International
Journal of Human Sciences
ISSN:2458-9489

Volume 14 Issue 3 Year: 2017

The effect of test instruction and risk taking tendency on psychometric properties of multiple choice tests[1]

Test yönergesi ve risk alma eğiliminin çoktan seçmeli testlerin psikometrik özelliklerine etkisi

Ercan Çoban[2]
Rahime Nükhet Demirtaşlı[3]

Abstract

The research is a quasi-experimental research which aims to investigate the effect of test instruction and risk taking tendency on the psychometric properties of multiple choice tests. The study group of the research comprised of 220 undergraduate students studying at different departments of Ankara University Faculty of Educational Sciences in 2014-2015 spring term. An achievement test of measurement and evaluation course consisting of 24 items was applied to students in the study group. The test was applied to four similar groups with different instructions. A risk taking scale was applied to determine the risk taking tendency of the students. The data were analyzed by using one way ANOVA, dependent sample t-test, independent sample t-test, Feldt test, Fisher's z test, Pearson correlation coefficient, Spearman's rank-order correlation coefficient and confirmatory factor analysis. The findings showed that corrected score means, uncorrected score means and reliability coefficients calculated from corrected scores of tests applied with different instructions were significantly different whereas reliability coefficients calculated from uncorrected scores were not significantly different. The validity of

Özet

Bu araştırma, test yönergesi ve risk alma eğiliminin çoktan seçmeli testlerin psikometrik özellikleri üzerindeki etkisini incelemeyi amaçlayan yarı deneysel bir araştırmadır. Araştırmanın çalışma grubu, 2014-2015 bahar döneminde Ankara Üniversitesi Eğitim Bilimleri Fakültesinde öğrenim gören 220 öğrenciden oluşmaktadır. Çalışma grubundaki öğrencilere ölçme ve değerlendirme dersi başarısını ölçen 24 maddelik çoktan seçmeli test uygulanmıştır. Başarı testi birbirine benzer dört öğrenci grubuna farklı yönergelerle uygulanmıştır. Öğrencilerin risk alma eğilimini belirlemek için risk alma ölçeği uygulanmıştır. Veriler; tek faktörlü varyans analizi, ilişkili örneklemler için t-testi, ilişkisiz örneklemler için t-testi, Feldt testi, Fisher z testi, Pearson korelasyon katsayısı, Spearman sıra farkları korelasyon katsayısı ve doğrulayıcı faktör analizi teknikleriyle analiz edilmiştir. Araştırmadan elde edilen bulgular, farklı yönergelerle uygulanan testlerin ham puan ortalamaları, düzeltilmiş puan ortalamaları ve düzeltilmiş puana göre hesaplanan güvenirlik katsayıları arasında manidar fark olduğunu; buna karşın ham puanlara göre hesaplanan güvenirlik katsayıları arasında manidar fark

[1] This study was summarized from the master thesis titled as "The Effect of Test Instruction and Risk Taking Tendency on the Psychometric Properties of Multiple Choice Tests and Guessing Tendency"

[2]PhD Student, Ankara University, Institute of Educational Sciences, coban.ercan@gmail.com

[3] Prof. Dr., Ankara University, Faculty of Educational Sciences, Department of Measurement and Evaluation, rnukhet@yahoo.com

Çoban, E., & Demirtaşlı, R. N. (2017). Test yönergesi ve risk alma eğiliminin çoktan seçmeli testlerin psikometrik özelliklerine etkisi. *Journal of Human Sciences*, *14*(3), 2949-2967. doi:10.14687/jhs.v14i3.4449

the test applied with the instruction stating that no correction for guessing would be made was the highest one. In general, risk taking tendency did not significantly change test scores.

Keywords: Correction formulas; test instruction; psychometric properties of test; validity and reliability; risk taking tendency.

(Extended English abstract is at the end of this document)

olmadığını göstermektedir. Şans başarısı için herhangi bir düzeltme yapılmayacağını bildiren yönergeyle uygulanan testin geçerliği, diğer yönergelerle uygulanan testlerin geçerliğinden yüksek çıkmıştır. Risk alma eğilimi genel olarak test puanlarında manidar değişikliğe yol açmamıştır.

Anahtar Kelimeler: Düzeltme formülleri; test yönergesi; testin psikometrik özellikleri; geçerlik ve güvenirlik; risk alma eğilimi.

GİRİŞ

Eğitimde öğrencilerin başarı düzeylerinin ölçülmesinde çeşitli ölçme araçları kullanılmaktadır. Bu ölçme araçlarından birisi çoktan seçmeli maddelerden oluşan testlerdir. Çoktan seçmeli madde, cevaplayıcıların üç ya da daha fazla seçenekten birisini seçmelerini gerektiren madde türüdür. (Popham, 2000, s. 242).

Çoktan seçmeli testlerin yaygın kullanılmasının bazı nedenleri bulunmaktadır. Çoktan seçmeli maddelerle, kısa sürede daha geniş bir kapsamı yoklamak mümkündür. Çoktan seçmeli testlerden elde edilen puanlar daha objektif ve güvenilirdir. Ayrıca bilgisayarla puanlama yapıldığı için, test sonuçlarını öğrencilere kısa sürede bildirme olanağı vardır (Haladyna, 1997, s. 65-66).Fakat çoktan seçmeli testlerin bazı sınırlılıkları da vardır. Çoktan seçmeli testlerin en önemli sınırlılıklarından birisi doğru yanıtların şansla bulunma olasılığının olmasıdır (Baykul, 2010, s. 393; Crocker ve Algina, 1986, s. 313; Zimmerman ve Williams, 1965).

Testi alan bireylerin şansla doğru yanıtı bulması sonucu kazandığı puana şans başarısı denilmektedir. Bir yanıtlayıcının testin tamamında şansla doğru yanıtladığı sorulardan kazandığı puana şans puanı denmektedir. Şans sonucu kazanılan şans puanları madde ve test istatistiklerine hata karışmasına neden olmaktadır. Bazı maddelerin şansla doğru cevaplanması sonucu ortaya çıkan bu hataya şans hatası denir. Bu hata, tesadüfî hatalardan farklıdır. Şans hatası, cevaplama sırasında alınan bazı önlemlerle azaltılabilen ve puanlama sırasında kullanılan bazı istatistiksel yöntemlerle düzeltilebilen bir hata türüdür (Baykul, 2010, s. 393). Test puanlarına hata karışmasına neden olan şans başarısı test güvenirliğini (Zimmerman ve Williams, 1965) ve geçerliğini (Baykul, 2010, s. 425-426) düşürmektedir.

Testin geçerliğini ve güvenirliğini düşüren şans başarısı ve tahmin davranışına karşı ne yapılacağı, test uygulayıcıların öncelikli sorunlarından birisi olmuştur. Bu sorunu çözmek için değişik düzeltme formülleri geliştirilmiştir (Crocker ve Algina, 1986, s. 399-409; Kurz, 1999).

Şans başarısı düzeltmek için kullanılan en yaygın düzeltme formüllerinden biri klasik düzeltme formülüdür. Bu formül, DP, düzeltilmiş test puanını; D, doğru yanıtlanan soru sayısını; Y, yanlış yanıtlanan soru sayısını ve k, seçenek sayısını göstermek üzere,

$$DP = D - \frac{Y}{k-1}$$

şeklindedir (Crocker ve Algina, 1986, s. 400; Reid, 1976). Tüm seçeneklerin seçilme olasılığının eşit olduğunu ve bütün yanlış yanıtların tahmin sonucu oluştuğunu varsayan bu formülün kullanılmasının amacı yanlış cevapları cezalandırarak, cevaplayıcıların tahminle yanıtlama davranışlarını engellemektir (Thorndike, 2005, s. 467). Bu düzeltme yöntemi yanıtlayıcıların bazen kısmi bilgilerini kullanarak ya da yanlış öğrenmelerden dolayı yanıt verebileceklerini dikkate

Çoban, E., & Demirtaşlı, R. N. (2017). Test yönergesi ve risk alma eğiliminin çoktan seçmeli testlerin psikometrik özelliklerine etkisi. *Journal of Human Sciences*, *14*(3), 2949-2967. doi:10.14687/jhs.v14i3.4449

almamaktadır (Rowley ve Traub, 1977). Bu nedenden dolayı, klasik düzeltme formülü bazı yanıtlayıcıların puanını gereğinden fazla düzeltirken (overcorrection), bazılarınınkini yeterince düzeltmemektedir (undercorrection) (Zimmerman ve Williams, 1965)

Tahminde bulunan cevaplayıcıları cezalandırmak yerine tahminde bulunmayan cevaplayıcıları ödüllendirmek, şans başarısını düzeltmek için kullanılan diğer bir yöntemdir. Bu yöntem tahmine gitmeyen cevaplayıcıların tahmine gitmeleri durumunda boş bıraktıkları maddelerin bir kısmını doğru cevaplayacakları varsayımı üzerine kuruludur. Seçenek sayısı k olmak üzere cevaplayıcıların rastgele tahminle soruların 1/k kadarını doğru cevaplayacaklarını varsayar. Bundan dolayı, tahminde bulunmayan cevaplayıcılara boş bırakılan soru sayısının seçenek sayısına oranı kadar puan ödül olarak verilir. Bu hesaplama yönteminde, DP, düzeltilmiş puan; D, doğru yanıtlanan soru sayısı; B, boş bırakılan soru sayısı ve k seçenek sayısı olmak üzere,

$$DP = D + \frac{B}{k}$$

formülü kullanılmaktadır (Ebel, 1965, s.224).

Düzeltilmiş test puanını hesaplamak için doğru sayısından yanlış sayısının çıkarıldığı düzeltme formülü, alan yazında kullanılan diğer bir yöntemdir. Bu yöntemde, DP, düzeltilmiş puan; D, doğru sayısı ve Y, yanlış sayısı olmak üzere,

$$DP = D - Y$$

formülüyle hesaplanmaktadır (Prieto ve Delgado, 1999a).

Düzeltme formülleri, rastgele tahmin davranışını ve tahmin davranışı sonucu test puanlarına karışan hatayı azaltmak için geliştirilmiştir. Bu formüllerle test puanlarının geçerlik ve güvenirliğini arttırmak amaçlanmıştır. Fakat yapılan çalışmalarda düzeltme formülü kullanmanın testin psikometrik özelliklerine etkisi hakkında farklı bulgulara ulaşılmıştır.

Alan yazındaki çalışmalara göre genel olarak, klasik düzeltme yöntemi kullanıldığında test puanlarının düzeltme yapılmayan duruma göre (number-right scoring) düşmektedir (Alnabhan, 2002; Betts, Elder, Hartley ve Trueman, 2009; Swineford ve Miller, 1953; Umay, 1998). Düzeltme uygulandığında puanlar azalmasına rağmen, bireylerin test puanları sıralamasında önemli bir değişim olmamaktadır. Ham ve düzeltilmiş puanlar yüksek korelasyon göstermektedir (Angoff ve Schrader, 1981; Çelen ve Çıkrıkçı-Demirtaşlı, 2006; Ebel, 1965, s. 227-229).

Düzeltme formülleri kullanmanın testin güvenirliği üzerindeki etkisine ilişkin çalışmalarda birbirinin zıttı olan sonuçlara ulaşılmıştır. Bazı araştırmalarda klasik düzeltme yöntemi kullanarak düzeltme yapıldığında, test puanları güvenirliğinin yükseldiği bulunmuştur (Alnabhan, 2002; Burton, 2002; Muijtjens, Mamaren, Hoogenboom, Evers ve Vleuten, 1999; Lord, 1975). Burton (2002) düzeltme kullanmanın güvenirlik üzerindeki bu etkisinin özellikle kısa testlerde daha açık bir şekilde görüleceğini belirtmiştir. Bazı araştırmalarda ise klasik düzeltme yöntemi kullanıldığında, test puanları güvenirliğinin düştüğü bulunmuştur (Glass ve Wiley, 1964; Socan, 2009). Socan (2009) güvenirliğin düşmesini, düzeltme formülü varsayımlarının kısmi bilgiyi dikkate almaması sonucunda puanlara hata karışmasına bağlanmıştır. Zimmerman ve Williams (1965), varsayımları sağlanmadığı durumda düzeltme formülü kullanıldığında, ölçmenin standart hatasının artacağını belirtmişlerdir. Ayrıca düzeltme yapmanın güvenirliği değiştirmediğini bulan araştırmacılar da vardır (Abu Sayf, 1975; Cross ve Frary, 1977; Frary, 1982; Sabers ve Feldt, 1968).

Düzeltme formülü kullanmanın test geçerliğine etkisi üzerine yapılan çalışmalarda da farklı sonuçlar bulunmuştur. Socan (2009) düzeltme yapılmadığı durumda testin geçerliğinin daha yüksek olduğunu bulmuştur. Prihoda, Pinckard, McMahan ve Jones (2006) ise düzeltme formülü uygulandığında geçerliğin arttığını bulmuşlardır. Bazı araştırmalarda ise düzeltme yapıldığında geçerliğin manidar olarak değişmediği bulunmuştur (Çelen ve Çıkrıkçı-Demirtaşlı, 2006; Frary, 1982; Sabers ve Feldt, 1968)

Alan yazında şans başarısına yol açan tahminle yanıtlama davranışını risk alma eğilimi açısından araştıran çalışmalar da bulunmaktadır. Espinosa ve Gardeazabal (2010), yaptıkları çalışmada riskten kaçınma eğilimi fazla olan yanıtlayıcıların yanıtını bilmedikleri maddeleri yanıtsız

Çoban, E., & Demirtaşlı, R. N. (2017). Test yönergesi ve risk alma eğiliminin çoktan seçmeli testlerin psikometrik özelliklerine etkisi. *Journal of Human Sciences*, *14*(3), 2949-2967. doi:10.14687/jhs.v14i3.4449

bırakma eğiliminin daha fazla olduğunu bulmuşlardır. Kubinger ve Wolfsbauer (2010) ise risk alma eğilimi ile yanıtlayıcıların bilmedikleri maddeleri yanıtsız bırakma eğilimi arasında manidar olmayan negatif korelasyon bulmuşlardır. Bu yüzden, risk alma eğilimi ile tahmin davranışı arasında ilişki olmadığı sonucuna ulaşmışlardır. Rubio, Hernandez, Zaldivar, Marquez ve Santacreu (2010), iki farklı risk alma ölçeğinin ölçüt geçerliğini belirlemek için tahminle yanıtlama eğilimini ölçüt olarak kullanmışlardır. Tahminle yanıt eğilimini matematiksel olarak belirlemek için, doğru ve yanlış sayılarını dikkate alarak bir model oluşturmuşlar ve risk alma ölçeklerinden elde edilen puanlarla, çoktan seçmeli testi tahminle yanıtlama eğilimi arasında manidar ilişki bulmuşlardır. Albanese (1988) karar verirken temkinli davranan yanıtlayıcıların daha fazla boş bırakma eğilimi gösterdiğini bulmuştur. Slakter (1968), düzeltme formülünün risk alma eğilimi düşük bireyleri, risk alma eğilimi yüksek bireylere göre daha fazla cezalandırdığını bulmuştur. Bu durumu risk alma eğilimi düşük bireylerin test yönergesini daha fazla dikkate almalarına bağlamıştır.

Sonuç olarak, tahminle yanıtlama davranışı sonucunda test puanlarına karışan hata puanlarının testin psikometrik özellikleri üzerindeki olumsuz etkisini azaltmak için düzeltme formülleri geliştirilmiştir. Fakat bu formülleri kullanmanın testin geçerlik ve güvenirliği üzerindeki etkisine yönelik yapılan çalışmalarda tartışmalı sonuçlara ulaşılmıştır. Türkiye'de YGS, LYS, KPSS ve ALES gibi sınavlarda düzeltme formülü kullanılırken; YDS, açık öğretim sınavları ve ehliyet sınavlarında düzeltme formülü kullanılmamaktadır. Türkiye'de farklı uygulamaların olduğu göz önüne alındığında, düzeltme formüllerinin testin psikometrik özellikleri üzerindeki etkisinin araştırılması önemli bulunmuştur. Ayrıca risk alma eğiliminin tahmin davranışını arttırdığını bulan çalışmalar olmasına rağmen risk alma eğiliminin testin psikometrik özellikleri üzerindeki etkisine yönelik çalışma bulunmamaktadır. Bu sebeple, bu araştırmada şans başarısı için farklı düzeltme formülleri kullanılacağını bildiren test yönergesi ve risk alma eğiliminin testin psikometrik özellikleri üzerindeki etkisini belirlemek amaçlanmıştır. Bu amaca dayalı olarak aşağıdaki soruların yanıtı aranmıştır:

1. Şans başarısı için düzeltme uygulanacağına dair farklı yönergeler içeren testlerden alınan ham puan ve düzeltilmiş puan ortalamaları arasında manidar farklılık var mıdır?
2. Şans başarısı için düzeltme uygulanacağına dair farklı yönergeler içeren testlerin geçerlik ve güvenirlik katsayıları arasında manidar fark var mıdır?
3. Ham ve düzeltilmiş test puanları, risk alma düzeyine göre manidar farklılık göstermekte midir?

YÖNTEM

Araştırma Modeli

Bu araştırmada, test yönergesinde puanlama yapılırken düzeltme formülleri kullanılacağına ilişkin yapılan açıklamaların testi psikometrik özellikleri üzerindeki etkisini belirlemek amaçlandığı için ölçme ve değerlendirme dersi başarı testi dört farklı yönergeyle uygulanmıştır. Yönergelerde sadece testin nasıl puanlanacağı hakkındaki açıklamalar farklı olup, diğer açıklamalar aynıdır. Yönergelerde puanlamaya yönelik açıklamalar şu şekildedir:

- **Yönerge 1 (Y_1):** Puanınız hesaplanırken, yanlış yanıtlarınız doğru yanıtlarınızı götürmeyecektir.
- **Yönerge 2 (Y_2):** Puanınız, dört yanlış yanıtınız bir doğru yanıtınızı götürecek şekilde hesaplanacaktır.
- **Yönerge 3 (Y_3):** Puanınız hesaplanırken, yanıtını boş bıraktığınız her beş soru bir doğru yanıt olarak sayılacaktır.
- **Yönerge 4 (Y_4):** Puanınız, doğru yanıtlarınızın sayısından yanlış yanıtlarınızın sayısı çıkarılarak hesaplanacaktır.

Çoban, E., & Demirtaşlı, R. N. (2017). Test yönergesi ve risk alma eğiliminin çoktan seçmeli testlerin psikometrik özelliklerine etkisi. *Journal of Human Sciences*, *14*(3), 2949-2967. doi:10.14687/jhs.v14i3.4449

Araştırma, son-test kontrol gruplu desende yarı deneysel bir çalışmadır (Creswell, 2012, s. 309). Testi yönerge 1 ile alan grup araştırmanın kontrol grubunu, diğer yönergelerle alan gruplar deney gruplarını oluşturmaktadır.

Yarı deneysel çalışmalar, yansız atamayı gerektirmediği için eğitim araştırmalarında sıklıkla kullanılmaktadır. Bu çalışmalarda yansız atama yapılmadığı için, araştırmacının olası dış etkiler (öğrenci başarı, yetenek, motivasyon vb.) üzerindeki kontrolü gerçek deneysel çalışmalara göre daha azdır (Creswell, 2012, s. 309-310).

Çalışma Grubu

Bu araştırmanın çalışma grubunu, 2014-2015 Eğitim Öğretim Yılı Bahar Dönemi'nde Ankara Üniversitesi Eğitim Bilimleri Fakültesinin lisans programlarında öğrenim gören 220 öğrenci oluşturmaktadır. Çalışma grubunu oluşturan öğrencilerin cinsiyet ve uygulanan yönergeye göre dağılımları Tablo 1'de verilmiştir.

Tablo 1. Çalışma Grubunun Uygulanan Yönerge ve Cinsiyete Göre Dağılımı

		Cinsiyet		
		Kadın	**Erkek**	**Toplam**
Yönerge	**Y_1**	46	9	55
	Y_2	48	7	55
	Y_3	43	12	55
	Y_4	42	13	55
	Toplam	179	41	220

Veri Toplama Araçları

Araştırmada test yönergesinin testin psikometrik özellikleri üzerindeki etkisini belirlemek için ölçme ve değerlendirme dersi başarı testi kullanılmıştır. Öğrencilerin risk alma eğilimini belirlemek için risk alma ölçeği kullanılmıştır.

Ölçme ve Değerlendirme Dersi Başarı Testi

Araştırma kapsamında kullanılan ölçme ve değerlendirme dersi başarı testi, araştırmacılar tarafından test geliştirme aşamalarına uygun olarak geliştirilmiştir. Testin kapsam geçerliğini sağlamak için belirtke tablosu hazırlanmış ve uzman görüşü alınmıştır. Testteki maddeler yazılırken, Ulutaş (2003) tarafından geliştirilen ölçme ve değerlendirme dersi başarı testindeki maddelerden yararlanılmıştır. Teste 5 seçenekli toplam 24 çoktan seçmeli madde bulunmaktadır. Teste bilgi, kavrama ve uygulama düzeylerini ölçen eşit sayıda madde bulunmaktadır. Testin A, B, C ve D olmak üzere her biri farklı bir yönerge içeren dört formu oluşturulmuş ve çalışma grubundaki öğrencilere uygulanmıştır.

Risk Alma Ölçeği

Öğrencilerin risk alma eğilimlerini belirlemek için Bayar ve Sayıl (2005) tarafından ergenler için geliştirilen Risk Alma Ölçeği kullanılmıştır. Ölçek ergenlerin günlük hayatta yaptığı riskli davranışları ölçmektedir. Ölçekteki maddeler bireylerin trafikte, sosyal ilişkilerde ve sağlığa zararlı maddelerin (sigara, alkol, uyuşturucu vb.) kullanımında ne kadar sıklıkta risk aldıklarını belirlemeyi amaçlamaktadır. Ölçeğin orijinal hali 5'li Likert tipi 25 maddeden oluşmaktadır. İç tutarlılık anlamındaki güvenirlik katsayısı 0.81'dir. Araştırmacılar daha sonra Türk kültüründe iyi sonuçlar vermeyen 7 maddeyi çıkarmışlardır ve ölçek 18 maddelik haliyle kullanılmaktadır. Bu haliyle ölçekten alınabilecek en düşük puan 18; en yüksek puan ise 90'dır. Ölçekten alınan puanın yüksek

Çoban, E., & Demirtaşlı, R. N. (2017). Test yönergesi ve risk alma eğiliminin çoktan seçmeli testlerin psikometrik özelliklerine etkisi. *Journal of Human Sciences*, *14*(3), 2949-2967. doi:10.14687/jhs.v14i3.4449

olması, risk alma eğiliminin yüksek olduğu anlamına gelmektedir. Ölçeğin geliştirilmesi sırasında uygulama yapılan grubun cinsiyet ve eğitim durumu Tablo 2'de verilmiştir.

Tablo 2. Ölçeğin Uygulandığı Grubun Cinsiyet ve Eğitim Durumuna Göre Dağılımı

		Eğitim Durumu			
		7.Sınıf	**9.Sınıf**	**11.Sınıf**	**Üniversite**
Cinsiyet	**Kız**	35	35	35	35
	Erkek	35	35	35	35

Tablo 2'de görüldüğü üzere, katılımcıların 70'i üniversite öğrencisidir. Ölçek geliştirilirken üniversite öğrencileri de araştırmaya dâhil edilmesine rağmen, ölçeğin sadece üniversite öğrencilerinden oluşan bir grupta geçerli olup olmadığını belirlemek için doğrulayıcı faktör analizi yapılmıştır. Ayrıca ölçeğin sadece üniversite öğrencilerinden oluşan grupta güvenirliğine dair kanıt toplanmıştır.

Ölçeğin üniversite öğrencileri için geçerli ve güvenilir ölçme yapıp yapmadığını belirlemek için Gazi Üniversitesi Eğitim Fakültesinde eğitim gören 204 kız ve 59 erkek olmak üzere toplam 263 üniversite öğrencisine uygulama yapılmıştır. Kayıp ve uç değerler çıkarıldıktan sonra 196 kadın ve 49 erkek öğrenciden elde edilen verilerle geçerlik ve güvenirlik analizleri yapılmıştır.

Öncelikle ölçeğin yapı geçerliğine kanıt toplamak üzere doğrulayıcı faktör analizi yapılmıştır. Veriler normallik varsayımını sağlamadığı için doğrulayıcı faktör analizinde, asimptotik kovaryans matrisi, Ağırlıklı En Küçük Kareler-AEKK (Weighted Least Squares-WLS) kestirim yöntemi ile birlikte kullanılmıştır (Jöreskog ve Sörbom, 2001).

Yapılan analiz sonucunda bütün t değerleri 0.01($t_{0.01} = 2.58$) düzeyinde manidar olduğu için, risk alma eğilimi örtük yapısının tüm maddeleri anlamlı bir şekilde yordadığı görülmüştür. Fakat bu durum tek başına modelin iyi uyum gösterdiğini anlamına gelmemektedir. Doğrulayıcı faktör analizinde p değerlerinin manidar çıkması örneklem büyüklüğünden etkilenebildiğinden (Çokluk, Şekercioğlu ve Büyüköztürk, 2012), modelin uyumu hakkında karar verebilmek için uyum indekslerine bakılmıştır. Modele ilişkin uyum indeksleri Tablo 3'te verilmiştir.

Tablo 3. Uyum İndeksleri

Uyum İndeksi	**İyi Uyum***	**Kabul Edilebilir Uyum***	**Hesaplanan Değer**	**Uyum**
GFI	.95≤ GFI<1.00	.90≤ GFI<.95	0.94	İyi
AGFI	.90≤ AGFI<1.00	.85≤AGFI<.90	0.93	İyi
CFI	.97≤ CFI<1.00	.95≤ CFI<.97	0.87	Kötü
SRMR	0≤SRMR≤.05	.05<SRMR≤.10	0.22	Kötü
RMSEA	0≤RMSEA≤.05	.05<RMSEA≤.08	0.08	Kabul Edilebilir
RMSEA Güven Aralığı	Güven Aralığı≤.10		0.07 - 0.09	İyi
sd			135	-
X^2			367.83	-
X^2/sd	0≤χ^2/sd<2	2≤χ^2/sd≤3	2.72	Kabul edilebilir

*Kaynak: Schermellen-Engel, Moosbrugger ve Müller, 2003; Schumacker ve Lomax, 2010

Tablo 3'te görüldüğü üzere; GFI ve AGFI değerleri iyi uyumun göstergesidir. RMSEA ve X^2/sd değerleri model uyumu için kabul edilebilir düzeydedir. CFI ve SRMR değerleri ise model uyumu için kabul edilebilir değerlerden düşük çıkmıştır.

Sonuç olarak, model uyum indekslerinin genel olarak kabul edilebilir düzeyde olması ve t değerlerinin manidar çıkması, risk alma ölçeğinin sadece üniversite öğrencilerinden oluşan bir örneklemde kullanılabileceğini göstermektedir.

Çoban, E., & Demirtaşlı, R. N. (2017). Test yönergesi ve risk alma eğiliminin çoktan seçmeli testlerin psikometrik özelliklerine etkisi. *Journal of Human Sciences*, *14*(3), 2949-2967. doi:10.14687/jhs.v14i3.4449

Ölçeğin güvenirliğine kanıt toplamak amacıyla Cronbach alfa iç tutarlık katsayısı ve test tekrar test güvenirliği belirlenmiştir. Ölçeğin Cronbach alfa katsayısı 0.83; test-tekrar test güvenirlik katsayısı ise 0.95 olarak hesaplanmıştır.

Verilerin Toplanması

Gerekli izinler alındıktan sonra veri toplama süreci başlamıştır. Veriler toplanırken, öğrencilere öncelikle başarı testinin yanıt kâğıdı ve risk alma ölçeği dağıtılmıştır. Öğrenciler yanıt kâğıdındaki demografik bilgileri doldurduktan sonra, risk alma ölçeğini doldurmuşlardır. Ardından başarı testlerinin dağıtımına başlanmıştır. Başarı testleri yanıtlama yönergeleri farklılaşan dört farklı form halinde (A, B, C ve D) katılımcılara seçkisiz olarak dağıtılmıştır. Test dağıtılırken A, B, C ve D formlarının eşit sayısında uygulanmasına dikkat edilmiştir. Test dağıtıldıktan sonra, öğrencilerden test yönergesini dikkatle okumaları istenmiş ve uygulama başlatılmıştır. Her uygulama yaklaşık 40 dakika sürmüştür.

Verilerin Analizi

Araştırma kapsamında toplanan veriler, SPSS 17.0, Lisrel 8.71 ve Microsoft Office Excel 2007 yazılımları ile analiz edilmiştir.

Farklı yönergelerle uygulanan testlerin ham puan ortalamaları arasındaki farkı belirlemek için tek faktörlü varyans analizi kullanılmıştır. Benzer şekilde düzeltilmiş puan ortalamaları arasındaki farklar da tek faktörlü varyans analiziyle karşılaştırılmıştır.

Test puanlarına düzeltme yapılacağını belirten Y_2, Y_3, ve Y_4 ile uygulanan testlerin ham ve düzeltilmiş test puanları arasında manidar farklılık olup olmadığını belirlemek için ilişkili örneklemler için t-testi kullanılmıştır.

Farklı yönergelerle uygulanan testlerin iç tutarlılık anlamındaki KR-20 güvenirlik katsayılarını karşılaştırmak için Alsawalmeh ve Feldt (1994) tarafından önerilen şu eşitlik kullanılmıştır:

$$W = \frac{(1-\alpha_2)}{(1-\alpha_1)}$$

α_1:1.teste ait güvenirlik katsayısı

α_2:2.teste ait güvenirlik katsayısı

Bu eşitlikten elde edilen W değeri, N testleri alan örneklem büyüklüğü olmak üzere, $F_{(N_1,N_2)}$ değeri ile karşılaştırılmıştır.

Farklı yönergelerle uygulanan testlerin ham ve düzeltilmiş test puanları arasında manidar bir ilişki olup olmadığını belirlemek için Pearson Momentler Çarpımı Korelasyon katsayısı kullanılmıştır. Ölçme ve değerlendirme dersi başarı notları ile ham ve düzeltilmiş puanlar arasındaki korelâsyonu belirlemek için Pearson Momentler Çarpımı Korelasyon Katsayısı ve Sperman Sıra Farkları Korelâsyon Katsayısı hesaplanmıştır. Korelâsyon katsayılarının büyüklüğü, mutlak değer 0.00-0.30 arasında iken düşük; 0.30-0.70 arasında iken orta ve 0.70-1.00 arasında iken yüksek olarak değerlendirilmiştir (Büyüköztürk, 2012).Ham ve düzeltilmiş test puanlarının risk alma eğilimine göre farklılaşıp farklılaşmadığını belirlemek için t-test kullanılmıştır.

Çoban, E., & Demirtaşlı, R. N. (2017). Test yönergesi ve risk alma eğiliminin çoktan seçmeli testlerin psikometrik özelliklerine etkisi. *Journal of Human Sciences*, *14*(3), 2949-2967. doi:10.14687/jhs.v14i3.4449

BULGULAR

Bu bölümde toplanan verilerin analizi sonucunda elde edilen bulgulara yer verilmiştir. Bulgular araştırmanın alt problemleriyle aynı sırada verilmiştir.

Test Puan Ortalamalarının Karşılaştırılmasına İlişkin Bulgular

Araştırmanın birinci amacı doğrultusunda, çoktan seçmeli test farklı yönergelerle uygulanıp test istatistikleri, hem ham hem de düzeltilmiş puanlar için belirlenmiş ve bu puanlardan hesaplanan test ortalamaları ve KR-20 güvenirlik katsayıları arasında manidar fark olup olmadığı sınanmıştır.

Test dört farklı yönergeyle uygulandığında elde edilen **ham** puanlardan hesaplanan test istatistikleri Tablo 4'te verilmiştir.

Tablo 4. Ham Puanlardan Elde Edilen Test İstatistikleri

	Test			
İstatistik	**Y_1**	**Y_2**	**Y_3**	**Y_4**
N	55	55	55	55
Ortalama	14.65	13.33	12.91	14.15
Ortanca	15	13	14	14
Tepe Değeri	15	16	15	13
Standart Sapma	3.30	3.23	3.20	3.13
Varyans	10.90	10.41	10.23	9.79
Çarpıklık	-0.18	-0.46	-0.43	-0.17
Basıklık	-0.24	-0.44	-0.40	-0.45
Ranj	15	13	14	14

Tablo 4'te yer alan ortalama değerlerinin uygulanan yönergeye göre manidar bir şekilde farklılaşıp farklılaşmadığı araştırılmıştır.

Farklı yönergelerle uygulanan testlerin **ham** puan ortalamaları arasındaki farkı belirlemek için tek faktörlü varyans analizi kullanılmıştır. Analiz sonucu elde edilen bulgular Tablo 5'te verilmiştir.

Tablo 5. Ham Puan Ortalamalarına Ait Varyans Analizi Sonuçları

Varyansın Kaynağı	Kareler Toplamı	sd	Kareler Ortalaması	F	p	Manidar Fark
Yönergeler Arası	103.18	3.00	34.39	3.34	**0.02**	Y_1-Y_3
Yönergeler İçi	2227.56	216.00	10.31			
Toplam	2330.75	219.00				

Tablo 5'te görüldüğü üzere test farklı yönergelerle uygulandığında elde edilen ham puan ortalamalarının en az ikisi arasında manidar bir fark bulunmuştur ($F_{(3,216)}$=3.34; p<0.05). Test sonucu hesaplanan etki büyüklüğü ($\eta^2 = 0.04$) bu farkın orta düzeyde olduğunu göstermektedir (Can, 2013). Yapılan Tukey HSD testi sonuçlarına göre, herhangi bir düzeltme yapılmayacağını belirten yönerge 1 ile uygulanan testin ortalamasının ($\bar{X} = 14.65$), boş bırakılan sorular için belirli bir oranda puan verileceğini belirten yönerge 3 ile uygulanan testin ortalamasından ($\bar{X} = 12.91$) manidar olarak yüksek olduğu bulunmuştur. Yanlış yanıtların cezalandırılacağını belirten yönerge 2 ($\bar{X} = 13.33$) ve yönerge 4 ($\bar{X} = 14.15$) ile uygulanan testlerin ham puanları ortalamaları ise yönerge 1 ile uygulanan testin ortalamasından düşük olmasına rağmen, aralarında manidar farklılık yoktur. **Düzeltilmiş** puanlardan hesaplanan test istatistikleri Tablo 6'da verilmiştir.

Çoban, E., & Demirtaşlı, R. N. (2017). Test yönergesi ve risk alma eğiliminin çoktan seçmeli testlerin psikometrik özelliklerine etkisi. *Journal of Human Sciences*, *14*(3), 2949-2967. doi:10.14687/jhs.v14i3.4449

Tablo 6. Düzeltilmiş Puanlardan Elde Edilen Test İstatistikleri

	Test			
İstatistik	**Y_1**	**Y_2**	**Y_3**	**Y_4**
N	55	55	55	55
Ortalama	14.65	11.62	13.90	7.87
Ortanca	15	12	14.6	9
Tepe Değeri	15	10.25	15.80	10
Standart Sapma	3.30	3.76	2.89	5.61
Varyans	10.90	14.11	8.35	31.45
Çarpıklık	-0.18	-0.65	-0.49	-0.46
Basıklık	-0.24	0.16	-0.34	0.26
Ranj	15	16.25	12.4	27

Tablo 6'da yer alan ortalama değerlerinin uygulanan yönergeye göre manidar bir şekilde farklılaşıp farklılaşmadığı araştırılmıştır. Düzeltilmiş test puan ortalamaları arasındaki farkın manidar olup olmadığını belirlemek için tek faktörlü varyans kullanılmıştır. Analize başlamadan önce varsayımların sağlanıp sağlanmadığı kontrol edilmiştir. Levene Testi sonucuna göre varyansların homojenliği varsayımı sağlanmamıştır. Bu sebeple post-hoc testlerinden varyansların homojen olmadığı durumda seçilebilen Dunnet's C kullanılmıştır (Büyüköztürk, 2012).

Farklı yönergelerle uygulanan testlerin **düzeltilmiş** puan ortalamalarının karşılaştırılması sonucu elde edilen bulgular Tablo 7'de verilmiştir.

Tablo 7. Düzeltilmiş Puan Ortalamalarına Ait Varyans Analizi Sonuçları

Varyansın Kaynağı	**Kareler Toplamı**	**sd**	**Kareler Ortalaması**	**F**	**p**	**Manidar Fark**
Yönergeler Arası	1552.85	3.00	517.62	32.16	**0.00**	Y_1-Y_2,Y_1-Y_4,Y_2-Y_3,
Yönergeler İçi	3477.00	216.00	16.10			Y_2-Y_4,Y_3-Y_4
Toplam	5029.85	219.00				

Tablo 7'de görüldüğü üzere test farklı yönergelerle uygulandığında elde edilen düzeltilmiş puan ortalamalarının en az ikisi arasında manidar bir fark bulunmuştur ($F_{(3,216)}$=32.16; p<0.05). Test sonucu hesaplanan etki büyüklüğü ($\eta^2 = 0.03$) bu farkın orta düzeyde olduğunu göstermektedir (Can, 2013). Yapılan Dunnet's C testi sonuçlarına göre, herhangi bir düzeltme yapılmayacağını belirten yönerge 1 ile uygulanan testin ortalamasının ($\bar{X} = 14.65$), yanlış yanıtların cezalandırılacağını belirten yönerge 2 ($\bar{X} = 11.62$) ve yönerge 4 ($\bar{X} = 7.87$) ile uygulanan testlerin ortalamasından manidar olarak yüksek olduğu bulunmuştur. Yönerge 3 ($\bar{X} = 13.90$) ile uygulanan testin ortalaması, yönerge 2 ve yönerge 4 ile uygulanan testlerin ortalamasından manidar olarak yüksektir. Ayrıca yanlış yanıtların cezalandırılacağını bildiren yönergelerden yönerge 2 ile uygulanan testin ortalaması, yönerge 4 ile uygulanan testin ortalamasından manidar olarak yüksek çıkmıştır.

Test puanlarına düzeltme yapılacağını belirten Y_2, Y_3 ve Y_4 ile uygulanan testlerin ortalamalarının düzeltme sonucunda manidar olarak farklılık gösterip göstermediğini belirlemek için **ilişkili örneklemler için t-testi** yapılmıştır. Yapılan t-testi sonuçları Tablo 8'de verilmiştir. Ayrıca ham ve düzeltilmiş test puanları arasında manidar bir ilişki olup olmadığını belirlemek için Pearson Momentler Çarpımı Korelasyon Katsayısı hesaplanmıştır. Hesaplanan korelasyon katsayıları Tablo 9'da verilmiştir.

Çoban, E., & Demirtaşlı, R. N. (2017). Test yönergesi ve risk alma eğiliminin çoktan seçmeli testlerin psikometrik özelliklerine etkisi. *Journal of Human Sciences*, *14*(3), 2949-2967. doi:10.14687/jhs.v14i3.4449

Tablo 8. Ham ve Düzeltilmiş Puan Ortalamalarının Karşılaştırılmasına İlişkin t-Testi Sonuçları

Test	Puan Türü	N	$\bar{X}$	S	sd	t	p
Y_2	Ham Puan	55	13.31	3.21	54	13.97	0.00
	Düzeltilmiş Puan	55	11.53	3.74			
Y_3	Ham Puan	55	12.91	3.20	54	-7.33	0.00
	Düzeltilmiş Puan	55	13.98	2.83			
Y_4	Ham Puan	55	14.15	3.13	54	14.26	0.00
	Düzeltilmiş Puan	55	7.87	5.61			

Tablo 8 incelendiğinde, yönerge 2 ile uygulanan testin ham puan ortalaması $(\bar{X} = 13.11)$ ile düzeltilmiş puan ortalaması $(\bar{X} = 11.53)$ arasında manidar fark olduğu görülmektedir ($t_{(54)}$=13.97; p<0.05). Bu yönergeyle düzeltme yapıldığında test puanları azalmıştır. Benzer şekilde, yönerge 3 ile uygulanan testin ham puan ortalaması $(\bar{X} = 12.91)$ ile düzeltilmiş puan ortalaması $(\bar{X} = 13.98)$ arasında manidar fark olduğu görülmektedir ($t_{(54)}$= -7.33; p<0.05). Bu yönergede boş yanıtlar için puan verildiği için ortalama yükselmiştir. Yönerge 4 ile uygulanan testin ham puan ortalaması $(\bar{X} = 14.15)$ ile düzeltilmiş puan ortalaması $(\bar{X} = 7.87)$ arasında manidar fark bulunmuştur ($t_{(54)}$=14.26; p<0.05). Bu yönerge yanlış yanıtları ağır bir şekilde cezalandırdığı için test puanları önemli derecede azalmıştır.

Tablo 9. Ham ve Düzeltilmiş Test Puanları Arasındaki İlişkiye Dair Korelâsyon Analizi Sonuçları

Test	N	r	r^2	p
Y_2	55	0.98	0.95	0.00
Y_3	55	0.94	0.89	0.00
Y_4	55	0.87	0.76	0.00

Tablo 9'da görüldüğü üzere, farklı düzeltme formülleri ile düzeltme yapılacağını belirten yönergelerle uygulanan testlerin tamamında, ham ve düzeltilmiş puanlar arasında pozitif yönde yüksek düzeyde manidar ilişki bulunmuştur (p<0.05).

Farklı Yönergelerle Uygulanan Testlerin Geçerlik ve Güvenirliğinin Karşılaştırılmasına İlişkin Bulgular

Ham puanlardan elde edilen KR-20 güvenirlik katsayılarının karşılaştırılması Feldt testi ile yapılmıştır. Feldt testi sonucunda hesaplanan W değerleri Tablo 10'da verilmiştir. Y_i'ler (i=1,2,3,4) sırasıyla yönerge 1, yönerge 2, yönerge 3 ve yönerge 4 uygulanan testleri simgelemektedir.

Tablo 10. Ham Puanlardan Elde Edilen Güvenirlik Değerlerinin Karşılaştırılmasına İlişkin Sonuçlar

		W İSTATİSTİĞİ			
Test	KR-20	Y_1	Y_2	Y_3	Y_4
Y_1	0.61	1			
Y_2	0.61	1	1		
Y_3	0.59	1.05	1.05	1	
Y_4	0.56	1.13	1.13	1.07	1

Çoban, E., & Demirtaşlı, R. N. (2017). Test yönergesi ve risk alma eğiliminin çoktan seçmeli testlerin psikometrik özelliklerine etkisi. *Journal of Human Sciences*, *14*(3), 2949-2967. doi:10.14687/jhs.v14i3.4449

Tablo 10'da görüldüğü üzere çoktan seçmeli ölçme ve değerlendirme dersi başarı testi, farklı düzeltme formülleri kullanılacağını belirten dört yönergeyle uygulandığında ham puanlardan hesaplanan güvenirlik katsayıları arasında manidar fark bulunmamıştır. Testler ikili olarak karşılaştırıldığında tüm W değerlerinin $F_{(54,54)} \cong 1.53$'ten küçük olduğu görülmektedir.

Düzeltilmiş puanlardan elde edilen KR-20 güvenirlik katsayılarının karşılaştırmak için yapılan Feldt testi sonucunda hesaplanan W değerleri Tablo 11'de verilmiştir.

Tablo 11. Düzeltilmiş Puanlardan Elde Edilen Güvenirlik Değerlerinin Karşılaştırılmasına İlişkin Sonuçlar

		W İSTATİSTİĞİ			
Test	**KR-20**	**Y_1**	**Y_2**	**Y_3**	**Y_4**
Y_1	0.61	1			
Y_2	0.72	1.39	1		
Y_3	0.48	0.75	0.54	1	
Y_4	0.89	**3.55***	**2.55***	**4.73***	1

*p<0.05

Tablo 11 incelendiğinde, yönerge 4 ile uygulanan testin güvenirlik katsayısının yönerge 1 (W=3.55; p<0.05), yönerge 2 (W=2.55; p<0.05) ve yönerge 3 (W=4.73; p<0.05) ile uygulanan testlerin güvenirlik katsayılarından manidar olarak **yüksek** olduğu görülmektedir. Diğer yönergelerle uygulanan testlerin düzeltilmiş puanlardan elde edilen güvenirlik katsayıları arasında manidar fark yoktur ($W<F_{(54,54)} \cong 1.53$).

Test puanlarına düzeltme yapılacağını bildiren yönerge 2, yönerge 3 ve yönerge 4 ile uygulanan testlerin KR-20 güvenirlik katsayılarının düzeltme yapıldığı durumda manidar olarak farklılık gösterip göstermediğini belirlemek için yönerge içi Feldt testi yapılmıştır. Elde edilen sonuçlar Tablo 12'de verilmiştir.

Tablo 12. Ham ve Düzeltilmiş Puanlara Göre Hesaplanan Güvenirlik Katsayılarının Yönerge İçi Karşılaştırılmasına İlişkin Sonuçlar

Test	Puan Türü	N	KR-20	W
Y_2	**Ham Puan**	55	0.61	1.39
	Düzeltilmiş Puan	55	0.72	
Y_3	**Ham Puan**	55	0.59	0.54
	Düzeltilmiş Puan	55	0.48	
Y_4	**Ham Puan**	55	0.56	**4.00***
	Düzeltilmiş Puan	55	0.89	

*p<0.05

Tablo 12'de görüldüğü üzere, yönerge 2 (W=1.39; p>0.05) ve yönerge 3 (W=0.54; p>0.05) ile uygulanan testlerin ham ve düzeltilmiş puanlarına göre hesaplanan güvenirlik katsayıları yönerge içi manidar farklılık göstermemektedir. Yönerge 4 ile uygulanan testin ise ham (0.56) ve düzeltilmiş (0.89) puanlara göre hesaplanan güvenirlik katsayıları arasında manidar fark bulunmuştur (W=4.00; p<0.05). Testin düzeltilmiş puanlarına göre hesaplanan güvenirlik katsayısı, ham puana göre hesaplanan güvenirlik katsayısına göre manidar olarak daha yüksek çıkmıştır.

Farklı yönergelerle uygulanan testlerin geçerlikleri arasında fark olup olmadığını belirlemek için ham ve düzeltilmiş test puanlarının ölçüt geçerliği hesaplanmıştır. Ölçüt geçerliği hesaplanırken, öğrencilerin ölçme ve değerlendirme dersi başarı notları ölçüt olarak kullanılmıştır. Normallik varsayımının sağlandığı durumlarda Pearson Momentler Çarpımı Korelasyon Katsayısı

Çoban, E., & Demirtaşlı, R. N. (2017). Test yönergesi ve risk alma eğiliminin çoktan seçmeli testlerin psikometrik özelliklerine etkisi. *Journal of Human Sciences*, *14*(3), 2949-2967. doi:10.14687/jhs.v14i3.4449

hesaplanmıştır. Normallik varsayımının sağlanmadığı durumlarda ise Sperman Sıra Farkları Korelâsyon Katsayısı hesaplanmıştır.

Öğrencilerin ölçme ve değerlendirme dersi başarı notlarının ham ve düzeltilmiş test puanlarıyla korelâsyonuna ait sonuçlar Tablo 13'te verilmiştir.

Tablo 13. Ölçme Değerlendirme Dersi Başarı Notunun Ham ve Düzeltilmiş Test Puanlarıyla İlişkiye Dair Korelâsyon Analizi Sonuçları

Puan Türü	Test	N	r	r^2	p
Ham	Y_1	55	0.55*	0.31	**0.00**
	Y_2	55	0.15*	0.02	0.27
	Y_3	55	0.48**	0.23	**0.00**
	Y_4	53	0.01**	0.00	0.96
Düzeltilmiş	Y_1	55	0.55*	0.31	**0.00**
	Y_2	55	0.03*	0.00	0.85
	Y_3	55	0.53**	0.28	**0.00**
	Y_4	53	0.08**	0.01	0.59

*Pearson Momentler Çarpımı Korelâsyon Katsayısı
**Sperman Sıra Farkları Korelâsyon Katsayısı

Tablo 13'te görüldüğü üzere, yalnızca yönerge 1 (r=0.55) ve yönerge 3 (r=0.48) ile uygulanan testlerin ham puanları ölçme ve değerlendirme dersi başarı notuyla manidar ilişki göstermektedir ($p<0.05$). Benzer şekilde yalnızca yönerge 1 (r=0.55) ve yönerge 3 (r=0.53) ile uygulanan testleri alan öğrencilerin düzeltilmiş puanları ölçme ve değerlendirme dersi başarı notuyla manidar ilişki göstermektedir ($p<0.05$). Ham ve düzeltilmiş test puanlarının ölçüt geçerliği dikkate alındığında, geçerliği en yüksek olan test herhangi bir düzeltme yapılmayacağını belirten yönerge 1 ile uygulanan testtir. Elde edilen bu geçerlik katsayıları arasında manidar fark olup olmadığını belirlemek için Fisher z dönüşümü yapılmıştır. Fakat Fisher z dönüşümü yalnızca Pearson korelasyon katsayısı için yapılabildiği için sadece yönerge 1 ve yönerge 2 ile uygulanan testlerin geçerlikleri arasındaki manidar fark olup olmadığı test edilmiştir. Düzeltme yapılmayacağını belirten Y_1 ile uygulanan testin ham puanlara göre hesaplanan geçerlik katsayısı (r=0.55), Y_2 ile uygulanan testin geçerlik katsayısından (r=0.15) manidar olarak yüksektir ($z=2.41>1.96$). Düzeltilmiş test puanlarına göre hesaplanan geçerlik katsayıları dikkate alındığında, Y_1 ile uygulanan testin geçerlik katsayısının (r=0.55), Y_2 ile uygulanan testin geçerlik katsayısından (r=0.03) manidar olarak yüksek olduğu görülmektedir ($z=3.05>1.96$). Klasik düzeltme formülüne göre düzeltme yapılacağını belirten Y_2ile uygulanan testin ham ve düzeltilmiş puanlara göre hesaplanan geçerlik katsayıları (sırasıyla 0.15 ve 0.03) arasında manidar fark bulunmamıştır ($z=0.63<1.96$).

Test Puanlarının Risk Alma Düzeyine Göre Değişimine İlişkin Bulgular

Risk alma eğilimini ham ve düzeltilmiş test puanlarında nasıl bir değişime yol açtığını belirlemek için risk alma eğilimi düşük ve yüksek olan iki grup oluşturulmuştur. Risk alma eğilimi belirlenirken, risk alma ölçeği puanlarının ortancası kesme noktası olarak kullanılmıştır. Risk alma ölçeği puanı ortancanın altında olan katılımcıların risk alma düzeyi **düşük**; üzerinde olan katılımcıların risk alma düzeyi ise **yüksek** olarak belirlenmiştir. Daha sonra bu iki grubun ham ve düzeltilmiş test puanları karşılaştırılmıştır.

Risk alma eğilimi farklı öğrencilerin ham puan ortalamaların karşılaştırılması sonucu elde edilen bulgular Tablo 14'te verilmiştir.

Çoban, E., & Demirtaşlı, R. N. (2017). Test yönergesi ve risk alma eğiliminin çoktan seçmeli testlerin psikometrik özelliklerine etkisi. *Journal of Human Sciences*, *14*(3), 2949-2967. doi:10.14687/jhs.v14i3.4449

Tablo 14. Ham Test Puanlarının Risk Alma Düzeyine Göre Karşılaştırılmasına İlişkin t-Testi Sonuçları

Test	Risk Alma Düzeyi	N	$\bar{X}$	S	sd	t	p
Y_1	**Düşük**	27	14.78	3.20	53	0.27	0.79
	Yüksek	28	14.54	3.45			
Y_2	**Düşük**	33	13.39	3.22	52	0.07	0.95
	Yüksek	21	13.33	3.28			
Y_3	**Düşük**	25	13.64	3.20	52	1.37	0.18
	Yüksek	29	12.48	3.01			
Y_4	**Düşük**	24	13.04	2.84	53	-2.40	**0.02**
	Yüksek	31	15.00	3.12			

Tablo 14'te görüldüğü üzere, **ham** test puan ortalaması **yalnızca** testi yönerge 4 ile alan grupta risk alma düzeyine göre manidar farklılık göstermektedir ($t_{(53)}$=-2.40; p<0.05). Yüksek risk alma düzeyindeki katılımcıların ham test puanları ($\bar{X} = 15.00$), düşük risk alma düzeyindeki katılımcıların ham test puanlarından ($\bar{X} = 13.04$) manidar olarak yüksektir.

Düşük ve yüksek risk alma düzeyine sahip öğrencilerin düzeltilmiş test puanlarının karşılaştırılması sonucu elde edilen bulgular Tablo 15'te verilmiştir.

Tablo 15. Düzeltilmiş Test Puanlarının Risk Alma Düzeyine Göre Karşılaştırılmasına İlişkin t-Testi Sonuçları

Test	Risk Düzeyi	N	$\bar{X}$	S	sd	t	p
Y_1	**Düşük**	27	14.78	3.20	53	0.27	0.79
	Yüksek	28	14.54	3.45			
Y_2	**Düşük**	33	11.78	3.75	52	0.47	0.64
	Yüksek	21	11.29	3.83			
Y_3	**Düşük**	25	14.73	2.66	52	1.63	0.11
	Yüksek	29	13.52	2.77			
Y_4	**Düşük**	24	6.92	5.17	53	-1.12	0.27
	Yüksek	31	8.61	5.90			

Tablo 15'te görüldüğü üzere, yönerge fark etmeksizin risk alma düzeyi farklı olan katılımcıların **düzeltilmiş** test puanları arasında manidar fark bulunmamaktadır (p>0.05).

TARTIŞMA

Farklı yönergelerle uygulanan testlerden yalnızca herhangi bir düzeltme yapılmayacağını belirten Y_1 ve boş bırakılan maddeler için puan verileceğini belirten Y_3 ile uygulanan testlerin ham puan ortalamaları arasında manidar fark vardır. Klasik düzeltme yöntemi ile düzeltme yapılacağını belirten Y_2 ile uygulanan testin ham puan ortalaması ile Y_1 ile uygulanan testin ham puan ortalaması arasında manidar fark yoktur. Bu bulgu Çelen ve Çıkrıkçı-Demirtaşlı (2006)'nın bulgusuyla paralellik göstermektedir.

Test puanlarına düzeltme uygulandığında, doğru sayısından yanlış sayısı çıkarılarak puanlama yapılacağını belirten Y_4 ile uygulanan testin ortalamasının diğer testlerin ortalamasından düşük olduğu görülmektedir. Bu durum yapılan düzeltme sonucunda test puanlarında önemli bir azalış olduğunu göstermektedir. Prieto ve Delgado (1999a) yaptıkları çalışmada benzer bir bulguya

Çoban, E., & Demirtaşlı, R. N. (2017). Test yönergesi ve risk alma eğiliminin çoktan seçmeli testlerin psikometrik özelliklerine etkisi. *Journal of Human Sciences*, *14*(3), 2949-2967. doi:10.14687/jhs.v14i3.4449

ulaşmışlardır. Klasik düzeltme formülüne dayalı düzeltme yapılacağını belirten Y_2 ile uygulanan testin ortalaması, Y_1 ve Y_3 ile uygulanan testlerin ortalamasından düşüktür. Bu testin düzeltilmiş puan ortalamasının herhangi bir düzeltme yapılmayacağını belirten Y_1 ile uygulanan testin ortalamasından düşük çıkması, alan yazındaki birçok çalışmanın sonucuyla tutarlıdır (Alnabhan, 2002; Betts, Elder, Hartley ve Trueman, 2009; Swineford ve Miller, 1953; Umay, 1998). Bu iki testin ham puanları arasında fark olmamasına rağmen böyle bir sonuç çıkması, Y_2 ile uygulanan testin düzeltilmiş puanlarının ham puanlardan manidar olarak düşük olmasından kaynaklanmaktadır. Düzeltme yapılacağını belirten Y_2, Y_3 ve Y_4 ile uygulanan testlere düzeltme uygulandığında, Y_2 ve Y_4 ile uygulanan testlerin ortalamaları manidar olarak azalırken Y_3 ile uygulanan testin ortalaması manidar olarak artmıştır. Düzeltme yapıldığında test puanlarında manidar farklılıklar oluşmasına rağmen, ham ve düzeltilmiş puanlar arasında yüksek korelasyon bulunmaktadır. Alan yazındaki çalışmalarla (Angoff ve Schrader, 1981; Çelen ve Çıkrıkçı-Demirtaşlı, 2006) tutarlı olan bu bulgu, düzeltme yapıldığında öğrencilerin başarı sırasının değişmediğini göstermektedir. En yüksek korelasyon, Y_2 ile uygulanan testin ham ve düzeltilmiş puanları arasındadır.

Farklı yönergelerle uygulanan testlerin ham puanlara göre hesaplanan güvenirlik katsayıları arasında fark yoktur. Prieto ve Delgado (1999a) yaptıkları çalışmada benzer bir sonuca ulaşmışlardır. Çelen ve Çıkrıkçı-Demirtaşlı (2006) ise Y_2 ile uygulanan testin ham puanlara göre hesaplanan güvenirliğini Y_1 ile uygulanan testin güvenirliğine göre yüksek bulmuşlardır. Güvenirlik katsayıları düzeltilmiş puanlara göre hesaplandığında, Y_2 ve Y_4 ile uygulanan testlerin güvenirlik katsayısı yükselirken Y_3 ile uygulanan testin güvenirlik katsayısı düşmüştür. Fakat yalnızca Y_4 ile uygulanan testin güvenirlik katsayısındaki değişim manidardır. Ayrıca bu yönergeyle uygulanan testin düzeltilmiş puana göre hesaplanan güvenirlik katsayısı diğer testlerin güvenirlik katsayılarından manidar olarak yüksektir. Prieto ve Delgado (1999a) ise benzer bir çalışmada Y_4 uygulanan testin düzeltilmiş puanlara göre hesaplanan güvenirlik katsayısını, diğer yönergelerle uygulanan testlerin güvenirlik katsayısından yüksek bulmuşlardır. Bu araştırmada farklı bir bulguya ulaşılması, Y_4 ile düzeltme yapıldığında test varyansının önemli derecede artmasından kaynaklanmıştır (bk. Tablo 4 ve Tablo 6). Testin 9.79 olan ham puan varyansı, puanlar düzeltildiğinde 31.45'e çıkmıştır. Alan yazında genellikle, güvenirlik katsayısının klasik düzeltme yöntemine göre düzeltme yapıldığında (Y_2) düzeltme yapılmayan duruma (Y_1) göre değişip değişmediğine yönelik çalışmalar yapılmış ve farklı sonuçlara ulaşılmıştır. Bazı araştırmalarda Y_2 ile uygulanan testin güvenirliği Y_1ile uygulanan teste göre daha yüksek bulunmuştur (Alnabhan, 2002; Lord, 1975). Diğer taraftan Y_1 ile uygulanan testin güvenirliğini Y_2ile uygulanan testten yüksek bulan araştırmacılar da bulunmaktadır (Glass ve Wiley, 1982; Socan, 2009). Alan yazında mevcut araştırmaya benzer bulgulara ulaşan çalışmalar da vardır. Frary (1982), bu çalışmada olduğu gibi, Y_2 ile uygulanan testin düzeltilmiş puanlara göre hesaplanan güvenirliğini Y_1 ile uygulanan testinkine göre biraz yüksek bulmuştur. Fakat aralarında manidar fark olmadığını belirtmiştir. Abu Sayf (1975), Y_1, Y_2 ve Y_3 ile uygulanan testleri karşılaştırdığı çalışmada, testlerin düzeltilmiş puanlara göre hesaplanan güvenirlikleri arasında manidar fark bulmamıştır.

Herhangi bir düzeltme yapılmayacağını belirten Y_1 ile uygulanan testin ham puanlara göre hesaplanan geçerlik katsayısı diğer testlerin geçerlik katsayılarından yüksektir. Düzeltme yapılacağını belirten yönergelerle uygulanan testlerden geçerliği en yüksek olan Y_3 ile uygulanan testtir. Benzer durum düzeltilmiş puanlara göre hesaplanan geçerlik katsayıları için de geçerlidir. Herhangi bir düzeltme yapılmayacağını belirten Y_1 ile uygulanan testin geçerlik katsayısının diğer testlerin geçerlik katsayısından yüksektir. Fakat yalnızca Y_2 ile uygulanan testin geçerliğinden manidar olarak yüksektir[4]. Klasik düzeltme yöntemine göre düzeltme yapılacağını belirten Y_2 ile uygulanan testin

[4]Uygun istatistiksel teknik olmadığı için diğer testlerin geçerliğinden manidar olarak yüksek olup olmadığı test edilememiştir.

Çoban, E., & Demirtaşlı, R. N. (2017). Test yönergesi ve risk alma eğiliminin çoktan seçmeli testlerin psikometrik özelliklerine etkisi. *Journal of Human Sciences*, *14*(3), 2949-2967. doi:10.14687/jhs.v14i3.4449

ham ve düzeltilmiş puanlara göre hesaplanan geçerlik katsayıları arasında manidar fark yoktur. Frary (1982) ve Çelen ve Çıkrıkçı-Demirtaşlı (2006)'nın bulguları bu bulguyu desteklemektedir. Diğer taraftan ham puanları (Socan, 2009) veya düzeltilmiş puanları (Prihoda, Pinckard, McMahan ve Jones, 2006) daha geçerli bulan çalışmalar da vardır.

Risk alma eğilimi genel olarak test puan ortalamalarını manidar olarak değiştirmemiştir. Yalnızca Y_4 ile uygulanan testin ham puan ortalaması risk alma eğilimine göre farklılık göstermektedir. Risk alma eğilimi yüksek olan grubun test puanları ortalaması daha yüksektir. Fakat düzeltme uygulandığında düşük ve yüksek risk alma eğilimine sahip gruplar arasındaki bu fark ortadan kalkmıştır.

SONUÇ VE ÖNERİLER

Dört farklı yönergeyle uygulanan testlerin ham puan ortalamaları uygulanan yönergeye göre farklılık göstermektedir. Test yönergesinde düzeltme yapılmayacağı belirtildiğinde test puanları daha yüksek çıkmaktadır. Test puanlarına düzeltme uygulandığında test puanlarında önemli derecede değişmektedir. Fakat ham ve düzeltilmiş test puanları arasındaki korelasyon yüksektir. Düzeltme yapıldığında test puanlarının değeri değişmesine rağmen sıralaması değişmemiştir. Test puanlarına düzeltme yapıldığında öğrencilerin başarı sıraları değişmediği için sıralamaya dayalı sınavlarda düzeltme formüllerinin kullanılmasında sakınca yoktur. Özellikle, klasik düzeltme yöntemiyle uygulanan testin ham ve düzeltilmiş test puanları arasındaki korelasyon çok yüksek olduğu için sıralamaya dayalı değerlendirme yapılan sınavlarda bu yöntem kullanılmalıdır. Fakat test puanları önemli derecede değiştiği için ölçüt dayanaklı değerlendirmelerde düzeltilmiş puanlar kullanılmamalıdır.

Farklı yönergelerle uygulanan testlerin ham puanlara göre hesaplanan güvenirlik katsayıları arasında fark yoktur. Düzeltilmiş puanlara göre hesaplanan güvenirlik katsayıları incelendiğinde, klasik düzeltme yöntemi ile düzeltme yapıldığında güvenirlik katsayısında artış olmasına rağmen bu artışın manidar olmadığı görülmektedir. Boş bırakılan maddeler için puan eklendiğinde güvenirlik katsayısında manidar olmayan bir düşüş gerçekleşmiştir. Herhangi bir düzeltme yapılmayacağını belirten yönerge ile uygulanan testin geçerliği daha yüksektir. Bu yüzden gerekli olmadıkça test yönergesinde düzeltme yapılacağına ilişkin açıklama konulmamalı ve düzeltme formülleri kullanılmamalıdır. Yönergeye konulan açıklamalarla, öğrencilerin tüm maddelere yanıt vermesi sağlanmalıdır.

Genel olarak ham test puanları katılımcıların risk alma düzeyine göre farklılık göstermemektedir. Sadece testi Y_4 ile alan grupta ham test puanları risk alma düzeyine göre farklılık göstermektedir. Düzeltilmiş test puanları da katılımcıların risk alma düzeyine göre farklılık göstermemektedir. Bu çalışmada kadın sayısının erkek sayısından fazladır. Morrongiello ve Rennie (1998) kadınların risk alma eğiliminin erkeklerden düşük olduğunu bulmuşlardır. Bu sebeple kadın ve erkek sayısının birbirine eşit ya da yakın olduğu gruplarda çalışma tekrarlandığında farklı sonuçlara ulaşılabilir.

KAYNAKLAR

Abu-Sayf, F. K. (1975). Relative effectiveness of the conventional formula score. *The Journal of Educational Research*, *69*(4), 160-162.

Albanese, M. A. (1988). The projected impact of the correction for guessing on individual scores. *Journal of Educational Measurement*, *25*(2), 149-157.

Alnabhan, M. (2002). An emprical investigation of the effects of three methods of handling guessing and risk taking on the psychometric indices of a test. *Social Behavior and Personality*, *30*(7), 645-652.

Çoban, E., & Demirtaşlı, R. N. (2017). Test yönergesi ve risk alma eğiliminin çoktan seçmeli testlerin psikometrik özelliklerine etkisi. *Journal of Human Sciences*, *14*(3), 2949-2967. doi:10.14687/jhs.v14i3.4449

Alsawalmeh, Y. M., and Feldt, L. S. (1994). A modification of Feldt's test of the equality of two dependent alpha coefficients. *Psychometrika*, *59*(1), 49-57.

Angoff, W. H., and Schrader, W. B. (1981). A study of alternative methods for equating rights scores to formula scores. *ETS Research Report Series*.

Bayar, N., and Sayıl, M. (2005). Brief report: risk taking behaviors in a non-western urban adolescent sample. *Journal of Adolescence*, *28*, 671-676.

Baykul, Y. (2010). *Eğitimde ve psikolojide ölçme: klasik test teorisi ve uygulaması* (2. Baskı). Ankara: Pegem Akademi Yayınları.

Betts, L. R., Elder, T. J., Hartley, J., and Trueman, M. (2009). Does correction for guessing reduce students' performance on multiple-choice examinations? Yes? No? Sometimes?. *Assessment &Evaluation in Higher Education*, *34*(1), 1-15.

Burton, R. F. (2002). Misinformation, partial knowledge and guessing in true/false tests. *Medical Education*, 36(9), 805-811.

Büyüköztürk, Ş., (2012). *Sosyal bilimler için veri analizi el kitabı*. (17. Baskı). Ankara: Pegem Akademi.

Can, A. (2013). *SPSS ile Bilimsel Araştırma Sürecinde Nicel Veri Analizi*. Ankara: Pegem Akademi Yayınları.

Creswell, J. W. (2012). *Educational research: Planning, conducting and evaluating qualitative research* (4th Edition*)*. Boston: Pearson Education.

Crocker, L., and Algina, J. (1986). *Introduction to classical ve modern test theory*. Belmont: Wadsworth Group.

Cross, L. H., and Frary, R. B. (1977). AN empirical test of Lord's theoretical results regarding formula scoring of multiple-choice tests. *Journal of Educational Measurement*, *14*(4), 313-321.

Çelen, Ü., ve Demirtaşlı, N. Ç. (2006). Düzeltme yönergesinin testin psikometrik özelliklerine etkisi. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi, 30*(2006), 82-91.

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

Ebel, R. L. (1965). *Measuring educational achievement*. New Jersey: Prentice-Hall.

Espinosa, M. P., and Gardeazabal, J. (2010). Optimal correction for guessing in multiple-choice tests. *Journal of Mathematical Psychology*, *54*(5), 415-425.

Frary, R. B. (1982). A simulation study of reliability and validity of multiple-choice test scores under six response-scoring modes. *Journal of Educational and Behavioral Statistics*, *7*(4), 333-351.

Glass, G. V., and Wiley, D. E. (1964). Formula scoring and test reliability. *Journal of Educational Measurement*, *1*(1), 43-49.

Haladayna, T. M. (1997). *Writing test İtems to evaluate higher order thinking*. USA: Ally ve Bacon.

Jöreskog, K. G., and Sörbom, D. (1996). *LISREL 8 user's reference guide*. Scientific Software International.

Kubinger, K. D., and Wolfsbauer, C. (2010). On the risk of certain psychotechnological response options in multiple-choice tests: Does a particular personality handicap examinees?. *European Journal of Psychological Assessment*, *26*(4), 302-308.

Kurz, T. B. (1999). *A review of scoring algorithms for multiple-choice tests*. San Antonio: Southwest Educational Research Association.

Lord, F. M. (1975). Formula scoring and number-right scoring. *Journal of Educational Measurement*, *12*(1), 7-11.

Morrengiello, B. A., and Rennie, H. (1998). Why do boys engage in more risk taking than girls? The role of attributions, beliefs, and risk appraisals. *Journal of Pediatric Psychology, 23*(1), 33-43.

Muijtjens, A. M. M, Mameren, H., Hoogenboom, R. J. I, Evers, J.L.H., and Vleuten, C. P. M. (1999), The effect of a 'don't know' option on test scores: number-right and formula scoring compared. *Medical Education*, *33*, 267-275.

Popham, W. J. (2000). *Modern educational measurement* (3th Edition). Boston: Ally and Bacon.

Çoban, E., & Demirtaşlı, R. N. (2017). Test yönergesi ve risk alma eğiliminin çoktan seçmeli testlerin psikometrik özelliklerine etkisi. *Journal of Human Sciences*, *14*(3), 2949-2967. doi:10.14687/jhs.v14i3.4449

Prieto, G., and Delgado, A. R. (1999a). The effect of instructions on multiple-choice test scores. *European Journal of Psychological Assessment, 15*(2), 143–150.

Prihoda, T. J., Pinckard, R. N., McMahan, C. A., and Jones, A. C. (2006). Correcting for guessing increases validity in multiple-choice examinations in an oral and maxillofacial pathology course. *Journal of Dental Education*, *70*(4), 378-386.

Reid, J. F. (1976). Scoring multiple- choice exams. The *Journal of Economic Education,* 8, 55-59.

Rowley, G. L., and Traub, R. E. (1977). Formula scoring, number right scoring, and test-taking strategy. *Journal of Educational Measurement, 14*(1), 15-22.

Rubio, V. J., Hernández, J. M., Zaldívar, F., Márquez, O., and Santacreu, J. (2010). Can we predict risk-taking behavior?: Two behavioral tests for predicting guessing tendencies in a multiple-choice test. *European Journal of Psychological Assessment*, *26*(2), 87-94.

Sabers, D., and Feldt, L. (1968). An empirical study of the effect of the correction for chance success on the reliability and validity of an aptitude test. *Journal of Educational Measurement, 5*(3), 251-258.

Schermelleh-Engel, K., Moosbrugger, H., and Müller, H. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research Online*, *8*(2), 23-74.

Scumacker, R. E., and Lomax, R. G. (2010). *A beginner's guide to structural equation modeling.*(3rd Editon). New York: Taylor And Francis Group.

Socan, G. (2009).Scoring of multiple choice items by means of internal weighting. *Review of Psychology*, *16(2)*, 77-85.

Swineford, F., and Miller, P. M. (1953). Effects of directions regarding guessing on item statistics of a multiple-choice vocabulary test. *Journal of Educational Psychology*, *44*(3), 129-139.

Thorndike, R. M. (2005). *Measurement and evaluation in psycholgy and education* (7th Edition). New Jersey: Pearson Education.

Ulutaş, S. (2003). *Genel liselerdeki öğretmenlerin ölçme ve değerlendirme alanındaki yeterlikleri ile ölçme ve değerlendirme ilkelerini uygulama düzeylerinin araştırılması* (Yayımlanmamış yüksek lisans tezi). Ankara Üniversitesi Eğitim Bilimleri Enstitüsü, Ankara.

Umay, A. (1998). Seçmeli derslerde yanıtlayıcı davranışları ve şans başarısının elimine edilmesi işlemlerine ilişkin bazı öneriler. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi*, *14*, 54-61.

Zimmerman, D. W., and Williams, R. H. (1965). Effect of chance success due to guessing on error of measurement in multiple-choice tests. *Psychological Reports*, *16*(3), 1193-1196.

Extended English Abstract

Introduction

Multiple choice tests are commonly used in educational measurement because of objective, reliable and easy scoring. However, there is a very important limitation of these tests. Test takers may be able to give correct answer to multiple choice items by random guessing (Baykul, 2010, p. 393; Crocker and Algina, 1986, p. 313 ; Zimmerman and Williams, 1965). As a result of this situation, reliability (Zimmerman and Williams, 1965) and validity (Baykul, 2010, p. 425-426) of test scores decrease. That is why; correction formulas such as classical correction formula have been developed to reduce the negative effect of random guessing on reliability and validity (Crocker and Algina, 1986, p. 399-409; Kurz, 1999). Nevertheless, there are conflicting findings about the use of correction formulas. Some studies have shown that correcting test scores increases the reliability (Alnabhan, 2002; Burton, 2002; Muijtjens, Mamaren, Hoogenboom, Evers and Vleuten, 1999; Lord, 1975) and validity (Prihoda, Pinckard, McMahan and Jones, 2006). On the other hand, there are studies that have found that reliability (Glass and Wiley, 1964; Socan, 2009) and validity (Socan,

Çoban, E., & Demirtaşlı, R. N. (2017). Test yönergesi ve risk alma eğiliminin çoktan seçmeli testlerin psikometrik özelliklerine etkisi. *Journal of Human Sciences*, *14*(3), 2949-2967. doi:10.14687/jhs.v14i3.4449

2009) decrease due to correction. Moreover, some studies have found that correction for guessing does not change reliability (Abu Sayf, 1975; Cross and Frary, 1977; Frary, 1982; Sabers and Feldt, 1968) and validity (Çelen and Çıkrıkçı-Demirtaşlı, 2006; Frary, 1982; Sabers and Feldt, 1968). Although there are many correction formulas such as the formula rewarding omissions, most of the studies have focused on classical formula scoring. In Turkey, multiple choice items are commonly used in different exams. There are both exams whose raw scores are corrected (university entrance exams and government official selection exam) and exams whose raw scores are not corrected (foreign language exam, open education exams and driving license exam). However, there are not enough studies about the effect of correction formulas on psychometric properties of multiple choice tests in Turkish literature. It is necessary to investigate the effect of correction formulas on psychometric properties of multiple choice tests. Moreover, there are some studies that have found that high risk-takers demonstrate a greater guessing tendency than low risk-takers (Espinosa and Gardeazabal, 2010; Kubinger and Wolfsbauer, 2010). However, these studies did not investigate the effect of risk taking tendency on test scores. That is why; current study aims to investigate the effect of both test instruction stating the use of different correction formulas against guessing and risk taking tendency on psychometric properties of multiple choice tests.

Method

The current research is a quasi experimental research. An achievement test of measurement and evaluation consisting of 24 items was applied to 220 undergraduate students studying at different departments of Ankara University Faculty of Educational Sciences. The test was applied to four similar groups with different instructions. Instructions only differed in terms of explanation about scoring, other explanations were same. Explanations about scoring were as follow:

- **Instruction 1 (Y_1)**: There will be no correction for wrong answers.
- **Instruction 2 (Y_2):** One correct answer will be deleted because of four wrong answers.
- **Instruction 3 (Y_3):** One correct answer will be added for five omitted items.
- **Instruction 4 (Y_4):** One correct answer will be deleted because of each wrong answer.

A risk taking scale was applied to determine the risk taking tendency of the students. The data were analyzed by using one way ANOVA, dependent sample t-test, independent sample t-test, Feldt test, Fisher's z test, Pearson correlation coefficient, Spearman's rank-order correlation coefficient and confirmatory factor analysis.

Results

The results indicated that the effect of test instruction on raw scores was statistically significant ($F_{(3,216)}$=3.34; p<0.05). Post hoc comparisons using the Tukey HSD test revealed that the raw score mean of the test applied with the instruction $Y_1(\bar{x} = 14.65)$ was significantly higher than the raw score mean of the test applied with instruction $Y_3(\bar{x} = 12.91)$. The corrected score means of four tests were statistically different ($F_{(3,216)}$=32.16; p<0.05). Post hoc comparisons using the Dunnet's C test showed that corrected score mean of the test applied with the instruction $Y_4(\bar{x} = 7.87)$ was significantly lower than the corrected score means of the tests applied with instruction $Y_1(\bar{x} = 14.65)$, $Y_2(\bar{x} = 11.62)$ and $Y_3(\bar{x} = 13.90)$. Moreover, corrected score mean of the test applied with Y_2 was significantly lower than the corrected score means of the tests applied with $Y_1$and Y_3. The difference between raw score mean and corrected score mean was statistically significant for all of the tests whose instructions stated that correction would be applied (Y_2, $Y_3$and Y_4). However, the correlations between raw scores and corrected scores of the tests

Çoban, E., & Demirtaşlı, R. N. (2017). Test yönergesi ve risk alma eğiliminin çoktan seçmeli testlerin psikometrik özelliklerine etkisi. *Journal of Human Sciences*, *14*(3), 2949-2967. doi:10.14687/jhs.v14i3.4449

applied with Y_2 $(r = 0.98)$, $Y_3(r = 0.94)$ and $Y_4(r = 0.87)$ were very high. The results of Feldt test indicated that reliability coefficients calculated from raw scores were not statistically significant. The reliability coefficients calculated from corrected scores of the test applied with Y_4 were statistically higher than the other tests' reliability coefficients calculated from corrected scores. The criterion validity of the test applied with Y_1 was the highest. After correcting for guessing, validities of the tests applied with Y_3 and Y_4 increased while the validity of the test applied with Y_2 decreased. In general, there were no significant differences between test score means of low and high-risk takers. Only significant change based on risk taking tendency was seen in the raw scores of the test applied with Y_4.

Discussion

According to results of the study, corrected score mean of the test applied with Y_4 was the smallest one. This finding, which is similar to the finding of Prieto and Delgado (1999a), shows that test scores significantly decrease when the number of wrong answers is subtracted from the number of right answers. Corrected score mean of the test applied with Y_2 was smaller than mean of the test applied with Y_1. This finding is consistent with the many studies (Alnabhan, 2002; Betts, Elder, Hartley and Trueman, 2009; Swineford and Miller, 1953; Umay, 1998). Means of the tests applied with Y_2 and Y_4 significantly decreased after correction whereas mean of the test applied with Y_3 increased. However, there were very high positive correlations between raw and corrected scores. This finding shows that correction for guessing does not significantly change the order of test scores. Hence, there is no drawback of using correction formulas in norm-based assessments. As parallel to finding of Prieto and Delgado (1999a), reliability coefficients calculated by raw scores did not significantly differ. Test reliability increased when raw scores were corrected with classical correction formula, but not significantly. This finding implies that correction for guessing does not change test reliability. Validity of the test applied with $Y_1$was highest, but was significantly higher than only the validity of the test applied with Y_2. This finding indicates that test scores are more valid when the test instruction stating that no correction for guessing will be applied is used. There was no significant difference between raw score validity and corrected score validity of the test applied with Y_2. In general, risk taking tendency does not change the test scores.

Conclusion and Recommendations

The highest raw score mean is obtained when the instruction stating that there will be no correction for guessing is used. Raw test scores significantly change when correction for guessing is applied. Nevertheless, raw scores and corrected scores are highly correlated, especially when classical correction formula is used. That is why, it is more appropriate to use this formula in norm-based assessments. Moreover, correction formulas should not be used in criterion based assessments since there are significant differences between raw scores and corrected scores. Validity of the test applied with the instruction stating that there will be no correction is the highest. That is why, correction formulas should not be used unless necessary. In general, risk taking tendency does not significantly affect the test scores. Since the number of girls in the study group of current research is larger than the number of boys, further researches whose study groups consist of equal number of girl and boy should be conducted.