Eğitim ve Bilim

Cilt 49 (2024) Sayı 220 239-257

Orijinal Makale

Türkçe D2 Alımlama Becerileri Testlerini Geçerleme: Standart Belirleme Yöntemi Aracılığıyla Argüman Temelli Bir Geçerlik Çalışması *

Yiğit Savuran [1], Zühal Çubukçu [2]

Öz

Bu makale, ikinci dil olarak Türkçe (Türkçe D2) öğrenen yetişkinlere yönelik daha büyük bir araştırma kapsamında geliştirilen dört seviye testi için ölçme-kullanım argümanına dayalı bir geçerleme çalışmasını rapor etmektedir. Test puanları, testlerin öğrenenleri amaçlanan Diller için Avrupa Ortak Başvuru Metni (D-AOBM) seviyelerine doğru bir şekilde sınıflandırdığına dair geçerlik iddiasını oluşturan veriler olarak ele alınmıştır. Her biri dinleme ve okuma bölümlerinden oluşan dört seviye testi (A1, A2, B1 ve B2), Ön-A1 ve C1 seviyesindekiler de dahil olmak üzere karışık öğrenci gruplarına uygulanmıştır. Dört testin her biri için kesme puanları Angoff yöntemiyle belirlenmiş ve geçerlik iddiası için destek olarak kabul edilmiştir. Gerekçe sağlamak amacıyla, her bir seviye için kesme puanının altında ve üstünde beklenen ve gözlenen öğrenci sayıları arasındaki istatistiksel anlamlılığı değerlendirmek için yapılan ki-kare uyum iyiliği testleri için bir test katılımcısının D-AOBM seviyesinde olma olasılığının %50 olduğu varsayılmıştır. Kabul edilebilir madde güçlüğü ve ayırt edicilik endeksleri eşliğinde öğrenci puanlarının dağılımı, kesme puanlarının bitişik seviyeler arasındaki aralıklarda bulunması ve dört teste ait ki-kare analizleri, testlerin amaçlanan öğrenci performansını geçerli bir şekilde gösterme potansiyeline sahip olduğu sonucuna varılmasını sağlamıştır. Yenilikçi tasarım, veri toplama ve analiz teknikleri sunan bu çalışma, Türkçe D2 alanı çalışanları için güçlü deneysel verilere dayanan teorik, yöntemsel ve uygulamaya dönük bilgiler sunmaktadır.

Anahtar Kelimeler

Türkçe D2
Test geçerleme
Ölçme-kullanım argümanı
Standart belirleme

Makale Hakkında

Gönderim Tarihi: 02.06.2023
Kabul Tarihi: 12.06.2024
Elektronik Yayın Tarihi: 27.10.2024

DOI: 10.15390/EB.2024.12959

Giriş

Uluslararası öğrenciler arasında Türkiye'de yükseköğrenim görmek için artan bir talep bulunmaktadır (Yükseköğretim Kurulu, 2023). Bu öğrencilerin çoğu Türkiye Bursları programından yararlanmaktadır. Burs, üniversite ve bölüm yerleştirmeleri, öğrenim ücretleri, konaklama, uçak biletleri, sağlık sigortası, aylık burs ve Türkçe dil kursu gibi unsurları içeren kapsamlı bir program

* Bu makale Yiğit Savuran'ın Zühal Çubukçu danışmanlığında yürüttüğü "Türkçenin yabancı dil olarak öğretiminde temel ve ara düzey betimleyicilerin geliştirilmesi" başlıklı doktora tezinden üretilmiştir.
Bu çalışmanın yazarlarından Zühal Çubukçu hocamız makalenin değerlendirmesi sırasında aramızdan ayrılmıştır. Makale, 1. yazar tarafından hocamızın anısına kendisine adanmıştır.

[1] Florida Üniversitesi, Sosyal ve Beşeri Bilimler Fakültesi, Dilbilim Bölümü, ABD; Anadolu Üniversitesi, Yabancı Diller Yüksekokulu, Türkiye, yigitsavuran@gmail.com
[2] Eskişehir Osmangazi Üniversitesi, Eğitim Fakültesi, Eğitim Bilimleri Bölümü, Türkiye, zuhal_cubukcu@hotmail.com

sunmakta ve bu da bursu uluslararası öğrenciler için cazip kılmaktadır. Program yılda yaklaşık 15000 öğrenciyi ağırlamaktadır ve bugüne kadar 150000'den fazla mezun vermiştir (Türkiye Bursları, t.y.). Ayrıca, son dönemde yaşanan siyasi gerilimler, sosyo-ekonomik çatışmalar ve *Erasmus+* (European Commission, t.y.) ve *Study in Turkey* (Türkiye Yükseköğretim Kurulu, 2024) gibi çeşitli kurumlar tarafından sunulan hareketlilik programları gibi nedenlerle, giderek artan sayıda öğrenci yükseköğrenim için Türkiye'yi tercih etmektedir. İster burslu ister kendi imkânlarıyla eğitim gören öğrenciler olsun, öğrenciler genellikle bir yıllık yoğunlaştırılmış Türkçe (D2) eğitim programı aracılığıyla Türkçe öğrenmek için ülkenin dört bir yanına dağılmış Türkçe Dil Merkezlerinde öğrenim görmektedir.

Türkiye'deki Türkçe Dil Merkezleri, müfredatlarını ve değerlendirme uygulamalarını tasarlarken genellikle D-AOBM ilkelerini ve terminolojisini takip etmektedir. Örneğin burslu öğrenciler, programlarının bir parçası olarak D-AOBM B2 veya C1 seviyesini başarıyla tamamlamak zorundadır. Bu gereklilik ve Türkçe D2 öğrenen öğrenci sayısındaki artış, araştırmacılar için D-AOBM ilkelerine dayalı iyi yapılandırılmış eğitim programları ve güvenilir ve geçerli ölçme araçları geliştirme fırsatı yaratmaktadır. D-AOBM'de (Council of Europe [CoE], 2001, 2020) belirtildiği üzere, bu tür araçlar geliştirmek için araştırmacıların çeşitli seviyelerde öğrenen performansını gösteren yerel olarak tasarlanmış ve uyarlanmış betimleyicilere sahip olması gerekir. Bu ihtiyaç dahilinde, ikinci dil olarak Türkçe öğrenenlerin ihtiyaçlarına yönelik alımlama becerileri için seviye testleri ve üretim becerileri için görevlerin uygulanmasıyla birlikte bu tür betimleyicilerin geliştirilmesine odaklanan daha geniş bir çalışma ortaya konulmuştur.

Bu geniş çalışma, özellikle yükseköğretim seviyelerine odaklanarak, Türkçe D2 öğrenenlerin ölçme ve değerlendirme ihtiyaçları için dört aşamada ('hazırlık', 'geliştirme', 'uygulama' ve 'geçerleme') betimleyiciler geliştirmeyi amaçlamıştır. Hazırlık aşamasında, ilk olarak betimleyicilerin uyarlanması üzerinde çalışılmıştır. Geliştirme aşamasında, değerlendirme amaçları için geliştirilecek test ve görevlere uygun betimleyicilerin seçimi gerçekleştirilmiştir. Uygulama aşamasında öğrenciler alımlama becerileri için dinleme ve okuma bölümlerinden oluşan seviye testlerine katılmışlardır. Benzer şekilde üretim becerileri için konuşma ve yazma görevlerine cevap vermişlerdir. Geçerleme aşamasında ise her bir test ve görev için kesme puanları belirlenerek standart belirleme çalışması yürütülmüştür.

Hazırlık ve geliştirme aşamaları bir önceki çalışmamızda ayrıntılı olarak raporlanmıştır (Savuran ve Çubukçu, 2021). Ancak bu çalışma, uygulama aşamasında öğrencilere uygulanan testlerin deneysel geçerliğini sağlamayı ve ölçme-kullanım argümanı doğrultusunda (Bachman ve Palmer, 2010; Papageorgiou ve Tannenbaum, 2016) D-AOBM seviyesinde olma olasılığının %50 olduğu varsayımı (De Jong ve Benigno, 2017; Harsch, 2019; North, 2000) aracılığıyla geçerleme aşamasında yürütülen standart belirleme sürecini raporlamayı amaçlamaktadır.

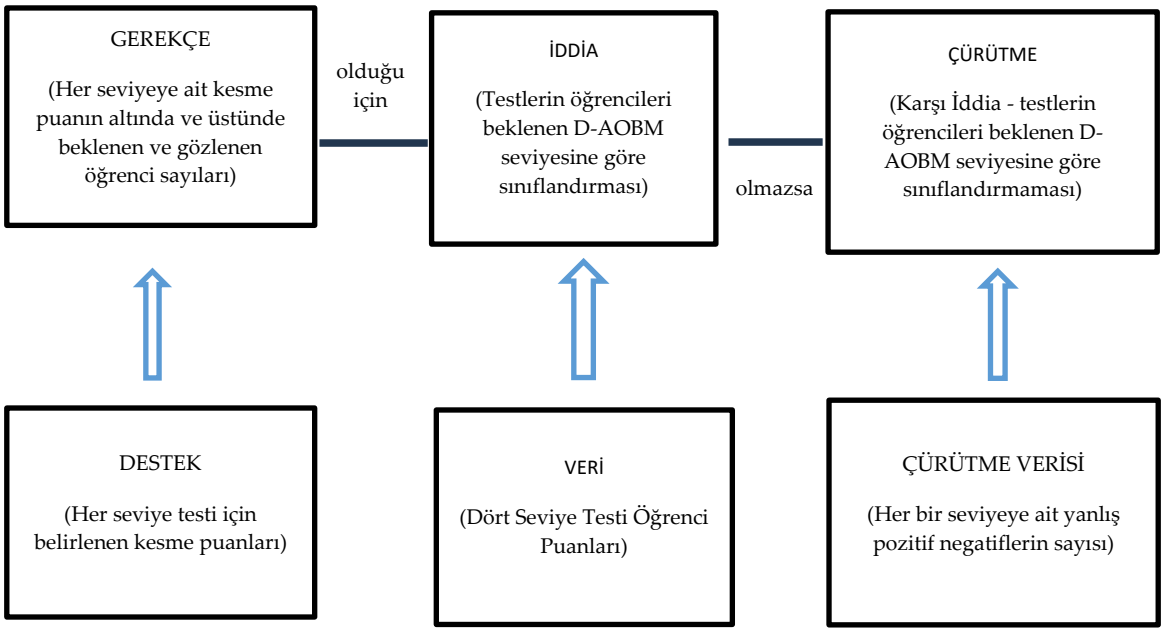
Dil testlerindeki iki beceri kategorisi - alımlama ve üretim - kullanılan madde türleri nedeniyle geçerleme sürecinde, özellikle de argümana dayalı geçerlemede uygulanan metodolojide bazı farklılıklara sahiptir. Alımlama becerilerini (dinleme ve okuma) ölçen testlerin çoğunda çoktan seçmeli maddeler kullanılırken, üretim becerilerini (yazma ve konuşma) ölçen testlerde katılımcıların genellikle açık uçlu yanıtlar vermesi gerekir. Bununla birlikte, bir test için geçerlik argümanı toplama amacıyla (Kane, 2006, 2013), bir araştırmacının değerlendirilen beceriye bağlı olarak farklı unsurları analiz etmesi gerekir. Alımlama becerilerinde, testin kullanımı ve puanların yorumlanmasına ilişkin bir geçerlik argümanı oluşturmak için toplanabilecek kanıtlar çoğunlukla testin kendisi, test puanları ve test katılımcısının her bir madde için yaptığı seçimlerdir. Öte yandan, üretim becerilerine yönelik testlerin geçerlik kanıtı için, bir araştırmacı derecelendirme ölçeğinin geçerliğinin yanı sıra maddeleri ve test katılımcılarının bu maddelere verdikleri açık uçlu yanıtları inceler (Chapelle ve Voss, 2014). Bu bağlamda, geçerlik için farklı bir yaklaşım gerektirdiğinden, üretim becerileri olan yazma ve konuşmanın geçerliği bu makalenin kapsamı dışında bırakılmıştır. Bu nedenle mevcut çalışma, ölçme-kullanım argümanı ve standart belirleme yoluyla dört D-AOBM seviyesinde (A1-B2) alımlama becerileri (dinleme ve okuma) testlerinin geçerliğine odaklanmaktadır.

Kuramsal Altyapı: Argüman Temelli Geçerleme

Geçerlik, eğitimde ölçme ve değerlendirme alan yazınında uzun süredir tartışılmaktadır (Cizek, 2012; Cureton, 1951; Messick, 1989) ve son zamanlarda testin kendisinin değil, test puanlarına dair yorumların ve onların kullanımlarının geçerli kılınabileceği fikri kabul edilmektedir (American Educational Research Association [AERA], 2014; Bachman, 2005; Fulcher, 2015; Kane, 1994). Bu fikir birliğine dayanarak Chapelle ve Voss (2014), aralarında 'kanıt toplama' ve 'argüman temelli' yaklaşımların son zamanlarda öne çıktığı dört geçerleme yaklaşımı önermektedir (Cheng ve Sun, 2015).

Geçerliğe yönelik argüman temelli yaklaşım, test bağlamına ve puanların kullanım ve yorumlarına dayalı argümantasyon oluşturmak için pratik bir kılavuz sağladığı için eğitimde ölçme araştırmacıları arasında büyük ilgi görmüştür (Bachman, 2005; Chapelle, Enright ve Jamieson, 2010; Kane, 2013; Knoch ve Chapelle, 2018). Bir geçerlik argümanı oluşturmak için araştırmacılardan, test puanlarının önerilen kullanımlarını ve yorumlarını destekleyen geçerlik kanıtları toplamaları beklenir (Chapelle, Enright ve Jamieson, 2008; Im, Shin ve Cheng, 2019); ancak, geçerlik sürecinin temel bileşenleri yalnızca kanıtın kendisinden ziyade mantık ve argümantasyondur (Lavery, Bostic, Kruse, Krupa ve Carney, 2020).

Argüman temelli yaklaşım, farklı amaçlar için iki tür argüman kullanır: Yorumlama/Kullanım Argümanı ve Geçerlik Argümanı (Kane, 2006, 2013). Argüman türünden bağımsız olarak Kane'in yaklaşımı, puanların yorumlanmasının nasıl planlanacağı, araştırmanın nasıl yürütüleceği, geçerlik argümanı için araştırma sonuçlarının nasıl düzenleneceği ve geçerlik argümanının nasıl sorgulanacağı gibi pratik geçerleme yöntemleri sunmaktadır (Chapelle vd., 2010). Bachman ve Palmer (2010), Kane'in çalışmasını temel alarak ve geçerlik argümanı türüyle ilişkili şekilde, Kane'in yaklaşımının üzerine inşa edildiği bir model olan Messick'in (1989) birleşik modelini dahil ederek testlerin kullanımının altını çizen Ölçme-Kullanım Argümanını (ÖKA) sunmuştur (Kane, 2013).



Şekil 1. Ölçme-Kullanım Argümanı Yapısı (Papageorgiou ve Tannenbaum, 2016'dan uyarlanmıştır)

ÖKA çerçevesi, Şekil 1'de gösterildiği gibi, kökleri verilere dayanan ve verilerle desteklenen iddialar üretilmesini önermektedir. Papageorgiou ve Tannenbaum'un (2016) ifade ettiği gibi, iddiaları ortaya koymak için "bir gerekçe belirtilir ve bu gerekçe de teorik veya deneysel kanıtlardan kaynaklanan desteklerle desteklenir" (s. 112). Bu çerçevede, çürütmeler 'karşı iddialar' olarak hizmet edebilir ve yeterli verilerle kanıtlandığı takdirde ilk iddialara karşı kullanılabilir (Bachman, 2005; Papageorgiou ve Tannenbaum, 2016). Papageorgiou ve Tannenbaum (2016), Bachman ve Palmer'dan (2010) yola çıkarak bir ÖKA çalışmasına entegre edilebilecek dört iddia sunmaktadır. İddia 1,

değerlendirmelerin kullanımının ve ardından alınan kararların paydaşlar için faydalı sonuçlar doğurduğunu iddia etmektedir. İddia 2, "değerlendirmeye dayalı yorumlar temelinde alınan kararların toplum değerlerini, ilgili yasal hükümleri dikkate aldığını ve bu kararlardan etkilenen paydaşlar için adil olduğunu" ileri sürmektedir (Bachman ve Palmer, 2010, s. 111). İddia 3, değerlendirilen becerilerin yorumlanmasının müfredat, teori ve hedef dil kullanım alanı ile uyumlu olması gerektiğini vurgular. İddia 4, 'değerlendirme kayıtları' olarak adlandırılan sayısal puanlar veya performans açıklamaları olarak ortaya çıkan değerlendirme sonuçlarının tutarlılığı ile ilgilidir. Her ne kadar bu iddiaların testlerin sonuçlarıyla başladığı ve çıkarımsal olarak bağlantılı olduğu öne sürülse de, bu çalışmanın odak noktası, test sonuçlarının yorumlanmasıyla ilgili olan 3. ve 4. İddialardır (Papageorgiou ve Tannenbaum, 2016).

Yukarıda özetlenen ÖKA çerçevesine dayanan bu çalışma, Türkçe D2 alımlama becerileri seviye testlerini, 'iddiaların' (yani, öğrencilerin test puanlarına göre amaçlanan D-AOBM seviyesine sınıflandırılması) formüle edildiği 'veriler' olarak görmektedir. Bu amaçla, karma öğrenci grupları fiziksel sınıf ortamlarında dört seviye testine katılmıştır. Örneğin, B1 seviye testi A2, B1 ve B2 öğrencilerine uygulanmış, böylece öğrenci performansının seviyeye göre derinlemesine analiz edilmesi ve dolayısıyla 'veri' sağlanması amaçlanmıştır. Farklı seviyelerdeki öğrencilerden alınan puan ortalamaları, Angoff standart belirleme yöntemiyle belirlenen kesme puanlarıyla karşılaştırılarak verilerden iddiaya yapılan çıkarım güçlendirilmiştir. Bu karşılaştırmalara daha fazla destek sağlamak ve geçerlik iddiasını desteklemek amacıyla, ki-kare uyum iyiliği testleri için literatürde tartışıldığı gibi (De Jong ve Benigno, 2017; Harsch, 2019), belirli bir seviyedeki maddeleri doğru çözme için %50 olasılık kavramı benimsenmiştir. Ki-kare analizleri, aynı seviye testini alan öğrencilerin gözlenen ve beklenen sayıları arasındaki farklarda istatistiksel anlamlılık olup olmadığını belirlemeye çalışmıştır. Testlerin öğrencileri amaçlanan seviyelere doğru bir şekilde sınıflandırmadığını ileri süren eden karşı iddialar, gözlemlenen ve beklenen rakamlar (yani, yanlış pozitifler ve yanlış negatifler) arasındaki farklılıkları açıklayan 'çürütücü veriler' ile desteklenen 'çürütmeler' olarak ele alınmıştır.

Bir D-AOBM Seviyesinde Olmak Ne Demek: %50 Olasılık Varsayımı

Değerlendirmelerin somut unsurlara dayandığı bilim dallarından farklı olarak, dil yeterliliğinde ölçülen yapılar soyuttur ve dil ölçme ve değerlendirme uzmanları için önemli zorluklar teşkil eder. Douglas'ın (2010) ölçüm birimlerinin doğasındaki potansiyel farklılıkları simgeleyen 'lastik cetvel' metaforu, dil yeterliliği kalibrasyonunda hassasiyetin gerekliliğinin altını çizmektedir. Bu hassasiyet, North'un (2000) D-AOBM'nin açıklayıcı ölçeklerinin ve tanımlayıcılarının temelini oluşturan çalışmasında örneklediği şekilde, ancak gizil dil yeterlilik özellikleri hem nicel hem de nitel yönlerle desteklenen açık özellikler haline geldiğinde elde edilebilir. North'a (2000) göre, çalışma bir öğrencinin bir seviyenin başlangıcında, o seviyedeki maddeleri ve görevleri çözme olasılığının %50 olduğunu varsayar ve bu olasılık modeline dayalı olarak seviye sınırlarını belirler. Bu varsayım, görevler basitleştikçe, bir öğrenenin maddeleri başarıyla çözme olasılığının arttığını ve bunun tersinin de geçerli olduğunu göstermektedir (Harsch, 2019). D-AOBM dışında yeterlilik çerçeveleri geliştiren çalışmalarda da benzer olasılık değerleri rapor edilmiştir. Zwick, Senturk, Wang ve Loomis (2001) madde sabitleme ve eşleştirme için .50 ile .80 arasında bir aralık önermiştir. Benzer şekilde, Gomez, Noah, Schedl, Wright ve Yolkut (2007), bir maddenin, belirli bir düzeydeki sınav katılımcılarının %50'sinden fazlasının doğru yanıt vermesi ve daha düşük bir düzeydeki sınav katılımcılarının %50'sinden azının doğru yanıt vermesi durumunda öğrenen yeteneğine işaret edebileceğini öne sürmüştür. De Jong ve Benigno (2017), İngilizce Küresel Ölçeği'ni (İng. 'GSE') diğer ölçeklerle uyumlaştırma çalışmalarında, "... B1'de olmanın, B1'deki tüm görevlerin %50'sini başarıyla yerine getirme beklentisi veya B1'deki herhangi bir görevi başarıyla yerine getirme şansının %50 olması anlamına geldiğini" belirtmektedir (s. 5). Tablo 1, De Jong ve Benigno'un (2016) North'un (2000) çalışmasını dayanak alarak geliştirdikleri farklı D-AOBM seviyelerindeki öğrenenler için sundukları varsayımsal olasılıkları göstermektedir.

Tablo 1. Farklı D-AOBM Seviyelerinde Öğrenenler için Olasılık Tahminleri (De Jong ve Benigno, 2016, s.11'den uyarlanmıştır)

		Seviyedeki Öğrenenler				
		A1	**A2**	**B1**	**B2**	**C1**
Seviyedeki Görev ve Maddeler	**C1**	.00	.00	.03	.17	**.50**
	B2	.00	.02	.12	**.50**	.83
	B1	.03	.12	**.50**	.88	.97
	A2	.18	**.50**	.88	.98	1.00
	A1	**.50**	.82	.97	1.00	1.00

Teorik değerlendirmelere dayandığı için %50 olasılık eşiğinin esnek olmadığı düşünülmemelidir. Belirli bir düzeydeki maddeleri ele almak için gereken bilgi, beceri ve yetenekleri tanımlamak ve beyan etmek test geliştiricilerin rolüdür (Papageorgiou, Xi, Morgan ve So, 2015). Sonuç olarak, Harsch'ın (2019) gözlemlediği gibi, belirli bir seviyede olmanın ne anlama geldiği sorusunun kesin bir cevabı yoktur: "...sınav katılımcılarının yeterlilik seviyelerine göre sınıflandırılması, 'katı' istatistiksel analizlere ek olarak insan yorumunu da gerektirmektedir" (s. 81). Bu nedenle, bu çalışmada geçerlik testlerini şekillendiren betimleyicilerin, bir D-AOBM seviyesine ulaşmak için gerekli bilgi, beceri ve yetenekleri niteliksel olarak açıkça belirttiği iddia edilmektedir. D-AOBM tanımlayıcılarının doğasında olmamasına rağmen, %50 olasılık varsayımı testin ölçmek için tasarlandığı D-AOBM seviyesindeki beklenen öğrenci sayısını belirlemek amacıyla kullanılmıştır. Bir başka deyişle, %50 olasılık varsayımı Rasch ölçeklemesindeki kökeni göz önünde bulundurularak, klasik test teorisi bağlamında ki-kare analizleri eşliğinde öğrenci performansını niceliksel olarak ortaya koymak için benimsenmiştir. Tablo 1'de görüldüğü gibi, belirli bir seviyedeki öğrencilerin daha yüksek seviyedeki görevlerle başarılı bir şekilde başa çıkma olasılıkları .12 ile .18 arasında değişmektedir; bu nedenle istatistiksel analizlerde de bu aralık ele alınmıştır.

Sonuç olarak, olasılık varsayımlarını ve argüman temelli geçerleme yaklaşımını temel unsurlar olarak alan bu çalışma, aşağıdaki araştırma soruları öncülüğünde dört seviyedeki Türkçe D2 alımlama becerileri testlerini geçerlemeyi amaçlamıştır:

1. Okuma ve dinleme seviye testleri öğrencileri hedeflenen D-AOBM seviyelerine göre sınıflandırmakta mıdır?
2. Her seviye testi için belirlenen teorik kesme puanları gerçek öğrenci puanlarını desteklemekte midir?
3. Kesme puanlarının altında ve üstünde beklenen ve gözlenen öğrenci sayıları geçerlik iddiası için gerekçe olarak kabul edilecek düzeyde istatistiksel olarak anlamlı mıdır?
4. Yukarıda belirtilen iddiaların geçerli olmadığı herhangi bir durum var mıdır? Varsa, karşı iddia olarak değerlendirilebilecek yeterli veri ile desteklenmekte midir?

Yöntem

Bu çalışma, A1, A2, B1 ve B2 seviyelerindeki dört beceri için Türkçe D2 yeterlilik betimleyicileri geliştirmeyi amaçlayan daha geniş bir çalışmanın parçasıdır. Bu kapsamlı çalışma hazırlık, geliştirme, uygulama ve geçerleme olmak üzere dört aşamada gerçekleşmiştir. Ancak mevcut çalışma, uygulama ve geçerleme aşamalarından elde edilen sonuçların raporlanmasına odaklanmakta ve çeşitli yenilikçi metodolojileri betimsel şekilde ele almaktadır.

Katılımcılar ve Materyaller

Uygulama aşamasındaki katılımcılar, Anadolu Üniversitesi Türkçe Dil Merkezi'ne kayıtlı 384 Türkçe D2 öğrencisidir. Üniversite öğrencilerinin seçilmesinin iki temel nedeni vardır: Birincisi, araştırmacıların yükseköğretime bağlı olmaları ve bu akademik düzeydeki öğrencilerle çalışma konusunda önemli deneyime sahip olmaları; ikincisi, D-AOBM tanımlayıcılarının özellikle genç yetişkinler ve yetişkin öğrenenler için uygun görülmesidir (Benigno ve De Jong, 2016). Öğrenciler Orta

Doğu, Doğu Avrupa, Afrika ve Orta Asya gibi bölgelerden gelen ve Türkiye'deki üniversitelerde lisans ve lisansüstü eğitimlerine kayıtlı çeşitli bir grubu temsil etmektedir. Katılımcıların çoğu ya Türkiye Bursları programından yararlanmış ya da kendi imkanları ile Türkiye'de okumaktadır. Katılımcılar arasında, Türki ülkelerden gelenler hariç, öğrencilerin çoğunun Türkçe konusunda herhangi bir eğitim geçmişi bulunmamaktadır; bu nedenle, genellikle bir seviye belirleme sınavına dayalı olarak A1 seviyesinde öğrenime başlamışlardır. Türkçe Hazırlık programı, her seviye için 7-8 haftalık modüllere bölünerek yıl boyunca haftada ortalama 25 saati içeren dil eğitimini kapsamaktadır. Program gereksinimi olarak, her bir seviyeyi başarılı olarak bitirebilmek için öğrencilerin dilbilgisi ve kelime bilgisinin yanı sıra dört beceriyi de içeren kapsamlı testlerden 70 veya daha yüksek bir puan almaları gerekmektedir; 70'in altındaki puanlar seviye tekrarı gerektirmektedir. Beş seviyeyi tamamladıktan sonra (A1'den başlayarak), ilgili üniversite programlarına başlamak için yıl sonuna kadar B2 ya da C1 seviyesine ulaşmaları beklenmektedir (Türkiye Bursları, t.y.). Bazı Türkçe dil merkezleri, C1 seviyesini başarıyla tamamlayan öğrenciler için C2 seviyesine denk gelen Akademik Türkçe kursu da sunmaktadır.

Standart belirlemeyi içeren geçerleme aşaması için dil merkezindeki öğretim görevlileri (veri toplama sırasında sekiz kişi) sürece dahil edilmiştir. Öğretim görevlilerinin yarısı lisans mezunu iken diğer yarısı da lisansüstü derecesine sahiptir. Öğretim deneyimleri 9 ila 23 yıl arasında değişmektedir. Bu deneyim seviyesi onları öğrencilerin yeterliliklerini ve Türkçe D2'de bir D-AOBM seviyesine ulaşmak için gereken yeterlilikleri değerlendirmek için gerekli olan uzmanlık ve pratik bilgi ile donatmıştır. Öğretim görevlileri, aşağıda süreç bölümünde özetlenen D-AOBM El Kitabında (CoE, 2009) ayrıntılı olarak açıklandığı üzere, belirleme, alıştırma ve standardizasyon eğitimine katılmıştır.

Daha önce de belirtildiği gibi, bu çalışma Türkçe D2 değerlendirmesinde dört beceri setini de ele alan daha büyük bir projenin bir bölümünü oluşturmaktadır. Ancak, geniş projedeki üretim ve alımlama becerileri arasındaki önemli metodolojik farklılıklar ve üretim becerileri görevlerini geçerlemek için benimsenen alternatif yaklaşımlar nedeniyle, bu çalışma alımlama becerilerine odaklanmaktadır. Bu nedenle, materyaller, her biri toplam 100 puan üzerinden değerlendirilen çoktan seçmeli formatta dinleme ve okuma bölümleri içeren dört seviye testinden (A1, A2, B1 ve B2) oluşmaktadır. Testler, okuma ve dinleme metinlerinin sayısı ve uzunluğu, madde sayıları ve puan değerleri gibi test özellikleri aracılığıyla testin betimleyiciler ile uyumunu gösteren belirtke tablolarına dayalı olarak geliştirilmiştir. Değerlendirme uzmanları bu özellikleri gözden geçirerek test içeriklerinin tanımlanan yapıları ne ölçüde yansıttığını değerlendirmiştir. Bu bulgular daha önceki çalışmamızda detaylı olarak ele alınmıştır (Savuran ve Çubukçu, 2021). Testlerin pilot uygulaması, aynı dil merkezinden ayrı bir grup öğrenci üzerinde önceki akademik dönemlerde gerçekleştirilmiş ve ön madde güçlük analizi yapılarak özellikle ayırt edicilik değerleri .10'un altında olan bazı maddeler üzerinde iyileştirilmeler gerçekleştirilmiştir. Aşağıdaki süreç bölümü, test uygulamalarına ilişkin ek ayrıntılar sunmaktadır.

Örneklem ve Süreç

Seviye testleri karma öğrenci gruplarına uygulanmıştır ve katılımcılar amaçlı örnekleme yoluyla seçilmiştir. Her seviye testi grubu, daha önce öğrenci seviyesinin belirlenmesine ilişkin 'Katılımcılar' bölümünde açıklandığı üzere, bir önceki seviyeden, hedef seviyeden ve bir üst seviyeden öğrencilerden oluşmuştur. Tablo 2, testlere katılan çeşitli seviyelerden öğrenci sayılarını göstermektedir.

Tablo 2. Dört Seviye Testlerine Giren Öğrenci Sayıları

Test/Öğrenci Sayısı	Ön-A1	A1	A2	B1	B2	C1
A1	25	66	23	-	-	-
A2	-	21	87	24	-	-
B1	-	-	25	88	27	-
B2	-	-	-	26	97	21

Tablo 2'de görüldüğü gibi, her seviye testi, çoğunluğu hedeflenen seviyedeki öğrencilerden oluşan üç grup tarafından alınmıştır. Örneğin, A2 seviye testi A1 seviyesinden 21, A2 seviyesinden 87 ve B1 seviyesinden 24 öğrenci olmak üzere toplam 132 öğrenciye uygulanmıştır. Örnekleme üst ve alt seviyelerin dahil edilmesinin iki temel nedeni vardır. Alt düzeydeki öğrenciler, o düzey için belirlenen kesme puanının altında ve üstünde kalan öğrenci sayısının gözlemlenebilmesi için eklenmiştir. Puan dağılımını dengelemek ve testin madde analizine yardımcı olmak amacıyla üst düzey öğrenciler dahil edilmiştir. Her bir teste katılım gönüllülük esasına dayanmaktadır ve öğrencilerden onay alınırken çalışma hakkında bilgi verilmiş ve testlerin yalnızca araştırma amaçlı olduğu, yani puanların resmi seviye tamamlama testlerinin puanlarıyla eşitlenmeyeceği ya da onların yerine kullanılmayacağı açıklanmıştır. Bununla birlikte, veri toplama aşamasında belirtilen dil merkezinde ders veren çalışmanın birinci yazarı, öğrencileri testlere girmeye teşvik ederek, testlerin seviye sonu sınavları için gerçek bir alıştırma işlevi göreceğini belirtmiştir. Dört testin uygulanması, bu koşullar altında, katılmaya ve kendi seviyelerindeki performanslarını değerlendirmeye istekli öğrencilerle yaklaşık iki yıl sürmüştür. Testler, A1 seviyesi testi için bir istisna olmak üzere, bir seviyenin öğretimi için belirlenen 8 haftalık modüllerin 6. haftasında uygulanmıştır. İlgili literatürde de belirtildiği üzere, A1 öncesi seviyenin belirlenmesi, D-AOBM Bölüm 3.5'te turistik nitelikte bir gerçek hayat yeterlilik seviyesi olarak kabul edildiği için zordur. (CoE, 2001). Bununla birlikte, Ön-A1 seviyesi 'A1'in yarısı' olarak yorumlanmaktadır (CoE, 2020, s. 243); bu nedenle, A1 seviyesi eğitimlerinin 3. haftasındaki öğrenciler Ön-A1 seviyesi öğrencileri olarak kabul edilmiş ve o öğrencilere A1 ve A2 seviyesi öğrencilerinden farklı zamanlarda test uygulanmıştır.

Amaç bireysel sınav katılımcısı performansından ziyade testlerin geçerliğini değerlendirmek olduğu için her seviye testi için bir kesme puanı belirlemek için, madde merkezli bir standart belirleme tekniği olan Angoff yöntemi tercih edilmiştir (CoE, 2009). Angoff yönteminin kullanılmasının bir diğer nedeni de kullanıcılara her seviye testi için ayrı ayrı kesme puanı atama esnekliği sağlamasıdır. Ayrıca Angoff yöntemi, sadece uygulamadaki basitliği ve kolaylığı nedeniyle değil, aynı zamanda sonuçlarının güvenilirliği nedeniyle de en yaygın kullanılan standart belirleme yöntemlerinden biridir (Buckendahl, Smith, Impara ve Plake, 2002; Plake ve Cizek, 2012). Standart belirleme süreci video konferans yöntemi ve çeşitli formlar aracılığıyla çevrimiçi olarak (COVID-19 kısıtlamaları nedeniyle) gerçekleştirilmiştir. Kane'in üç tür geçerliğine (iç, dış ve prosedürel) (1994) dayanarak, D-AOBM Kılavuzunda (CoE, 2009) belirtilen adımlar özenle takip edilmiştir. Panelistler, birkaç teknik sorun dışında çevrimiçi standart belirleme sürecinde önemli bir zorluk yaşamadıklarını bildirmişlerdir. Ayrıca, uzman değerlendirmelerinin çevrimiçi olarak toplanması, ilgili literatürde önerildiği gibi (örneğin, Katz ve Tannenbaum, 2014) verilerin çevrimiçi tablolara aktarılması ve analizlerin yapılması sırasında büyük ölçüde zaman kazandırdığı için pratiklik sağlamıştır.

D-AOBM betimleyicileri beklenen öğrenen performansını tanımladığından, seviye tanımları bir standart belirleme toplantısında performans seviyesi tanımlayıcıları olarak işlev görür (Tannenbaum ve Cho, 2014). Alıştırma aşaması için, panelistler arasında D-AOBM seviye tanımları, ölçekleri ve tanımlayıcılarının ortak bir şekilde anlaşılmasını sağlamak amacıyla D-AOBM betimleyicilerinin nitel bir analizi ile birlikte hazırlık ve tanıtım faaliyetleri yürütülmüştür. Test geliştirme sürecinde hazırlanan belirtke tabloları panelistlere verilmiştir. Panelistlere değerlendirme uzmanlarının betimleyicilerin test içeriğiyle uyumunu onaylama ve pilot test süreci gibi işlemleri kapsayan test hazırlama aşaması hakkında bilgi verilmiştir. Ayrıca, panelistlere seviyeler arasında görevlerin zorluk derecesini göz önünde bulundurmaları gerektiği hatırlatılmıştır. Örneğin, bir metindeki anaforik ilişkileri belirlemenin A2 seviyesinde basit ve tanıdık metinlerdeki temel kelimelerle başladığı; ancak bunun B2 seviyesinde genişletilmiş metinlerde ölçüldüğü gibi bilgilere yer verilmiştir. Panelistlerden her bir test için kesme puanlarını belirlerken 'sınır öğrenci' (Cizek ve Bunch, 2007) kavramını göz önünde bulundurmaları istenmiştir. Seviyeler arasında basit görevlerden karmaşık görevlere bu tür geçişlerin vurgulanması çalışma açısından önem arz etmiştir. Her seviyede 'asgari düzeyde yeterli bir öğrenciyi' neyin oluşturduğunu tartıştıktan sonra, 'sınır öğrenci 'kavramını içselleştirebilmeleri için zayıf, orta ve iyi olmak üzere üç kademedeki sınav katılımcısının seviyeler arasında belirtilen görevleri ne kadar iyi yapabileceklerini değerlendirmeleri istenmiştir. Standardizasyon eğitimi için örnek bir standart

belirleme uygulaması, iki turda yapılan dört testin her biri için gerçek standart belirleme çalışmasından önce yürütülmüştür. Panelistlere "100 tane sınır öğrencisinin bu maddeyi yanıtladığını düşünün; kaç tanesi doğru yanıt verir?" sorusu yöneltilmiştir (CoE, 2009, s. 63). İlk turdan sonra sonuçlar panelistlerle paylaşılarak kendi olasılık tahminlerini diğerlerininki ile karşılaştırmaları sağlanmıştır. Bu turdaki tartışmaların ardından, sonuçları kabul edilecek olan ikinci değerlendirme turuna geçilmiştir.

Veri Analizi

Madde tepki kuramı (MTK) son zamanlarda yapılan birçok test geçerleme çalışmasında kullanılmasına rağmen (McNamara ve Knoch, 2012), bu çalışmada iki temel nedenden dolayı veri analizlerinde klasik test kuramına (KTK) bağlı kalınmıştır. Birincisi, KTK gözlenen puanların gerçek puanlar ve hatanın bir kombinasyonu olduğunu varsayar ve ölçmenin standart hatasını gruba bağlı olarak değerlendirirken, MTK bunu gruptan bağımsız olarak değerlendirir (Magno, 2009). İkinci olarak, MTK'nın daha büyük örneklemler için daha uygun olduğu düşünülmektedir (Fulcher ve Davidson, 2013). Her test üç farklı öğrenci grubuna uygulandığından ve örneklem büyüklüğü nispeten küçük olduğundan, grupların puan varyanslarını düzeylere göre karşılaştırırken ve madde analizlerinde MTK'ya bağlı kalınmıştır.

Geçerlik iddialarını güçlü bir şekilde desteklemek için, öncelikle her bir test çarpıklık ve basıklık değerleri ve genel puanlar üzerinde Shapiro-Wilk testleri aracılığıyla puan dağılımının normalliğine göre analiz edilmiştir. Daha sonra, testlerdeki her bir maddenin güçlük (p) ve ayırt edicilik (d) indeksleri incelenmiştir. Testlerdeki maddeler dört seçenekli olduğu için ideal p değeri .63 civarında kabul edilmiştir (Thompson ve Levitov, 1985). Madde ayırt edicilik indeksi (d değeri) için, .40 ve üzerindeki maddeler 'çok iyi'; .30 ile .39 arasındaki maddeler 'iyi'; .11 ile .29 arasındaki maddeler 'orta'; .10 ve altındaki maddeler 'zayıf' olarak kategorize edilmiştir (Hopkins, 1998, s. 260). Madde bazlı güçlük ve ayırt edicilik değerlerine ek olarak, bir testteki her bir maddenin p ve d değerlerinin ortalamaları aracılığıyla her bir test için genel madde güçlüğü ve ayırt ediciliği hesaplanmıştır.

Her test için bir kesme puanı belirlemek amacıyla, panelistlerin çevrimiçi olarak toplanan olasılık tahminleri elektronik tablolara aktarılmıştır. İki turun her birinde her bir madde için verdikleri yanıtlar Tablo 3'te gösterildiği gibi birbirinin altına kaydedilmiştir.

Tablo 3. Kesme Puanı Belirleme için Örnek Veri Analiz Tablosu

	Panelist									
Madde/ Tur	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**Med.**	**Ss.**
1 / 1	.2	.25	.4	.54	.12	.27	.3	.35	.3	.13
1 / 2	.4	.38	.42	.52	.45	.32	.42	.45	.4	.09
2 / 1	.24	.28	.38	.41	.32	.58	.42	.75	.4	.17
2 / 2	.35	.45	.52	.4	.47	.45	.48	.39	.4	.06
Toplam/ **Medyan**	17.1	16.7	18.7	19.1	15.6	16.5	14.7	21.4	**16.7**	**2.09**

Med: Medyan, Ss.: Standard Sapma

Tablo 3, kesme puanlarını belirlemek için kullanılan örnek bir veri analiz sayfasını göstermektedir. İkinci turda panelistlerin yanıtları toplanmış ve her panelistin o test için bireysel kesme puanını gözden geçirmesi için alt satırda gösterilmiştir (Cizek ve Bunch, 2007). Bu bireysel kesme puanlarını tek bir ortak puanda birleştirmek için, Cizek ve Bunch'ın (2007) 'sınır öğrenci'' kavramının kullanıldığı standart belirleme çabaları için önerdiği gibi medyan değerleri tercih edilmiştir. Sekiz ayrı kesme puanının medyanı her bir test için önerilen kesme puanı olarak kabul edilmiş ve daha sonra testler bu ölçekte puanlandığı için 100 üzerinden bir puana dönüştürülmüştür. Bu puanların küsuratlı tutulduğu ve yuvarlanmadığı unutulmamalıdır, çünkü kesme puanı belirlemede amaç geçme/kalma kararları almak değil, o puanları öğrencilerin sınav puanlarıyla karşılaştırmaktır.

İddiaları geçerlemek için, kesme puanının ayırdığı varsayılan iki bitişik seviyedeki öğrencilerin test puanlarını karşılaştırmak üzere her bir test için belirlenen kesme puanları kullanılmıştır. Örneğin, B1 seviyesi testi için önerilen kesme puanı, Tablo 1'de sunulan olasılık değerlerine dayalı olarak gözlenen ve beklenen sayılar ışığında, bu testi alan A2 ve B1 seviyesi öğrencilerinin ortalama test

puanlarını karşılaştırmak için kullanılmıştır. Bu yaklaşımla, gözlenen dağılım ile teorik olarak beklenen dağılım arasında istatistiksel olarak anlamlı bir fark olup olmadığını analiz eden ki-kare uyum iyiliği testi yapılmıştır (Frey, 2018). Serbestlik derecesi 1 ve anlamlılık da $p < .05$ olarak belirlenmiştir. Bir kesme puanının hem altında hem de üstünde iki bitişik düzeyde beklenen ve gözlenen öğrenci sayısı arasındaki tutarsızlığı gösteren artık değerler, çürütücü veriler olarak ele alınmış ve yanlış pozitif ve yanlış negatif örnekleri olarak kabul edilmiştir. Standart belirleme sürecinin iç geçerliği, D-AOBM El Kitabında (CoE, 2009) önerildiği gibi, Cronbach alfa (panelistlerin yargılarının güvenilirliği için) ve sınıf içi korelasyon ölçümleri (panelistler arasındaki uyum tutarlılığı için) yoluyla teyit edilmiştir.

Bulgular

Bu bölümde üç tür veri analizi sonucu sunulacaktır; puan dağılımları ve testlerin psikometrik özellikleri, kesme puanları ve bunlarla ilgili güvenilirlik ve sınıf içi korelasyon katsayısı ölçümleri ve ki-kare uyum iyiliği testlerinin sonuçları.

Test Puanları ve Testlerin Psikometrik Özellikleri

Her bir seviye testi için üç grup test katılımcısının puanlarını incelemeden önce, genel puanlar üzerinde normallik testleri gerçekleştirilmiştir. A1'den B2'ye kadar dört testin her biri için çarpıklık değerleri sırasıyla -.23, -.22, -.13 ve -.1 iken basıklık değerleri sırasıyla -1.21, -.89, -.91 ve -.98'dir. Shapiro-Wilk testleri için .05 düzeyinde (α) anlamlılık (p) değerleri sırasıyla .000, .003, .004 ve .002'dir. Dört test için de negatif çarpıklık ve basıklık değerleri, dağılımların sola çarpık ve kısa kuyruklu olduğunu göstermektedir. Shapiro-Wilk testleri için .05 anlamlılık düzeyinden küçük olan bu p değerleri, puan dağılımlarının normal olmadığını farklı bir şekilde doğrulamaktadır. Bu normal olmama durumu, dört seviye testinin her birine giren üç öğrenci grubunun puanlarının istatistiksel olarak anlamlı düzeylerde farklılık gösterdiği sonucuna varılmasını sağlamıştır.

Bu istatistiksel analizler dağılım normalliği hakkında fikir oluşmasını sağlamıştır. Bu adımın ardından, her bir seviye testini alan üç grubun puanlarının birbirlerinden nasıl farklılaştığını gözlemlemek hedeflenmiştir. Tablo 4 her bir seviye testine giren üç farklı seviyedeki öğrenci sayıları, ortalama puanları ve standart sapmaları bütüncül bir şekilde ortaya koymaktadır.

Tablo 4. Seviyelere Göre Puan Dağılımları

A1 Seviyesi Testi				
	Ön-A1	**A1**	**A2**	**Genel**
Öğrenci Sayısı	25	66	23	114
Puan Ortalaması	38.8	64.65	80.1	62.1
Ss.	11.83	18.78	10.32	21.06
A2 Seviyesi Testi				
	A1	**A2**	**B1**	**Genel**
Öğrenci Sayısı	21	87	24	132
Puan Ortalaması	45.23	62.95	79.89	63.21
Ss.	9.93	19.07	13.15	19.67
B1 Seviyesi Testi				
	A2	**B1**	**B2**	**Genel**
Öğrenci Sayısı	25	88	27	140
Puan Ortalaması	41.15	60.71	74.25	59.69
Ss.	13.08	21.9	16.79	22.05
B2 Seviyesi Testi				
	B1	**B2**	**C1**	**Genel**
Öğrenci Sayısı	26	97	21	144
Puan Ortalaması	42.07	63.4	83.09	62.67
Ss.	11.2	19.15	10.39	20.41

Ss.: Standart Sapma

Tablo 4, öğrencilerin testlerden aldıkları puanların dağılımını (100 üzerinden) göstermektedir. Dört testte de görülen önemli bir bulgu, belirli bir test için ortalama puanların seviyeye göre tutarlı bir şekilde artmasıdır. Örneğin, A2 seviye testini alan A1, A2 ve B1 seviyesindeki öğrencilerin ortalama puanları sırasıyla 45.23, 62.95 ve 79.89'dur. Bu sınava giren üç seviyedeki toplam öğrenci sayısı 132 olup, genel ortalama puan 63.21'dir ve A2 seviyesindeki öğrencilerin ortalama puanıyla (62.95) neredeyse aynıdır. Dört seviye testinde de ortalama puanların seviyeye göre açık bir şekilde yükselmesi, testlerin amaçlanan öğrenci performansını etkili bir şekilde yansıttığını göstermektedir.

Üç seviyedeki öğrencilerin puanlarının standart sapması (Ss.) da bu iddiayı desteklemektedir. Örneğin, B1 seviyesindeki test için genel standart sapma 22.05 iken, A2 için 13.08, B1 için 21.9 ve B2 seviyesindeki öğrenciler için 16.79'dur. Hedeflenen seviyedeki öğrenciler için daha yüksek bir Ss. değeri, testin o seviyedeki düşük ve yüksek performanslı öğrenciler arasında ayrım yaparak teste katılanların yapılarını daha güvenilir bir şekilde ölçtüğünün kanıtı olarak görülebilir (B1 seviyesi testi için B1 seviyesi öğrencileri). Bu bulgu, testin becerilerini ölçmeyi amaçlamadığı öğrencilerin (B1 seviyesi testi için A2 ve B2 öğrencileri) puanları için daha düşük Ss. değerleri ile desteklenmekte ve testin onlar için daha az uygun olduğunu gösterebilecek daha düzgün bir dağılıma işaret etmektedir (bir başka deyişle, B1 seviyesi testinin A2 ve B2 seviyesi öğrencileri için uygun olmaması durumu).

Tablo 5. Test Güvenilirliği ve Madde Analizi

	KR-20	Madde Güçlüğü (p) Aralığı	Test Geneli Madde Güçlüğü (p)	Madde Ayırt Ediciliği Aralığı (d)	Test Geneli Madde Ayırt Ediciliği (d)
A1	.79	.35 - .78	.58	.11 - .58	.26
A2	.90	.31 - .82	.56	.11 - .51	.25
B1	.84	.33 - .86	.55	.12 - .53	.27
B2	.78	.16 - .84	.49	.16 - .55	.30

Puan dağılımı ve normallik dışında, dört seviye testine dair güvenilirlik analizi için KR-20 tekniği ve madde analizi için de madde güçlük (p) ve ayırt edicilik (d) indeksleri hesaplanmıştır. A1 ve B2 düzeyindeki testler için KR-20 değerleri A2 ve B1 düzeyindeki testlere göre görece düşük olsa da tüm testlerin güvenilirliği kabul edilebilir düzeyde bulunmuştur (Frey, 2018). Her bir maddenin p ve d indekslerine dayalı madde analizinin ortalama değerleri Tablo 5'te görülebilir. Diğerlerinden oldukça farklı olan B2 düzeyindeki testte yer alan maddelerin p değerleri .16 ile .86 arasında değişmektedir. A1, A2 ve B1 düzeyindeki testler için p aralığı .31 ile .86 arasındadır. Bununla birlikte, dört test için genel p değerleri sırasıyla .58, .58, .55 ve .49'dur; bu da ortalama olarak sınava girenlerin yaklaşık yarısının tüm testlerde maddeleri doğru yanıtladığını göstermektedir. Dört testteki maddeler için d değerleri .11 ile .58 arasında değişmektedir. Bununla birlikte, her bir test için genel ortalama d değerleri sırasıyla .26, .27, .25 ve .30'dur; bu da testlerin iyi ve kötü performans gösteren öğrenciler arasındaki ayırt edici etkisinin ilgili literatürde belirtildiği gibi 'ortalama' aralıkta olduğunu göstermektedir. Bu yorum, testlerin, test katılımcılarını kendi seviyelerindeki alımlama becerilerine göre ayırt etmek için oldukça kullanışlı olduğunu göstermektedir.

Seviyelere göre puan dağılımları ve testlerin madde analizleri, testlerin amaçlanan öğrenci performansını yansıtma potansiyeline ilişkin önemli sonuçlar sunmaktadır. Çarpıklık, basıklık ve Shapiro-Wilk analizleri yoluyla elde edilen puan dağılımları, üç farklı seviyedeki öğrenciler tarafından uygulanan dört testin tamamında kabul edilebilir güvenirlik ve madde analizi değerleriyle birleştiğinde, testlerin öğrencilerin dinleme ve okuma becerilerini ölçmede etkili olduğunu göstermektedir. Dolayısıyla, testlerin Ölçme-Kullanım Argümanı (ÖKA) çerçevesinde işe konulmasının, etkililiğe ilişkin ilk izlenimleri sağladığı sonucuna varılabilir. Bu güçlü psikometrik özellikler sayesinde, test puanları geçerlik iddiasının dayandırıldığı 'veriler' (Şekil 1'de belirtildiği gibi) olarak değerlendirilebilmiştir. İddiayı desteklemek üzere verilerden daha fazla destek sağlamak için, her test için Angoff standart belirleme prosedürleri uygulanmıştır.

Öğrenci Puanları ile Karşılaştırmalı Olarak Standart Belirleme Sonuçları

Örnek standart belirleme oturumlarını da içeren alıştırma, belirleme ve standartlaştırma eğitimi aşamalarından geçtikten sonra, sekiz panelistten bir testteki her bir madde için olasılık önerilerini bildirmeleri istenmiştir. İkinci turdaki tahminlerinin analizi Tablo 3'te gösterildiği gibi gerçekleştirilmiştir. Böylece her bir test için önerilen kesme puanları, raunt 1 (R1) ve raunt 2 (R2) olarak tanımlanan iki tur için Cronbach alfa ve sınıf içi korelasyon değerleri ile birlikte Tablo 6'da sunulmuştur. Tablo 6'nın sol tarafında bir seviye testine giren öğrencilerin puanları (100 üzerinden), sağ tarafında ise söz konusu test için kesme puanı, Cronbach alfa ve sınıf içi korelasyon katsayısı değerleri gösterilmektedir.

Tablo 6. Test Sonuçları ile Karşılaştırmalı Olarak Standart Belirleme Sonuçları

Test Sonuçları							**Standart Belirleme Sonuçları**				
	Öğrenci Skor Ortalamaları						**Kesme Skoru**	**Cronbach alfa**		**Sınıf içi korelasyon katsayısı**	
	Ön-A1	**A1**	**A2**	**B1**	**B2**	**C1**		**R1**	**R2**	**R1**	**R2**
A1	38.8	64.65	80.1				41.97	.71	.77	.64	.72
A2		45.23	62.95	79.89			50.6	.79	.91	.74	.83
B1			41.15	60.71	74.25		48.4	.83	.94	.76	.86
B2				42.07	63.4	83.09	46.7	.75	.89	.71	.84

Her test için üç grup test katılımcısının ortalama puanlarını o test için belirlenen kesme puanı ile karşılaştırmadan önce, standart belirleme süreçlerinin iç geçerliği incelenmiştir. Her bir testin her bir turu için Cronbach alfa ve sınıf içi korelasyon değerleri kabul edilebilir aralıkta (.70'in üzerinde) bulunmuş, ikinci turlarda gözlemlenen artış panelistlerin yargılarının daha güvenilir ve tutarlı hale geldiğini göstermiştir. Öte yandan, öğrencilerin puanlarına genel bir perspektiften bakıldığında, panelistler tarafından Angoff yöntemi ile önerilen kesme puanlarının, testin hitap etmesi amaçlanan öğrencilerin puanları ile kesme puanının ayırt etmesi beklenen alt düzey öğrencilerin puanları arasındaki aralıkta kaldığı açıkça görülmektedir. Örneğin, B1 düzeyindeki test için panelistlerin belirlediği kesme puanı 100 üzerinden 48.4'tür; A2, B1 ve B2 düzeyindeki öğrencilerin aynı test için gerçek puanları ise sırasıyla 41.15, 60.71 ve 74.25'tir. Bu bulgular, testlerin değerlendirmek üzere tasarlandıkları öğrencilerin yeteneklerini etkili bir şekilde ölçtüğü yorumunun yapılmasını sağlamaktadır. Ayrıca, panelistlerin madde bazlı önerilerine dayanan kesme puanları test sonuçlarını desteklemektedir (örneğin, B1 seviyesi testi için kesme puanı 48.4, A2 seviyesi için 41.15, ve B1 seviyesi 60.71 olan öğrenci puanları arasında yer almaktadır). Bu nedenle, kesme puanlarının geçerlik iddiası (testlerin öğrencileri amaçlanan D-AOBM seviyelerine göre sınıflandırma becerisi) için verileri (test puanlarını) 'destekleyen' kanıtlar olarak görülebileceği sonucuna varılabilir.

Kesme puanları ile öğrenci puanları karşılaştırıldığında dikkat çeken bir diğer bulgu da A1 seviyesi testinin kesme puanının A2, B1 ve B2 seviyelerine göre oldukça düşük olmasıdır ve bu da öğrenci puanları ile uyumludur. A1 seviyesi teorik ve pratik olarak D-AOBM'nin temel seviyesidir, bu da panelistlerin asgari düzeyde yeterli bir adayın (varsayımsal sınırda bir öğrenci) performansını dört seviye arasında en düşük olarak değerlendirmesini mantıklı bir beklenti haline getirmektedir. Bu örüntü, sınava üçüncü haftalarında giren Ön-A1 seviyesindeki öğrencilerin, 12 öğrenci puan kategorisi içerisinden (dört seviye testinin her biri için üç farklı seviye öğrencisinin puanları) en düşük puanı almasıyla doğrulanmıştır. Bu durumun nedeni, eğitimin üçüncü haftasında Türkçe diline henüz yeterince aşina olmamaları ve dolayısıyla testte iyi performans gösterecek yetkinlikten yoksun olmaları gerçeği ile açıklanabilir.

Kesme Puanlarına Dayalı Uyum İyiliği Testlerinin Sonuçları

Bir seviye testine giren öğrencilerin ortalama puanları o testin kesme puanı ile karşılaştırılarak analiz edilmiştir. Böylece, testlerin sınav katılımcılarının performansını değerlendirmedeki etkisine dair

net bir açık bir görünüm sağlamıştır. Ancak, bitişik seviyelerdeki öğrenciler arasında ayrım yaptığı varsayılan kesme puanının altında ve üstünde puan alan öğrencilerin sayısını incelemek için daha ileri istatistiksel analizler ele alınmıştır. Bu amaçla, sınava giren gerçek öğrenci sayısı ve Tablo 1'de özetlenen %50 olasılık varsayımıyla belirlenen beklenen öğrenci sayısı temel alınarak her seviye için ki-kare uyum iyiliği testleri yapılmıştır. Serbestlik derecesi (sd) 1 olarak belirlenmiş ve anlamlılığı belirlemek için eşik olarak <.05 anlamlılık düzeyi (p) kullanılmıştır. A1, A2, B1 ve B2 testleri için bu analizlerin sonuçları Tablo 7'de sunulmuştur.

Tablo 7. Seviye Testleri için ki-kare Uyum İyiliği Testleri

	Ön-A1			**A1**		
	Gözlenen Değer	**Beklenen Değer**	**Artık değer**	**Gözlenen Değer**	**Beklenen Değer**	**Artık değer**
Kesme Puanı Altında	14	22.0	-8.0	12	33.0	-21.0
Kesme Puanı Üstünde	11	3.0	8.0	54	33.0	21.0
Toplam		25			66	
ki-kare		24.242			26.727	
Anlamlılık		.000			.000	
	A1			**A2**		
	Gözlenen Değer	**Beklenen Değer**	**Artık değer**	**Gözlenen Değer**	**Beklenen Değer**	**Artık değer**
Kesme Puanı Altında	15	17.2	-2.2	20	43.5	-23.5
Kesme Puanı Üstünde	6	3.8	2.2	67	43.5	23.5
Toplam		21			87	
ki-kare		40.614			25.391	
Anlamlılık		.000			.000	
	A2			**B1**		
	Gözlenen Değer	**Beklenen Değer**	**Artık değer**	**Gözlenen Değer**	**Beklenen Değer**	**Artık değer**
Kesme Puanı Altında	17	22	-5.0	25	44.0	-19.0
Kesme Puanı Üstünde	8	3	5.0	63	44.0	19.0
Toplam		25			88	
ki-kare		12.593			16.409	
Anlamlılık		.000			.000	
	B1			**B2**		
	Gözlenen Değer	**Beklenen Değer**	**Artık değer**	**Gözlenen Değer**	**Beklenen Değer**	**Artık değer**
Kesme Puanı Altında	16	22.0	-6.0	20	48.5	-28.5
Kesme Puanı Üstünde	9	3.0	6.0	77	48.5	28.5
Toplam		26			97	
ki-kare		13.636			33.495	
Anlamlılık		.000			.000	

Tablo 7, uyum iyiliği testlerine dayalı olarak bitişik seviyelerdeki öğrencilerin seviye testlerinden aldıkları puanlar arasındaki önemli farklılıkları ayrıntılı bir şekilde göstermektedir. Analizi daha derinlemesine incelemek için örnek olarak B1 seviye testi incelenebilir. Bu test için kesme puanı 48,4 olarak belirlenmiştir. Ki-kare analizi için A2 seviyesinden 25 ve B1 seviyesinden 88 olmak üzere toplam 113 öğrenci B1 seviyesi testine dahil edilmiştir. Tablo 1'de verilen olasılık değerlerine dayanarak - A2 seviyesindeki öğrencilerin bir üst seviyedeki görevleri çözme olasılığı %12 - A2 seviyesindeki 25 öğrenciden 3'ünün 48.4'ün üzerinde puan alması, kalan 22'sinin ise altında puan alması beklenebilirdi. Kendi seviyelerindeki görevleri çözme olasılığı %50 olan B1 seviyesindeki öğrenciler için 88 öğrenciden 44'ünün 48,4'ün üzerinde puan alması beklenirdi. Ancak, A2 seviyesindeki öğrencilerin kesme puanının

altında ve üstünde puan almalarına ilişkin gözlemlenen sayılar sırasıyla 17 ve 8'dir ve bu da ± 5.0'lık bir artık değerle sonuçlanmaktadır. B1 seviyesindeki öğrenciler için beklenen sayılar kesme puanının altında ve üstünde 44 iken, gözlenen sayılar altında 25 ve üstünde 63'tür ve ± 19.0'lık artık değerler elde edilmiştir. A2 ve B1 seviyeleri için uyum iyiliği testlerinin sonuçları sırasıyla 12.59 ve 16.4 (sd: 1) olup, gözlenen ve beklenen öğrenci sayıları arasındaki fark her iki seviye için de istatistiksel olarak anlamlıdır ($p < .05$). Panelistler tarafından test maddelerine dayalı olarak belirlenen kesme puanlarının, öğrencileri gerçek test puanlarına göre doğru seviyelere anlamlı bir şekilde kategorize ettiği yorumu yapılabilir. Bu da bizi, kesme puanlarına dayalı ki-kare analizlerinin geçerlik iddiası için (testlerin öğrencileri amaçlanan D-AOBM seviyelerine başarılı bir şekilde sınıflandırdığı) gerekçe olarak kullanılabileceği sonucuna götürmektedir.

Özetle, bu araştırmada ikinci dil olarak Türkçe için dinleme ve okuma becerileri testlerini argüman temelli bir geçerlik çerçevesinde geçerlemek için Ölçme Kullanım Argümanı (ÖKA) ve standart belirleme yöntemleri kullanılmıştır. Bu bölümde sunulan bulgular, bir seviye testi için üç farklı seviyedeki öğrenci puanlarına dayanarak, testlerin önerilen öğrenen performansını gösterdiği ve öğrencileri amaçlanan D-AOBM seviyelerine doğru bir şekilde sınıflandırdığı geçerlik iddiasını desteklemektedir. Her bir test için Angoff standart belirleme süreci aracılığıyla belirlenen kesme puanları geçerlik iddiası için 'destek' sağlamaktadır. Benzer şekilde, her seviyede beklenen ve gözlenen öğrenci sayıları arasındaki farkların istatistiksel anlamlılığı da 'gerekçe' işlevi görmektedir. Öte yandan, artıklar yanlış pozitifler ve yanlış negatifler olarak yorumlanabilir ve bunlar 'çürütücü veriler' olarak hizmet eden somut sayılarla 'çürütme' olarak düşünülebilir. Verilere ve geçerlik iddiasına ilişkin olası neden ve açıklamalara, gerekçe ve çürütmelerle birlikte aşağıdaki tartışma ve sonuç bölümlerinde yer verilmektedir.

Tartışma

Papageorgiou ve Tannenbaum (2016), Toulmin'ın (1958, 2003) argüman yapısı ve Bachman ve Palmer'ın (2010) Ölçme-Kullanım Argümanı (ÖKA) hakkındaki ufuk açıcı çalışmalarına dayanarak, ilgili gerekçeler ve çürütmelerle birlikte dört geçerlik iddiası önermiştir. Bu geçerlik iddiaları çıkarımsal ve birbiriyle bağlantılı olmasına rağmen, bu çalışmada ölçülen yeteneklerin ve değerlendirme kayıtlarının yorumlanmasıyla ilgili olan 3. ve 4. iddialara odaklanılmıştır. Amaç test sonuçlarına göre karar vermek olmadığı için, test sonuçları ve kararlarıyla ilgili olan iddia 1 ve 2 değerlendirilmemiştir. Literatürde vurgulandığı üzere, araştırmacılar test sonuçlarına ilişkin öne sürmek istedikleri iddia türlerini seçebilir ve seçtikleri iddialarını destekleyen kanıtları toplayabilirler (Im vd., 2019; Kane, 2013). Dolayısıyla, iddia 3 ile ilgili olarak, öğrencilerin dinleme ve okuma becerilerinin değerlendirilmesine ilişkin yorumların amaçlanan öğrenen performansının göstergesi olduğunu ve testlerin öğrenen performansının temelini oluşturan tanımlayıcılara dayalı olarak tasarlanıp belirlendiğinden, büyük ölçüde hedef dil kullanım alanına, bizim durumumuzda Türkçe D2'ye, atfedilebileceği ve bu alan içinde genellenilebileceği savunulabilir. İddia 4 bağlamında, dört test için üç seviyedeki öğrenciler arasındaki puanların karşılaştırılmasından da anlaşılacağı üzere, öğrencilerin farklı seviyeler ve görev türleri arasındaki puanlarının tutarlı olduğu iddia edilebilir.

Öğrenen performansından elde edilen kanıtlar ÖKA için bir ön koşul olarak kabul edildiği (Im vd., 2019; Kane, 2013) ve dil testlerindeki birçok geçerleme çalışması, öğrenen performansını argüman temelli bir geçerleme çerçevesinde değerlendirmeye dayandığı (Becker, 2018; Chapelle vd., 2010; Knoch ve Chapelle, 2018; Mendoza ve Knoch, 2018; O'Loughlin, 2011) için geçerleme sürecinde test puanları ve bunlara dair yorumlar kullanılmıştır. Bununla birlikte, argümantasyon için standart belirlemeyi benimseyen (Cizek ve Bunch, 2007; Lavery vd., 2020; Papageorgiou ve Tannenbaum, 2016; Shin ve Lidster, 2017) veya argüman temelli geçerlik için standart belirleme prosedürlerini tanımlayan (Kenyon, 2012; Kenyon ve Römhild, 2013) çalışmaların sayısı nispeten azdır. Bu kısıtlılık, araştırmacılara nispeten yeni çerçeveler içinde belirli geçerlik argümanları oluşturmaları için alan bırakmaktadır (Davies, 2012; Knoch ve Chapelle, 2018). ÖKA çerçevesindeki Angoff standart belirleme prosedürlerinden türetilen kesme puanlarından faydalanan bu çalışma, bir D-AOBM seviyesinde olma kavramına dair olasılık varsayımlarını kullanarak geçerleme çabaları için pratik çıkarımlar sunmaktadır.

Dört seviyeli testler için belirlenen kesme puanlarına ilişkin istatistiksel kanıtlar sunmak için Madde Tepki Kuramı ölçeklendirmesine dayanan %50 D-AOBM seviyesinde olma olasılığı varsayımı (De Jong ve Benigno, 2017) kullanılmıştır. Her ne kadar bu konu tartışmalara konu olmuş (Harsch, 2019; Harsch ve Hartig, 2015; Hulstijn, 2007) ve kesin cevaplar sağlanmamış olsa da, D-AOBM seviyelerinin belirlenmesine ilişkin son araştırmalar bazı nicel bilgiler ortaya koymuştur (De Jong ve Benigno, 2016, 2017). İstatistiksel anlamlılık aramak için öğrenenin kendi seviyesindeki görevler için %50 olasılık ve bir üst seviyedeki görevler için %12-18 olasılık kullanılmasına rağmen, bu rakamların sayısal çıkarımlar sağladığı ve kesin olarak kabul edilmemesi gerektiği unutulmamalıdır. Hulstijn'in belirttiği gibi, "D-AOBM'de sunulan dil yeterliliği kavramı iç içe geçmiş iki sütun üzerine oturmaktadır: nicelik ve nitelik" (2007, s. 663), bu da olasılıkların bir öğrencinin bilgi açısından yeterlilik düzeyinin niceliksel yönlerini yansıtabileceğini göstermektedir. D-AOBM ile uyumlu değerlendirme amaçları için, araştırmacılar ve test geliştiriciler, benimsenen yeterlilik yaklaşımının nitel yönünü ele alarak belirli ihtiyaçları ve hedefleri karşılamak için beceri ve yetenekleri tanımlamalı (Harsch, 2014; Papageorgiou vd., 2015) ve geniş bant seviyeleri yerine sağlam belgeler ve daha hassas puanlama raporları sunmalıdır (Harsch, 2019). Bununla birlikte, bu alandaki başarılarının kapsamını ve belirli bağlamlarda bu tür çabaları etkileyen genel sorunları deneysel verilerle raporlayan çok az çalışma vardır (Brunfaut ve Harding, 2020). Türkçe D2 değerlendirmesine ilişkin olarak bu çalışma, titizlikle tasarlanmış ve standart belirleme yoluyla ÖKA yaklaşımı içinde doğrulanmış testleriyle yeni ve öncü bir çalışmayı temsil etmektedir.

Eğitimde ölçme alanında 1960'lardan bu yana çok sayıda standart belirleme çalışması belgelenmiştir (Harsch ve Kanistra, 2020) ve 'altın' veya 'gerçek' bir standart olmadığı kabul edilmektedir (Cizek, 1993; Kane, 1994); ayrıca, D-AOBM gibi evrensel bir çerçeve olmadan, bitişik seviyeler veya sınıflar arasındaki sınırlar standart belirleme sürecinin dışındakiler için anlamsız olabilir (Harsch ve Hartig, 2015). Bununla birlikte, altı D-AOBM seviyesinde (Ön-A1 - C1) dört test (A1 - B2) için kesme puanları belirledikten ve El Kitabında (CoE, 2009) belirtilen yönergelere bağlı kaldıktan sonra, kesme puanlarının, öğrencileri amaçlanan D-AOBM seviyelerine göre doğru bir şekilde sınıflandırma potansiyeline sahip olduğu iddia edilebilir. Literatürde aktarıldığı üzere, yüksek sayıda yanlış pozitif ve yanlış negatife sahip olmanın riskleri göz önünde bulundurulduğunda çalışma bağlamı oldukça önemli hale gelmektedir (Papageorgiou ve Cho, 2014; Xi, 2007) ve bu riskleri en aza indirmek için kesme puanlarını optimize etmek, onların kesinliğinden daha önemlidir (Cizek ve Bunch, 2007; Eckes, 2017). Bu nedenle, amaç öğrencileri sınıflandırmak ya da seviyelerine karar vermek değil, belirlenen kesme puanlarının altında ve üstünde puan alan öğrencilerin sayısını deneysel olarak karşılaştırmaktır. Bu bağlamda, ki-kare analizleri, dört testin hiçbiri için yanlış pozitif ve negatiflere atıfta bulunan istatistiksel olarak anlamlı artık değerler göstermemiştir ve bu da çürütmelerin ÖKA çerçevesi dahilinde karşı iddialar oluşturmadığı sonucuna varılmasına yol açmıştır.

Sonuç ve Öneriler

Bu çalışma, standart belirleme yoluyla argüman temelli bir geçerleme yaklaşımı çerçevesinde ikinci dil olarak Türkçe testlerini geçerlemeyi amaçlamıştır. Bu amaçla da "Geçerli kılınan testin kendisi değil, testin puanları ve kullanımlarıdır" (AERA, 2014; Kane, 1994) perspektifini benimsemektedir. Bu araştırma, test puanlarını kesme puanlarıyla yorumlayarak ve karşılaştırarak, bir D-AOBM seviyesindeki görevleri yerine getirme veya çözme olasılıklarına dayanan istatistiksel kanıtlarla desteklenen verilere, desteklere, gerekçelere ve çürütmelere dayanan geçerlik iddiaları ile alımlama beceri testlerini geçerlemek için yenilikçi bir yöntem sunmaktadır.

Bu çalışmanın çeşitli sınırlılıkları bulunmaktadır. İlk olarak, test puanları çalışmaya katılan öğrencilerden oluşan ve nispeten küçük bir örneklemden elde edilen sonuçları temsil etmektedir. Katılımcılar ilk yazar tarafından gerçek sınavlar için pratik yapmak üzere testlere girmeye motive edilmiştir. İkinci olarak, çalışma analizleri yalnızca klasik test kuramına dayanmakta ve genellenebilirlik kuramı veya madde tepki kuramını içermemektedir. Üçüncü olarak, katılımcılar tek bir dil merkezindeki öğrenciler ve panelistlerle sınırlıdır. Bu durum, özellikle standart belirlemede 'sınır öğrenci' kavramı için aşinalık ve tutarlılığı artırırken, dış geçerliği güçlendirmek için farklı öğrenci

gruplarını yeniden test ederek ve diğer uzman panelleriyle veya alternatif standart belirleme yöntemleriyle yeni kesme puanları belirleyerek bulguları tekrarlamak için gelecekteki araştırmalara yol açmaktadır. Bir diğer kısıtlama da testlerin yetişkin öğrencilerle ve A1-B2 seviye aralığıyla sınırlandırılması, C1 ve C2 seviyelerinin hariç tutulmasıdır.

Bu çalışma bir geçerleme sürecini raporladığı için, spesifik bulgulardan ziyade yaklaşım ve metodolojiyi vurgulamaktadır. Çalışma, ÖKA çerçevesinde Türkçe D2 değerlendirmeleri bağlamında testlerin geçerlenmesinde öncü bir adım teşkil etmektedir. Erişilebilir alan yazında bulgularımızla karşılaştırma adına Türkçe D2 için yapılan bir çalışma bulunmamaktadır. Bazı kapsamlı sistematik inceleme çalışmaları (Chapelle ve Voss, 2021; Im vd., 2019; Lavery vd., 2020), geçerlik çalışmalarının çoğunun Türkçe gibi daha az yaygın olarak öğretilen dillerden ziyade ana akım dillere odaklandığını vurgulamaktadır.

Sonuç olarak, bu çalışma bir başlangıç çalışması olarak görülmelidir. Türkçe D2 için farklı yaş ve yeterlilik seviyelerini hedefleyen betimleyiciler ve testler geliştirmeye ve geçerlemeye yönelik gelecekteki araştırmalar, gelişmekte olan literatüre katkıda bulunacaktır. Bu araştırma Türkçe D2 alanındaki araştırmacıları benzer çalışmalar yapmaya ve yöntemlerini belgelemeye teşvik ederek bu alandaki ölçme ve değerlendirme uygulamaları üzerindeki etkiyi artırmaya davet etmektedir. Bu yaklaşım, daha güvenilir biçimlendirici ve özetleyici değerlendirme süreçlerinin uygulanmasını sağlayacak ve böylece sayıları giderek artan Türkçe D2 öğrencilerin ihtiyaçlarına cevap verilebilecektir.

Kaynakça

American Educational Research Association. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Bachman, L. F. (2005). Building and supporting a case for test use. *Language Assessment Quarterly*, *2*(1), 1-34. doi:10.1207/s15434311laq0201_1

Bachman, L. F. ve Palmer, A. S. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world.* Oxford: Oxford University Press.

Becker, A. (2018). Not to scale? An argument-based inquiry into the validity of an L2 writing rating scale. *Assessing Writing*, *37*, 1-12. doi:10.1016/j.asw.2018.01.001

Benigno, V. ve De Jong, J. (2016). The "global scale of English learning objectives for young learners": A CEFR-based inventory of descriptors. M. Nikolov (Ed.), *Assessing young learners of English: Global and local perspectives* içinde (s. 43-64). New York: Springer. doi:10.1007/978-3-319-22422-0_3

Brunfaut, T. ve Harding, L. (2020). International language proficiency standards in the local context: Interpreting the CEFR in standard setting for exam reform in Luxembourg. *Assessment in Education: Principles, Policy & Practice*, *27*(2), 215-231. doi:10.1080/0969594X.2019.1700213

Buckendahl, C. W., Smith, R. W., Impara, J. C. ve Plake, B. S. (2002). A comparison of angoff and bookmark standard setting methods. *Journal of Educational Measurement*, *39*(3), 253-263. http://www.jstor.org/stable/1435081 adresinden erişildi.

Chapelle, C. A. ve Voss, E. (2014). Evaluation of language tests through validation research. A. J. Kunnan (Ed.), *The companion to language assessment* içinde (s. 1079-1097). New York: Wiley. doi:10.1002/9781118411360.wbcla110

Chapelle, C. A. ve Voss, E. (Ed.). (2021). *Validity argument in language testing: Case studies of validation research*. Cambridge: Cambridge University Press.

Chapelle, C. A., Enright, M. K. ve Jamieson, J. (Ed.). (2008). *Building a validity argument for the test of English as a foreign language*. New York: Routledge.

Chapelle, C. A., Enright, M. K. ve Jamieson, J. (2010). Does an argument-based approach to validity make a difference?. *Educational Measurement: Issues and Practice, 29*(1), 3-13. doi:10.1111/j.1745-3992.2009.00165.x

Cheng, L. ve Sun, Y. (2015). Interpreting the impact of the Ontario Secondary School Literacy Test on second language students within an argument-based validation framework. *Language Assessment Quarterly*, *12*(1), 50-66. doi:10.1080/15434303.2014.981334

Cizek, G. J. (1993). Reconsidering standards and criteria. *Journal of Educational Measurement, 30*(2), 93-106. doi:10.1111/j.1745-3984.1993.tb01068.x

Cizek, G. J. (2012). Defining and distinguishing validity: Interpretations of score meaning and justifications of test use. *Psychological Methods, 17*(1), 31-43. doi:10.1037/a0026975

Cizek, G. J. ve Bunch, M. B. (2007). *Standard setting*. Thousand Oaks, CA: Sage. doi:10.4135/9781412985918

Council of Europe. (2001). *Common European Framework of References for Languages.* Retrieved from: https://rm.coe.int/1680459f97

Council of Europe. (2009). *Relating language examinations to the common European framework of reference for languages: Learning, teaching, assessment (CEFR)*. https://rm.coe.int/CoERMPublicCommonSearchServices/DisplayDCTMContent?documentId=0900001680667a2d

Council of Europe. (2020). *Common European framework of references for languages - companion volume.* Retrieved from: https://rm.coe.int/common-european-framework-of-reference-for-languages-learning-teaching/16809ea0d4

Cureton, E. E. (1951). Validity. E. F. Lindquist (Ed.), *Educational measurement* içinde (pp. 621-694). Washington: American Council on Education.

Davies, A. (2012). Kane, validity and soundness. *Language Testing, 29*(1), 37-42. doi:10.1177/0265532211417213

De Jong, J. ve Benigno, V. (2016). The CEFR in higher education: Developing descriptors of academic English. Language Testing Forum 2016 - University of Reading'de sunulan bildiri. https://ukalta.org/wp-content/uploads/2016/10/DeJongBenigno_LTF2016.pdf adresinden erişildi.

De Jong, J. ve Benigno, V. (2017). Alignment of the global scale of English to other scales: The concordance between PTE Academic, IELTS, and TOEFL. Pearson: Global Scale of English Research Series. https://www.pearson.com/content/dam/one-dot-com/one-dot-com/english/TeacherResources/GSE/GSE-Alignment-other-scales.pdf adresinden erişildi.

Douglas, D. (2010). *Understanding language testing* (1. bs.). New York: Routledge. doi:10.4324/9780203776339

Eckes, T. (2017). Setting cut scores on an EFL placement test using the prototype group method: A receiver operating characteristic (ROC) analysis. *Language Testing*, *34*(3), 383-411. doi:10.1177/0265532216672703

European Commission. (t.y.). *Erasmus+ EU programme for education, training, youth and sport.* https://erasmus-plus.ec.europa.eu/ adresinden erişildi.

Frey, B. (Ed.). (2018). *The SAGE encyclopedia of educational research, measurement, and evaluation*. Thousand Oaks, CA: Sage. doi:10.4135/9781506326139

Fulcher, G. (2015). *Re-examining language testing: A philosophical and social inquiry*. New York: Routledge. doi:10.4324/9781315695518

Fulcher, G. ve Davidson, F. (Ed.). (2013). *The Routledge handbook of language testing*. New York: Routledge.

Gomez, P. G., Noah, A., Schedl, M., Wright, C. ve Yolkut, A. (2007). Proficiency descriptors based on a scale-anchoring study of the new TOEFL iBT reading test. *Language Testing, 24*(3), 417-444. doi:10.1177/0265532207077209

Harsch, C. (2014). General language proficiency revisited: Current and future issues. *Language Assessment Quarterly*, *11*(2), 152-169. doi:10.1080/15434303.2014.902059

Harsch, C. (2019). What it means to be at a CEFR level. Or why my Mojito is not your Mojito - on the significance of sharing Mojito recipes. A. Huhta, G. Erickson ve N. Figueras (Ed.), *Developments in language education: A memorial volume in honour of Sauli Takala* içinde (s. 76-93). Jyväskylä: University of Jyväskylä Centre for Applied Language Studies. https://www.ealta.eu.org/documents/resources/Developments%20in%20Language%20Education%20A%20Memorial%20Volume%20in%20Honour%20of%20Sauli%20Takala.pdf adresinden erişildi.

Harsch, C. ve Hartig, J. (2015). What are we aligning tests to when we report test alignment to the CEFR?. *Language Assessment Quarterly*, *12*(4), 333-362. doi:10.1080/15434303.2015.1092545

Harsch, C. ve Kanistra, V. P. (2020). Using an innovative standard-setting approach to align integrated and independent writing tasks to the CEFR. *Language Assessment Quarterly*, *17*(3), 262-281. doi:10.1080/15434303.2020.1754828

Hopkins, K. D. (1998). *Educational and psychological measurement and evaluation* (8. bs.). Boston: Allyn & Bacon.

Hulstijn, J. H. (2007). The shaky ground beneath the CEFR: Quantitative and qualitative dimensions of language proficiency. *The Modern Language Journal, 91*(4), 663-667. http://www.jstor.org/stable/4626094 adresinden erişildi.

Im, G. H., Shin, D. ve Cheng, L. (2019). Critical review of validation models and practices in language testing: Their limitations and future directions for validation research. *Language Testing in Asia, 9*. doi:10.1186/s40468-019-0089-4

Kane, M. T. (1994). Validating the performance standards associated with passing scores. *Review of Educational Research*, *64*(3), 425-461. doi:10.3102/00346543064003425

Kane, M. T. (2006). Validation. R. L. Brennan (Ed.), *Educational measurement* içinde (4. bs, s. 17-64). Washington: American Council on Education.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*(1), 1-73. doi:10.1111/jedm.12000

Katz, I. R. ve Tannenbaum, R. J. (2014). Comparison of web-based and face-to-face standard setting using the Angoff method. *Journal of Applied Testing Technology*, *15*(1), 1-17.

Kenyon, D. M. (2012). Using Bachman's assessment use argument as a tool in conceptualizing the issues surrounding linking ACTFL and CEFR. E. Tschirner (Ed.), *Aligning frameworks of reference in language testing: The ACTFL proficiency guidelines and the Common European Framework of Reference for Languages* içinde (s. 23-34). Almanya: Stauffenburg Verlag.

Kenyon, D. M. ve Römhild, A. (2013). Standard setting in language testing. A. J. Kunnan (Ed.), *The companion to language assessment* içinde (s. 944-961). Hoboken, NJ: John Wiley & Sons.

Knoch, U. ve Chapelle, C. A. (2018). Validation of rating processes within an argument-based framework. *Language Testing*, *35*, 477-499. doi:10.1177/0265532217710049

Lavery, M. R., Bostic, J. D., Kruse, L., Krupa, E. E. ve Carney, M. B. (2020). Argumentation surrounding argument-based validation: A systematic review of validation methodology in peer-reviewed articles. *Educational Measurement: Issues and Practice*, *39*(4), 116-130. doi:10.1111/emip.12378

Magno, C. (2009). Demonstrating the difference between classical test theory and item response theory using derived test data. *The International Journal of Educational and Psychological Assessment, 1*(1), 1-11. https://files.eric.ed.gov/fulltext/ED506058.pdf adresinden erişildi.

McNamara, T. ve Knoch, U. (2012). The Rasch wars: The emergence of Rasch measurement in language testing. *Language Testing, 29*(4), 555-576. doi:10.1177/0265532211430367

Mendoza, A. ve Knoch, U. (2018). Examining the validity of an analytic rating scale for a Spanish test for academic purposes using the argument-based approach to validation. *Assessing Writing, 35*, 41-55. doi:10.1016/j.asw.2017.12.003

Messick, S. (1989). Validity. R. L. Linn (Ed.), *Educational measurement* içinde (3. bs., s. 13-103). New York, NY: American Council on education and Macmillan.

North, B. (2000). *The development of a common framework scale of language proficiency.* New York: Peter Lang.

O'Loughlin, K. (2011). The interpretation and use of proficiency test scores in university selection: How valid and ethical are they?. *Language Assessment Quarterly, 8*(2), 146-160. doi:10.1080/15434303.2011.564698

Papageorgiou, S. ve Cho, Y. (2014). An investigation of the use of TOEFL® JuniorTM Standard scores for ESL placement decisions in secondary education. *Language Testing, 31*(2), 223-239. doi:10.1177/0265532213499750

Papageorgiou, S. ve Tannenbaum, R. J. (2016). Situating standard setting within argument-based validity. *Language Assessment Quarterly*, *13*(2), 109-123. doi:10.1080/15434303.2016.1149857

Papageorgiou, S., Xi, X., Morgan, R. ve So, Y. (2015). Developing and validating band levels and descriptors for reporting overall examinee performance. *Language Assessment Quarterly*, *12*(2), 153-177. doi:10.1080/15434303.2015.1008480

Plake, B. S. ve Cizek, G. J. (2012). Variations on a theme: The Modified Angoff, Extended Angoff, and Yes/No standard setting methods. G. J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* içinde (2. bs., s. 181-199). New York: Routledge.

Savuran, Y. ve Çubukçu, Z. (2021). Yabancı dil olarak Türkçe öğretiminde performans betimleyicileri geliştirme: Temel ve ara düzeyler. *Türk Eğitim Bilimleri Dergisi, 19*(2), 831-856. doi:10.37217/tebd.876422

Shin, S.-Y. ve Lidster, R. (2017). Evaluating different standard-setting methods in an ESL placement testing context. *Language Testing*, 34(3), 357-381. doi:10.1177/0265532216646605

Tannenbaum, R. J. ve Cho, Y. (2014). Critical factors to consider in evaluating standard-setting studies to map language test scores to frameworks of language proficiency. *Language Assessment Quarterly*, *11*(3), 233-249. doi:10.1080/15434303.2013.869815

Thompson, B. ve Levitov, J. E. (1985). Using microcomputers to score and evaluate items. *Collegiate Microcomputer*, *3*(2), 163-168.

Toulmin, S. (1958). *The uses of argument.* Cambridge: Cambridge University Press.

Toulmin, S. (2003). *The uses of argument* (Güncellenmiş bs.). Cambridge: Cambridge University Press.

Türkiye Bursları. (t.y.). Hakkımızda. https://www.turkiyeburslari.gov.tr/about adresinden erişildi.

Xi, X. (2007). Validating TOEFL® iBT Speaking and setting score requirements for ITA screening. *Language Assessment Quarterly*, *4*, 318-351. doi:10.1080/15434300701462796

Yükseköğretim Kurulu. (2023). *Statistics*. https://istatistik.yok.gov.tr/ adresinden erişildi.

Yükseköğretim Kurulu. (2024). *Scholarships for international students*. https://www.studyinturkiye.gov.tr/StudyinTurkey/ShowDetail?rID=KlqzJ6l8YDQ=&&cId=PE4Nr0mMoY4= adresinden erişildi.

Zwick, R., Senturk, D., Wang, J. ve Loomis, S. C. (2001). An investigation of alternative methods for item mapping in the National Assessment of Educational Progress. *Educational Measurement: Issues and Practice*, *20*(2), 15-25.