Hacettepe Üniversitesi Eğitim Fakültesi Dergisi (H. U. Journal of Education) 32(1): 168-184 [2017]
doi: 10.16986/HUJE.2016023168

Uzaktan Eğitimde Çoktan Seçmeli Soruların Güçlük ve Ayırt Edicilik Değerlerinin Soru Türlerine Göre İncelenmesi*

Examining Difficulty and Discrimination Indices of Multiple Choice Questions according to Item Types in Distance Education

Serpil Koçdar**, Nejdet Karadağ***, Murat Doğan Şahin****, Abdulkadir Karadeniz*****

• ***Geliş Tarihi:*** *05.08.2016* • ***Kabul Tarihi:*** *24.11.2016* • ***Yayın Tarihi:*** *31.01.2017*

ÖZ: Bu araştırmanın amacı, Türkiye'de yükseköğretimde uzaktan eğitim bağlamında, çoktan seçmeli soruların güçlük ve ayırt edicilik değerlerinin soru türlerine göre farklılık gösterip göstermediğini belirlemek ve öğrencilerin soru türlerine ilişkin görüşlerini almaktır. Karma olarak desenlenen araştırmada hem nicel hem nitel veriler toplanmıştır. Nicel veriler için 905 soru üzerinde çalışılmış ve madde analizi raporlarından yararlanılmıştır. Nitel veriler ise öğrencilerin soruların güçlük değerlerine ilişkin görüşlerini ortaya koymak amacıyla 20 uzaktan öğrenciyle yapılan yarı-yapılandırılmış görüşmelerle gerçekleştirilmiştir. Araştırmada soruların olumlu soru kipindeki sorular, olumsuz soru kipindeki sorular, cevabı işlem gerektiren sorular ve K Tipi sorular olmak üzere 4 türde olduğu saptanmıştır. Araştırma sonuçlarına göre olumlu ve olumsuz soru kipindeki sorularla cevabı işlem gerektiren soruların madde güçlükleri arasında istatistiksel olarak manidar bir fark bulunurken, soruların ayırt edicilik değerleri arasında manidar bir fark bulunmamıştır. Öğrencilerin soruların güçlük değerlerine ilişkin görüşleri ve tercihlerinin farklılaştığı tespit edilmiştir.
Anahtar sözcükler: çoktan seçmeli sorular, güçlük değeri, ayırt edicilik değeri, soru türü, uzaktan eğitim

ABSTRACT: The purpose of this study is to determine whether difficulty and discrimination indices of multiple choice questions differ according to item types in Turkey in higher education in a distance education setting and to identify the opinions of students on item types. In this mixed study, both quantitative and qualitative data were collected. Quantitative data were collected from Item Analysis reports including 905 items whereas qualitative data were collected from 20 students via semi-structured interviews to reveal the opinions of students on difficulty levels of questions. It was identified that items were in 4 types which were positive, negative, problem-based and K-Type questions. As a result, there is a significant difference between the difficulty indices of positive questions and problem-based questions as well as negative questions and problem-based questions. On the other hand, no significant difference was found between the discrimination indices. In addition, it was found that opinions of students on the difficulty levels of questions showed variety.
Keywords: multiple choice questions, difficulty index, discrimination index, item type, distance education

1. GİRİŞ

Eğitimde ölçme ve değerlendirme, öğretim tasarımı sürecinin önemli bir öğesidir; öğrenme-öğretme süreçleri ile ilgili geribildirim sağlayarak tüm sürecin gözden geçirilmesini ve geliştirilmesini sağlar. Öğrenci başarısının ölçülmesinde çeşitli araçlar kullanılmaktadır. Bu araçların geleneksel ve tamamlayıcı ölçme araçları olmak üzere iki başlık altında toplandığı söylenebilir (Popham, 2003). Sık kullanılan geleneksel araçların çoktan seçmeli testler, doğru-yanlış testleri, eşleştirme testleri, klasik yazılı sınavlar, sözlü sınavlar ve ödevler; tamamlayıcı araçların ise portfolyo, rubrikler, tartışma forumları, kavram haritaları, proje, akran

* Bu çalışma Anadolu Üniversitesi Bilimsel Araştırma Projeleri Komisyonu tarafından kabul edilen 1406E308 nolu proje kapsamında desteklenmiştir.
** Yrd. Doç. Dr., Anadolu Üniversitesi, Açıköğretim Fakültesi, Uzaktan Öğretim Bölümü, Eskişehir-Türkiye, skocdar@anadolu.edu.tr
*** Yrd. Doç. Dr., Anadolu Üniversitesi, Açıköğretim Fakültesi, Yaygın Öğretim Bölümü, Eskişehir-Türkiye, nkaradag@anadolu.edu.tr
**** Anadolu Üniversitesi, Test Araştırma Birimi, Eskişehir-TÜRKİYE, muratdogansahin@gmail.com
***** Arş. Gör., Anadolu Üniversitesi, Açıköğretim Fakültesi, Uzaktan Öğretim Bölümü, Eskişehir-Türkiye, abdulkadirkaradeniz@anadolu.edu.tr

değerlendirme ve öz-değerlendirme olduğu görülmektedir. Son yıllarda yükseköğretimde alternatif ölçme araçlarına daha fazla ilgi gösterilmesiyle birlikte çoktan seçmeli testlerin gerek düzey belirleme gerekse seçme amaçlı olarak halen yaygın olarak kullanıldığı gözlenmektedir (Haladyna, 2004; Karadağ, 2014). Özellikle öğrenme etkinliklerinin fiziksel olarak farklı mekânlarda olan öğretici ve öğrenciler arasında çeşitli iletişim teknolojilerinin kullanılmasıyla gerçekleştirildiği uzaktan eğitimde, öğrencilerin bilgilerinin ölçülmesinde testler önemli bir role sahiptir (Lindler,1998; Puspitasari, 2010). Uzaktan eğitimde ölçme ve değerlendirme faaliyetleri yüz yüze eğitim ortamlarına göre sınırlılıklar içermektedir (Puspitasari, 2010). Öğrenci sayısının yüksek olduğu uzaktan eğitim uygulamalarında öğreticilerin, yüz yüze öğrenme ortamlarında olduğu gibi öğrencileri yazılı çalışmalarının yanında derse katılımlarına ve sordukları soruların kalitesine göre değerlendirme olanakları bulunmamaktadır. McIsaac ve Gunawardena (1996) kalabalık sınıflarda öğrencilerden fiziksel olarak ayrı olan öğreticilerin, onların performanslarını ölçme konusunda çok az seçeneğe sahip olduklarını belirtmektedir. Buna ek olarak, uzaktan öğrencilerin farklı çeşitlilikte öğrenme materyallerine sahip olmaları, farklı yaş ve meslek grupları içinde bulunmaları, programlara farklı amaçlarla kayıt yaptırmaları ve başarı için farklı ölçütlerinin olması ve benzeri nedenlerle uzaktan eğitimde öğrenci başarısının değerlendirilmesi, sorunlu bir süreç olarak nitelendirilmektedir (Thorpe, 1988). Kullanım kolaylığı sağlaması ve objektif sonuçlar sunması açısından çoktan seçmeli testler yaygın olarak kullanılmaktadır (Sanderson, 2010; Simonson, Smaldino, Albright ve Zvacek, 2012; Zhang, 2002). Çoktan seçmeli testler, farklı türde çoktan seçmeli sorulardan oluşmaktadır. Çoktan seçmeli soruların en çok kullanılan türleri; soru kökü olumlu veya olumsuz soru kipinde olan sorular, soru kökü eksik cümle olan sorular, tek bir doğru cevap istenen sorular, en doğru cevap istenen sorular, cevabı işlem gerektiren sorular, alternatif seçmeli, doğru-yanlış, çoklu doğru-yanlış, eşleştirme, karmaşık çoktan seçmeli (K Tipi) sorular, ortak seçenekli sorular ve ortak köklü sorular olarak sıralanabilir (Haladyna, Downing ve Rodriguez, 2004; Tekin, 1994).

Çoktan seçmeli testler, uygulandıktan sonra çeşitli yöntemlerle analiz edilmekte ve analiz sonuçlarından yararlanılarak testler geliştirilmektedir (Özçelik, 1989). Bu yöntemlerden biri de her bir test maddesi (sorusu) için yapılan madde puanları analizidir. Bu tür analizde madde güçlüğü (p) ve madde ayırt ediciliği (r) değerleri hesaplanmaktadır. Madde güçlüğü, bir maddeye verilen doğru cevap sayısının tüm cevaplayıcıların sayısına oranıdır. Bir maddeyi cevaplayıcılardan büyük bir kısmı doğru cevaplamışsa bu madde kolay bir maddedir. Böyle bir maddenin güçlüğü 1'e yakın olur. Bir maddenin ayırt ediciliği ise, yoklanan davranışa sahip olan cevaplayıcıları bu davranışa sahip olmayanlardan ayırma gücüdür ve -1 ile 1 arasında değer alır. Bu değer 1'e ne kadar yakınsa madde, testten yüksek puan alanlarla düşük puan alanları birbirlerinden o kadar iyi ayırmış demektir. Her bir maddenin analiz edilmesi, öğrencilerin ne öğrendikleri hakkında geribildirim verir ve kusurlu soruların belirlenerek düzeltilmesini sağlar. Başka bir deyişle, soruların iyi işleyip işlemediğini ortaya koyarak testlerin geçerlik ve güvenirliğinin artırılmasına katkıda bulunur (Özçelik, 1989).

Alanyazında soru türlerinin güçlük ve ayırt edicilik indeksleri üzerindeki etkisini inceleyen araştırmalarda farklı bağlamlarda farklı soru türlerinin yer aldığı ve farklı sonuçların ortaya çıktığı görülmektedir. Örneğin, genel olarak olumsuz soruların öğrenciler tarafından zor yanıtlandığı ve bu tür sorulardan mümkün olduğunca kaçınılması gerektiği (Haladyna, Downing ve Rodriguez, 2002; Tekin, 1994) ileri sürülse de olumlu ve olumsuz soruların güçlük ve ayırt edicilik değerleri konusunda yapılan araştırmalarda farklı bulgulara rastlanmıştır. Bazı araştırmalarda olumlu ve olumsuz soru türü arasında güçlük indeksi açısından fark bulunmazken (Downing, Dawson-Saunders, Case ve Powell, 1991; Rachor ve Gray, 1996; Tamir, 1993; Varughese ve Glencross, 1997) bazı araştırmalarda olumsuz soruların öğrenciler tarafından daha zor yanıtladığı (Johnstone 1983, Haladyna, Downing ve Rodriguez 2002; Haladyna 2004, Sanderson, 2010) belirtilmektedir. Buna karşılık, bir çalışmada Harasym, Price, Brant, Violato

ve Lorscheider (1992) olumsuz soruların daha az zor olduğuna dair bulgular elde etmiştir. Tamir (1993) ise, alt bilişsel düzeydeki sorularda fark bulamazken, üst bilişsel düzeyde yer alan olumsuz soruların daha zor olduğu bulgusuna ulaşmıştır. Bunlara ek olarak, Downing, Dawson-Saunders, Case ve Powell (1991) ile Rachor ve Gray (1996), olumlu ve olumsuz sorularda ayırt edicilik indeksi açısından bir fark olmadığı sonucuna ulaşmıştır. Sanderson (2010), Güney Afrika Açık Üniveristesi'nde uzaktan verilen İngilizce Dilbilim dersinde 160 soru üzerinde yaptığı çalışmasında olumsuz soruların daha zor, ancak daha ayırt edici olduğu bulgusuna ulaşmıştır. Nnodim (1992), İnsan Anatomisi testinde K Tipi soruların klasik çoktan seçmeli sorulardan daha zor olduğunu ancak ayırt edicilik açısından farklı olmadığını; benzer şekilde Albanese (1993) K Tipi soruların iyi çalışmadığını belirtmiştir. Kaptan (1985), fizik konusunda kök maddesi uzun ve kısa olan soruların yanı sıra, madde kökünün sadece sözle, şekil ve sözle ve de tablo, grafik ve sözle ifade edildiği 257 test maddesini incelemiştir. Kök maddesi uzun ve kısa olan soruların güçlük ve ayırt edicilikleri arasında istatistiksel olarak manidar bir fark olmadığı görülmüştür. Diğer yandan, soru kökünün sadece sözle, şekil ve sözle ve de tablo, grafik ve sözle oluşturduğu üç farklı soru türü, güçlük ve ayırt edicilik açısından ikili olarak t–testi kullanılarak karşılaştırılmıştır. Elde edilen sonuçlar, bu üç soru türünün güçlükleri ve ayırt edicilikleri arasında istatistiksel olarak manidar farklar bulunduğunu ortaya koymuştur. Öte yandan, bu farklılığın hangi yönde olduğu bilinmemektedir. Kaya (1991), araştırmasında Türkçe dersinde eksik köklü, ortak seçenekli, ortak köklü soru türlerinin soru ve test istatistiklerine etkisini incelemiştir. Elde edilen sonuçlara göre; eksik köklü ve ortak seçenekli soruların güçlükleri arasında manidar bir fark olmadığı, ayırt edicilikleri ve soru güvenirlikleri arasındaki manidar farkın ortak seçenekli sorular lehine olduğu, test güvenirlikleri arasındaki manidar farkın ise eksik köklü sorular lehine olduğu görülmüştür. Eksik köklü ve ortak köklü sorular arasında yapılan karşılaştırmada ise, soru güçlükleri ve ayırt edicilikleri arasında istatistiksel olarak manidar bir fark bulunmamıştır. Madde güvenirlikleri ve test güvenirlikleri arasındaki manidar farkın ise eksik köklü sorular lehine olduğu görülmüştür. Buna ek olarak, ortak seçenekli ve ortak köklü soruların, madde ve test parametreleri arasında yapılan karşılaştırmada madde güçlükleri, madde ayırt edicilikleri, madde güvenirlikleri ve test güvenirlikleri arasında istatistiksel olarak manidar bir fark olmadığı görülmüştür. Bir diğer araştırmada Karaca (2004), ortaokul düzeyindeki sınavlarda seçme gerektiren, kısa cevaplı ve doğru-yanlış test sorularının madde güçlükleri ve ayırıcılık gücü indeksleri arasında manidar farklılıkların olup olmadığını araştırmıştır. Çalışmada kısa cevaplı teste ait ortalama madde güçlüğünün, doğru-yanlış ve seçme gerektiren teste ait ortalama madde güçlüklerinden daha küçük olduğu; doğru-yanlış testinin seçme gerektiren ve kısa cevaplı testten daha çok ayırıcı soru içerdiği tespit edilmiştir.

Alanyazındaki araştırmaların az sayıda soru ile sınırlı olduğu, konularının birbirinden farklılık gösterdiği ve Sanderson (2010) tarafından yapılan çalışma dışında tüm çalışmaların örgün eğitim bağlamında gerçekleştirildiği görülmektedir. Buna ek olarak, Türkiye'de çoktan seçmeli testler yaygın olarak kullanılmasına rağmen bu tür çalışmaların çok az sayıda olduğu görülmektedir. Alanyazındaki çalışmalardan farklı olarak bu araştırmanın amacı yükseköğretimde uzaktan eğitim bağlamında, öğrenci sayısının en fazla olduğu İşletme bölümü sınavlarında kullanılan çoktan seçmeli soruların güçlük ve ayırt edicilik değerlerinin soru türlerine göre farklılık gösterip göstermediğini belirlemek ve öğrencilerin soru türlerine ilişkin görüşlerini almaktır. Bu kapsamda, aşağıdaki sorulara yanıt aranmıştır:

1. İşletme Bölümü sınavlarında yer alan çoktan seçmeli soruların güçlük (p) değerleri soru türlerine göre farklılık göstermekte midir?
2. İşletme Bölümü sınavlarında yer alan çoktan seçmeli soruların ayırt edicilik (r) değerleri soru türlerine göre farklılık göstermekte midir?
3. İşletme Bölümünde kayıtlı öğrencilerin sınavlarda sorulan farklı soru türlerine ilişkin görüşleri nelerdir?

2. YÖNTEM

2.1. Araştırma Modeli

Araştırma, uzaktan eğitim bağlamında sınavlarda kullanılan çoktan seçmeli soruların güçlük ve ayırt edicilik değerlerinin soru türlerine göre incelenmesini ve öğrenci görüşlerinin alınmasını içerdiğinden, var olan durumu ortaya koymaya yönelik karma bir çalışmadır. Karma araştırmalar, ortaya koyulan sorunu daha iyi anlayabilmek amacıyla nicel ve nitel yaklaşımların bir arada kullanıldığı, araştırmanın bazı aşamalarında hem nicel hem de nitel verilerin toplandığı, analiz edildiği veya bütünleştirildiği çalışmalardır (Creswell, 2008; Tashakkori ve Teddlie 2003). Karma araştırmaların temel ilkesi, bir yöntemin zayıf yönlerinin diğer yöntemin kullanımı ile en aza indirilmesidir. Bu araştırmada da nicel verilere tamamlayıcı olarak öğrencilerin soruların güçlük değerleri konusundaki görüşleri alınmıştır.

2.2. Çalışma Grubu ve Katılımcılar

Araştırmanın çalışma grubunu, Türkiye'de bir devlet üniversitesinde uzaktan eğitimle İşletme Bölümü'nde okutulan 11 alan dersinde 2011-2012 Güz ve Bahar dönemi ara, yılsonu ve bütünleme sınavlarında kullanılan 905 soru oluşturmaktadır. İşletme Bölümü sorularının seçilmesinin nedeni, öğrenci sayısının en fazla olduğu bölüm olmasıdır.

Araştırmanın katılımcılarını İşletme Bölümü'nde öğrenim gören 20 gönüllü öğrenci oluşturmaktadır. Öğrenciler kolay ulaşılabilir durum örneklemesi yöntemiyle seçilmiştir. Öğrencilerin demografik özellikleri Tablo 1'de verilmiştir. Kimlik bilgilerinin gizli tutulması açısından öğrenciler Ö1, Ö2, Ö3, … olarak adlandırılmıştır.

Tablo 1: Görüşmeye katılan öğrencilerin demografik bilgileri

Öğrenciler	Cinsiyeti	Yaşı	Yaşadığı İl	İş durumu
Ö1	Kadın	40	Eskişehir	Çalışıyor
Ö2	Kadın	22	Eskişehir	Çalışmıyor
Ö3	Erkek	23	Eskişehir	Çalışmıyor
Ö4	Kadın	22	Eskişehir	Çalışıyor
Ö5	Erkek	27	Eskişehir	Çalışmıyor
Ö6	Erkek	30	Eskişehir	Çalışıyor
Ö7	Erkek	23	Eskişehir	Çalışmıyor
Ö8	Kadın	38	Eskişehir	Çalışıyor
Ö9	Kadın	21	Eskişehir	Çalışıyor
Ö10	Kadın	23	Eskişehir	Çalışıyor
Ö11	Kadın	22	Eskişehir	Çalışmıyor
Ö12	Kadın	29	Eskişehir	Çalışmıyor
Ö13	Erkek	27	Eskişehir	Çalışıyor
Ö14	Kadın	40	Ankara	Çalışıyor
Ö15	Kadın	27	İstanbul	Çalışmıyor
Ö16	Kadın	21	Uşak	Çalışıyor
Ö17	Kadın	32	Yalova	Çalışmıyor
Ö18	Erkek	26	Kahramanmaraş	Çalışıyor
Ö19	Kadın	27	Eskişehir	Çalışmıyor
Ö20	Erkek	35	Eskişehir	Çalışıyor

2.3. Veri Toplama Araçları

Araştırmada çoktan seçmeli soruların güçlük ve ayırt edicilik değerlerinin soru türlerine göre değişip değişmediğini belirlemek için nicel veriler, öğrencilerin görüşlerinin belirlenmesi için nitel veriler toplanmıştır.

2.3.1. Nicel veri toplama araçları

Araştırmada kullanılan 902 soruya ilişkin p ve r değerlerinin tespit edilmesi amacıyla her bir ders için hazırlanan madde analizi dökümlerinden yararlanılmıştır. Madde analizi dökümleri her sınavdan sonra Bilgisayar Araştırma ve Uygulama Merkezi Müdürlüğü tarafından bilgisayar ortamında hazırlanmaktadır. Madde analizi, bir testten alınan puanların büyükten küçüğe doğru sıralandıktan sonra yüksek puan alan %27'lik grup ile düşük puan alan %27'lik grubun her bir soruya verdikleri cevapların karşılaştırılmasıyla hesaplanır. Madde analizlerinin elde edilmesi için öncelikle sorulara verilen yanıtlar puanlanır; her öğrencinin testin tümünden ve varsa bölümlerinden kaç doğru yanıtı olduğu sayılır. Doğru yanıt sayısı puan olarak alınır. Puanlama bittikten sonra eldeki cevap kâğıtları, puanı en yüksek olandan en düşük olana doğru sıralanır. En üstte en yüksek puanlı cevap kâğıdı bulunacak şekilde konur. Daha sonra, puan sırasında en üst ve en alt %27'lik bölümler içinde kalanların cevapları analiz edilir (Özçelik, 1989). Araştırmada madde analizi raporlarının kullanıldığı sınavlarda en az 1.998, en fazla 71.210 öğrenci yer almaktadır.

2.3.2. Nitel veri toplama araçları

Nitel veriler öğrencilerle yapılan yarı yapılandırılmış görüşmelerle toplanmıştır. Görüşme soruları oluşturulduktan sonra 3 uzmanın görüşüne sunularak gerekli düzeltmeler yapılmıştır.

2.4. Verilerin Toplanması

2.4.1. Nicel verilerin toplanması

Soruların hangi kategoride yer aldığı 3 ölçme ve değerlendirme uzmanı tarafından kodlanarak belirlenmiş ve kodlama güvenirliği hesaplanmıştır. Uzman 1, ölçme-değerlendirme alanında yüksek lisans yapmış olup aynı alanda doktora eğitimine devam etmektedir ve 5 yıldır bir devlet üniversitesinin ölçme ve değerlendirme biriminde çalışmaktadır. Uzman 2, uzaktan eğitim alanında doktora derecesine sahip olup, uzaktan eğitimde ölçme ve değerlendirme konusunda bir doktora tezi hazırlamıştır. Buna ek olarak, 19 yıldır bir devlet üniversitesinin ölçme ve değerlendirme biriminde çalışmaktadır. Uzman 3, uzaktan eğitimde doktora derecesine sahip olup, 12 yıl bir devlet üniversitesinin ölçme ve değerlendirme biriminde görev yapmıştır. Soru türlerinin kodlanmasında kodlayıcılar arasında 905 sorudan 9'unda görüş ayrılığı yaşanmıştır. Buna göre kodlama güvenirliği Miles ve Huberman'ın (1994) önerdiği formüle göre hesaplanmış ve %99 olarak bulunmuştur. Yıldırım ve Şimşek (2008), kodlama güvenirliğinin en az %70 olması gerektiğini belirtmektedir. Kullanılan formül aşağıda verilmiştir:

Güvenirlik= görüş birliği sayısı / toplam görüş birliği sayısı + görüş ayrılığı sayısı

Görüş ayrılığı yaşanan 9 soru kodlayıcılar tarafından birlikte yeniden gözden geçirilmiş ve soruların hangi kategoriye girdiği konusunda uzlaşmaya varılmıştır. İşlemli ve sıralı soru türünde olup aynı zamanda olumsuz olan sorular, işlemli ve sıralı soru türü kategorisinde değerlendirilmiştir.

Buna göre, soru türlerinin olumlu soru kipindeki sorular, olumsuz soru kipindeki sorular, cevabı işlem gerektiren sorular ve K Tipi sorular (sıralı sorular olarak da adlandırılmaktadır) olmak üzere 4 kategoride olduğu tespit edilmiştir. Soruların türlerine göre dağılımı Tablo 2'de verilmiştir.

Tablo 2: Soruların soru türüne göre dağılımı

Soru Türleri	Sayı	Yüzde
Olumlu	556	61,4
Olumsuz	251	27,7
İşlemli	59	6,5
Birleşik	39	4,3
Toplam	**905**	**100,0**

Soruların türleri belirlendikten sonra, her bir sorunun p ve r değeri madde analizi dökümlerinden bulunarak analiz edilmek üzere tablo haline getirilmiştir.

2.4.2. Nitel verilerin toplanması

Yarı yapılandırılmış bireysel görüşmeler için İşletme Bölümü öğrencilerine telefon ve sosyal medya aracılığıyla ulaşılarak araştırmanın konusu ve kapsamıyla ilgili bilgi verilmiştir. Katılımcılara, kimliklerinin gizli tutulacağı ve üçüncü kişilerle paylaşılmayacağı belirtilmiştir. Araştırmaya gönüllü olarak katılmak isteyen öğrencilerle belirlenen tarihte çevrimiçi olarak Skype veya telefon aracılığıyla görüşmeler gerçekleştirilmiştir. Görüşmenin sesli olarak kaydedilmesi konusunda öğrencilerden izin alınmıştır.

2.5. Verilerin Analizi

2.5.1. Nicel verilerin analizi

Araştırmada, çoktan seçmeli soruların güçlük ve ayırt edicilik değerlerinin soru tipine göre farklılaşıp farklılaşmadığını incelemek amacıyla Tek Yönlü MANOVA kullanılmıştır. Tek Yönlü MANOVA, bağımlı değişken sayısı birden fazlayken, birden fazla kategori sayısına sahip tek bir bağımsız değişkenin olduğu durumlarda kullanılan çok değişkenli bir istatistik analiz yöntemidir (Tabachnick and Fidell, 2007). Tek Yönlü MANOVA sonucu elde edilen farkın manidar çıkması halinde ise bu farkın bağımlı değişken setindeki değişkenlerin hangisinden kaynaklandığını belirlemek amacıyla Tek Yönlü ANOVA uygulanmıştır. Tek Yönlü ANOVA sonucunda, varyansların homojenliği sayıltısının sağlandığı durumlarda, ANOVA tablolarına göre manidar farkın görülmesi durumunda Scheffe; varyansların homojenliği sayıltısının sağlanmadığı durumlarda Brown-Forsythe ve Welch test sonuçlarına bakılmış, manidar sonuçlara ulaşıldığı durumlarda ise Tamhane's T2 testleri ile ikili karşılaştırmalar yapılmıştır. Söz konusu analizlerin tamamı, SPSS 22.0 paket programı kullanılarak gerçekleştirilmiştir.

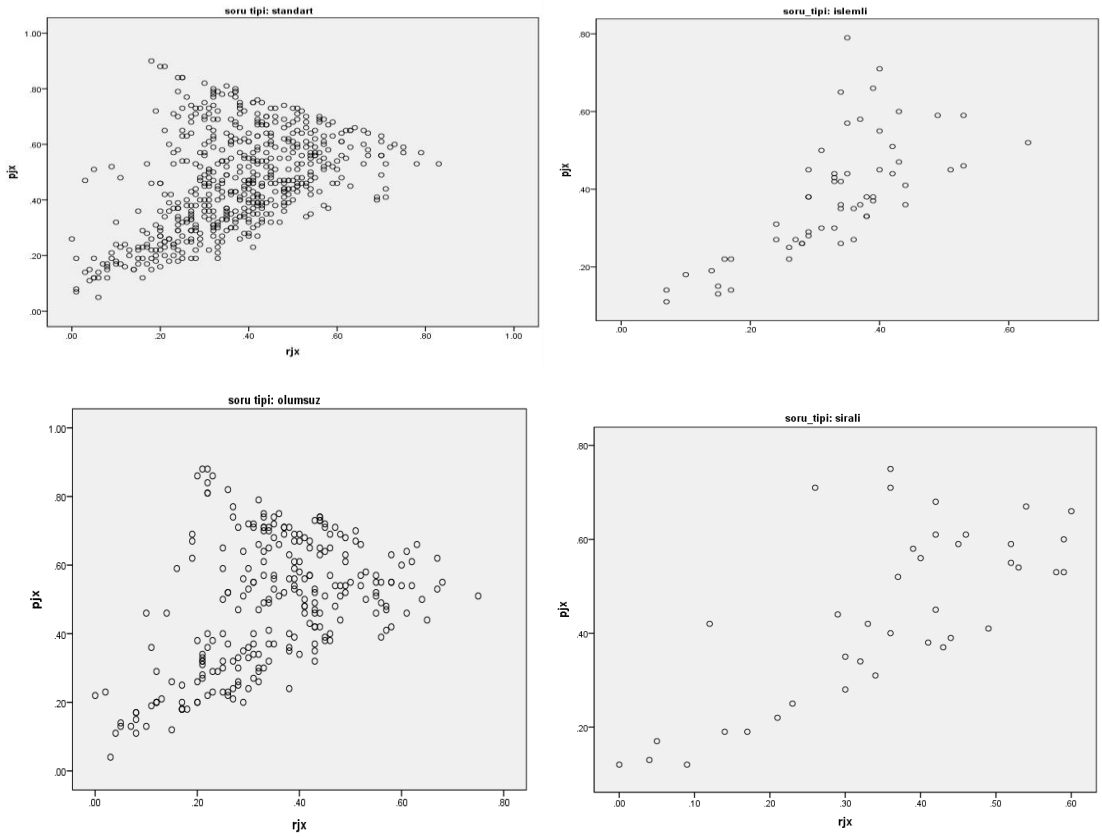
Soruların güçlük ve ayırt edicilik değerlerinin, soru türüne göre farklılık gösterip göstermediğini Tek Yönlü MANOVA ile test etmek için, öncelikle MANOVA uygulaması için gerekli olan varsayımların test edilmesi gerekmektedir. MANOVA, bağımlı değişkenler arasındaki ilişkilerin dikkate alınması (Field, 2005) ve yapılan analizlerin I. tip hatadan arınıklığının sağlanması (Bray ve Maxwell, 1982; Stevens, 2009; Stangor, 2010) avantajlarının yanı sıra birçok varsayımı da beraberinde getirmektedir. Tek ve çok değişkenli normallik, uç değerler, doğrusallık, çoklu doğrusal bağıntı ve tekillik, varyans-kovaryans matrisinin homojenliği varsayımlarının sağlanması, MANOVA'nın uygulanabilmesi için ön koşuldur (Pallant, 2005). Bu nedenle, Tek Yönlü MANOVA analizlerine geçilmeden önce varsayımların karşılanıp karşılanmadığı test edilmiştir.

2.5.1.1. Varsayımların test edilmesi

İlk olarak, bağımlı değişkenlerin tek değişkenli normallik varsayımını karşılama durumlarına bakılmak üzere uygulanan Kolmogorov-Smirnov (K-S) testi sonuçlarının istatistiksel olarak manidar çıktığı görülmüştür ($p<.01$). Oysa ki, verilerin normal dağıldığının kabul edilebilmesi için K-S sonuçlarının manidar çıkmaması gerekmektedir. Bununla birlikte, araştırmadaki veri sayısının çok olması durumunda normallikten çok küçük sapmaların bile manidar çıktığı bilinmektedir (Çokluk, Şekercioğlu ve Büyüköztürk, 2012). Verilere ait çarpıklık katsayısının +- 1 aralığında olmasının puanların normalden kayda değer bir sapma göstermediği şeklinde yorumlanabileceği de (Büyüköztürk, 2015) göz önüne alınarak, son kararı vermek için bağımlı değişkenlere ait çarpıklık katsayıları incelenmiştir. Elde edilen çarpıklık değerleri, soruların güçlüğü ve ayırt ediciliği için sırasıyla -.019 ve .054 olarak bulunmuştur. Dolayısıyla bağımlı değişkenlerin tek değişkenli normallik şartını karşıladığı sonucuna ulaşılmıştır.

Tek değişkenli normallik şartının sağlanmasının ardından, araştırma verilerinin çok değişkenli normallik varsayımını karşılayıp karşılamadığını test etmek üzere Mahalanobis uzaklık değerleri hesaplanmıştır. Pearson ve Hartley (1958), iki bağımsız değişkenin bulunduğu birçok değişkenli analiz verisinde, Mahalanobis uzaklığı için kritik değerin 13.82 olduğunu belirtmişlerdir. Söz konusu kritik değerin üzerindeki Mahalanobis değerleri uç değer olarak kabul edilmektedir (Pallant, 2005). Araştırmada, elde edilen Mahalanobis değerlerinden üçünün (14.77, 14.12 ve 13.95) kritik değer olan 13.82'yi aştığı görülmüştür. Söz konusu bu uç değerlerin, araştırma sonuçlarını olumsuz etkileyecekleri düşünülerek silinmelerine karar verilmiş ve araştırmaya 902 çoktan seçmeli soruya ait veri setiyle devam edilmiştir.

MANOVA uygulamasına geçilmeden önce incelenecek diğer bir varsayım ise, bağımlı değişkenler arasında doğrusal bir ilişkinin bulunup bulunmadığının belirlenmesidir. Öyle ki, bağımlı değişkenlerin ikili tüm kombinasyonları arasında doğrusal bir ilişkinin bulunması gerekmektedir. Araştırmanın bağımsız değişkeni olan soru türünün (olumlu soru kipindeki sorular, olumsuz soru kipindeki sorular, cevabı işlem gerektiren sorular, K Tipi sorular) tüm alt kategorilerine göre bağımlı değişkenlerin (madde güçlük ve madde ayırt edicilikleri) tüm ikili kombinasyonlarının doğrusallığına dair elde edilen grafikler Grafik 1'de verilmiştir.



Grafik 1. Soru türüne göre bağımlı değişkenler arasındaki doğrusal ilişki

Yapılan analizler sonucu elde edilen grafikler, bağımlı değişkenler arasında doğrusallık şartını ihlal edecek herhangi bir durum olmadığını göstermiştir.

MANOVA, bağımlı değişkenler arasında ancak orta derecede bir korelasyon olduğunda en iyi sonuçları vermektedir. Korelasyonun düşük olması durumunda ise tek değişkenli varyans analizinin uygulanması önerilmektedir. Bağımlı değişkenlerin arasında .80 veya .90'ın üzerinde korelasyon olması ise çoklu doğrusal bağıntı olması anlamına gelmekte ve MANOVA'da sorun yaşanmasına neden olmaktadır. (Pallant, 2005). Yapılan korelasyon analizi sonucu bağımlı değişkenler arasında .49'luk orta derecede bir ilişki olduğu görülmüş, çoklu doğrusal bağıntı sorununun olmadığı sonucuna ulaşılmıştır.

MANOVA'nın uygulanabilmesi için, varyans-kovaryans matrislerinin homojenliği şartının da yerine getirilmesi gerekmektedir. Buna göre, elde edilen Box's M testi sonuçlarının manidar olması, söz konusu varsayımın karşılanmadığı anlamına gelmektedir. Yapılan analiz sonucu, soru türleri ve soruların bilişsel düzeyine göre yapılan Box's M testleri manidar bulunmuş, dolayısıyla varyans-kovaryans matrislerinin eşitliği varsayımının ihlal edildiği görülmüştür. Değişkenlerdeki kategorilerde gözlem sayılarının eşit olmadığı ve Box's M testinin .001 düzeyinde manidar olduğu durumlarda sağlamlık (robustness) sağlanamaz. Ancak, Box's M testinin manidar çıkmasıyla varsayımlarda biri ihlal edilmiş gibi görünse de, bu testin geniş örneklem gruplarındaki çok küçük değişiklikler nedeniyle bile manidar çıkabileceği göz ardı edilmemelidir (Tabachnick ve Fidell, 2007). Böyle durumlarda MANOVA'da genel olarak yorumlanan Wilk's Lambda yerine değerlendirme ölçütü olarak Pillai's Trace'nin kullanılması

uygun olmaktadır. Dolayısıyla, MANOVA'nın uygulanmasına engel olabilecek herhangi bir durum söz konusu olmadığı sonucuna ulaşılmıştır.

Elde edilen sonuçlar, soruların güçlük ve ayırt edicilik değerlerinden oluşan bağımlı değişken setinin, soru türü bağımsız değişkenine göre test edilmesinde kullanılacak Tek Yönlü MANOVA'nın uygulanabilmesi için tüm şartları sağladığını göstermektedir.

2.5.2. Nitel verilerin analizi

Öğrencilerle yapılan görüşmeler sesli olarak kaydedilmiş ve deşifre edildikten sonra veriler betimsel analiz yöntemi ile çözümlenmiştir. Yıldırım ve Şimşek'e (2008) göre betimsel analiz, içerik analizine göre daha yüzeysel olup, daha çok araştırmanın kavramsal yapısının önceden açık biçimde belirlendiği araştırmalarda kullanılmaktadır. Veriler, araştırma sorularının ortaya koyduğu temalara veya görüşme ve gözlem süreçlerinde kullanılan sorulara göre düzenlenebilmektedir. Başka bir deyişle, veriler daha önceden belirlenen temalara göre özetlenerek yorumlanmaktadır. Bu anlamda araştırma soruları bağlamında betimsel analizin çalışma için yeterli olduğu düşünülmüştür. Elde edilen verilerin analizinde kodlama güvenirliğinin sağlanması için 2 araştırmacının kodlama yapması sağlanmıştır. Kodlama güvenirliği %89 olarak hesaplanmıştır. Farklı kodlanan ifadeler üzerinde tartışılarak kodlamada görüş birliği sağlanmıştır.

3. BULGULAR

Araştırma kapsamında öncelikle madde güçlüğü ve madde ayırt ediciliği değerlerinin soru türü bağımsız değişkenine göre betimsel istatistikleri elde edilmiştir. Söz konusu istatistikler Tablo 3'te görülmektedir.

Tablo 3: Bağımsız değişkene göre kategorilenen soruların betimsel istatistikleri

Bağımsız değişken	Kategori[1]	Madde sayısı	Bağımlı değişken	Ortalama	Std. sapma
Soru türü	Olumlu	554	p_{jx}	.467	.175
			r_{jx}	.372	.151
	Olumsuz	250	p_{jx}	.484	.187
			r_{jx}	.349	.144
	İşlemli	59	p_{jx}	.381	.153
			r_{jx}	.329	.113
	Birleşik	39	p_{jx}	.445	.182
			r_{jx}	.355	.162
Toplam		**902**			

Araştırmanın bağımlı değişkenleri olan madde güçlüğü (p_{jx}) ve ayırt ediciliği (r_{jx}) değerlerinin ortalama ve standart sapmaları, araştırmanın bağımsız değişkeni olan soru türünün tüm kategorilerine göre hesaplanmış ve bulgular Tablo 3'te sunulmuştur. Buna göre, 554 olumlu soru kipindeki soruların, madde güçlüğü ortalaması .467 ve ayırt edicilik gücü ortalaması .372 bulunmuştur. Bu parametrelerin standart sapmaları ise sırasıyla .175 ve.151'dir. Olumsuz soru kipindeki soruların sayısının 250 olduğu görülmektedir. Bu soruların madde güçlüğü ortalaması .484, ayırt edicilik gücü ortalaması ise .349'dur. Standart sapmalarının ise sırasıyla .187 ve .144 olduğu görülmektedir. Cevabı işlem gerektiren madde sayısının ise 59 olduğu görülmektedir. Bu soruların madde güçlüğü ve ayırt edicilik gücü ortalamaları sırasıyla .381 ve .329 olarak hesaplanmış, standart sapmalarının ise yine sırasıyla .153 ve .113 olduğu görülmüştür. Madde soru türüne göre gruplandığında son alt kategori olan K Tipi soruların sayısı 39'dur. Bu soruların

[1] Tabloya sığmadığı için bağımsız değişkenin kategorileri kısaltılarak yazılmıştır.

madde güçlüğü ortalaması .445, ayırt edicilik gücü ortalaması ise .355'tir. Bu soruların standart sapmaları ise sırasıyla .182 ve .162 olarak hesaplanmıştır. Tüm bu veriler birlikte değerlendirildiğinde, soru türüne göre madde güçlük indeksi en yüksek olan (en kolay) soruların olumsuz soru kipindeki sorular olduğu, ayırt edicilik gücü en yüksek olan soruların ise olumlu soru kipindeki sorular olduğu görülmektedir.

Betimsel istatistiklerin değerlendirilmesinin ardından, araştırma amaçları doğrultusunda madde güçlük ve madde ayırt edicilik gücü değişkenlerinin soru türü bağımsız değişkenine göre manidar fark gösterip göstermediğini test etmek üzere Tek Yönlü MANOVA yapılmıştır.

Varyans-kovaryans matrislerinin homojenliğini test etmek için yapılan Box's M testinin manidar çıkmasının ardından Wilk's Lambda yerine hesaplanan Pillai's Trace değeri Tablo 4'te verilmiştir.

Tablo 4: Soru türlerine göre MANOVA testi sonuçları

Test	F	Manidarlık
Pillai's Trace	1324.711	.000*

p<.001

Araştırma sonucu elde edilen bulgulara göre, Pillai's Trace katsayısı istatistiksel olarak manidar bulunmuştur (p<.01). Buna göre, araştırmanın bağımlı değişkenleri olan test sorularının güçlük ve ayırt edicilik değerlerinin oluşturduğu doğrusal bileşen, bağımsız değişken olan soru türünün en az iki kategorisi arasında manidar şekilde farklılık göstermektedir. Söz konusu farklılığın hangi kategoriler arasında olduğunu tespit etmek amacıyla, bağımlı değişkenlere ayrı ayrı Tek Yönlü ANOVA uygulanmalıdır. Tek Yönlü ANOVA uygulanabilmesi için, varyansların homojenliği sayıltısının yerine getirilmesi gerektiğinden, bu amaçla yapılan Levene Testi sonuçları ve yorumu Tablo 5'te verilmiştir.

Tablo 5: Soru türü bağımsız değişkeni için Levene testi sonuçları

	F	Sd	Manidarlık
Madde güçlüğü	2.275	2	.078
Madde ayırt ediciliği	3.061	2	.027*

*p<.05

Varyansların homojenliği sayıltısının karşılanması için, Levene testinin .05 düzeyinde manidar olmaması gerekmektedir. Levene testi sonucunda elde edilen sonuçlara göre madde güçlüğü değişkeni için soru türü açısından varyansların homojenliği sağlanmışken, madde ayırt edicilik değişkeni için varyans homojenliği şartının sağlanmadığı görülmektedir. Bu nedenle, varyansların homojenliği şartını sağladığı için madde güçlüğü değişkeninin soru türüne göre Tek Yönlü ANOVA analizlerinin manidarlığı için ANOVA tablosu, varyans homojenliği sayıltısının karşılanmadığı madde ayırt edicilik gücü değişkeninin soru türüne göre Tek Yönlü ANOVA analizlerinin manidarlığı içinse Brown-Forsythe ve Welch testi sonuçlarına bakılmıştır. Elde edilen değerler Tablo 6 ve 7'de verilmiştir.

Tablo 6: Madde güçlüğü için ANOVA sonuçları

	Sd	F	Manidarlık
Gruplar arası	2	5.623	.001*

*p<.025

Tablo 7: Madde ayırt ediciliği için Welch ve B-F sonuçları

	Sd	Manidarlık
Welch	2	.026
Brown-Forsythe	2	.041

*p<.025

Madde güçlüğü ve madde ayırt ediciliği bağımlı değişkenleri için, aynı bağımsız değişkenin etkisini inceleyen birbirinden ayrı iki analizin söz konusu olduğu görülmektedir. Bu gibi durumlarda I. tip hatayı önlemek için Bonferroni düzeltmesi yapılmalıdır (Pallant, 2005). En kolay anlatımıyla Bonferroni düzeltmesi, alfanın (genelde kullanılan değer .05) aynı bağımsız değişkenle yapılacak analiz sayısına, bir diğer ifadeyle bağımlı değişken sayısına bölünmesiyle bulunur (Tabachnick ve Fidell, 2007). Bu çalışmada bağımlı değişken sayısı iki olduğundan, .05 alfa değeri 2'ye bölünerek yeni alfa değeri .025 olarak elde edilmiştir (.05/2=.025). Bu düzeltmenin yapılmasıyla elde edilen yeni manidarlık değeri göz önüne alınarak yapılan analizlere göre, madde güçlüğü değişkeninin soru türüne göre değişip değişmediğini belirlemek için yapılan ANOVA sonuçları istatistiksel olarak manidardır (p=.001<.025). Madde ayırt ediciliği için uygulanan Welch ve Brown-Forsythe test sonuçlarının ise, elde edilen manidarlık düzeyine göre (.025) manidar olmadıkları görülmektedir (p>.025). Dolayısıyla, bu adımdan sonra sadece manidar değerler veren madde güçlüğü değişkeni için Scheffe testi yapılarak, manidar farklılığın bağımsız değişkendeki hangi düzeylerden kaynaklandığı araştırılmıştır. Scheffe testi sonucunda elde edilen değerler Tablo 8'de verilmiştir.

Tablo 8: Madde Güçlüğü için Soru Türüne göre Scheffe Çoklu Karşılaştırma Testi Sonuçları

Soru türü (I)	Soru türü (J)	Ortalamalar farkı (I – J)	Standart hata	Manidarlık
Olumlu	**Olumsuz**	-.018	.014	,636
	İşlemli	.086	.024	,006*
	Birleşik	.022	.029	,906
Olumsuz	**Olumlu**	.018	.014	,636
	İşlemli	.104	.026	,001*
	Birleşik	.040	.031	,641
İşlemli	**Olumlu**	-.086	.024	,006*
	Olumsuz	-.104	.026	,001*
	Birleşik	-.064	.037	,383
Birleşik	**Olumlu**	-.022	.029	,906
	Olumsuz	-.040	.031	,641
	İşlemli	.064	.037	,383

Scheffe çoklu karşılaştırma testinden elde edilen sonuçlara göre, cevabı işlem gerektiren sorularla olumlu soru kipindeki soruların madde güçlükleri arasında istatistiksel olarak manidar bir fark bulunmuştur (p=.006<.05). Buna göre; olumlu soru kipindeki soruların madde güçlüğü, cevabı işlem gerektiren soruların güçlüğüne göre manidar şekilde yüksektir. Yine cevabı işlem

gerektiren sorularla olumsuz soru kipindeki soruların güçlükleri arasındaki fark da istatistiksel olarak manidardır (p=.001<.05). Buna göre; olumsuz soru kipindeki soruların madde güçlüğü, cevabı işlem gerektiren soruların güçlüğüne göre manidar şekilde daha yüksektir. K Tipi soruların madde güçlükleri diğer hiçbir soru türünün madde güçlük değerleriyle manidar bir farklılık göstermezken (p>.05), olumlu soru kipindeki sorularla ve olumsuz soru kipindeki soruların madde güçlükleri arasındaki fark da istatistiksel olarak manidar bulunmamıştır (p=.636>.05). Araştırmada elde edilen nicel bulguların özeti aşağıdaki gibidir:

- Madde güçlük indeksi en yüksek olan (en kolay) sorular olumsuz soru kipindeki sorulardır.
- Ayırt edicilik gücü en yüksek olan sorular olumlu soru kipindeki sorulardır.
- Olumlu soru kipindeki soruların madde güçlüğü, cevabı işlem gerektiren soruların güçlüğüne göre manidar şekilde yüksektir, başka bir ifadeyle olumlu sorular daha kolaydır.
- Olumsuz soru kipindeki soruların madde güçlüğü, cevabı işlem gerektiren soruların güçlüğüne göre manidar şekilde daha yüksektir, başka bir ifadeyle, olumsuz sorular daha kolaydır.
- Olumlu soru kipindeki sorularla ve olumsuz soru kipindeki soruların madde güçlükleri arasında istatistiksel olarak manidar bir fark bulunmamıştır.
- K Tipi soruların madde güçlükleri ile diğer hiçbir soru türünün madde güçlük değerleri arasında manidar bir farklılık yoktur.
- Soru türlerinin madde ayırt ediciliği indeksleri arasında manidar bir farklılık bulunmamıştır.

Nicel verilere tamamlayıcı olarak öğrencilerle bireysel görüşmeler yapılmış ve öğrencilerin soruların güçlük düzeyleriyle ilgili görüşleri alınmıştır. K tipi sorulara ilişkin bazı görüşler şöyledir:

Ö14: *Sıralı sorular da olumsuz sorulardan farklı değil, zor. Olumsuz ve sıralı soruları konuyu çok iyi bilmelisiniz ki cevaplayasınız. Sınav ortamında stres, zaman sınırı nedeniyle böyle. Olumsuz ve sıralıları tekrar okumak gerekiyor, daha uzun. 1-1.5 dakikada çözdüğüm soruyu 3 dakikada çözebiliyorum. Tekrar tekrar okumak gerekiyor.*
Ö15: Sıralı soruları zor yanıtlıyorum. Kafamı karıştırıyor I,II,III hepsi olunca.
Ö19: Yalnız I, yalnız II olanlar şaşırtmacalı oluyor. Zorlanıyorum, yanlış çıkıyor genelde.
Ö3: Sıralı soru tipindeki soruları yanıtlamak benim için daha kolay.
Ö4: Sıralı sorularsa bana daha kolay geliyor çünkü genelde cevabı "E" seçeneği oluyor.
Ö13: Sıralı sorular en kolay sorulardır.

Olumsuz soruların güçlük değerlerine ilişkin bazı görüşler aşağıdaki gibidir:

Ö14: Olumsuz soruları yanıtlamakta zorlanıyorum. Olumsuz ve sıralı soruları konuyu çok iyi bilmelisiniz ki cevaplayasınız. Sınav ortamında stres, zaman sınırı nedeniyle böyle. Olumsuz ve sıralıları tekrar okumak gerekiyor, daha uzun. 1-1,5 dakikada çözdüğüm soruyu 3 dakikada çözebiliyorum. Tekrar tekrar okumak gerekiyor.
Ö4: Kökü olumsuz ifadeli olan soruların daha zor olduğunu düşünürüm.
Ö11: Olumsuz soru tipi en zor ve yanıltıcı soru tipidir.
Ö17: Zorlanmadan cevapladım, hiçbirinde zorlanmadım.

İşlemli sorulara ilişkin öğrenci görüşleri aşağıdaki gibidir:

Ö10: İşlemli sorular bence en zor ve zaman alıcı sorular.

Ö12: En çok zorlandığım sorular ise işlemli sorular...
Ö7: Standart soru tipi en kolay ve kısa sürede yaptığım soru tipidir.
Ö8: Olumlu ifadeli soruları daha kolay ve kısa sürede yanıtlıyorum.
Ö9: Standart soru tipi daha kolay ve daha az zaman alıcı.

Sınavlarda başka soru türlerine de yer verilmesi konusunda öğrencilerin görüşlerinden bazıları aşağıdaki gibidir:

Ö10: Bu sorular gayet iyi, başka sorular karıştırarak sistemi zorlaştırmaya gerek yok.
Ö11: Yeni soru türlerine gerek olduğunu düşünmüyorum.
Ö3: Özellikle boşluk doldurmalı ve seçenek olmayan (kısa cevap) sorular olması gerektiğini düşünüyorum. Bu durum kaliteyi arttırır. İyi çalışanla çalışmayanı ortaya koyar.
Ö8: Farklı türde soru maddesi tercih etme şansım olsa doğru-yanlış türündeki soruları tercih ederdim. Daha kolay olurdu benim için.

Yapılan bireysel görüşmelerde, öğrencilerin sınavda sorulan soruların güçlük değerlerine ve soru türlerine ilişkin görüşleri farklılık göstermektedir. Öğrencilerin bir bölümü (Ö7, Ö8, Ö8, Ö14, Ö15, 19) K Tipi soruları yanıtlamakta zorlandıklarını belirtmişlerdir. Öte yandan, bazı öğrenciler (Ö3, Ö4, Ö5, Ö11, Ö12, Ö13 ve Ö18) K Tipi soruları kolayca yanıtladıklarını belirtmişlerdir. Olumsuz sorularla ilgili olarak bazı öğrenciler (Ö4, Ö11, Ö14, Ö15, Ö17, Ö18) bu tür sorularda zorlandıklarını ve daha uzun sürede yanıtladıklarını belirtmişlerdir. İşlemli sorularla ilgili olarak öğrencilerin bazıları (Ö10, Ö12, Ö13, Ö19) işlemli sorularda zorlandıklarını ve bu sorular için daha fazla zaman harcadıklarını belirtmişlerdir. Öte yandan, bazı öğrenciler (Ö5, Ö14) işlemli soruları daha kolay yanıtladıklarını söylemişlerdir. Olumlu soru kipindeki sorulara ilişkin olarak bazı öğrenciler (Ö6, Ö7, Ö8, Ö9, Ö10, Ö15, Ö16) bu tür soruları daha kolay yanıtladıklarını belirtmişlerdir. Öğrencilere sınavlarda sorulan soru türlerinin yanı sıra farklı soru türlerinin yer almasını isteyip istemedikleri sorulduğunda, öğrencilerin birçoğu farklı soru türlerinin kullanılmasını istemediklerini belirtmişlerdir. Öte yandan, bazı öğrenciler (Ö3, Ö6, Ö7, Ö12) sınavlarda boşluk doldurmalı soruların yer almasını önermişlerdir. Bir öğrenci (Ö8) ise doğru-yanlış sorularını tercih ettiğini belirtmiştir.

4. TARTIŞMA, SONUÇ ve ÖNERİLER

Bu araştırmada, uzaktan eğitim bağlamında İşletme Bölümü derslerinin sınavlarında sorulan çoktan seçmeli soruların güçlük ve ayırt edicilik değerlerinin soru türlerine göre farklılık gösterip göstermediğinin belirlenmesi ve öğrencilerin soru türlerine ilişkin görüşlerinin alınması amaçlanmıştır. Yapılan araştırmada soruların olumlu soru kipindeki sorular, olumsuz soru kipindeki sorular, cevabı işlem gerektiren sorular ve K Tipi sorular olmak üzere 4 türde olduğu saptanmıştır. Araştırma sonuçlarına göre olumlu ve olumsuz soru kipindeki sorularla cevabı işlem gerektiren soruların madde güçlükleri arasında istatistiksel olarak manidar bir fark bulunmuştur; başka bir ifadeyle, öğrenciler olumlu ve olumsuz soruları işlemli sorulara göre daha kolay yanıtlamaktadır. Bu durumun, öğrencilerin sayısal ya da işlem becerilerinin zayıf olması nedeniyle ortaya çıkmış olabileceği ileri sürülebilir. Öğrencilerin sayısal ve sözel becerileri konusunda yapılmış araştırmalar olmamakla birlikte, sınav sonuçları ve öğrencilerden alınan geribildirimler ışığında öğrencilerin sayısal derslerde daha çok zorlandıkları ve düşük başarı gösterdikleri gözlenmektedir. Bu sonuca göre, öğrencilerin işlemli soruları çözme becerilerinin geliştirilmesi ve sınav başarılarının artırılması amacıyla gerek kitaplarda, gerek e-öğrenme malzemelerinde işlem gerektiren soruların yer aldığı etkinlikler ve sorular çoğaltılabilir.

Araştırma sonuçlarına göre, Varughese ve Glencross (1997), Downing ve diğerleri (1991), Rachor ve Gray (1996) ve Tamir (1993) tarafından yapılan çalışmaların bulgularına

benzer bir şekilde olumlu ve olumsuz soru türü arasında madde güçlüğü açısından manidar bir fark çıkmamıştır. Bununla birlikte, Haladyna Downing ve Rodriguez (2002) ve Tekin (1994) olumsuz soruların kullanımında dikkatli olunmasını; gerekmedikçe kullanılmamasını; olumlu sorunun birkaç doğru cevabı bulunduğu ya da olumlu köke çeldirici bulmanın çok zor olduğu durumlarda tercih edilmesini önermektedir. Bu görüşe paralel bir şekilde, yapılan görüşmelerde bazı öğrenciler olumsuz sorularda zorlandıklarını ve bu tür soruları daha uzun sürede yanıtladıklarını belirtmişlerdir. Öğrencilerden biri, olumsuz soru türüne yer verilmesi halinde soru kökündeki ifadelerin uzun olmamasını istemiştir. Bir diğer öğrenci ise seçeneklerin eşit dağılmasından çok yararlandığını belirtmiştir. Bu açıdan, şans başarısını önlemek için seçeneklerin eşit dağıtılmaması gerekmektedir. Çoktan seçmeli testlerin etkililiğini artırmanın bir yolu da yanıtların yanı sıra öğrencilerin verdiği yanıtın doğruluğundan emin olma düzeyini işaretlemesidir (Gardner-Medwin ve Curtin, 2007). Bir diğer yöntem de İngiliz Açık Üniversitesi tarafından yıllardır başarılı bir biçimde kullanılan; seçme yanıt yerine "yapılandırılmış yanıt" soru türünün kullanılması olabilir (Jordan ve Mitchell, 2009).

Araştırmada K Tipi sorular ile diğer soru türleri arasında güçlük değeri açısından manidar bir fark bulunmamıştır. Bu sonuç, Nnodim (1992) tarafından yapılan araştırmada insan anatomisi testinde K Tipi soruların klasik çoktan seçmeli sorulardan daha zor olduğu sonucu ile paralellik göstermemektedir. Öte yandan, Downing ve diğerleri (1991), Nnodim (1992), Rachor ve Gray (1996) ve Tamir (1993) tarafından yapılan çalışmaların bulgularını destekler nitelikteki bu araştırmada K Tipi soruların ayırt edicilik değerleri arasında manidar bir farklılık çıkmamıştır. Bunun bir sebebi öğrencilerden bazılarının görüşmelerde belirttiği gibi bu tür soruların doğru yanıtlarının genellikle E seçeneğine koyulması olabilir. Bu soruların yanıtlarının genellikle E seçeneğine koyulup koyulmadığı kontrol edilebilir ve doğru yanıtın diğer seçeneklere de yerleştirilmesi sağlanabilir. Buna karşılık, Haladyna Downing ve Rodriguez (2002) K Tipi soruların kesinlikle kullanılmaması gerektiğini belirtmektedir. Benzer şekilde, Albanese (1993) K Tipi soruların iyi çalışmadığını belirtmiştir. K Tipi sorular yerine Harasym ve diğerleri (1992) çoklu doğru-yanlış (multiple true false) sorularının kullanılmasını önermektedir. Karadağ (2014) tarafından yapılan; ölçme araçları konusunda uzaktan öğrencilerin görüşlerinin alındığı bir araştırmada, öğrencilerin en çok doğru-yanlış soru türünü tercih ettikleri tespit edilmiştir. Buna göre, ileride yapılacak sınavlarda K Tipi sorular yerine çoklu doğru-yanlış sorularına yer verilebilir.

Alanyazında farklı bağlamlarda yapılan araştırmalar farklı sonuçlar ortaya koymuştur. Sonuçların genellenebilmesi için, bu araştırmanın İşletme bölümünde farklı yıllara ilişkin sorular kullanılarak tekrar edilmesi ve sonuçların karşılaştırılması önerilebilir. Buna ek olarak, araştırma Klasik Test Kuramı çerçevesinde yapılmış olup klasik test kuramında madde parametreleri gruba bağımlıdır. Soruların Madde Tepki Kuramına göre ölçeklenip analiz edilmesi farklı sonuçlar ortaya koyabilir. Madde Tepki Kuramı sonucunda elde edilecek güçlük ve ayırt edicilik parametreleriyle Klasik Test Kuramından elde edilen madde parametreleri arasındaki ilişki incelenebilir. Ayrıca, Madde Tepki Kuramıyla elde edilen madde parametrelerinin soru türüne göre farklılaşıp farklılaşmadığı incelenerek elde edilen bulgular bu çalışmanın bulgularıyla karşılaştırılabilir.

Öğrencilerin değerlendirilmesi, öğrenme tasarımı sürecinde önemle üzerinde durulması gereken bir konudur. Değerlendirme faaliyetleri, öğrencilere ve öğreticilere geribildirim sağlayarak sistemin geliştirilmesine ve kalitenin artırılmasına olanak verir. Kalite güvence ve akreditasyon kuruluşları için ölçme-değerlendirme sistemlerinin geçerlik ve güvenirliği akredite olmanın temel koşullarından biridir. Özellikle öğrenci sayısının fazla olduğu uzaktan eğitim programlarında bazen tek ölçme aracı çoktan seçmeli sorular olabilmektedir. Bu tür soruların bilişsel düzeyleriyle güçlük ve ayırt edicilik değerleri arasındaki ilişkiyi ortaya koyan

çalışmaların farklı alanlarda ve farklı bağlamlarda yapılması, değerlendirmede geçerlik ve güvenirliğin sağlanması ve sorularda kalitenin artırılması açılarından önem taşımaktadır. Bu bağlamda, araştırma sonuçlarının eğitim alanındaki test geliştiricilerine, ölçme-değerlendirme uzmanlarına, öğreticilere, alınacak kararlar açısından yöneticilere katkı sağlayacağı umulmaktadır.

5. KAYNAKLAR

Albanese, M. (1993). Type K and other complex multiple-choice items: an analysis of research and item properties. *Educational Measurement: Issues and Practice, 12*(1), 28–33.

Bray, J.H. ve Maxwell, S.E. (1982). Analyzing and interpreting significant MANOVAs. *Review of Educational Research, 52*, 340–367.

Büyüköztürk, Ş., Çokluk, Ö. ve Köklü, N. (2015). *Sosyal bilimler için istatistik.* Ankara: Pegem Akademi.

Creswell, J. W. (2008). *Educational research: planning, conducting and evaluating quantitaive and qualitative research* (3. baskı). New Jersey:Pearson Education, Inc. Upper Saddle River.

Çokluk, Ö., Şekercioğlu, G. ve Büyüköztürk, Ş. (2012). *Sosyal bilimler için çok değişkenli istatistik: SPSS ve LISREL uygulamaları.* Ankara: Pegem Akademi.

Downing, S. M., Dawson-Saunders, B., Case, S. M., ve Powell, R. D. (1991, Nisan). *The psychometric effects of negative stems, unfocused questions, and heterogeneous options on NBME Part I and Part II item characteristics.* National Council on Measurement in Education Konferansı'nda sunulan bildiri, Chicago.

Field, A. (2009). *Discovering statics using SPSS*. London: SAGE Publications Ltd.

Johnstone, A.H. (1983). Training teachers to be aware of student learning difficulties. İçinde Tamir, P., Hofstein, A. & Ben Peretz, M. (eds) *Preservice and inservice education of science teachers.* Rehovot, Israel & Philadelphia: Balaban International Science Services: 109-116.

Jordan, S. ve Mitchell, T. (2009). e-Assessment for learning? The potential of short-answer free-text questions with tailored feedback. *British Journal of Educational Technology, 40*(2), 371-385.

Haladyna, T. M., Downing, S. M. ve Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education*, 15(3), 309-309.

Haladyna, T.M. (2004). *Developing and validating multiple-choice test items.* (3 ed.) Mahwah, NJ: Lawrence Erlbaum Associates.

Harasym, P. H., Price, P. G., Brant, R., Violato, C. ve Lorscheider, F. L. (1992). Evaluation of negation in stems of multiple-choice items. *Evaluation and the Health Professions, 15*, 198–220.

Johnstone, A.H. (1983). Training teachers to be aware of student learning difficulties. İçinde Tamir, P., Hofstein, A. & Ben Peretz, M. (eds) *Preservice and inservice education of science teachers.* Rehovot, Israel & Philadelphia: Balaban International Science Services: 109-116.

Kaptan, F. (1985). ÖSYS *Fizik soruları üzerine bir nitelik araştırması.* Yayımlanmamış yüksek lisans tezi, Hacettepe Üniversitesi Sosyal Bilimler Enstitüsü, Ankara.

Karaca, E. (2004). Seçme gerektiren, kısa cevaplı ve doğru-yanlış testlerinin madde ve test özelliklerinin karşılaştırılması. *Dumlupınar Üniversitesi Sosyal Bilimler Dergisi*, 10.

Karadağ, N. (2014) *Açık ve uzaktan eğitimde ölçme ve değerlendirme: mega üniversitelerdeki uygulamalar*. Yayımlanmamış doktora tezi, Anadolu Üniversitesi Sosyal Bilimler Enstitüsü, Eskişehir.

Kaya, A. (1991). *Eksik köklü – ortak seçenekli – ortak köklü madde türlerinin madde ve test istatistiklerine etkisi.* Yayımlanmamış yüksek lisans tezi, Hacettepe Üniversitesi Sosyal Bilimler Enstitüsü, Ankara.

Lindler, P. (1998). Assessment tools for distance learning: A review of the literature. Washington State Board for Comunity and Technical Colleges, Olympia. [Çevrim-içi: http://files.eric.ed.gov/fulltext/ED426725.pdf], Erişim tarihi: 09.02.2016.

McIsaac, M.S. ve Gunawardena, C.N. (1996). Distance education. *Handbook of research for educational communications and technology* (Ed: D. Jonassen). New York: Simon and Schuster Macmillan, ss. 403.

Miles, M. B. ve Huberman, A.M. (1994). *Qualitative data analysis: An expanded sourcebook.* Thousand Oaks, California: SAGE.

Gardner-Medwin, T. ve Curtin, N. (2007). *Certainty-based marking (CBM) for reflective learning and proper knowledge assessment*. REAP Int. Online Conf. on Assessment Design for Learner Responsibility Konferansında sunulan bildiri.

Nnodim, J. O. (1992). Multiple-choice testing in anatomy. *Medical Education, 26*, 301–309.

Özçelik, D.A. (1989). *Test hazırlama kılavuzu*. Ankara: ÖSYM Yayınları 8.

Pallant, J. (2005). *SPSS survival manual: A step by step guide to data analysis using SPSS for Windows*. Australia: Allen & Unwin.

Pearson, E. S., Pearson, K. ve Hartley, H.O. (1958). *Biometrika tables for statisticians*. New York: Cambridge University Press.

Popham, J. (2003). *Test better, teach better: The instructional role of assessment*. Alexandria, VA: ASCD.

Puspitasari, K.A. (2010). Student assessment. *Policy and Practice in Asian Distance Education* (Ed: T. Belawati ve J. Baggaley). New Delhi: SAGE, pp.60-65.

Rachor, R. E. ve Gray, G. T. (1996, Nisan). *Must all stems be green? A study of two guidelines for writing multiple choice stems.* American Educational Research Association Konferansında sunulan bildiri, New York.

Sanderson, P.J. (2010). *Multiple-choice questions: a linguistic investigation of difficulty for first-language and second-language students.* Yayınlanmamış doktora tezi. University of South Africa, Güney Afrika.

Simonson, M., Smaldino, S., Albright, M. ve Zvacek, S. (2012). *Teaching and learning at a distance: Foundations of distance education*. Boston: Allyn & Bacon.

Stevens, J.P. (2009). *Applied multivariate statistics for the social sciences*. New York: Routledge.

Strangor, C. (2010) *Research methods for the behavioral sciences*. Boston, MA: Houghton Mifflin.

Tabachnick, B. G. ve Fidell, L.S. (2007). *Using Multivariate Statics.* Boston: Pearson.

Tamir, P. (1993). Positive and negative multiple choice items: Howdifferent are they? *Studies in Educational Evaluation, 19,* 311–325.

Tashakkori, A. ve Teddlie, C. (2003). *Handbook of mixed methods in social and behavioral research.* Thousand Oaks, CA: Sage.

Tekin, H. (1994). Eğitimde ölçme ve değerlendirme. Ankara: Yargı Yayınları.

Thorpe, M. (1988). *Evaluating open and distance learning*. Great Britain: Biddles Ltd.

Varughese, K.V. ve Glencross, M.J. (1997). The effect of positive and negative modes of multiple choice items on students' performance in Biology. *South African Journal of Higher Education, 11* (1), 177-179.

Yıldırım, A. ve Şimşek, H. (2008). *Sosyal bilimlerde nitel araştırma yöntemleri* (7. baskı). Ankara: Seçkin.

Zhang, W., Tsui, C., Jedege, O., Ng, F. ve Kowk, L. (2002). *A comparison of distance education in selected Asian open universities.* 14th Annual Conference of Asian Association of Open Universities Konferansında sunulan bildiri, Manila, Philippines. [Available online at: http://www.ouhk.edu.hk/cridal/gdenet/Management/Governance/EAM11A.html], Retrieved on 23.02.2016.

Extended Abstract

Assessment in education is one of the most crucial components of an instructional design that provides feedback on learning and teaching processes and it also enables to review and improve the whole process. For assessing student learning, a variety of tools are used to assess student learning in higher education. Among those tools, multiple-choice items (questions) are still being used widely in higher education (Haladyna, 2004; Karadağ, 2014). Especially in distance education settings, in which learning activities are carried out via various telecommunication technologies as the learners, teachers and learning materials are separated in terms of time and/or place, multiple-choice tests have a crucial role in the assessment of student learning (Lindler,1998; Puspitasari, 2010).

Multiple-choice tests are improved after being analyzed by various methods (Özçelik, 1989). One of these methods is called "the item analysis" that is conducted for each item in a test in which difficulty (p) and discrimination (r) indices are calculated. In literature, research on the relationship between item types and difficulty and discrimination indices reveal diverse results based on the subject, context and different types of questions. It is observed that existing research has been conducted in various subject

areas covering a limited number of questions and most of those studies have been conducted in relation with traditional education except for the one which has been conducted by Sanderson (2010) in a distance education context. In addition, although multiple-choice tests have been commonly used for years in Turkish higher education, there is limited research on the relationship between item types and difficulty and discrimination indices of items. In this regard, this study aims to determine whether the difficulty and discrimination indices of the multiple-choice questions show differences according to item types, which are asked in the exams of the courses in a business administration bachelor's degree program offered through open and distance learning in a public university in Turkey, and to obtain the opinions of the learners on the item types. No related studies were found in literature regarding the questions of business administration programs, which is one of the most common programs with a large number of learners, offered both in Turkey and the world. Below are the research questions of this study:

1. Do the difficulty indices (p) of multiple-choice questions show a significant difference according to item types?

2. Do the discrimination indices (r) of multiple-choice questions show a significant difference according to item types?

3. What are the distance learners' opinions about the item types used in the exams?

In this descriptive mixed study, both quantitative and qualitative data were collected. Quantitative data were collected from item analysis reports including 905 items whereas qualitative data were collected via semi-structured interviews with 20 students. In the first step, item types were coded by three assessment experts, and the inter-coder reliability was calculated by using the formula (Inter-coder reliability= Agreement / Agreement + Disagreement) of Miles and Huberman (1994) and found to be 99%. It was identified that the items included four types which were positive, negative, problem-based and K-Type questions. After determining the item types, the p and r indices of each item was identified from item analysis documents and tabulated to be analyzed by the SPSS 22.0 program. One-way MANOVA Test was used. When a significant difference was found in One-way MANOVA results, the One-way ANOVA was used to determine the dependent variables that caused the difference. When a significant difference was found as a result of One-way ANOVA, Scheffe was used in cases where the homogeneity of variances assumption was ensured, and the Brown-Forsythe and Welch Test was used in cases where the homogeneity of variances was not ensured. Pairwise comparisons were made using Tamhane's T2 tests if significant results were found.

In order to collect qualitative data, the learners in the Department of Business Administration were accessed through phone and social media for semi-structured individual interviews and were informed of the subject and scope of the study. It was explained to the participants that their identities would be kept confidential and would not be shared with any third parties. The learners who volunteered to participate in this study were interviewed through Skype or phone on a scheduled date. The permission of the learners was obtained to record the interview. Interview questions were reviewed by three experts. The interviews were recorded, decoded, and analyzed using the descriptive analysis method. Yıldırım and Şimşek (2008) stated that descriptive analysis is more superficial than content analysis and is used in studies where the conceptual structure of the study is clearly previously determined. The data can be organized according to the themes set by the study questions or by the questions used during the interviews and observations. In this respect, the data were summarized and interpreted according to the interview questions. Two researchers coded the data for the reliability in data analysis. The inter-coder reliability was found to be at 89%. Agreement was ensured by discussing on the items that were coded differently.

As a result, this study revealed a significant difference between the difficulty indices of positive items and problem-based items as well as negative items and problem-based items. In other words, positive and negative items were answered more easily than the problem-based items by the students. On the other hand, no significant difference was found between the discrimination indices. Finally, it was found that opinions of students on the item types showed variety.