ISSN: 2146 - 9466
www.ijtase.net
International Journal of New Trends in Arts, Sports &Science Education – 2024, volume 13, issue 1

DÖRT AŞAMALI KİMYA TANI TESTİNİN YANITLAMA SÜRESİ VE YANITLAMA PERFORMANSLARININ İNCELENMESİ

EXAMINATION OF RESPONSE TIME AND PERFORMANCE OF THE FOUR TIER CHEMISTRY DIAGNOSTIC TEST

Canan BAŞTÜRK
Dokuz Eylül Üniversitesi, İzmir, Türkiye
ORCID: https://orcid.org/0000-0002-1184-7181
cnn.basturk.18@gmail.com

Suat TÜRKOGUZ
Prof.Dr., Dokuz Eylül Üniversitesi, Buca Eğitim Fakültesi, İzmir, Türkiye
ORCID: https://orcid.org/0000-0002-7850-2305
suat.turkoguz@gmail.com

Received: October 01, 2023 **Accepted:** January 16, 2024 **Published:** January 31, 2024

Suggested Citation:
Baştürk, C., & Türkoguz, S. (2024). Dört aşamalı kimya tanı testinin yanıtlama süresi ve yanıtlama performanslarının incelenmesi. *International Journal of New Trends in Arts, Sports &Science Education (IJTASE), 13*(1), 31-43.

Öz

Kavram yanılgılarını belirlemek için birçok ölçme yöntemi geliştirilmiş ve bu yöntemler ölçme değerlendirme süreçlerinde kullanılmıştır. Bu yöntemler; mülakatlar, açık uçlu sorular, çoktan seçmeli testler, iki aşamalı, üç aşamalı, dört aşamalı testlerdir. Katılımcılar test maddesiyle karşılaştıklarında iki şekilde cevap verir: hızlı tahmin ve çözüm davranışı. Katılımcının hızlı tahmin davranışı veya çözüm davranışı gösterip göstermediğini anlamak için yanıtlama süresine bakmak önemlidir. Bu çalışmada dört aşamalı kimya tanı testi ve çoktan seçmeli kimya testinin yanıtlama süresi ve yanıtlama performansının incelenmesi amaçlanmıştır. Bu çalışmada betimsel tarama yöntemi kullanılmıştır. Veri toplama aracı olarak, gaz basıncı ile ilgili 9 maddelik **D**ört **A**şamalı **D**iagnostik **K**imya **T**esti (**DADKT**) ve **Ç**oktan **S**eçmeli **K**imya **T**esti (**ÇSKT**) kullanılmıştır. DADKT, Ünsal (2019) tarafından geliştirilmiştir. ÇSKT ise aşamalı testten uyarlanmış olup DADKT'nin I. aşaması test maddeleri olarak kullanılmıştır. Bu çalışmada DADKT için bilimsel bilgi güvenirliği KR-20 hesaplanmış ve 0,460 bulunmuştur. Aynı zamanda DADKT için kavram yanılgısı güvenirliği KR-20 hesaplanmış ve 0,570 bulunmuştur. Bu çalışmada ÇSKT'nin KR-20 güvenirlik katsayısı 0,520 bulunmuştur. Çalışma 2020-2021 öğretim yılında Dokuz Eylül Üniversitesi, Buca Eğitim Fakültesi'nde öğrenim gören fen bilgisi öğretmen adaylarının katılımıyla gerçekleştirilmiştir. DADKT, 75 kişiye ve ÇSKT 74 kişiye uygulanmıştır. Çalışma sonucunda DADKT'nin I. aşamasının ortalama yanıtlama performansı 3,84 olarak hesaplanırken III. aşamasının ortalama yanıtlama performansı 3,11 olarak bulunmuştur. ÇSKT'nin ortalama yanıtlama performansı ise 3,45'dir. Yanıtlama süreleri ise DADKT'nin I. Aşaması ve III. Aşaması için sırasıyla 10:47 ve 7:15; ÇSKT için 12:52 saniyedir. Bu çalışmada katılımcılara test maddelerine geri dönüş hakkı verilmediğinden dolayı her katılımcı test maddelerini bir kez cevaplamak zorunda kalmış ve cevabı değiştirme hakkı olmamıştır. İleriki çalışmalarda test uygulama biçimine göre çalışmanın kapsamı genişletilebilir.

Anahtar Terimler: Kavram yanılgısı, diagnostik test, yanıtlama süresi, yanıtlama performansı.

Abstract

Many measurement methods have been developed to identify misconceptions and have been used in measurement and evaluation processes. These methods include; interviews, open-ended questions, multiple-choice tests, two-tier, three-tier, and four-tier tests. When participants encounter the test item, they respond in two ways: quick guessing and solution behavior. It is important to examine the response time to understand whether the participant exhibits quick guessing behavior or solution behavior. This study aimed to determine the response time and response performance of the **F**our-**T**ier **C**hemistry **D**iagnostic **T**est (**FTCDT**) and the **M**ultiple-**C**hoice **C**hemistry **T**est (**MCCT**). In this study, a descriptive survey method was used. As the data collection tool, the 9-item FTCDT and MCCT related to gas pressure were used. The FTCDT was developed by Ünsal (2019). On the other hand, the MCCT was adapted from the tiered test, and the first tier of the FTCDT was used as a test item. In this study, scientific information reliability KR-20 was calculated for the FTCDT and was found to be 0.460. At the same time, the misconception reliability KR-20 for the FTCDT was calculated and found to be 0.570. In this study, the KR-20 reliability coefficient of the MCCT was found to be 0.520. The study was conducted with the participation of science preservice teachers studying at Dokuz Eylul University, Buca Faculty of Education in the 2020-2021 academic year. The FTCDT was applied to 75 people, and the MCCT was applied to 74 people. Because of this study, the mean response performance of the first tier of FTCDT was calculated as 3.84, while the third tier was calculated as 3.11. In

addition, the mean response performance of MCCT was found to be 3.45. The response times for the first and third tier of FTCDT are 10:47 and 7:15, respectively. Additionally, the response time for MCCT is 12:52. In this study, because the participants were not given the right to return the test items, each participant had to answer the test items once and did not have the right to change the answer. In future studies, the scope of the study can be expanded according to the test application form.

Keywords: Misconception, diagnostic test, response time, response performance.

GİRİŞ

Ölçme değerlendirme yöntemlerinde genellikle çoktan seçmeli testler ve açık uçlu sınavlar kullanılır. Bu testlerle öğrencilerin bilgi düzeyi belirlenebilir, ancak öğrenme süreci ile ilgili bilgi elde edilemez. Öğrencilerde yanlış veya eksik öğrenilmiş bilgilerin belirlenmesi doğru bilgiye ulaşmaları için önem teşkil eder. Öğrencilerin kavramları anlama sürecini ve önceden var olan ya da süreç boyunca oluşan kavram yanılgılarını belirleyebilmek için çoktan seçmeli testlerin kullanımı yaygındır, ancak çoktan seçmeli testlerde şans faktörü olmasından dolayı dezavantajlıdır. Bu yüzden çoktan seçmeli testlerin dezavantajları göz önünde bulundurularak aşamalı testler geliştirilmiştir. Aşamalı testler sırayla iki aşamalı, üç aşamalı ve dört aşamalı testler olarak geliştirilmiştir. Dört aşamalı testlerin kavram yanılgılarını belirlemede etkili olduğu görülmüştür (Önsal, 2016). Dört aşamalı testler, öğrencilerin bilgiye ulaşma sürecinin takibi ile kavram yanılgılarının ortaya çıkarılmasında diğer aşamalı testlere göre daha güvenilir sonuçlar göstermiştir (Kaltakçı, 2012). Aşamalı testler öğrenci değerlendirilmesinde hızlı sonuç ürettiği halde çoktan seçmeli testlerle karşılaştırıldığında aşamaların puana dönüştürme zorluğu ortaya çıkmış, geçerlik ve güvenirlik puanları için testin aşamaları tartışılmıştır. Özellikle aşamalı testlerde aşamaların kendi içinde nasıl puanlanacağı veya birbiriyle olan ilişkileri, güvenirlik ve geçerlik puanlarının nasıl hesaplanacağı ile ilgili problemler ortaya çıkmıştır.

Bir test maddesinin yanıtlama süresinden yararlanarak katılımcıların yanıtlama performansı veya motivasyonu ile ilgili durumlar belirlenebilir. Alanyazın incelendiğinde yanıtlama süresine bağlı olarak yanıtlama performansının incelendiği çalışmalar genellikle çoktan seçmeli testlerde yürütülmüş aşamalı testler kapsamında fazla örneğe rastlanılmamıştır (DeMars & Wise, 2010; DeMars, Wise, & Smith; 2009). Aşamalı testler kapsamında Türkoguz (2020), eşik değerden yararlanarak yanıtlama süresi ve madde yanıtlama performansını incelemiştir. Ancak bu araştırma dört aşamalı test kullanımıyla gerçekleşmiştir. Alanyazında dört aşamalı test kullanılarak yanıtlama süresi ve yanıtlama performansının incelendiği çalışmalara rastlanılmamaktadır. Ayrıca literatürde olan çalışmaların çoğu katılımcıların yüz yüze katılımıyla gerçekleşmiştir. Aşamalı testlerin online yürütüldüğü çalışmalar çok azdır. Örneğin; Yang ve Sianturi (2019), üç aşamalı matematik testini online olarak yürütmüşlerdir. Ancak bu çalışma üç aşamalı bir çalışma olup matematik alanıyla ilgilidir. Alanyazın incelendiğinde dört aşamalı kimya tanı testinin kullanıldığı ve online yürütülen yanıtlama süresi ve yanıtlama performansının beraber incelendiği araştırmalara rastlanılmamıştır.

Aşamalı testlerin ve çoktan seçmeli testlerin puanlanması, geçerlik ve güvenirliklerinin sağlanması konusunda günümüzde de tartışmalar devam etmektedir (Taber, 2017, Bademci, 2006; Bademci, 2007). Aşamalı testin I. aşamasının nasıl puanlanacağı II. aşamasının nasıl puanlanacağı ve birbirlerini nasıl etkilediği hususunda belirsizlikler hala devam etmektedir. Ayrıca bu testlerin güvenirlik katsayısı hesaplamasında problemler bulunmaktadır (Gürel ve Eryılmaz ve McDermott, 2015; Taber, 2017; Romine, Schaffer & Barrow, 2015; Wang, 2004; Peşman & Eryılmaz; 2010; McClary & Lowery Bretz, 2012; Stankov, Lee, Luo, Hogan; 2012).

Eğitim testlerinde yanıtlama süresi ile yanıtlama doğruluğu arasında ilişkilerin incelendiği araştırmalar 1990 yılına kadar dayanmaktadır (Schnipke & Scrams, 1997; Yamamoto, 1995; Wise ve Kong; 2005). Madde yanıtlama süresi ile madde yanıt doğruluğundan faydalanarak eşik değer belirlemesi yapılan çalışmalar ve bu eşik değerden yaralanarak madde yanıtlama performansını belirleyen çalışmalar bulunmaktadır (Guo, Rios, Haberman, Liu, Wang & Paek, 2016; Setzer, Wise, Heuvel & Ling, 2013; Weeks, Davier & Yamamoto, 2016; Meyer, 2010; Wise & DeMars, 2010; Wise & Gao, 2017; Bulut, 2015; Wright, 2016; Wise & DeMars; 2006; Lee & Chen, 2011; Ferrando & Lorenzo-Seva, 2007). Madde yanıtlama süresi ile yanıtlama doğruluğu arasında yapılan görsel grafiklerde eşik değerleri bulunmaya çalışılarak performans belirlemelerine gidilmiştir. Lojistik analizlerle madde yanıt

doğruluğu, madde yanıtlama süresi ve eşik değere göre yanıtlama performansları kişisel parametre inclemelerinde kullanılmaktadır (Ranger & Kuhn; 2012; Lee & Chen, 2011; Wise& DeMars; 2006; Setzer, Wise, Heuvel & Ling, 2013; Ferrando & Lorenzo-Seva, 2007; Bulut, 2015; Elbaum, Golan, Lupu, Wagner & Braw, 2019; Bolsinova, Boeck & Tijmstra; 2015; Van Der Linden, Klein Entink & Fox; 2010; Van Der Linden; 2006; Van Der Linden; 2007; Wright; 2016; Weeks, Davier &Yamamoto; 2016). Madde yanıt doğruluğu ile madde yanıtlama süresinin ilişkileri üzerine Wise, DeMars Smith'in ortaklaşa yaptıkları çalışmalar yaygındır (DeMars & Wise, 2010; DeMars, Wise & Smith; 2009; Wise, 2014; Kong,Wise & Bhola, 2007; Wise & Gao, 2017; Setzer, Wise, Heuvel & Ling, 2013; Wise, 2019; Wise & DeMars, 2010; Wise, Pastor & Kong, 2009; Wise, Bhola, & Yang, 2006; Wise & Kingsbury, 2016; Wise, 2019; Wise, 2014; Wise, Bhola, , & Yang, 2006; Wise & Kong, 2005). Alanyazında belirtilen çalışmaların çoğunluğu çoktan seçmeli testler kapsamında gerçekleştirilmiştir. Aşamalı diagnostik testler veya kavramsal anlama testler kapsamında madde yanıtlama süresiyle madde yanıtlama performanslarının incelendiği çalışmalara rastlanılmamaktadır.

Bu maksatla aşamalı testlerde yanıtlama performansı ve yanıtlama süresinin incelenmesi ilerideki çalışmalara rehber olabileceği düşünülmüştür. Bu kapsamda **D**ört **A**şamalı **D**iagnostik **K**imya **T**estiyle (**DADKT**) **Ç**oktan **S**eçmeli **K**imya **T**estinin (**ÇSKT**) madde yanıtlama performansları ve madde yanıt süreleri belirlenerek öğrencinin yanıtından emin olma teyiti ve aşamalı diagnostik testinin aşamaları da dikkate alınarak karşılaştırılmıştır.

Problem Cümlesi

Bu çalışmanın problem cümlesi "Dört Aşamalı Kimya Tanı Testi ve Çoktan Seçmeli Kimya Testinin yanıtlama süreleri ve buna bağlı olarak yanıtlama performansları arasında fark mıdır?" olarak belirlenmiştir. Ana problemin çözümüne ilişkin iki tane alt problem belirlenmiştir. Bunlar:

1-DADKT ile ÇSKT'nin yanıtlama performansları arasında fark var mıdır?

2-DADKT ile ÇSKT'nin yanıtlama süreleri arasında fark var mıdır?

YÖNTEM

Araştırmanın Modeli

Bu araştırmada, nicel araştırma yöntemlerinden biri olan genel tarama modellerinden ilişkisel tarama modeli kullanılmıştır. Tarama modeli, geçmişte veya şu an olan bir durumu olduğu haliyle tanımlayan, bireylerdeki istendik davranış biçimlerinin gelişmesi için uygulanan süreçlerdir. Genel tarama modellerinde, evren hakkında genel bir yorum yapabilmek amacıyla evrenden alınan bir çalışma grubu veya örneklem üzerinde tarama yapılmaktadır. İki ya da daha fazla değişken arasındaki değişimi ve değişim seviyesini belirlemeyi amaçlayan araştırma modeline ilişkisel tarama modeli denir (Karasar,2005). Bu tarama modelinde değişkenler arasında değişim varsa bunun nasıl olduğu bulunmaya çalışılır.

Çalışma Grubu / Katılımcılar

Araştırma; 2020-2021 öğretim yılında Dokuz Eylül Üniversitesi, Buca Eğitim Fakültesi fen bilgisi öğretmenliği bölümünde öğrenim görmekte olan 149 öğretmen adayının katılımıyla gerçekleştirilmiştir. Katılımcıların; 24'ü erkek, 125'i bayandır. Çalışmaya katılan öğretmen adaylarının 14'ü 2.sınıf, 49'u 3.sınıf, 86'sı 4.sınıf seviyesinde öğrenim görmektedir. Çalışmaya katılan öğretmen adaylarının 83 (%55,71)'ü Ege, 25(%16,78)'i Marmara, 22(%14,77)'si Akdeniz, 8(%5,37)'i İç Anadolu, 6(%4,03)'sı Güneydoğu Anadolu, 4(%2,69)'ü Doğu Anadolu, 1(%0,67)'i Karadeniz Bölgesinde ikamet etmektedir.

Veri Toplama Aracı

Dört Aşamalı Diagnostik Kimya Testi (DADKT)

Ünsal (2019), gaz basıncı konusuyla ilgili 9 maddeden oluşan Dört Aşamalı Diagnostik Kimya Testini (DADKT) fen bilgisi öğretmen adaylarıyla geliştirmiştir. Ünsal (2019), öğretmen adaylarına testi yanıtlamaları için 27 dk vermiştir. Ünsal (2019) testi geliştirirken açık uçlu soru oluşturma, çoktan

seçmeli formata dönüştürme ve aşamalı formata dönüştürme aşamalarını izlemiştir. Öncelikle testi geliştirirken öğretmen adaylarına açık uçlu 52 soru sorulmuş, daha sonra yanıtlara göre açık uçlu sorular çoktan seçmeli teste dönüştürülmüştür. Sonrasında çoktan seçmeli testin maddelerine verilen cevaplardaki seçtikleri seçeneklerin seçilme gerekçesini açıklamalarını isteyen bir boşluk bırakılarak öğretmen adaylarından (n:88) yanıtlarına gerekçeler istenmiştir. Öğretmen adaylarının test maddelerine verdikleri yanıtların gerekçeleri incelenerek çoktan seçmeli teste gerekçelerin olduğu II. aşama eklenmiştir. Öğretmen adaylarının verdikleri cevaplara güven düzeyini belirlemek için test maddesinin soru kökünün olduğu I. aşamadan sonra "Verdiğiniz yanıttan emin misiniz?" ve I. aşamanın soru köküne verilen yanıtın gerekçesinin seçildiği II. aşamadan sonra da yine "Verdiğiniz yanıttan emin misiniz?" soruları yöneltilmiştir. Böylelikle gaz basıncıyla ilgili çoktan seçmeli test dört aşamalı diagnostik kimya testine (DADKT) dönüştürülmüştür.

Tablo 1. DADKT'nin puanlama tablosu.

	I. Aşama	**II. aşama**	**III. aşama**	**IV.aşama**
	Cevap	**Güven düzeyi**	**Cevap**	**Güven güzeyi**
Bilimsel bilgi	Doğru	Eminim	Doğru	Eminim
	Doğru	Eminim	Doğru	Emin Değilim
I.tip bilgi eksikliği	Doğru	Emin Değilim	Doğru	Eminim
	Doğru	Emin Değilim	Doğru	Emin Değilim
Yanlış pozitif	Doğru	Eminim	Yanlış	Eminim
	Doğru	Eminim	Yanlış	Emin Değilim
II.tip bilgi eksikliği	Doğru	Emin Değilim	Yanlış	Eminim
	Doğru	Emin Değilim	Yanlış	Emin Değilim
Yanlış negatif	Yanlış	Eminim	Doğru	Eminim
	Yanlış	Eminim	Doğru	Emin Değilim
III.tip bilgi eksikliği	Yanlış	Emin Değilim	Doğru	Eminim
	Yanlış	Emin Değilim	Doğru	Emin Değilim
Kavram yanılgısı	Yanlış	Eminim	Yanlış	Eminim
	Yanlış	Eminim	Yanlış	Emin Değilim
IV.tip bilgi eksikliği	Yanlış	Emin Değilim	Yanlış	Eminim
	Yanlış	Emin Değilim	Yanlış	Emin Değilim

DADKT'nin Güvenirlik ve Geçerlik Bilgileri:

Bu çalışmada kullanılan DADKT'nin bilimsel bilgi güvenirliği KR-20 (tüm aşamaların doğru olması durumunda 1 puan alma şartına göre) 0,460; kavram yanılgısı güvenirliği KR-20 (I. ve III. Aşamaya yanlış cevap verilmesi, II. ve IV. Aşamada emin olunması şartına göre) 0,570 olarak hesaplanmıştır. 9 maddelik DADKT'nin yapı geçerliği için açıklayıcı ve doğrulayıcı faktör analizleri gerçekleştirilmiştir. DADKT'nin 4 faktörlü yapı sergilediği bu analizler doğrultusunda doğrulanmıştır. DADKT'nin s1, s6 ve s9 maddeleri İdeal Gaz kavramını, s3 ve s7, maddeleri Gaz-sıvı basınçları ilişkisini, s4 ve s5 maddeleri kapalı kaplardaki gaz sistemlerini s2 ve s8 maddeleri barometre kavramlarını içermektedir. DADKT'nin test maddeleri için fen eğitimi alanında 4 uzman görüşüne başvurulmuştur. Uzmanlar, 9 maddelik DADKT'de yer alan test maddelerini öğretmen adaylarına uygunluğuna ve kimya ders içeriğine 'uygun', 'uygun değil, düzeltilmesi gerekiyor' ve 'uygun değil' şeklinde incelemiştir. DADKT'nin tüm maddeleri için I. ve III. aşamanın kapsam geçerlilik indeksi (KGI) sırasıyla 0,98 ve 0,95 olarak hesaplanmıştır (Lawshe, 1975). Ayrıca bu çalışmanın içerik geçerliliği için yanlış pozitif (YP) ve yanlış negatif (YN) değerleri de hesaplanmıştır. Hestenes ve Halloun (1995), aşamalı diagnostik testlerde dış geçerliliğin kanıtı olarak YP ve YN tanımını önermiştir. Hestenes ve Halloun (1995) YP'yi yanlış bir nedene dayalı olarak kendinden emin bir

tutumla test maddesine doğru yanıt olarak tanımlarken, YN'yi doğru nedene dayalı olarak kendinden emin bir tutumla test maddesine verilen yanlış yanıt olarak tanımlamışlardır. Aşamalı diagnostik testlerde dış geçerlilik için YN, yüzde 10'dan az olmalıdır (Gürçay ve Gülbaş, 2015). Ancak, aşamalı diagnostik testlerde YP'yi azaltmak zordur. Bilgi eksikliği olan öğrenciler çoktan seçmeli testlerde doğru cevabı tahmin etme şansına sahip olurlar ve test maddesinin çeldiricilerinden doğru seçeneği seçmeleri olasıdır (Peşman ve Eryılmaz 2010). Tablo 1'in puanlamasında, 9 maddelik DADKT'nin YP ve YN oranları sırasıyla %17,9 ve %11,3 olarak hesaplanmıştır. Yanlış negatif oranlar, Hestenes ve Halloun'un (1995) görüşlerine göre 9 maddelik DADKT'nin geçerli bir araç olduğunu kabul edilebilir. DADKT'nin yapı geçerliğinin çalışmanın örnekleminde doğrulanıp doğrulanmadığının incelenmesi amacıyla doğrulayıcı faktör analizi (DFA) gerçekleştirilmiştir. DFA, AMOS 16 (Arbuckle, 2008) programında en yüksek olabilirlik yöntemiyle (Maximum Likelihood) gerçekleştirilmiştir. Faktöriyel yapının gözlemlenen değerlerle uyum derecesinin belirlenebilmesi amacıyla CMIN/df <5, RMSEA<0,08 ve RMR<0,08 uyum indekslerinin değerleri hesaplanmıştır (Çokluk, Şekercioğlu ve Büyüköztürk, 2012; Kline, 2011). DFA, DADKT'nin dört faktörlü DFA modeli çalışma kapsamında elde edilen verilerle iyi bir uyum gösterdiği saptanmıştır (CMIN/DF = 1,490; RMSEA = 0,079; RMR = 0,021).

ÇSKT'nin Güvenirlik ve Geçerlik Bilgileri:

Bu çalışmada gaz basıncıyla ilgili Çoktan Seçmeli Kimya Testi de kullanılmıştır. Aslında ÇSKT, Ünsal (2019) tarafından geliştirilen gaz basıncıyla ilgili 9 maddelik DADKT'nin I. aşamasıdır. ÇSKT, DADKT'den diğer aşamalar çıkartılarak kullanılmıştır. ÇSKT'nin KR-20 güvenirlik analizi için doğru yanıtlara 1 diğer yanıtlara 0 verilmiştir. Bu çalışma için ÇSKT'nin KR-20 güvenirlik katsayısı 0,520 bulunmuştur. ÇSKT'nin kapsam geçerliği için uzman görüşlerine başvurulmuştur. Bu nedenle, Yüksek Öğretim Kurumu'nun Eğitim Fakültelerinin Fen Bilgisi Öğretmenliği Programının Kimya ders içeriğine göre 9 maddelik ÇSKT'nin kapsam ve görünüş geçerliliği dokuz uzman (Fen eğitiminden dokuz yüksek lisans öğrencisi) tarafından yeniden değerlendirilmiştir. Uzmanlar, 9 maddelik ÇSKT'de yer alan test maddelerini öğretmen adaylarına uygunluğuna ve kimya ders içeriğine 'uygun', 'uygun değil, düzeltilmesi gerekiyor' ve 'uygun değil' şeklinde incelemiştir. Varsa ek görüşlerini test maddesinin yanında bırakılan boş alana yazmışlardır. ÇSKT 'nin her bir maddesi için madde kapsam geçerlilik oranları (KGOi) ve tanı testinin tüm maddeleri için test kapsam geçerlilik indeksi (KGI), Lawshe (1975) formülleri kullanılarak hesaplanmıştır: ÇSKT'nin tüm maddeleri için kapsam geçerlilik indeksi (KGI) 0,98 hesaplanmıştır (Lawshe, 1975). ÇSKT'nin yapı geçerliğinin çalışmanın örnekleminde doğrulanıp doğrulanmadığının incelenmesi amacıyla doğrulayıcı faktör analizi (DFA) gerçekleştirilmiştir. DFA, AMOS 16 (Arbuckle, 2008) programında en yüksek olabilirlik yöntemiyle (Maximum Likelihood) gerçekleştirilmiştir. Faktöriyel yapının gözlemlenen değerlerle uyum derecesinin belirlenebilmesi amacıyla CMIN/df <5, RMSEA<0,08 ve RMR<0,08 uyum indekslerinin değerleri hesaplanmıştır (Çokluk, Şekercioğlu ve Büyüköztürk, 2012; Kline, 2011). DFA ÇSKT'nin dört faktörlü DFA modeli çalışma kapsamında elde edilen verilerle iyi bir uyum gösterdiği saptanmıştır (CMIN/DF = 1,344; RMSEA = 0,069; RMR = 0,020).

Verilerin Toplanması ve Analizi

Verilerin Toplanma Süreci

Veri toplama, Dokuz Eylül Üniversitesi Buca Eğitim Fakültesi Fen Bilgisi Öğretmenliği bölümünde öğrenim görmekte olan öğrencilere uygulanan DADKT ve ÇSKT ile gerçekleştirilmiştir. Araştırmanın veri toplama işleminin yapılacağı sırada COVID-19 pandemisi meydana gelmiş ve dersler online olarak bilgisayar üzerinden yürütülmüştür. Bu yüzden bu çalışmanın verileri bilgisayar ile online test ortamlarında toplanmıştır. Online test ortamlarında yanıtlanma süresi toplam testin yanıtlanmasına göre belirlenmekte her test maddesi için kayıt tutulmamaktadır. Bu nedenle online test ortamlarında her test maddesinin yanıtlanma süresinin belirlenmesi önemli bir problem teşkil etmektedir. Online test ortamları olarak Microsoft Office uygulaması olan Teams-Form ortamı ve Dokuz Eylül Üniversitesi'nin SAKAI uzaktan eğitim portalı kullanılmıştır. Her test maddesi için ayrı ayrı online test formu oluşturulmuş ve bu test formları SAKAI uzaktan eğitim portalının test uygulama ortamında tek tek sırayla öğrencilere sunulmuştur. Böylelikle her test maddesi için yanıtlama performansı ve

süresi kaydedilmiştir. Katılımcıların geri dönüt olarak tekrar test maddelerini görme ve yanıtlama şansı olmamıştır. Her iki test türü için testin ilerleme yönü doğrusal olup geriye dönük değildir. Son olarak SAKAİ programında her test maddesi için yanıtlama süresine ulaşılabildiğinden dolayı bu programla toplanan veriler araştırma için kodlama ve analiz edilme aşamasına geçilmiştir.

Verilerin Analiz Süreci:

İlk olarak çalışmanın verileri MS Office Excel'de düzenlenmiştir. DADKT ve ÇSKT'nin güvenirliği, MS Office Excel'de KR-20 formülü uygulanarak hesaplanmıştır. DADKT ve ÇSKT'nin yapı geçerliği SPSS istatistik programında açıklayıcı faktör analizi yapılarak test edilmiştir. DADKT ve ÇSKT uzman görüşlerine sunulmuş ve uzman görüşlerinin uyum oranları Lawshe'nin (1975) formülleri MS Office Excel'de uygulanarak hesaplanmıştır. Yine benzer şekilde DADKT'nin I. ve III. aşama arasındaki yanıt uyumları Hestenes ve Halloun'un (1995) "Bilimsel Bilgi", "Yanlış Pozitif", Yanlış Negatif" ve "Kavram Yanılgısı" kodlamasına göre MS Office Excel'de hesaplanarak dış geçerliği test edilmiştir. Son olarak AMOS programı yardımıyla açıklayıcı faktör analizi ile ortaya çıkan DADKT ve ÇSKT'nin yapısı doğrulanmıştır. Çalışmanın ana problemin çözümünde MS Office Excel'de yanıtlama performansları hesaplanmış ve bu değerler DADKT ve ÇSKT test grupları arasında t testi ile SPSS istatistik programında 0,05 anlamlılık düzeyinde karşılaştırılmıştır. DADKT ve ÇSKT testleri için yanıtlama performanslarının normal dağılım sağlayıp sağlamadığı Kolmogorov-Smirnov ve Shapira Wilks normal dağılım testleriyle kontrol edilmiştir. Ayrıca DADKT ve ÇSKT test grupları arasındaki karşılaştırmalarda ortalama yanıtlama performanslarına ve standart sapmalarına da bakılmıştır. Çalışmanın ikinci alt problemi, SPSS istatistik programında DADKT ve ÇSKT için yanıtlama sürelerinin ortalama ve standart sapmaları t testi ile hesaplanarak karşılaştırılarak çözülmüştür. DADKT ve ÇSKT testleri için yanıtlama sürelerinin normal dağılım sağlayıp sağlamadığı Kolmogorov-Smirnov ve Shapira Wilks normal dağılım testleriyle kontrol edilmiştir.

BULGULAR

Bu çalışmanın ilk problemi "DADKT ile ÇSKT'nin yanıtlama performansları arasında fark var mıdır?" şeklinde ifade edilmiş ve bu alt problemde dört aşamalı diagnostik kimya testi (DADKT) ile çoktan seçmeli kimya testinin (ÇSKT) yanıtlama performansları arasındaki fark incelenmiştir. DADKT'nin soru maddesini belirten I. aşama ve neden sorusunun sorulduğu III. aşamanın yanıtlama performansları eşleştirilerek t testi ile karşılaştırılmış Tablo 4'de sayısal değerlerle verilmiştir. Daha sonra DADKT'nin I. aşamasıyla ÇSKT'nin yanıtlama performansları bağımsız t testi ile karşılaştırılmış ve Tablo 5'de sayısal değerleri gösterilmiştir. Daha sonra DADKT'nin III. aşamasıyla ÇSKT'nin yanıtlama performansları bağımsız t testi ile karşılaştırılmış ve Tablo 6'da sayısal değerleri verilmiştir.

Tablo 4. DADKT'nin I. aşama ve III. aşama ortalama yanıtlama performanslarının karşılaştırması

	Ortalama Yanıtlama Performansı	Standart Sapma	Eşleştirilmiş t testi	p
DADKT I. AŞAMA n:75)	3,84	1,71	6,009	0,000
DADKT III. AŞAMA (n:75)	3,11	1,62		

Tablo 4'de DADKT'nin I. aşaması ile DADKT'nin III. Aşaması ortalama yanıtlama performansına göre eşleştirilmiş t testi ile karşılaştırıldığında anlamlı bir sonucun olduğu görülmüştür ($t_{(75,2)}$=6,009; p<0,05). DADKT'nin I. aşamasının ortalama yanıtlama performansının (M=3,84) III. aşamasının ortalama yanıtlama performansından (M=3,11) yüksek olduğu anlaşılmıştır.

Tablo 5. DADKT'nin I. aşama ve ÇSKT'nin ortalama yanıtlama performansının karşılaştırması

	Ortalama Yanıtlama Performansı	Standart Sapma	Bağımsız t testi	p
DADKT I. AŞAMA n:75)	3,84	1,70	1,333	0,185
ÇSKT (n:74)	3,45	1,90		

Tablo 5'de DADKT'nin I. aşaması ile ÇSKT'nın ortalama yanıtlama performanslarına göre bağımsız t testi ile karşılaştırıldığında anlamlı bir sonuca ulaşılamamıştır ($t_{(159,2)}$=1,333; p>0,05). DADKT'nin I. aşamasının ortalama yanıtlama performansının (M=3,84) ÇSKT'nin ortalama yanıtlama performansından (M=3,45) yüksek olmasına rağmen istatistiki olarak farklı olmadığı şeklinde yorumlanmıştır.

Tablo 6. DADKT'nin III. aşama ve ÇSKT'nin ortalama yanıtlama performansının karşılaştırması

	Ortalama Yanıtlama Performansı	Standart Sapma	Bağımsız t testi	p
DADKT III. AŞAMA n:75)	3,11	1,62	-1,171	0,185
ÇSKT (n:74)	3,45	1,90		

Tablo 6'da DADKT'nin III. aşaması ile ÇSKT'nın ortalama yanıtlama performansına göre bağımsız t testi ile karşılaştırıldığında anlamlı bir sonuca ulaşılamamıştır ($t_{(159,2)}$=-1,171; p>0,05). DADKT'nin III. aşamasının ortalama yanıtlama performansının (M=3,11) ÇSKT'nin ortalama yanıtlama performansından (M=3,45) düşük olmasına rağmen istatistiki olarak farklı olmadığı şeklinde yorumlanmıştır.

Bu çalışmanın ikinci problemi "DADKT ile ÇSKT'nin yanıtlama süreleri arasında fark var mıdır?" şeklinde ifade edilmiş ve bu alt problemde öncelikle DADKT'nin I. aşaması ve III. aşamasının yanıtlama süreleri arasındaki fark eşleştirilmiş t testi ile incelenmiş Tablo 7'de sayısal olarak ifade edilmiştir. Daha sonra DADKT'nin I. aşaması ve ÇSKT'nin yanıtlama süreleri arasındaki fark bağımsız t testi ile hesaplanmış ve Tablo 8'de gösterilmiştir. Daha sonra DADKT'nin III. aşaması ile ÇSKT'nin yanıtlama süreleri arasındaki fark gözlemlenmiş Tablo 9'da gösterilmiştir.

Tablo 7. DADKT'nin I. aşama ve III. aşama yanıtlama sürelerinin karşılaştırması

	Ortalama (dd:ss)	Standart Sapma (dd:ss)	Eşleştirilmiş t testi	p
DADKT I. AŞAMA n:75)	10:47	05:53	7,206	0,000
DADKT III. AŞAMA (n:75)	07:15	05:16		

Tablo 7'de DADKT'nin I. aşaması ile DADTKT'nin III. aşaması yanıtlama sürelerine göre eşleştirilmiş t testi ile karşılaştırıldığında anlamlı bir sonucun olduğu görülmüştür (t(75,2)=7,206; p<0,05). DADKT'nin I. aşamasının yanıtlama süresi ($M_{süre}$=10:47 sn.) III. aşamasının yanıtlam süresinden ($M_{süre}$=07:15 sn.) yüksek olduğu anlaşılmıştır.

Tablo 8. DADKT'nin I. aşama ve ÇSKT'nin yanıtlama sürelerinin karşılaştırması

	Ortalama (dd:ss)	Standart Sapma (dd:ss)	Bağımsız t testi	p
DADKT I. AŞAMA n:75)	10:47	05:53	-1,928	0,056
ÇSKT (n:74)	12:52	07:10		

Tablo 8'de DADKT'nin I. aşaması ile ÇSKT'nın yanıtlama sürelerine göre bağımsız t testi ile karşılaştırıldığında anlamlı bir sonuca ulaşılamamıştır ($t_{(159,2)}$=-1,928; p>0,05). DADKT'nin I. aşamasının yanıtlama süresi ($M_{süre}$=10:47 sn.) ÇSKT'nin yanıtlama süresinden ($M_{süre}$=12:52 sn.) düşük olmasına rağmen istatistiki olarak farklı olmadığı şeklinde yorumlanmıştır.

Tablo 9. DADKT'nin III. aşama ve ÇSKT'nin yanıtlama sürelerinin karşılaştırması

	Ortalama (dd:ss)	Standart Sapma (dd:ss)	Bağımsız t testi	p
DADKT III. AŞAMA n:75)	07:15	05:16	-5,444	0,000
ÇSKT (n:74)	12:52	07:10		

Tablo 9'da DADKT'nin III. aşaması ile ÇSKT'nın yanıtlama sürelerine göre bağımsız t testi ile karşılaştırıldığında anlamlı bir sonuca ulaşılmıştır ($t_{(159,2)}$=-5,444; p<0,05). DADKT'nin III. aşamasının yanıtlama süresi ($M_{süre}$=07:15 sn.) ÇSKT'nin yanıtlama süresinden ($M_{süre}$=12:52 sn.) oldukça düşük olduğu anlaşılmıştır.

TARTIŞMA, SONUÇ ve ÖNERİLER

Bu bölümde ilk etapta çalışmanın alt problemlerine göre sırasıyla ilgili literatür desteğiyle karşılaştırma yapılarak tartışma gerçekleştirilmiştir. Daha sonra genel olarak sonuçlar sunulmuş ve bu sonuçlara ilişkin olarak öneriler getirilmiştir.

Bu çalışmanın ilk alt probleminde DADKT ile ÇSKT'nin yanıtlama performansları arasındaki fark incelenmiştir. Tablo 5'de görüldüğü üzere DADKT'nin I. aşamasında ortalama yanıtlama performansı 3,84 ve III. aşamasının ortalama yanıtlama performansı 3,11 olarak bulunmuştur. Çoktan seçmeli kimya testi için ortalama yanıtlama performansı 3,45 olarak hesaplanmıştır. Bu çalışma, aşamalı diagnostik testleri için önemli sonuçlar vermektedir. DADKT'nin III. aşaması nedensel ilişkinin kurulacağı kavram yanılgılarından oluşan bir bölümdür. Bu bölümde seçenekler kısa cümlelerden oluşmasına rağmen ortalama yanıtlama performansı, çözüm içeriğinin olduğu I. aşamaya göre daha düşük çıkmıştır. Bu durum katılımcıların kavramsal anlama ve nedensel ilişkilerde zorlandıklarını ortaya koymuştur. Katılımcılara soru maddelerine tekrar cevap verme hakkı verilmemiştir. Çünkü daha önce yapılan çalışmalarda test maddesine tekrar cevap verme hakkı tanındığında ortalama yanıtlama performanslarının olumsuz etkilendiği gözlemlenmiştir (Türkoguz, 2020).

Bu çalışmanın ikinci alt probleminde DADKT ile ÇSKT'nin yanıtlama süreleri arasındaki fark incelenmiştir. Tablo 7'ye bakıldığında DADKT'nin I. aşaması için yanıtlama süresi 10:47 sn., III. aşaması için 7:15 sn. bulunmuştur. Çoktan seçmeli kimya testi için yanıtlama süresi 12:52 sn. olarak hesaplanmıştır. Bu sonuçlar DADKT ve ÇSKT için yanıtlama sürelerinin farklılık gösterdiğini kanıtlamıştır. Katılımcılar I. aşamaya III. aşamaya göre daha fazla zaman harcamışlardır. Bunun sebebi katılımcıların DADKT'nin I. aşamasını yanıtlarken III. aşama ile ilgili bilgi edinebilmeleri, yorum yapabilmeleri, diğer aşamaya daha kısa zamanda geçmek istemeleridir.

Bu sonuçlardan hareketle aşamalı testler ve çoktan seçmeli testlerdeki yanıtlama süresini karşılaştırdığımızda test tasarımının ve test yönetiminin katılımcıların yanıtlama sürelerini ve motivasyonlarını etkileyebildiğini gözlemleyebiliriz. Demars (2000), düşük riskli testler için yaptığı araştırmada yapılandırılmış cevap maddelerini içeren testlerde çoktan seçmeli testlere göre katılımcıların motivasyonlarının daha düşük olduğunu tespit etmiştir. Wise (2014) araştırmasında test maddelerinin uzunluğunun ve maddelerin konumunun katılımcıların motivasyonlarını etkilediğini belirtmiş ancak madde zorluğunun ve maddelerdeki şekil, tablo, görsellerin motivasyonu etkilemediğini belirtmiştir. Dört aşamalı testte testin III. aşamasındaki test maddelerinin seçenekleri kısa cümleler ve kavram yanılgılarından oluşmuştur. Ancak bu durumda katılımcıların test maddelerine yanıt verme süresinde anormal bir durum gözlemlenmemiştir. Buradan hareketle katılımcıların kavram yanılgısı yaşadıkları soru maddelerinde kendilerine güvendiklerini söyleyebiliriz. Katılımcılar dört aşamalı testin III. aşamasına istikrarlı ve makul yanıtlar vermişlerdir.

Sonuç

Bu çalışmada DADKT ve ÇSKT için öğretmen adaylarının yanıtlama performansları ve yanıtlama süreleri incelenmiştir. Bu amaçla çalışmanın ilk alt probleminde DADKT ile ÇSKT'nin yanıtlama performansları karşılaştırılmıştır. DADKT'nin I. aşamasının ortalama yanıtlama performansı 3,84 ve DADKT'nin III. aşamasının ortalama yanıtlama performansı 3,11 olarak bulunmuştur. ÇSKT'nin ortalama yanıtlama performansı 3,45 olarak hesaplanmıştır. Bu değerlere göre öğretmen adayları DADKT'nin I. ve III. aşamalarını yanıtlarken neden bölümünün olduğu III. aşamada çözüm içeriğinin olduğu I. aşamaya göre yanıtlama ortalama yanıtlama performansları daha düşük çıkmıştır. Bunun sebebi öğretmen adaylarının nedensel ilişki kurarken zorlanmasıdır.

Çalışmanın ikinci alt probleminde DADKT ve ÇSKT' nin yanıtlama süreleri karşılaştırılmıştır. Öğretmen adaylarının DADKT'nin I. aşaması için yanıtlama süresi 10:47 sn; III. aşaması için 7:15 sn olarak bulunmuştur. Öğretmen adayları nedensel ilişkinin incelendiği aşamada daha kısa sürede çözüm davranışı göstermişlerdir. Öğretmen adaylarının ÇSKT için ise yanıtlama süresi 12:52 sn olarak belirlenmiştir. Yapılan hesaplamalardan elde edilen sonuçlarda öğretmen adaylarının DADKT'nin I. aşaması ve ÇSKT'nin yanıtlama süreleri birbirlerine yaklaşık değerler gösterdiği; ancak DADKT'nin

I. aşamasının ve ÇSKT'nin yanıtlama sürelerinin DADKT'nin III. aşamasının yanıtlama sürelerinden farklı olduğu ve daha uzun sürede yanıtlandığı gözlemlenmiştir.

Öneriler

Bu çalışmada kullanılan testler öğretmen adayları için düşük riskliydi. Bu durumun öğretmen adaylarının motivasyonlarını etkilediği düşünülmektedir. İleriki çalışmaların yüksek riskli testler için yapılması önerilmektedir. Bu çalışma bilgisayar ortamında online test olarak yapılmıştır. Bu yüzden öğretmen adaylarının testi nerede, nasıl, ne şekilde cevaplandığının gözlemlenebilme imkânı olmamıştır. İleriki çalışmalarda yanıtlama süresi ve yanıtlama performansını gözlemlemek için akıllı telefon, tablet, akıllı kalemler gibi daha teknolojik cihazlar kullanılabilir (Moharkan, Choudhury, Gupta ve Raj, 2017; Edgecomb, Schaack ve Marggraff, 2014; Mehlhorn, Parrott, Mehlhorn ve Burcham, 2011). Bu çalışmada öğretmen adaylarına soru maddelerine tekrar dönme ve cevaplama şansı verilmemiştir. Ancak yapılan çalışmalarda tekrar cevaplama hakkı verildiğinde iç geçerlik oranlarının arttığı kanıtlanmıştır (Türkoguz, 2020). İleriki çalışmalar için tekrar cevaplama hakkının verilmesi kavram yanılgıları için farklı sonuçlar ortaya çıkarabileceğinden önerilmektedir.

Etik ve Çıkar Çatışması

Bu çalışma, Dokuz Eylül Üniversitesi Eğitim Bilimleri Enstitüsü Matematik ve Fen Bilimleri Eğitimi Anabilim Dalı Fen Bilgisi Öğretmenliği proramında yapılan "Yanıtlama süresine göre dört aşamalı kimya tanı testinin iç geçerlik oranlarının eşik değerinin belirlenmesi" başlıklı tezden üretilmiştir. Bu araştırmanın 03.05.2021 tarih ve E-87347630-640.99-52264 sayılı kararıyla Dokuz Eylül Üniversitesi Etik Kurulu'ndan Etik İzni bulunmaktadır. Araştırmanın yazarları olarak, verilerin toplanması, analizi ve araştırmanın tüm süreçlerinde etik kurallara uygun davrandığımızı beyan ederiz. Yazarlar arasında herhangi bir çıkar çatışması olmadığını bildirmişlerdir.

KAYNAKÇA

Arbuckle, J. (2008). *AMOS 17.0 user's guide*. SPSS Inc.

Bademci, V. (2006). Tartışmayı sonlandırmak: Cronbach'ın alfa katsayısı, iki değerli [0,1] ölçümlenmiş maddeler için kullanılabilir. *Kazım Karabekir Eğitim Fakültesi Dergisi, 13,* 438-446.

Bademci, V. (2007). *Ölçme ve araştırma yöntembiliminde paradigma değişikliği: Testler Güvenilir Değildir.* Ankara: Yeniyap Yayınları.

Bolsinova, M., De Boeck, P. ve Tijmstra, J. (2017). Modelling conditional dependence between response time and accuracy. *Psychometrika*, *82*(4), 1126-1148. DOI: 10.1007/s11336-016-9537-6.

Bulut, O. (2015). An empirical analysis of gender-based DIF due to test booklet effect. *European Journal of Research on Education, 3*(1), 7-16.

Bulut, O. (2015). Madde tepki kuramının akademik personel ve lisansüstü eğitimi giriş sınavı'na uyarlanması: uygulamadaki sorunlar ve öneriler. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi, 6*(2), 313-330.

Çokluk, Ö., Şekercioğlu, G. ve Büyüköztürk, Ş. (2012). *Sosyal bilimler için çok değişkenli istatistik SPSS ve LISREL uygulamaları (2. baskı).* Ankara: Pegem.

DeMars, C.E. (2000). Test stakes and item format interactions. *Applied Measurement in Education, 13*(1), 55-77. DOI: 10.1207/s15324818ame1301_3.

Edgecomb, T. L., Van Schaack, A., & Marggraff, J. (2014). *U.S. Patent No. 8,638,319*. Washington, DC: U.S. Patent and Trademark Office.

Elbaum, T., Golan, L., Lupu,, T., Wagner , M. ve Braw, Y. (2019) Establishing supplementary response time validity indicators in the word memory test (WMT) and directions for future research, Applied Neuropsychology: Adult. DOI: 10.1080/23279095.2018.1555161.

Ferrando, P.J. ve Lorenzo-Seva, U. (2007). An ıtem response theory model for ıncorporating response time data in binary personality ıtems. *Applied Psychological Measurement*, *31*(6), 525–543. DOI: 10.1177/0146621606295197.

Guo, H., Rios, J.A., Haberman, S., Liu, O.L., Wang, J. ve Paek, I. (2016) A new procedure for detection of students' rapid guessing responses using response time. *Applied Measurement in Education*, *29*(3), 173-183. DOI:10.1080/08957347.2016.1171766.

Gürçay, D. ve Gülbaş, E. (2015). Development of three-tier heat, temperature and internal energy diagnostic test. *Research in Science and Technological Education, 33*(2), 197-217. DOI:10.1080/02635143.2015.1018154.

Gürel, D.K., Eryılmaz, A. ve McDermott, L.C. (2015). A review and comparison of diagnostic ınstruments to ıdentify students' misconceptions in science. *Eurasia Journal of Mathematics, Science & Technology Education*, *11*(5), 989-1008.

Hestenes, D. ve Halloun, I. (1995). Interpreting the force concept inventory: A response to March 1995 critique by Huffman and Heller. *The Physics Teacher*, *33*, 502-506. DOI: 10.1119/1.2344278.

Kaltakçı, D. (2012). *Fizik öğretmen adaylarının geometrik optik ile ilgili kavram yanılgılarını ölçmek amacıyla dört basamaklı bir testin geliştirilmesi ve uygulanması (Yayınlanmamış Doktora Tezi).* Orta Doğu Teknik Üniversitesi, Fen Bilimleri Enstitüsü, Ortaöğretim Fen ve Matematik Alanları Eğitimi Ana Bilim Dalıı Ankara.

Karasar, N. (2005). *Bilimsel araştırma yöntemi (17. Baskı).* Ankara: Nobel yayın dağıtım.

Kline, R.B. (2011). *Principles and practice of structural equation modeling (3rd. Edition).* New York, NY: Guilford.

Kong, X.J., Wise, S.L. ve Bhola, D.S. (2007). Setting the Response Time Threshold Parameter to Differentiate Solution Behavior From Rapid-Guessing Behavior. *Educational and Psychological Measurement*, *67*(4), 606–619. DOI:10.1177/0013164406294779.

Lawshe, C.H. (1975). A quantitative approach to content validity. *Personnel psychology*, *28*(4), 563-575. DOI: 10.1111/j.1744-6570.1975.tb01393.x.

Lee, Y.H. ve Chen, H. (2011). A review of recent response-time analyses in educational testing. *Psychological Test and Assessment Modeling*, *53*(3), 359-379.

McClary, L.M. ve Bretz, S.L. (2012). Development and assessment of a diagnostic tool to ıdentify organic chemistry students' alternative conceptions related to acid strength. *International Journal of Science Education*, *34*(15), 2317-2341. DOI: 10.1080/09500693.2012.684433.

Mehlhorn, S., Parrott S., Mehlhorn, J., Burcham, T., Roberts, J., & Smartt, P. (2011, February). *Using digital learning objects to improve student problem solving skills.* Paper presented at the meeting of the Southern Agricultural Economics Association Annual Meeting, Corpus Christi, Texas. Retrieved on 26 November 2014 from http://ageconsearch.umn.edu/bitstream/98763/2/LivescribeSAEAPaperFINAL.pdf

Meyer, J.P. (2010). A Mixture rasch model with ıtem response time components. *Applied Psychological Measurement*, *34*(7), 521–538Meyer, J.P.(2010). A mixture rasch model with ıtem response time components. *Applied Psychological Measurement*, *34*(7), 521–538.

Moharkan, Z.A., Choudhury, T., Gupta, S.C., and Raj, G. (2017). *Internet of Things and its applications in E-learning*. In Proceedings of the 3rd International Conference on Computational Intelligence and Communication Technology (CICT). IEEE, Ghaziabad India, 1–5. DOI: 10.1109/CIACT. 2017.7977333.

Önsal, G. (2016). *Özel görelilik kuramıyla ilgili kavram yanılgılarını belirlemeye yönelik dört aşamalı bir testin geliştirilmesi ve uygulanması (Yüksek lisans tezi).* Gazi Üniversitesi, Eğitim Bilimleri Enstitüsü, Ortaöğretim Fen ve Matematik Alanları Eğitimi Ana Bilim Dalı, Fizik Öğretmenliği Bilim Dalı, Ankara.

Peşman, H. ve Eryılmaz, A. (2010) Development of a three-tier test to assess misconceptions about simple electric circuits. *The Journal of Educational Research*, *103*(3), 208-222. DOI: 10.1080/00220670903383002.

Ranger, J. ve Kuhn, J.T. (2012). A flexible latent trait model for response times in tests. *Psychometrika*, *77*, 31–47.

Romine, W.L., Schaffer, D.L. ve Barrow, L. (2015) Development and application of a novel rasch-based methodology for evaluating multi-tiered assessment ınstruments: validation and utilization of an undergraduate diagnostic test of the water cycle. *International Journal of Science Education*, *37*(16), 2740-2768. DOI:10.1080/09500693.2015.1105398.

Schnipke, D.L. ve Scrams, D.J. (1997). Modeling item response times with a two-state mixture model: A new method of measuring speededness. *Journal of Educational Measurement*, *34*, 213–232. DOI: 10.1111/j.1745-3984.1997.tb00516.x.

Setzer, J.C., Wise, S.L., Van Den Heuvel , J.R. ve Ling, G. (2013) An ınvestigation of examinee test-taking effort on a large-scale assessment. *Applied Measurement in Education*, *26*(1), 34-49. DOI:10.1080/08957347.2013.739453.

Stankov, L., Lee, J., Luo, W. ve Hogan, D.J. (2012). Confidence: A better predictor of academic achievement than self-efficacy, self-concept and anxiety? *Learning and Individual Differences*, *22*, 747–758. DOI: 10.1016/j.lindif.2012.05.013.

Swerdzewski, P.J., Harmes, J.C. ve Finney, S.J. (2011). Two approaches for identifying low-motivated students in a low-stakes assessment context. *Applied Measurement in Education*, *24*(2), 162-188. DOI: 10.1080/08957347.2011.555217.

Taber K.S. (2017). The Use of Cronbach's Alpha when developing and reporting research instruments in science education. *Res. Sci. Educ*., 1–24. Doi:10.1007/s11165-016- 9602-2.

Türkoguz, S. (2020). Comparison of Threshold Values of Three-Tier Diagnostic and Multiple-Choice Tests Based on Response Time. *Anatolian Journal of Education*, *5*(2), 19-36. DOI: 10.29333/aje.2020.522a.

Ünsal, A.A. (2019). *Fen bilgisi öğretmen adaylarının gaz basıncı konusundaki kavram yanılgılarının belirlenmesi (Yayımlanmamış yüksek lisans tezi)*. Hacettepe Üniversitesi Eğitim Bilimleri Enstitüsü, Ankara.

Van Der Linden, W.J. (2006). A lognormal model for response times on test items. *Journal of Educational and Behavioral Statistics, 31,* 181–204. DOI: 10.3102/10769986031002181.

Van Der Linden, W.J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika, 72,* 287–308. DOI: 10.1007/s11336-006-1478-z.

Van Der Linden, W.J., Klein Entink, R.H. ve Fox, J.P. (2010). IRT parameter estimation with response times as collateral information. *Applied Psychological Measurement*, *34*(5), 327–347. DOI:10.1177/0146621609349800.

Wang, J.R. (2004). Development and validation of a two-tier ınstrument to examine understanding of internal transport in plants and the human circulatory system. *Int J Sci Math Educ*, *2*(2), 131-157. DOI:10.1007/S10763-004-9323-2.

Weeks, J.P., Von Davier, M. ve Yamamoto, K. (2016). Using response time data to inform the coding of omitted responses. *Psychological Test and Assessment Modeling*, *58*(4), 671.

Weeks, J.P., Von Davier, M. ve Yamamoto, K. (2016). Using response time data to inform the coding of omitted responses. *Psychological Test and Assessment Modeling, 58*(4), 671-701.

Wise, S. ve Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education*, *18*(2), 163-183. DOI: 10.1207/s15324818ame1802_2.

Wise, S.L. (2014). The utility of adaptive testing in addressing the problem of unmotivated examinees. *Journal of Computerized Adaptive Testing, 2,* 1-17.

Wise, S.L. Bhola, D. ve Yang, S. (2006). Taking the time to improve the validity of low-stakes tests: The effortmonitoring CBT. *Educational Measurement: Issues and Practice*, *25*(2), 21–30. DOI: 10.1111/j.1745-3992.2006.00054.x.

Wise, S.L. ve DeMars, C.E. (2006). An application of item response time: The effort-moderated IRT model. *Journal of Educational Measurement, 43,* 19–38. DOI: 10.1111/j.1745-3984.2006.00002.x.

Wise, S.L. ve DeMars, C.E. (2010) Examinee noneffort and the validity of program assessment results. *Educational Assessment*, 15:1, 27-41. DOI: 10.1080/10627191003673216.

Wise, S.L. ve Gao, L. (2017) A General approach to measuring test-taking effort on computer-based tests. *Applied Measurement in Education*, 30(4), 343-354. DOI: 10.1080/08957347.2017.1353992.

Wise, S.L. (2019) Controlling construct-irrelevant factors through computer-based testing: disengagement, anxiety, & cheating. *Education Inquiry*, *10*(1), 21-33. DOI:10.1080/20004508.2018.1490127.

Wise, S.L., Pastor, D.A. ve Kong, X.J. (2009) Correlates of rapid-guessing behavior in low-stakes testing: ımplications for test development and measurement practice. *Applied Measurement in Education*, *22*(2), 185-205. DOI:10.1080/08957340902754650.

Wright, D.B. (2016). Treating all rapid responses as errors (TARRE) improves estimates of ability (slightly). *Psychological Test and Assessment Modeling, 58,* 15–31.

Wright, D.B. (2016). Treating all rapid responses as errors (TARRE) improves estimates of ability (slightly). *Psychological Test and Assessment Modeling, 58,* 15–31.

Yamamoto, K. (1995). *ETS Research Report Series: Estimating the effects of test length and test time on parameter estimation using the HYBRID model (TOEFL Technical Report TR-10).* Princeton, NJ: Educational Testing Service. DOI: 10.1002/j.2333-8504.1995.tb01637.x.

Yang, D.C. ve Sianturi, I.A.J. (2019). Assessing students' conceptual understanding using an online three-tier diagnostic test. *Journal of Computer Assisted Learning, p. 678-689.* DOI:10.1111/jcal.12368.

EXTENDED ABSTRACT

By using the response time of a test item, situations related to the participants' response performance or motivation can be determined. In the literature, studies examining responding performance depending on answering time were generally conducted in multiple-choice tests, but not many examples were found within the scope of diagnostic tests (DeMars & Wise, 2010; DeMars, Wise, & Smith, 2009). Within the scope of diagnostic tests, Türkoguz (2020) examined the response time and item response performance by using the threshold value. However, this study was conducted using a four-tiered diagnostic test. In the literature, there are no studies examining response time and response performance using the four-tiered diagnostic test. In addition, most of the studies in the literature were

conducted with face-to-face participation of participants. There are very few studies where tiered diagnostic tests are conducted online. For example; Yang and Sianturi (2019) conducted the three-tiered mathematics test online. However, this study is a three-tiered diagnostic test and is related to the field of mathematics. In the literature, no studies were found that used the four-tiered chemistry diagnostic test and examined online response time and response performance together. Research examining the relationships between response time and response accuracy in educational tests has begun since 1990. There are reserach that determine threshold values by making use of item response time and item response accuracy and accordingly, determine item response performance. The joint studies of Wise and DeMars on the relationship between item response accuracy and item response time are quite common. The majority of the studies mentioned in the literature were conducted within the scope of multiple-choice tests. There are no studies examining item response time and item response performances within the scope of tiered diagnostic tests or conceptual understanding tests. The problem statement of this study was stated as 'Is there a difference between the response times and therefore the response performances of the Four-Tiered Chemistry Diagnostic Test (FTCDT) and the Multiple Choice Chemistry Test (MCCT)?'. Two sub-problems were identified for the solution of the main problem. These: 1-Is there a difference between the response performances of FTCDT and MCCT?, and 2-Is there a difference between the response times of FTCDT and MCCT?**.** In this study, the relational survey model, one of the general survey models, which is one of the quantitative research methods, was used. The study model that aims to determine the change and level of change between two or more variables is called the relational survey model (Karasar, 2005). The study was conducted with the participation of 149 pre-service teachers studying in the science education department of Dokuz Eylül University, Buca Faculty of Education in the 2020-2021 academic year. Participants consist of 24 men and 125 women. ***Four-Tiered Diagnostic Chemistry Test (FTCDT) and Multiple Choice Chemistry Test (MCCT):*** FTCDT developed by Ünsal (2019) was used in the study. The test consists of 9 items on gas pressure and was developed with a pre-service science teacher. The measurement tool consists of four tiers, the first tier measures the level of knowledge and the third tier inquires about the justification of the answer given in the first tier. The second tier measures the confidence in the first tier and the fourth tier measures the confidence level in the third tier. The scientific information reliability of the FTCDT used in this study is KR-20 (according to the condition of receiving 1 point if all tiers are correct) 0.460; Misconception reliability KR-20 (according to the condition of giving wrong answers in the First and Third Tiers and being sure in the Second and Fourth Tiers) was calculated as 0.570. Explanatory and confirmatory factor analyses were performed for the construct validity of the 9-item FTCDT. It was confirmed in these analyses that FTCDT exhibits a 4-factor structure. MCCT regarding gas pressure was also used in this study. MCCT is the first tier of the 9-item FTCDT related to gas pressure developed by Ünsal (2019). MCCT was used by subtracting other tiers from FTCDT. For the KR-20 reliability analysis of MCCT, correct answers were given 1 and other answers were given 0. For this study, the KR-20 reliability coefficient of MCCT was found to be 0.520. Experts' opinions were taken for the content validity of the measurement tools. Data were collected by FTCDT and MCCT applied to students studying in the Science Education Department at Dokuz Eylül University Buca Faculty of Education. During the data collection process of the study, the COVID-19 pandemic occurred and the lessons were conducted online via computer. Therefore, the data were collected in computer and online test environments. Participants did not have the opportunity to see or answer the test items again as feedback. For both types of tests, the direction of test progression is linear and not retrospective. In the first sub-problem of this study, the responding performances and response times of pre-service teachers for FTCDT and MCCT were examined. The mean response performance of the first tier of FTCDT was found to be 3.84, and the mean response performance of the third tier of FTCDT was 3.11. The mean response performance of MCCT was calculated as 3.45. According to these values, while pre-service teachers answered the first and third tiers of FTCDT, their mean response performance was lower in the third tier, which included the why section, compared to the first tier, which contained the solution content. In the second sub-problem of the study, the response times of FTCDT and MCCT were compared. The response time of pre-service teachers for the first tier of FTCDT was 10:47 seconds; It was found to be 7:15 seconds for the third tier. Pre-service teachers showed solution behaviour in a shorter time at the

tier where the causal relationship was examined. The response time of pre-service teachers for MCCT was determined as 12:52 seconds. According to the results obtained from the calculations, the response times of the pre-service teachers for the first tier of FTCDT and MCCTT showed approximate values to each other; However, it was observed that the response times of the first tier of FTCDT and MCCT were different from the response times of the third tier of FTCDT and were answered in a longer time. The tests used in this study were low-risk for pre-service teachers. It is thought that this situation affects the motivation of pre-service teachers. It is recommended that future studies be conducted for high-risk tests. This study was conducted as an online test in a computer environment. Therefore, it was not possible to observe where, how and in what way the pre-service teachers answered the test. In future studies, more technological devices such as smartphones, tablets, and smart pens can be used to observe response time and response performance (Moharkan, Choudhury, Gupta, & Raj, 2017; Edgecomb, Schaack, & Marggraff, 2014; Mehlhorn, Parrott, Mehlhorn, & Burcham, 2011). In this study, pre-service teachers were not given the chance to return to the question items and answer them. However, studies have proven that internal validity rates increase when the right to answer again is given (Türkoguz, 2020). It is recommended to give the right to answer again for future studies, as it may lead to different results for misconceptions.