmıştır.

Genellenebilirlik çalışmalarında çaprazlanmış ve yuvalanmış olmak üzere iki tür desen vardır (Shavelson ve Webb, 1991). Çaprazlanmış desende, bir ölçümün bütün koşullarının diğer değişkenlik kaynağının bütün koşullarıyla birlikte gözlemlenmektedir (Shavelson ve Webb, 1991). Diğer bir ifadeyle çaprazlanmış desende, bütün öğrenciler testte yer alan maddelerin hepsini cevaplarken, bütün puanlayıcılar maddelerin ve bireylerin tümünü puanlamaktadır.

Yuvalanmış desende ise bir yüzeyin iki veya daha fazla koşulu diğer yüzeyin bazı koşulları ile gözlemlenmektedir (Shavelson ve Webb, 1991). Bu çalışmada çaprazlanmış desen kullanılmıştır.

Öğrencilerin tüm gösterimdeki maddeleri cevapladığı ve bütün puanlayıcıların da her öğrenciyi dereceli puanlama anahtarı ile puanlamasıyla elde edilen veriler tümüyle çaprazlanmış desende ana etki ve etkileşim etkisine sahip değişkenlik kaynakları için kareler ortalamaları, varyans bileşenleri, yüzdeleri ve G ve Phi katsayıları hesaplanmıştır. G katsayısı gerçek puan varyansının, bağıl (göreli) puan varyansı ve gerçek puan varyansının toplamına oranıyla ve Phi katsayısı gerçek puan varyansının mutlak hata varyansı ile gerçek puan varyansının toplamına oranıyla elde edilir (Brennan, 2001, s.13). G katsayısı 0-1 aralığında değer alır ve puanların güvenirliğinin veya genellenebilirliğinin düzeyini belirtir (Shavelson ve Webb, 1991).

Araştırmada öğrenci (o), gösterim (g) ve puanlayıcı (p) değişkenin çaprazlandığı o x p x g deseninin genellenebilirlik kuramı kapsamında, Edu G 6.1 programı (Cardinet, Johnson ve Pini, 2010) kullanılmıştır. Ayrıca puanlayıcıların dereceli puanlama anahtarı ile verdikleri puanlar arasındaki ilişkinin belirlenmesi için Pearson momentler çarpımı korelasyon katsayısı [PMÇKK] hesaplanmıştır.

Puanlayıcıların dereceli puanlama anahtarı kullanarak 16 maddeye verdikleri puanlara ait betimsel istatistikler Tablo 4'de yer almaktadır.

Tablo 4 Betimsel İstatistikler

Puanlayıcı	Min.	Mak.	Ort.	Standart Sapma
P1	11,00	47	28,83	7,52
P2	1,00	46	22,39	10,81
P3	7,00	45	25,57	9,46
P4	2,00	45	23,59	10,28
P5	4,00	44	23,00	9,33

Tablo 4 incelendiğinde puanlama anahtarına göre yapılan puanlamalardan en yüksek ortalamanın 1. puanlayıcıya (P1), en düşük ortalamanın ise 2. puanlayıcıya (P2) ait olduğu görülmektedir. Ayrıca, puanlayıcıların puanlama anahtarı kullanarak öğrencilere verdikleri puanlar arasındaki ilişkiye ait korelasyon değerleri incelendiğinde 0,789 ile 0,939 arasında değiştiği ve yüksek düzeyde pozitif yönlü ilişkiye sahip olduğu bulunmuştur ($p<.01$). Her bir puanlayıcının verdiği puanların güvenirliği için Cronbach Alfa iç tutarlılık katsayısı kullanılmıştır. Elde edilen güvenirlik katsayıları 0,82 ile 0,86 arasında değiştiği görülmüştür.

Bir başka ifadeyle, puanlamadan elde edilen verilerin iç tutarlığının yeterli olduğu şeklinde yorum yapılabilir.

Bulgular

54 öğrencinin beş puanlayıcı tarafından dört farklı gösterimdeki maddeleri puanlandığı desene ilişkin G çalışması sonucunda elde edilen sonuçlar Tablo 5'de yer almaktadır. G kuramı ile öğrenci (o), puanlayıcı (p), gösterim (g), öğrenci-puanlayıcı etkileşimi (oxp), öğrenci-gösterim etkileşimi (oxp), puanlayıcı- gösterim etkileşimi (pxg) ve son olarak öğrenci-puanlayıcı-gösterim etkileşimi (oxpxg,e) (artık varyansı) hata kaynakları olarak değerlendirilmektedir.

Birinci alt probleme ilişkin bulgular

Tabloda öğrenci "o", puanlayıcı "p" ve gösterim "g" sembolleri ile gösterilmiştir. Gösterim yüzeyine ilişkin dört durum olduğundan bu yüzey sabit olarak alınmıştır.

Tablo 5 Tümüyle Çaprazlanmış o x p x g Deseni için Elde Edilen Varyans Bileşenleri

varyans kaynağı	kareler toplamı	sd	kareler ortalaması	varyans	varyans yüzdesi
o	5375,47130	53	101,42399	4,26578	45,6
p	372,42037	4	93,10509	0,40342	4,3
g	312,40278	3	104,13426	0,32374	3,5
op	697,07963	212	3,28811	0,51557	5,5
og	2233,34722	159	14,04621	2,56407	27,4
pg	46,86111	12	3,90509	0,04962	0,5
opg,e	779,63889	636	1,22585	1,22585	13,1
Toplam	9817,22130	1079			100

Çaprazlanmış desen için elde edilen varyans bileşenlerine ait bulgular tüm etkiler (ana etki, etkileşim etkisi ve artık) dikkate alınarak yorumlanmıştır. Bu araştırmada ölçme objesi öğrencilerdir ve bu nedenle öğrencilerden kaynaklanan değişkenliğin fazla olması beklenir. Tablo incelendiğinde, öğrenci (o) ana etkisi için kestirilen varyans bileşeninin (4,26) toplam varyans içindeki en büyük orana sahip olduğu ve toplam varyansın %45,6'sını açıkladığı

görülmektedir. Öğrenci için kestirilen bu varyans değeri, yapılan ölçümlerin öğrenciler arasındaki farklılıkları belirleyebildiği anlamına gelmektedir. Puanlayıcılar için kestirilen varyans bileşeninin (0,403) toplam varyansın %4,3'ünü açıkladığı görülmektedir. Puanlayıcı ana etkisi için kestirilen varyansın, toplam varyansı açıklama oranının çok büyük olmadığından puanlayıcı farklılığından kaynaklanan etkilerin çok yüksek olmadığı şeklinde yorumlanabilir. Bir başka ifadeyle, puanlayıcıların öğrencileri puanlarken katılık veya cömertlik düzeylerinin çok az farklılık gösterdiği anlamına gelmektedir. Gösterimlere (g) ait kestirilen varyans bileşeninin (0,323) toplam veryansın %3,5'ini açıkladığı görülmektedir. Elde edilen varyans bileşeni diğer ana etkilere ait varyans bileşenleri ile kıyaslandığında daha düşük olması, gösterimlerin güçlük bakımından düzeylerinin farklılaşmasının az olduğunun göstergesidir. Diğer ifadeyle bu bulgu; farklı gösterim biçimleri ile sunulan maddelerin kendi içlerinde heterojenliğinin az olduğu şeklinde yorumlanabilir.

Öğrenci gösterim (og) etkileşme etkisi için kestirilen varyans bileşeninin (2,56) toplam varyans içindeki en büyük ikinci orana sahip olduğu ve toplam varyansın %27,4'ünü açıkladığı görülmektedir. Elde edilen bu bulgu, öğrencilerin gösterim türlerindeki performansın gösterim türlerine göre farklılaştığı anlamına gelmektedir. Öğrenci puanlayıcı etkileşimi (op) için kestirilen varyans bileşeninin (0,51) toplam varyansın %5,5'ini açıkladığı görülmektedir. Bu bulgu puanlayıcıların öğrencileri değerlendirirken öğrenciden öğrenciye çok az farklılık gösterdiği ve öğrencilerin puanlarının puanlayıcıdan puanlayıcıya çok az miktarda değiştiği şeklinde yorumlanabilir. Puanlayıcı- gösterim (pg) etkileşimine ait kestirilen varyans bileşeni (0,049) toplam varyansın %0,5'ini açıklamaktadır. Kalan etkiye (artık varyans) ait varyans bileşeninin toplam varyansın içindeki üçüncü en büyük orana sahip olduğu ve %13'ünü açıkladığı görülmektedir. Bu değerin büyük olması desende olmayan başka tesadüfi hata kaynaklarının da deseni etkilediğini göstermektedir.

Tümüyle çaprazlanmış desende öğrenci-gösterim (og) ortak etkisinin toplam varyansın %27,4 gibi büyük bir kısmını oluşturduğundan, gösterimdeki puanlayıcı farklılaştığını göstermektedir. Bu nedenle sabitlenmiş yüzey olan gösterim için ayrı ayrı çaprazlanmış desen yapılmıştır. g1 sözel, g2 denklem, g3 tablo ve g4 grafik ile ilgili olmak üzere elde edilen sonuçlar Tablo 6'da yer almaktadır.

Tablo 6 Sabitlenmiş Yüzey İçin Ayrı Ayrı o x p Deseni İçin Kestirilen Varyans Bileşenleri

Varyans Bileşeni	g1		g2		g3		g4	
	Varyans	*Varyans Yüzdesi*	*Varyans*	*Varyans Yüzdesi*	*Varyans*	*Varyans Yüzdesi*	*Varyans*	*Varyans Yüzdesi*
O	7,670	78,6	7,571	78,9	4,941	51,7	4,470	66,3
P	0,359	3,7	0,505	5,3	0,433	4,5	0,507	7,5
OP	1,735	17,8	1,517	15,8	4,181	43,8	1,7628	26,1

Her bir gösterim için ayrı ayrı yapılan analiz sonucunda bütün gösterimlerde puanlayıcıdan kaynaklı bir farklılık olmakla birlikte, en düşük puanlayıcı varyansı sözel gösteriminde ve en yüksek ise grafik gösteriminde olduğu görülmektedir. Bir başka ifadeyle, grafik gösterimindeki soruların puanlanmasında puanlayıcılar arası tutarsızlıkların olduğu ve puanlayıcıların aynı öğrenciyi puanlarken katılık/cömertlik düzeylerinin daha fazla değiştiği anlamına gelmektedir. Ayrıca kalan etkiye ait varyans bileşeni yüzdesinin en düşük g2'de (%15,8) ve en yüksek g3'te (%43,8) olduğu bulunmuştur.

İkinci alt probleme ilişkin bulgular

Tablo 4'de yer alan tümüyle çaprazlanmış desen için hesaplanan güvenirlik katsayıları Tablo 7'de yer almaktadır.

Tablo 7 o x p x g desenine ait G ve Phi katsayıları

n_{birey}	*54*
n_{madde}	*20*
$n_{puanlayıcı}$	*5*
G katsayısı	*0,84*
Phi katsayısı	*0,81*

Tümüyle çaprazlanmış desen için elde edilen G ve Phi katsayıları incelendiğinde, mutlak değerlendirmeler için hesaplanan Phi katsayısı 0,81 ve göreli (bağıl) değerlendirmeler için hesaplanan G katsayısı ise 0,84 olarak bulunmuştur. Her iki güvenirlik katsayısı 0,80'den büyük çıktığı için puanlamanın güvenilir olduğu söylenebilir (Brennan, 2001).

Sonuç, Tartışma ve Öneriler

Araştırma kapsamında öğrencilerin dört farklı gösterimde soruların yer aldığı bir matematik testinde tümüyle çaprazlanmış desen için elde edilen varyans bileşenleri

incelenmiştir. Araştırma kapsamında elde edilen ilk bulguya göre, öğrenci ana etkisi için kestirilen varyans bileşeninin, toplam varyans içindeki en büyük orana sahip olmasıdır. Bu da yapılan ölçümlerin öğrenciler arasındaki farklılıkları belirlediğini açıklamaktadır. Bu sonuç alan yazınla benzerlik göstermektedir (Kaya, 2011; Polat Demir, 2016). Ayrıca birey değişkenlik kaynağının ölçme objesi olduğunda yüksek varyansa sahip olması istenilen bir durumdur (Güler, Uyanık ve Teker, 2012).

Ayrıca gösterim türlerine göre sunulan problemlerin güçlük bakımından düzeylerinin farklılaşmadığı da elde edilen bir diğer bulgudur. Bunun yanında, öğrencilerin gösterim türlerindeki performanslarının gösterimden gösterime farklılaştığı, bir başka ifade ile öğrencilerin bir gösterim türünde, diğer gösterim türlerine göre daha başarılı oldukları belirlenmiştir. Alan yazında öğrencilerin kimi zaman bir problem için kullanılan bir gösterim türünden diğerine geçmekte zorlanabildikleri belirtilmiştir (Yerushalmy, 1997). Bu durum öğrencilerin tek bir tür gösterim türüne yoğunlaşmalarına neden olabilir. Bu durum da bir gösterim türünde, diğerlerine göre daha başarılı olma durumunu açıklayan bir neden olarak gösterilebilir.

Çalışmadan elde edilen bir başka bulgu da puanlayıcı farklılığından kaynaklanan etkilerin yüksek düzeyde olmadığı, ancak puanlayıcıların öğrencileri değerlendirirken az da olsa farklılık gösterdiği belirlenmiştir. Swartz ve diğerleri (1999) genellenebilirlik kuramı çerçevesinde gerçekleştirdiği puanlayıcıların iyi eğitilmiş olduğu taktirde ciddi bir değişkenlik kaynağı olmadığını belirtmişlerdir. Bu noktadan bakıldığında çalışmanın bu bulgusuyla benzerlik gösterdiği söylenebilir.

Dikkat çeken bulgulardan biri de, puanlayıcıların grafik gösterimi kullanılarak sunulan problemlerin çözümlerinin puanlanması sırasında katılık/cömertliklerinin daha fazla farklılaşmasıdır. En yüksek düzeyde farklılık grafik gösteriminin puanlanmasında yaşanmıştır. Bu durumun ortaya çıkması bu gösterim türünün daha çok görselliğe dayanmasından ve puanlayıcıların görsel sunumları ve çözümlere bakış açısının farklılaşabileceğinden kaynaklandığı düşünülmektedir. Öte yandan bu bulgu Atmaz (2009) gerçekleştirdiği çalışmanın bulgusu ile çelişmektedir. Atmaz (2009), grafik yorumlama becerisinin ölçüldüğü dört açık uçlu maddeyi dereceli puanlama anahtarı kullanarak farklı puanlayıcılara puanlatmış, puanlayıcıların verdikleri puanlar ortalamaları arasındaki fark anlamlı bulunmadığını ve söz konusu puanlar arasında pozitif yönde yüksek bir ilişki olduğunu belirlenmiştir. Ancak Atmaz (2009)'un çalışmasında fark bulunamamasının, tek bir gösterim türü kullanılmasından

kaynaklandığı düşünülmektedir. Gerçekleştirilen bu çalışma da dört farklı gösterim türünün kullanılmış olması, böyle bir farklılığın ortaya çıkmasına neden olmuş olabilir.

Alan yazında puanlayıcı güvenirliği ile ilgili gerçekleştirilen çalışmaların büyük bir çoğunluğunun farklı ülkelerde uygulanan testlerin güvenirliğini belirlemek (Evans-Hampton, Skinner, Henington, Sims ve McDaniel, 2002; Güler ve Gelbal, 2010; Stecker ve Fuchs, 2000; Thurber, Shinn ve Smolkowski, 2002) amacıyla gerçekleştirildiği gözlenmiştir. Bunun yanında, puanlayıcılar arası güvenirliği belirlemede farklı yöntemleri inceleyen araştırmalar da göze çarpmaktadır (Güler ve Teker, 2015; Goodwin, 2001). Ancak matematik eğitiminde önemli bir yere sahip olan çoklu gösterimler ile puanlayıcı güvenirliğini birlikte ele alan çalışmaya rastlanamamıştır.

Bu noktadan bakıldığında çoklu gösterimlerin kullanıldığı açık uçlu problemlerin öğrenci başarısını belirlemede güvenilir sonuçlar ortaya koyduğu görülmektedir. Ancak öğretmenlerin gösterimler türleri arasındaki katılık/cömertlikleri farklılaştığından (grafik gösteriminde), açık uçlu problemleri puanlamadan önce zümre öğretmenleriyle birlikte ölçütleri net olarak belirlenmesi yoluna gidilebilinir. Böylece puanlama güvenirliği yükselmiş olur. Bu bağlamda öğretmenlere sınıf ortamlarında öğrencilere farklı gösterim türlerini kullanmaya teşvik etmek ve bir gösterim türünden diğerine rahatlıkla geçiş yapabilmelerini sağlayacak öğrenme ortamları oluşturması önerilmektedir. Çünkü, öğrenme ortamlarında tek tip gösterim türü kullanmak, öğrencilerin farklı gösterim türlerini kullanmalarına engel teşkil edebilmektedir (Greeno & Hall, 1997). Bu nedenle farklı gösterim türlerinin öğrenme ortamlarında sunulmasıyla, öğrenciler bir gösterim türünden diğerine rahatlıkla geçiş yapabilmeyi de öğrenebilirler.

Bu çalışmada kullanılan bütüncül dereceli puanlama anahtarının yanı sıra genel izlenimle puanlama ve analitik puanlama anahtarı kullanılarak, çoklu gösterim içeren matematik problemlerinde puanlayıcı güvenirlikleri incelenebilir. Ayrıca daha genelleyici sonuçlar elde etmek için, farklı konularda çoklu gösterimler kullanılabilir. Benzer çalışmalarda farklı madde ve puanlayıcı sayıları ile Genellenebilirlik kuramında farklı desenler veya Çok Değişkenli Rasch Modeli de kullanılabilir.

Kaynakça

Aiken, L. R. (2000). *Psychological testing and assessment*. Boston: Allyn and Bacon.

Atmaz, G. (2009). *Puanlama yönergesi (rubrik) kullanılması durumunda puanlayıcı güvenirliğinin incelenmesi*. Yayınlanmamış Yüksek Lisans Tezi. Mersin Üniversitesi Sosyal Bilimler Enstitüsü.

Birgin, O., & Gürbüz, R. (2008). Sınıf öğretmeni adaylarının ölçme ve değerlendirme konusundaki bilgi düzeylerinin incelenmesi. *Selçuk Üniversitesi Sosyal Bilimler Enstitüsü Dergisi*, *20,* 163-179.

Brennan, R. L. (1992). Generalizability theory. *Educational Measurement: Issues and Practice, 11*(4), 27-34.

Brennan, L. R. (2001). *Generalizability theory, statistics for social science and public policy.* New York: Springer-Verlag.

Bresciani, M. J., Zelna C. L., & Anderson. J. A. (2004). *Assessing student learning and development: A handbook for practitioners.* Washington, DC: National Association of Student Personnel Administrators.

Büyükkıdık, S., & Anıl, D. (2015). Investigation of reliability in generalizability theory with different designs on performance based assessment. *Education and Science, 40*(177), 285–296.

Büyüköztürk, Ş., Çakmak, K.E., Akgün, E. Ö., Karadeniz, Ş., & Demirel, F. (2012). *Bilimsel araştırma yöntemleri*(11. Baskı). Ankara: Pegem Akademi Yayınları

Cardinet, J., Johnson, S., & Pini, G. (2010). *Applying generalizability theory using EduG.* New York, NY: Routledge.

Cleaves, W. P. (2008). Promoting mathematics accessibility through multiple representations jigsaws. *Mathematics Teaching in the Middle School, 13*(8), 446-452.

Cobb, P., Yackel, E., & Wood, T. (1992). A constructivist alternative to the representational view of mind in mathematics education. *Journal for Research in Mathematics Education, 23*(1), 2-33.

Crocker, L., & Algina, J. (2008). *Introduction to classical and modern test theory Mason*. OH: Cengage Learning.

Doğan, C. D., & Anadol, H. Ö. (2017). Genellenebilirlik kuramında tümüyle çaprazlanmış ve maddelerin puanlayıcılara yuvalandığı desenlerin karşılaştırılması. *Kastamonu Eğitim Fakültesi Dergisi*, *25*(1), 361-372.

Dörnyei, Z. (2007). *Research methods in applied linguistics*. New York: Oxford University Press.

Dufour-Janvier, B., Bednarz, N. & Belanger, M. (1987). Pedagogical considerations concerning the problem of representation. In C. Janvier (Ed.), *Problems of representations in the learning and teaching of mathematics* (pp. 109-123). Hillsdale, NJ: Lawrence Erlbaum Associates.

Erbilgin, E., (2003). *Effects of spatial visualization and achievement on students' use of multiple representations.* Unpublished Master Thesis, Florida State University.

Evans-Hampton, T. N., Skinner, C. H., Henington, C., Sims, S., & McDaniel, C. E. (2002). An investigation of situational bias: Conspicuous and covert timing during curriculum-based measurement of mathematics across African American and Caucasian students. *School Psychology Review*, *31*, 529–539.

Goldin, G., & Shteingold, N. (2001). Systems of representations and the development of mathematical concepts. In A. A. Cuoco, & F. R. Curcio (Eds.), *The roles of representation in school mathematics* (pp. 1-24). Reston: NCTM Publications.

Goodwin, L. D. (2001). Interrater agreement and reliability. *Measurement in Physical Education and Exercise Science, 5* (1), 13-14.

Gravetter, F.J. and Forzano, L.B. (2012) *Research Methods for the Behavioral Sciences (*4th edn), Wadsworth, Cengage Learning, Belmont, CA

Greeno J. G., & Hall R. P. (1997). Practicing representation: Learning with and about representational forms. http://www.pdkintl.org/kappan/k_v78/k9701gre.htm adresinden alınmıştır.

Güler, N. (2008). *Klasik test kuramı, genellenebilirlik kuramı ve Rasch modeli üzerine bir araştırma*. Yayınlanmamış doktora tezi. Hacettepe Üniversitesi, Ankara.

Güler, N., & Gelbal, S. (2010). Studying Reliability of Open Ended Mathematics Items According to the Classical Test Theory and Generalizability Theory. Educational Sciences: *Theory and Practice, 10(*2), 1011-1019.

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

Güler, N, Kaya-Uyanık, G., & Taşdelen-Teker, G. (2012). *Genellenebilirlik kuramı*. Pegem Akademi, Ankara, Türkiye

Güven, B., & Eskitürk, M. (2007). Sınıf Öğretmenlerinin Ölçme ve Değerlendirmede Kullandıkları Yöntem ve Teknikleri. *XVI. Eğitim Bilimleri Kongresi Bildiri Kitabı,* Cilt 3, (504-509), Ankara: Detay Yayıncılık.

Herbel-Eisenmann, B. A. (2002). Using student contributions and multiple representations to develop mathematical language. *Mathematics Teaching in the Middle School, 8*(2), 100-105.

Hitt, F. (1999). Representations and mathematical viualization. In F. Hitt, & M. Santos (Eds.), *Proceedings of the twenty-first annual meeting of the North American chapter of the third international group of Psychology of Mathematics Education,* (pp. 131-138). Mexico.

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, 31*(2), 346-368.

Kan, A. (2005). The effect of using grading scale and response key to (same) grader's reliability. *Eurasian Journal of Educational Research, 19*, 166–167.

Kaya, G. (2011). *Genellenebilirlik kuramının doldurma kavram haritası değerlendirme çalışmasına uygulanması.* Yayınlanmamış yüksek lisans tezi. Hacettepe Üniversitesi, Ankara.

Keller, B. A. & Hirsch, C. R. (1998). Student preferences for representations of functions. *International Journal in Mathematics Education Science Technology, 29*(1), 1-17.

Mertler, C. A. (2000). Designing scoring rubrics for your classroom. *Practical assessment, research, and evaluation, 7*(1), 25.

Moskal, B., M. (2000). Scoring rubrics: what, when, how? *Pratical Assessment, Research and Evaluation,8*(14).https://scholarworks.umass.edu/cgi/viewcontent.cgi?article=1086&context=pare adresinden alınmıştır.

Moskal, B. M., & Leydens, J. A. (2000). Scoring rubric development: Validity and reliability. *Practical assessment, research & evaluation*, *7*(10), 71-81.

NCTM (2000). *Principles and Standards for School Mathematics.* Reston, VA: NCTM Publications.

Nitko, A. J. (2001). *Educational assessment of students* (3rd ed.). Upper Saddle River, NJ: Merrill

Ömür, S., & Erkuş, A. (2013). Dereceli puanlama anahtarıyla, genel izlenimle ve ikili karşılaştırmalar yöntemiyle yapılan değerlendirmelerin karşılaştırılması. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi, 28*(2), 308-320.

Polat Demir, B. (2016). Vee diyagramından elde edilen puanların güvenirliğinin klasik test kuramı ve genellenebilirlik kuramına göre incelenmesi. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi, 7*(2), 419-431.

Schultz, J. E., & Waters, M. S. (2000). Why representations? *Mathematic teacher, 93*(6). 448-453.

Shavelson RJ, & Webb NM. (1991). *Generalizability theory: a primer*. Newbury Park, CA: Sage.

Stecker, P. M., & Fuchs, L. S. (2000). Effecting superior achievement using curriculum-based measurement: The importance of individual progress monitoring. *Learning Disabilities Research and Practice, 15*, 128–134.

Struyven, K., Dochy, F., & Janssens, S. (2005). Students' perceptions about evaluation and assessment in higher aducation: A review. *Assessment and Evaluation in Higher Education, 30*(4), 325-341.

Swartz, C. W., Hooper, S. R., Montgomery, J. W., Wakely, M. B., De Kruif, R. E., Reed, M.,

... & White, K. P. (1999). Using generalizability theory to estimate the reliability of writing scores derived from holistic and analytical scoring methods. *Educational and Psychological Measurement, 59*(3), 492-506.

Tekindal, S. (2000). Klasik Yazılı Sınavla ve Çok Sorulu Testle Elde Edilen Ölçümlerin Güvenirlik ve Geçerliği. *Pamukkale Üniversitesi Eğitim Fakültesi Dergisi, 8*(8), 38-46.

Thurber, R. S., Shinn, M. R., & Smolkowski, K. (2002). What is measured in mathematics tests? Construct validity of curriculum-based mathematics measures. *School Psychology Review, 31,* 498–513.

Yerushalmy, M. (1997). Designing representations: Reasoning about functions of two variables. *Journal for Research in Mathematics Education*, *28*(4), 431-466.

Yılmaz, F. N., & Başusta, B. (2015). Genellenebilirlik kuramıyla dikiş atma ve alma becerileri istasyonu güvenirliğinin değerlendirilmesi. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, *6*(1), 107-116.