Ülper, H. (2026). Öğrencilerin alıcı sözcük dağarcığının sıklık dilimlerine göre belirlenmesine yönelik ölçme aracı geliştirme. *Ana Dili Eğitimi Dergisi*, *14*(2), 329-353. https://doi.org/10.16916/aded.1873846

Ana Dili Eğitimi Dergisi
Journal of Mother Tongue Education
www.anadiliegitimi.com

Geliş/Received: 28.01.2025 Kabul/Accepted: 31.03.2026
Araştırma Makalesi / Research Article

Öğrencilerin Alıcı Sözcük Dağarcığının Sıklık Dilimlerine Göre Belirlenmesine Yönelik Ölçme Aracı Geliştirme

Hakan ÜLPER[*]

Öz

Bu çalışmada anadili Türkçe olan öğrencilerin alıcı sözcük dağarcığını ölçmek amacıyla geliştirilen Sıklık Odaklı Alıcı Sözcük Dağarcığı Ölçme Aracı (SASÖ) tanıtılmış ve aracın Klasik Test Kuramı (KTK), Madde Tepki Kuramı (MTK) ve Genellenebilirlik Kuramına (GK) göre geçerlik ile güvenirliğine ilişkin elde edilen kanıtlar sunulmuştur. Araç, sözcük sıklığı temel alınarak toplam 195 madde olacak biçimde yapılandırılmıştır. Araştırmanın verileri Burdur'da öğrenim gören 5-8. sınıf düzeyindeki toplam 549 öğrenciden toplanmıştır. KTK sonuçları, aracın çok yüksek iç tutarlılığa iye olduğunu göstermiştir. MTK kapsamında uygulanan iki parametreli lojistik model (2PL) sonucunda, maddelerin büyük çoğunluğunun modele uyum sağladığı belirlenmiştir. Küresel uyum göstergelerine (SRMR, SRMSR, 100×MADCOV) göre model–veri uyumu çok iyi düzeydedir. Genellenebilirlik katsayısı ve KR20 güvenirlik katsayısı aracın güvenirliğinin ve genellenebilirliğinin çok yüksek olduğunu göstermektedir. Ayrıca, SASÖ puanları ile okuma başarısı arasında anlamlı ilişkiler saptanmış ve bu bulgular aracın yapı geçerliğini destekleyici niteliktedir. Sonuç olarak SASÖ, Türkçede alıcı sözcük dağarcığını ölçmek için geliştirilen geçerliği ve güvenilirliği çok yönlü olarak kanıtlanmış bir ölçme aracı görünümündedir. Bu ölçme aracı hem süreç hem de sonuç odaklı değerlendirmede kullanılabilecek çağdaş bir ölçme aracı niteliğindedir.

Anahtar Kelimeler: Sözcük bilgisi, sözcük dağarcığı, sözcük ölçme aracı, madde tepki kuramı

Development of A Measurement Tool for Determining Students' Receptive Vocabulary Profiles Based on Frequency Bands

Abstract

This study introduces the Frequency-Focused Receptive Vocabulary Test (FFRVT), developed to assess students' receptive vocabulary knowledge in Turkish. It presents evidence for its validity and reliability based on Classical Test Theory (CTT), Item Response Theory (IRT), and Generalizability Theory (GT). The instrument comprises 195 frequency-based items. Data were collected from 549 students enrolled in Grades 5-8 in the province of Burdur. Results from CTT analyses indicated very high internal consistency. Analyses conducted using the two-parameter logistic (2PL) model within the IRT framework showed that the vast majority of items demonstrated adequate model fit. Global fit indices (SRMR, SRMSR, and 100 × MADCOV) further confirmed an excellent model–data fit. Both the generalizability coefficient and the KR-20 reliability coefficient indicated that the instrument had very high reliability and generalizability. In addition, significant relationships were observed between FFRVT scores and reading achievement, providing support for the test's construct validity. In conclusion, FFRVT appears to be a psychometrically sound instrument with robust evidence of validity and reliability for assessing receptive vocabulary knowledge in Turkish. As a contemporary measurement tool, it can be effectively employed in both process-oriented and outcome-based assessments.

Keywords: Vocabulary knowledge, vocabulary, vocabulary test, item response theory

[*] Prof. Dr., Burdur Mehmet Akif Ersoy Üniversitesi, Eğitim Fakültesi, Türkçe ve Sosyal Bilimler Eğitimi Bölümü, Burdur, hakanulper@mehmetakif.edu.tr

Giriş

Sözcük olgusu henüz kesin çizgileriyle tanımlanabilen bir olgu olmasa da dil öğretimi bağlamında, "*Metin üretme sürecinde anlamlı bir iletişimsel yapı ve bağlam oluşturmak ya da üretilmiş olan bir metni okuma ve dinleme sürecinde bağlamsal bir anlama ulaşmak için işlevleri ve anlamları temel alınarak işlemlenen sesbirimsel bireşimlerin oluşturduğu yapı ya da öbek*" olarak tanımlanabilir (Ülper, 2023). Bu tanımın da gönderimde bulunduğu gibi, bir sözcüğü biliyor olmak dört beceri bağlamında sözcük konusunda yeterli bilgiye iye olmayı gerektirir. Bu bağlamda sözcük bilmek biçim, anlam ve kullanım boyutlarından oluşur. Biçim sözcüğün yazım ve sesletimini, anlam çeşitli anlamsal yönlerini ve kullanım sözcüğün nerede, nasıl ve hangi sözcükler bir arada kullanılacağını içerir (Nation, 2001). Ancak her dil becerisini sergileyebilmek bu boyutların tümünü de bilmeyi gerektirmez. Örneğin okuma ve dinleme gibi alıcı beceriler için genel olarak biçim ve anlam bilgisi yeterli iken yazma ve konuşma gibi üretici beceriler içinse kullanım bilgisi de özellikle gereklidir. Bu bağlamda sözcük dağarcığını alıcı ve üretici olarak düşünmek yerinde olacaktır. Bu ayrım, sözcük dağarcığını ölçmek amacıyla bir ölçme aracı geliştirme sürecinde göz önünde bulundurulması gereken önemli bir konudur. Öyle ki Schmitt (2014) tarafından da belirtildiği gibi, ölçme aracının alıcı mı, üretici mi yoksa her iki bilgi türünü mü ölçeceğini açıkça belirtmek gerekir. Çünkü ölçme aracının olası kullanıcıları ölçülen sözcük dağarcığıyla dil becerileri bakımından neler yapılabileceğine yani sözcük dağarcığının hangi dil becerilerine yönelik olduğuna da bakarlar. Ayrıca araştırmalar da alıcı olarak bilinen sözcüklerin tamamının üretici olarak kullanılamadığını göstermektedir. Alıcı sözcük dağarcığının üretici olarak kullanılabilme oranı %16-50 arasında değişmektedir (Fan, 2000; Laufer, 2005). Dolayısıyla alıcı ve üretici ayrımı yapmadan belirlenecek olan sözcük dağarcığı yanıltıcı olacaktır. Bu nedenle bu çalışmada geliştirilecek olan ölçme aracı alıcı sözcük dağarcığı ve okuma becerisi ile ilgilidir. Çünkü uygulama, öğrencilerin soruları dinleyerek değil yalnız okuyarak yanıtlamaları ilkesine dayanmaktadır.

Alanyazında üretici ve alıcı ayrımı belirgin bir biçimde yapılmaktadır (Laufer ve Goldstein, 2004; Nation, 2001; Webb, 2005). Bunun altında yatan en temel varsayım da sözcük bilgisinin derinliği ile ilgilidir. Bu bağlamda alıcı sözcük bilgisinin derinliği üretici sözcük bilgisine göre daha azdır. Bu bakımdan önce alıcı sonra da üretici sözcük bilgisi gelişir (Read, 2000). Bu durumun doğal bir sonucu olarak da kişilerin alıcı sözcük dağarcıkları daha geniştir ve alıcı sözcük dağarcığındaki her sözcük üretici olarak kullanılmaz ya da kullanılamaz. Bu nedenle de sözcük ölçme araçları geliştirilirken bu ayrımlar göz önünde bulundurulur ve araçların düzenleri ona göre oluşturulur.

Alıcı ve üretici sözcük dağarcığı kavramlarına ayrıntılı bir biçimde bakınca, üretici sözcük dağarcığının dildeki diğer sözcükler tarafından, alıcı sözcük dağarcığının ise dış uyaranlar tarafından etkinleştirildiği görülecektir. Bu bakımdan üretici sözcük dağarcığının tersine alıcı sözcük dağarcığı kişilerin yazılı ya da sözlü bir metin üretmelerini gerektirmeyen buna karşın üretilmiş metinleri anlamlandırırken kullandıkları sözcük dağarcığıdır. Bu nedenle hem okuma hem de dinleme becerileriyle yakından ilişkilidir (Meara, 1990). Bu ilişki alanyazındaki çalışmalarla da desteklenmektedir (bkz. Henriksen vd., 2004; Qian, 2002; Zhang ve Anual, 2008). Ortaya konan bu ilişki öğrencilerin okuma ve dinleme becerilerini verimli bir biçimde gerçekleştirmeleri bakımından alıcı sözcük dağarcığının önemine gönderimde bulunmaktadır. Buna koşut olarak alanyazında kişilerin okuma ve dinleme becerilerini yerine getirebilmeleri için bilmeleri gereken sözcük sayıları da belirlenmiştir. Buna göre gazete ve romanları etkili bir biçimde okuyabilmek için İngilizcede 8000-9000 sözcük ailesine (Nation, 2006), Türkçede 14000 sözcüğe (Ülper ve Kiraz, 2020) gereksinim vardır. Dinleme içinse 7000 sözcük gerekmektedir (van Zeeland ve Schmitt, 2013). Bu durumda öğrencilerin kaç sözcük bildiklerinin saptanması son derece önemlidir. Ancak bu saptama yapıldıktan sonra öğrencilerin okuma becerilerine ilişkin içinde bulundukları düzey sağlıklı bir biçimde betimlenebilir, olası okuma sorunlarına daha nesnel açıklamalar getirilebilir.

Alanyazındaki çalışmalara göre İngilizcede eğitimli kişiler yaklaşık olarak 17.000 sözcük ailesi bilmektedir. Bu da günde 2-3 dolayında yılda ise yaklaşık 1000 sözcük ailesi öğrenimine gönderimde bulunmaktadır. Her sözcük ailesine giren bireysel sözcüklerin de sayıma katılması durumunda bu sayı daha da yükselecektir (D'Anna vd.,1991; Goulden vd., 1990). Bu bulgular temel alınan sözlüklerin, seçilen yöntemin, sözcük seçme ilkelerinin ve ölçme araçlarının ayrı olmasından dolayı diğer pek çok

çalışmanın bulgularıyla çelişebilmektedir. Bu nedenle alanyazında ayrı sayılar veren çok sayıda araştırmayla karşılaşmak olanaklıdır (Anderson ve Nagy, 1993; Milton ve Treffers-Daller, 2013).

Özellikle İngilizce alıcı sözcük dağarcığı genişliğinin ve gelişiminin ölçülmesinde Vocabulary Level Test (VLT) olarak bilinen ve ayrı bir dizgeye iye olan ölçme aracının çok kullanıldığı ve süreç içinde daha da geliştirildiği göze çarpmaktadır. Buna karşın bu tür bir ölçme aracı Türkçenin ne anadili ne de yabancı dil olarak öğretim süreçlerinde sözcük dağarcığı gelişimini saptamak için alanyazına kazandırılmamış ve buna koşut olarak da hem anadili hem de yabancı dil öğretim süreçlerinde öğrencilerin bildikleri toplam alıcı sözcük sayısına ve gelişimine ilişkin bir araştırma yapılmamıştır. Alıcı sözcük dağarcığının dil becerileri, okuma anlama ve sınav başarısı bakımından, yukarıda da kısaca değinildiği gibi, ne denli önemli olduğu göz önünde tutulunca böyle bir çalışmanın sonucunda ortaya konacak bir ölçme aracının önemi daha da belirginleşecektir. Bunun yanında böyle bir ölçme aracını alanyazına kazandırarak Türkiye ölçeğinde öğrencilerin binlik dilimlere göre bildikleri alıcı sözcük sayılarını belirleyebilmek ve bir norm ortaya koyabilmek yine benzer biçimde öğrencilerin gelişim hedeflerini de saptayabilmek olanaklı olacaktır. Bunun yanında ortaya çıkan sonuçlara göre özellikle sözcük öğretimine nasıl ve nereden (hangi binlik dilimden) başlanacağı konusunda alanyazın için önemli veriler sağlanabilecektir. Bunlar da araştırmanın önemini ortaya koyan diğer olası kazanımlar olarak görülmelidir.

Bu çalışmanın amacı sıklık dilimlerine ve sınıf düzeylerine (5-8. sınıf) göre ortaokul öğrencilerinin okuma süreçlerinde işlevsel olan binlik dilimlere göre kaç sözcük bildiklerini belirlemeye yönelik bir ölçme aracı geliştirmektir. Buna göre aşağıdaki araştırma sorularının yanıtları aranacaktır.

1. Sıklık Odaklı Alıcı Sözcük Dağarcığı Ölçme Aracı (SASÖ) ölçümleri geçerli sonuçlar üretiyor mu?
2. Sıklık Odaklı Alıcı Sözcük Dağarcığı Ölçme Aracı (SASÖ) güvenilir sonuçlar üretiyor mu?

Yöntem

Evren/Örneklem

Sözcük dağarcığı ölçme araçları belirli bir hedef kitleye uygun olarak geliştirilmelidir. Çünkü hiçbir sözcük dağarcığı ölçme aracı tüm düzeylerdeki bireyler için geçerli ve uygun değildir; bu durum elde edilen puanların yanıltıcı olmasına yol açabilir (Schmitt vd., 2020). Bu bakımdan öncelikle aracın yönelik olduğu hedef kitlenin belirlenmesi ve içeriğinin buna göre düzenlenmesi gerekmektedir. Bu araştırmada katılımcılar "orantısız tabakalı yansız örnekleme yöntemi" (Christensen vd., 2015) kullanılarak Burdur il merkezindeki üst, orta ve alt sosyoekonomik düzeydeki okullarda öğrenim gören 5–8. sınıf öğrencileri içinden her bir sınıf düzeyine göre seçilmiştir.

Araştırmaya yalnızca gönüllü olarak katılmak isteyen öğrenciler alınmıştır. Toplamda 598 öğrenciden veri toplanmış, ancak bu verilerin 549'unun geçerli olduğu belirlenmiş ve analizler 549 öğrenci verisi üzerinden yürütülmüştür. Örneklemin 288'i kız, 261'i erkektir. Öğrencilerin Türkçe dersi başarı notları 41 ile 100 arasında değişmektedir; bunların 14'ü 41–60, 102'si 61–80 ve 433'ü 81–100 aralığındadır. Sınıf düzeylerine göre dağılım 5. sınıf 115, 6. sınıf 116, 7. sınıf 127 ve 8. sınıf 191 öğrenci biçimindedir. Öğrencilerin yaş ortalaması 12,6'dır (SS = 1,26; aralık = 10–15).

Araştırmanın Modeli ve Deseni

Bu araştırma, nicel araştırma yaklaşımı çerçevesinde yürütülen ve hem klasik test kuramı hem de madde tepki kuramına (MTK) dayalı olarak geliştirilen bir ölçme aracı geliştirme çalışmasıdır. Araştırmada veriler tek bir uygulama sürecinde toplanmış olup çalışma kesitsel desen temelinde yürütülmüştür (Christensen vd., 2015).

Verilerin Çözümlenmesi

Bir ölçme aracının geliştirilmesinde göz önünde bulundurulması gereken en temel iki kavram geçerlik ve güvenirliktir. Geçerlik, bir ölçme aracının ölçmek istediği özelliği ne ölçüde ölçtüğüyle ilişkiliyken; güvenirlik, ölçme aracının bir özelliği tutarlı ve doğru biçimde ölçebilme düzeyini anlatır (Milton, 2009). Hem geçerlik hem de güvenirlik durumunun belirlenebilmesi için ayrı istatistiksel ve yöntemsel yaklaşımlar kullanılabilmektedir.

Geçerlik için başvurulan en önemli yöntemlerden biri *içerik geçerliği*dir. Bu, ölçme aracının sağlıklı bir ölçüm için gerekli ve uygun içeriğe iye olma derecesine gönderimde bulunur. Bu tür ölçme araçlarının iyi bir içerik geçerliğine iye oldukları ileri sürülebilir. Çünkü bu araçlar öğrencilerin bilgi iyesi olmalarının olasılık içinde olduğu sözcükleri sınamak için sözcüklerin sıklık bilgilerinden yararlanır. Dolayısıyla sözcük dağarcığı ölçme araçları için içerik oluşturulurken sözcüklerin sıklık bilgisini kullanmak içerik geçerliğini sağlamak için yararlanılan önemli bir stratejidir. Bir diğeri *yapı geçerliği*dir. İçerik geçerliliği ile yakından ilişkili olan bu geçerlik ölçme aracının ölçmesi gereken yapıyı ölçüp ölçmediğine odaklanır. Alıcı sözcük dağarcığı ölçümleri oluşturulurken araştırılacak sözcükler seçebilir ve burada bir yapıdan söz edilebilir. Bunun için de ölçme aracının yapısal özelliklerinin uygun yöntemlerle belirlenmesi gerekir. Son olarak işe koşulabilecek olan geçerlik *görünüş geçerliliği*dir. Bu, ölçme aracının ölçmesi gereken şeyi ölçen bir araç olarak kullanıcılara inandırıcı gelip gelmediği ile ilgili bir geçerliktir. İyi, yapı ve içerik geçerliliğine iye araçlar bile bazen öğrenciler tarafından sorgulanabilir. Bu bakımdan öğrencilerin görüşlerine başvurulması da önemli görünmektedir (Dallers vd.2007; Milton, 2009). Bununla birlikte dille ilgili konularda geliştirilen ölçme araçları için geçerlilik siyah beyaz bir konu olarak ele alınamaz. Yani bir ölçme aracı için geçerli ya da geçersiz demekten çok güven dereceleri ve yorumlama incelikleri de göz önünde tutulmalıdır. Bu bağlamda puan yorumu üzerinde durmak yararlı olacaktır. Puan yorumu puanların gerçekte ne anlama geldiği ile ilgilidir ve önemli bir doğrulama yöntemidir. Bunun için de sözcük dağarcığı ölçme aracı performansının okuma, yazma, konuşma ve dinleme gibi uygun düşen dil becerileriyle ilişkili olması gerekir. Burada gereksinim duyulan şey yanıtların söz konusu hedef sözcüğün ne denli iyi kullanılabileceğini bir dış ölçütle doğrudan karşılaştırmaktır (Schmitt vd., 2020).

Ölçme aracı güvenilirliği ise, bir aracın ölçmesi gereken şeyi aynı yeteneğe iye kişiler için tutarlı bir biçimde ölçme doğruluğudur. Bu bağlamda sözcük dağarcığı değişmeyen bir kişi için birkaç kez ölçüm yapılınca aynı sonuçların alınması beklenir. Bu beklentiyi karşılayan bir ölçme aracı güvenirliği olan bir araçtır (Dallers vd., 2007). Bu bağlamda seçeneklerden biri test-tekrar test yöntemini kullanmaktır. Çünkü sözcük bilgisi madde temellidir bu nedenle sözcükler sıklık, biçim, anlam tema bakımından benzer olsa bile yalnız iç tutarlıklarının belirlenmesi yeterli değildir. Bununla birlikte birçok sözcük dağarcığı ölçme aracında yalnız alfa, KR20 ve KR21 değerleri de hesaplanmaktadır (Schmitt vd., 2020).

Öğrencilerden toplanan veriler kodlanarak doğru yanıtlara 1, yanlış yanıtlara ise 0 puan verilmiştir. Elde edilen veriler Klasik Test Kuramı (KTK) bağlamındaki çözümlemeler için SPSS ve Excel programları; Madde Tepki Kuramı (MTK) bağlamındaki çözümlemeler için ise R programı kullanılarak geçerlik ile güvenirlik analizleri gerçekleştirilmiştir. Bu bağlamda yukarıda sunulan kuramsal açıklamalardan ve alanyazındaki benzer nitelikteki alıcı sözcük bilgisi ölçme aracı geliştirme çalışmalarından (bkz. Çetinkaya vd., 2023; McLean vd., 2015; Schmitt vd., 2001; Qi vd., 2024; Webb vd., 2017) yönelimle, ölçme aracının geçerliğini belirlemek amacıyla şu yöntemlere başvurulmuştur: uzman görüşü alma, içerik geçerliği için uygun sözcük seçimi, Klasik Test Kuramı ve Madde Tepki Kuramına göre madde güçlük ve ayırtedicilik değerlerinin hesaplanması, eşdizimlik testi ve okuma anlama testi ile karşılaştırma, sınıf düzeylerine göre karşılaştırma, sıklık dilimlerine göre karşılaştırma, açımlayıcı faktör analizi ve katılımcılarla görüşme. Ölçme aracının güvenirliğini belirlemek için ise KR20 ve McDonald's Omega katsayısının hesaplanması ve Genellenebilirlik Kuramı bağlamında genellenebilirlik katsayılarının elde edilmesi yöntemleri kullanılmıştır. Böylece geliştirilen ölçme aracının ürettiği puanlara ilişkin geçerlik ve güvenirlik kanıtları bütüncül biçimde ortaya konmuştur.

Taslak Ölçme Aracının Oluşturulması

Bu çalışma bir sözcük dağarcığı ölçme aracı geliştirme çalışması olduğundan, güncel alanyazın doğrultusunda böyle bir araç geliştirme sürecinde uyulması gereken şu temel ilkelere bağlı kalınmıştır:

1. Belirlenen hedeflere ulaşmada hangi madde düzenlerinin uygun olduğuna ilişkin eleştirel bir çözümleme yapılması,
2. Belirli madde düzenlerinin sorunlarının ve sınırlılıklarının ayırt edilmesi,
3. Ölçme aracı için sözcük seçkilerinin oluşturulmasında uygun derlem(ler)in kullanılması,

4. Sözcük seçkilerinin oluşturulmasına ilişkin ilkeler, sayım birimi, nelerin sözcük olarak sayılacağı ve nelerin sayılmayacağı ile sıklık verilerinin kullanımına ilişkin açık tanımlamaların yapılması,
5. Sözcük evreninden örnekleme yönteminin belirlenmesi (Schmitt, Nation ve Kremmel, 2020).

Bununla birlikte, Cameron'a (2002) göre Sözcük Düzeyleri Testi (Vocabulary Level Test) ortaokul öğrencilerinin sözcük bilgilerini ölçmede oldukça etkili bir araçtır. Bu nedenle öğrencilerin sıklık dilimlerine göre bilinen sözcük sayılarını kestirmede kullanılacak (Webb vd., 2017) bir ölçme aracı geliştirmek için bu aracın geliştirme ilkelerine benzer ilkeler temel alınacaktır. Ancak araştırmamızın amacı ve hedef kitlesi bağlamında bazı değişiklikler ve düzenlemeler yapılacaktır. Örneğin VLT'den ayrı olarak hedef kitle anadili öğrenicileridir ve yine ölçme aracında 1000 ve 4000 sıklık diliminden sözcükler de yer alacaktır. Bu değişiklikler araştırmamız açısından önemli ve gereklidir çünkü Dağhan ve Ülper'e (2022) göre 1000'lik dilimdeki sözcükler ders kitapları metinlerindeki sözcüklerin yaklaşık %64'ünü, 2000'lik dilim %75'ini karşılamaktadır. Gazeteler için de benzer oranlardan söz edilebilmektedir (Ülper ve Kiraz, 2020).

SASÖ geliştirme süreçlerinin yaşamsal aşamalarından biri, ölçülecek sözcüklerin hedef dili örnekseyecek biçimde seçilmesidir. Bu seçme işleminin uygun bir biçimde yapılması araçların ölçme gücüyle doğrudan ilgilidir. Bu bağlamda hedef kaynaktan elde edilen geniş ve örnekleyici derlemlerden oluşturulan sözcük seçkileri, ölçme aracı geliştirme sürecinde güvenilir örnekleme havuzları sunmaktadır. Bu tür seçkilerin yöntemsel olarak sağlam biçimde tasarlanması, bu seçkilerden yararlanılarak geliştirilen araçların geçerliliğini artırmaktadır. Tersi durumda, ölçme araçlarının hedeflenen sözcük bilgisini doğru biçimde yansıtması olanaklı olmayacaktır (Dang ve Webb, 2025).

Sözcükler, ölçme aracının güncelliğini sağlamak amacıyla en güncel sıklık sözlüğü durumundaki Türkçe Ulusal Derlemi Tabanlı Sözcük Sıklık Sözlüğü'nden (Aksan vd., 2017) seçilmiştir. Sözcüklerin seçiminde orantılı tabakalı yansız örnekleme yöntemine başvurulmuştur (Christensen, Johnson ve Turner, 2015). Buna göre sözlükteki 5x1000'lik dilimlere göre ad, eylem ve sıfat türünün her birinin yaklaşık %39'u olanaklı olduğunca her yüzlük dilimden de sözcük olacak biçimde seçilmiştir. Bunun için sözcükler ilk aşamada türlerine (ad, eylem, sıfat) ve 100'lük dilimlere göre ayrıştırılmış sonrasında her 1000'lik dilimdeki sözcük türlerinin oranları hesaplanmış ve bu oranlara göre seslem sayısı da göz önünde bulundurularak sözcükler seçilmiştir. Bu süreçte bilimsel terminoloji, coğrafya adları, özel adlar, argo, ünlem, bağlayıcı, belirteç, birden çok sözcüklü öbekler, aynı kökten türeyen birden fazla sözcük seçilmemiştir. Aynı sözcük ailesinden yalnız bir sözcük seçilmiştir. Ayrıca seslem sayısının da sözcük zorluğu üzerindeki etkisinden dolayı (Çetinkaya ve Uzun, 2018) seslem sayısı bakımından da dengeli bir dağılım sağlanmaya çalışılmıştır. Diğer yandan da her binlik dilimdeki sözcüklerin ödünçleme olup olmadıklarına bakılmış ve sözcük seçiminde bu konu da göz önünde bulundurulmuştur. Bu yaklaşım önemlidir; çünkü Laufer ve McLean (2016) ödünçleme sözcüklerin sonuçları etkilediğini bulmuştur.

Seçilen sözcükler kendi içinde her kümede 6 sözcük ve üç açıklama olacak biçimde ad, sıfat ve eylem kümesi olarak ayrıştırılmıştır. Bu bağlamda ölçme aracının ön uygulaması için 1x1000 dilimindeki sözcüklerin tür dağılım oranlarına göre 7 ad, 2 sıfat 4 eylem kümesi; 2x1000 dilimindeki sözcüklerin tür dağılım oranlarına göre 8 ad, 2 sıfat, 3 eylem; 3x1000, 4x1000 ve 5x1000 dilimindeki sözcüklerin tür dağılım oranlarına göre 8 ad, 3 sıfat 2 eylem kümesi belirlenmiştir.

Bu ölçme aracında çoklu eşleştirme dizgesi kullanılmıştır. Bu dizgenin bir örneği Tablo 1'de gösterilmektedir. Bu dizge İngilizce öğrenen öğrencilerin sözcük bilgisini ölçmek amacıyla Webb, Sasao ve Ballance (2017) tarafından geliştirilen dizgeden esinlenerek oluşturulmuştur.

Tablo 1
SASÖ Ad Kümesi Örneği

	fener	mağaza	okul	radyo	tarla	yolcu
Eğitim yeri			x			
Ekilen toprak					x	
Işık veren alet	X					

Çizelgede de görüleceği gibi kümelerdeki sözcükler abecesel olarak, açıklamalar ise uzunluklarına göre dizilmiştir. Açıklamalar elden geldiğince ilk 2000 dilimine giren sözcüklerle yapılmaya çalışılmış, olanaklı olmayan durumlarda ise sorulan sözcüğün yer aldığı binlik dilimin daha alt dilimlerindeki sözcüklerle yapılmıştır. Örneğin 5x1000'lik dilimdeki bir sözcüğün açıklaması 1-4x1000'lik dilimdeki sözcüklerle yapılmıştır. Yalnız bu durum 2x1000'lik dilim için geçerli değildir. Bu dilimdeki sözcüklerin açıklamaları yine ilk 2000'lik dilimdeki sözcükler içinden yapılmıştır.

Veri Toplama Süreci

SASÖ, uygulama için kâğıt-kalem düzenine göre tasarlanmıştır. Bu nedenle fotokopilerle çoğaltılarak 5-8. sınıf düzeyindeki öğrencilere araştırmacı tarafından uygulanmıştır. Uygulama öncesinde öğrencilerin ebeveynlerinden imzalı onam formları alınmıştır. Uygulama sürecinin ilk aşamasında uygulamanın ne için yapıldığı ve ders notlarını etkilemeyeceği öğrencilere açıklanmış ve ardından araştırmacı tarafından uygulamanın nasıl yapılacağı uygulamalı olarak gösterilmiştir. Yine öğrencilerden anlamını bilmedikleri sözcükler için herhangi bir im koymamaları, boş bırakmaları istenmiştir. Herkesin nasıl yapacağını anladığından emin olduktan sonra süre başlatılmış ve bir ders saati sonrasında toplanmıştır. Bir hafta sonra ise eşdizimlik ve okuma anlama testleri uygulanmıştır.

Araştırma ve Yayın Etiği

Bu çalışmada "Yükseköğretim Kurumları Bilimsel Araştırma ve Yayın Etiği Yönergesi" kapsamında uyulması belirtilen tüm kurallara uyulmuştur. Yönergenin ikinci bölümü olan "Bilimsel Araştırma ve Yayın Etiğine Aykırı Eylemler" başlığı altında belirtilen eylemlerden hiçbiri gerçekleştirilmemiştir.

Etik Kurul İzni

Kurul adı=Mehmet Akif Ersoy Üniversitesi Girişimsel Olmayan Klinik Araştırmalar Etik Kurulu
Karar tarihi= 18.06.2025 Çarşamba
Belge sayı numarası= GO 2025/1722

Bulgular

Ön Uygulamalara İlişkin Bulgular

Ön uygulamanın ilk aşamasında üniversite öğrencilerine yönelinmiştir. Türkçe öğretmenliğinde öğrenim gören 100 üniversite öğrencisine yapılan bu ön uygulamada öncelikle ölçme aracının açıklamalar kısmına ilişkin öğrencilerden yorumlar alınmıştır. Buna göre iki ayrı kümedeki birer açıklamanın iki yanıtının olabileceği ve iki açıklamanın da anlamının yeterince anlaşılır olmadığı belirlenmiştir. Ayrıca 1x1000 diliminde yer alan bir sözcüğün (gönder) %90 oranında bilinmediği saptanmış ve bu sözcüğün çıkarılmasına, yine sorunlu açıklamaların değiştirilmesine karar verilmiştir. Ölçme aracındaki ilk 2000 sözcüğün %100, diğer tüm sözcüklerin ortalama %98-100 arasında bilindiği görülmüştür. Bu durum da seçilen sözcüklerin 5-8. sınıf düzeyindeki öğrenciler için uygun olacağı biçiminde yorumlanmıştır.

Ön uygulamanın ikinci aşamasında 5-8. sınıf düzeyinden ve ayrı sosyo-ekonomik bölgelerdeki ayrı okullardan 100 öğrenci ile bir ön uygulama yapılmıştır. Ardından SASÖ'nün açıklamalar kısmında yer alan sözcükler öğrencilere dağıtılmış ve öğrencilerin bu sözcükler için biliyorum ya da bilmiyorum seçeneklerinden birini imlemeleri istenmiştir. Bunun sonucunda öğrencilerin açıklama kısmındaki tüm sözcükleri bildikleri görülmüştür. Bu durum SASÖ'nün uygulanacağı hedef kitle için açıklamalar kısmındaki sözcüklerin tamamının bilindiği biçiminde yorumlanmıştır. Bunun üzerine açıklamalar kısmındaki tüm sözcükler olduğu gibi bırakılmıştır. Yine ölçme aracının uygulama süresi, punto büyüklüğü ve düzeninin anlaşılabilir olup olmadığı, uygulama ve imleme açsından bir sorun olup olmadığı sorulmuş ve öğrencilerin tamamının anlaşılabilir ve uygun bulması üzerine ölçme aracında 10 punto ve bu düzenin kullanılmasına ayrıca uygulamanın bir ders saati içinde yapılabilirliğine karar verilmiştir. Sonraki aşamada ise öğrencilere bu ölçme aracının öğrencilerin sözcük dağarcıklarını ölçmek için yararlı ve kullanışlı olup olmayacağı yönündeki görüşleri sorulmuş ve öğrencilerin hiçbirisinden de bu konuda olumsuz bir yanıt alınmamıştır. Tam tersine öğrenciler böyle bir aracın

gerekli olduğunu belirtmiş ve bu uygulamanın kendi düzeylerini görmek açısından önemli olduğunu ve kendilerini güdülediğini söylemişlerdir.

SASÖ'yü asıl uygulamaya hazır duruma getirmek amacıyla gerçekleştirilen son işlem ise Türkçe eğitimi alanında görev yapan üç akademisyenden uzman görüşü almaktır. Bu işlem için uzmanlara ölçme aracının amacı, dizgesi ve sözcük seçme ilkeleri anlatılmış ve öğrencilerle yapılan işlemler açıklanmıştır. Üç uzman önce SASÖ'yü seçilen sözcükler ve sözcüklerin açıklamaları bağlamında sonra da sözcük seçme ilkeleri bakımından değerlendirmişlerdir. Sonrasında ise uzmanlarla toplantı yapılarak tartışılmış ve bazı sözcüklerin açıklamalarının düzeltilmesinin yerinde olacağı kararlaştırılmıştır. Bu düzeltmeyle birlikte uzmanlar arasında tam bir görüş birliği sağlanmıştır. Bu son aşamadan sonra SASÖ öğrencilere asıl uygulama için hazır duruma gelmiştir.

Ölçümlerin Geçerlik ve Güvenirlik Bulguları

Yapı Geçerliğinin İçerik Boyutu

Yapı geçerliliğinin içerik boyutunun temel göstergelerinden biri örneksenebilirlik (temsil edilebilirlik)tir (Messick, 1989). Değerlendirme sürecinde görevlerin seçimi, ölçülmek istenen yapının örneksenmesini gerektirir. Yapının edime (performansa) gerçekçi biçimde yansıtılabilmesi, değerlendirme görevlerinin hem kapsamlı hem de aslına uygun olmasına bağlıdır. Bu durum, eğitimsel ve psikolojik ölçümlerde, özellikle de performans değerlendirmelerinde belirleyici bir öneme iyedir. Nitekim Messick'e (1994) göre yapı geçerliliğinin içerik yönü açısından yaşamsal önemdeki bir diğer konu, değerlendirilecek yapı alanının sınırlarının açık biçimde belirlenmesidir. Yukarıda sözcük seçimi bölümünde, ayrıntılı bir biçimde açıklandığı gibi, sözcüklerin seçiminde sıklık dilimleri, temel alınmış ve sözcükler, sözcük türlerine göre ve derlemdeki oranlarına göre belirlenmiştir. Bu belirlemede ayrıca sözcüklerin seslem sayılarının ve ödünçleme olup olmadıklarının da göz önünde tutulması önemlidir. Bundan dolayı içerik bakımından içinden seçildiği 5000 sözcükten oluşan derlemi örneksediği düşünülmektedir.

Gyllstad, McLean ve Stewart (2021), sözcük dağarcığı ölçme araçlarında 1000 sözcüklük dilimlerin çoğu kez yalnızca 5-10 maddeyle örneksenmesini geniş güven aralıkları üretmesi nedeniyle eleştirmektedir. Buna karşın binlik dilimler başına 30 ve üzeri sözcüğün seçilmesi durumunda güven aralığının daraldığını ve ölçümlerin çok daha kararlı duruma geldiğini böylece de belirsizliğin açık biçimde azaldığını ortaya koymaktadır. Bu durumda çalışmamızda 5000 sözcük evreni olduğu için en az 150 sözcüğün olması gerektiği açıktır. Bununla birlikte hedef kitlenin anadili öğrencileri olması dolayısıyla bu sayıyı artırmanın uygulama açısından bir sorun yaratmayacağı düşünüldüğünden dolayı güven aralığını daha da daraltarak örnekseme gücünü daha da yükseltmek için binlik dilim başına 39 toplamda ise 195 sözcük seçilmiştir. Bunun yanında sözcüklerin kuramsal olarak zorluk düzeyleri de sözcük seçiminde dikkate alınmış ve ölçme açısından uygun sözcükler seçilmiştir. Bu da içerik boyutu açısından geçerlik için önemli bir kanıt sunmaktadır.

İçerik uygunluğu, maddelerin ölçülmekte olan yapıyla yani sözcüklerin biçim-anlam ilişkilerine ilişkin alıcı bilgisiyle ne ölçüde ilişkili olduğunu gösterir (Qi vd., 2024). Bu bağlamda katılımcıların uygulama sürecinde bir sözcük ile o sözcüğün anlamını eşleştirmeleri gerektiği için SASÖ dizgesinin, verilen biçim-anlam bağlantıları kurma yani alıcı sözcük bilgisini ölçme bakımından uygun olduğu açıktır. Hem uzman görüşleri hem de benzer amaç güdümünde hazırlanmış olan birçok çalışmanın (McLean vd., 2015; Schmitt vd., 2001; Webb vd., 2017) da bu tür çoklu eşleştirme dizgesini kullanmış olması önemli bir gösterge olarak sunulabilir.

Değerlendirme görevlerinin hem içeriğe uygunluğu hem de örnekseme yeteneği geleneksel olarak uzman görüşü ile değerlendirilir. Bu, yapı geçerliliğinin içerik yönünü ele almaya katkı sağlar (Messick, 1994). Kapsamın tanımlanması, sınırlandırılması ve örnekseyiciliği okullarda kullanılan ölçme araçlarının geçerliği için çok önemli kanıtlardır (Kelecioğlu ve Göçer Şahin, 2014). Bu bağlamda uzman görüşlerine de başvurulmuş olduğu için bu aşamaya dek yapılan tüm uygulamaların sonucunda içerik boyutunun uygun olduğu düşünülmektedir.

Klasik Test Kuramına Göre Madde Analizleri Bulguları

Tüm maddeler (195 madde) Excel programına aktarılarak her bir madde için doğru yanıtlanma oranları (madde güçlüğü, *p*) ve toplam puan ile olan ilişkileri (madde ayırtediciliği, *r*) bağlamında analiz edilmiştir. Bu analizin sonuçları Tablo 2'de sunulmaktadır.

Tablo 2

Madde Güçlük (MG) ve Madde Ayırtedicilik (MA) Değerleri

Madde	(p)	(r)	Madde	(p)	(r)	Madde	(p)	(r)	Madde	(p)	(r)
Taraf	0,75	0,33	Tutku	0,89	0,54	Sentez	0,29	0,31	Marjinal	0,27	0,43
Koru	0,68	0,33	Otorite	0,74	0,65	Grev	0,61	0,59	Ham	0,66	0,56
Yat	0,94	0,42	Bünye	0,83	0,42	Bileşen	0,65	0,63	Harbi	0,38	0,39
Aş	0,96	0,33	Kuram	0,31	0,45	İzah	0,57	0,60	Kıvranmak	0,77	0,56
Ölçü	0,65	0,31	Tanık	0,83	0,56	Buçuk	0,84	0,41	Yadırgamak	0,66	0,57
Kişilik	0,93	0,40	Kitle	0,77	0,57	Esnek	0,38	0,35	Yeğlemek	0,52	0,57
Nitelik	0,72	0,54	Konut	0,83	0,54	Nesnel	0,33	0,33	Zedelemek	0,69	0,66
Kamu	0,52	0,36	Sergi	0,82	0,43	Asgari	0,50	0,36	İlişmek	0,58	0,51
Yatırım	0,91	0,36	Kategori	0,67	0,45	Meşru	0,23	0,34	Yadsımak	0,55	0,53
Yaklaşım	0,48	0,42	Yazın	0,60	0,51	Sanal	0,42	0,30	Korsan	0,78	0,59
Kur	0,53	0,49	Ufuk	0,67	0,62	Reel	0,31	0,41	Esin	0,37	0,32
Boyut	0,89	0,35	Yazık	0,79	0,46	İlkel	0,56	0,58	İleti	0,64	0,49
Basın	0,74	0,40	Olasılık	0,89	0,55	Sözde	0,61	0,52	Us	0,21	0,32
Tepki	0,93	0,35	Rejim	0,62	0,50	Sinmek	0,47	0,45	Ödün	0,37	0,50
Gündem	0,56	0,39	Dinamik	0,71	0,51	Paralamak	0,57	0,53	Islah	0,20	0,33
Uğraş	0,68	0,43	Sıradan	0,76	0,53	Türemek	0,66	0,56	Tedariki	0,43	0,58
Yetki	0,82	0,50	Standart	0,70	0,54	Bürümek	0,58	0,50	Sitem	0,42	0,48
Yorum	0,88	0,39	Alternatif	0,80	0,54	Donatmak	0,69	0,50	Ezgi	0,55	0,59
Yerel	0,86	0,45	Yüce	0,77	0,56	Bezmek	0,68	0,51	Gayrimenkul	0,34	0,33
Boz	0,71	0,39	Gözde	0,76	0,53	Yeti	0,57	0,58	Kota	0,46	0,52
Modern	0,86	0,48	Anımsamak	0,80	0,52	Trend	0,46	0,37	Tescil	0,36	0,44
Yanıt	0,92	0,37	Yığmak	0,71	0,57	Figür	0,75	0,65	Özveri	0,24	0,40
Aydın	0,54	0,57	Sakınmak	0,87	0,56	Zat	0,29	0,42	Dizge	0,42	0,40
Öneri	0,81	0,45	Arınmak	0,89	0,51	Belde	0,52	0,49	Tabu	0,24	0,33
Tabii	0,50	0,46	Özenmek	0,76	0,50	Söylenti	0,75	0,63	Azami	0,45	0,39
Resmi	0,49	0,32	Dizmek	0,89	0,55	Sezgi	0,66	0,69	Muhit	0,26	0,38
Asıl	0,60	0,44	Sarsmak	0,74	0,48	Profil	0,75	0,59	Ekol	0,21	0,34
Karmak	0,79	0,43	Derlemek	0,74	0,62	Rezerv	0,58	0,56	Temenni	0,37	0,54
Sağlamak	0,79	0,48	Sızmak	0,85	0,57	Komut	0,71	0,55	Sağduyu	0,35	0,49
Dönüşmek	0,88	0,53	Kanaat	0,46	0,40	Yaban	0,56	0,57	Kıvılcım	0,39	0,47
Dinmek	0,79	0,51	Taslak	0,74	0,61	Önerge	0,28	0,32	Katalog	0,81	0,41
Eşmek	0,90	0,52	Estetik	0,69	0,54	Sur	0,70	0,71	Sansür	0,45	0,38
Saçmak	0,91	0,48	Coşku	0,86	0,53	Azim	0,57	0,49	Tufan	0,53	0,52
Yaymak	0,89	0,34	Odak	0,56	0,48	Telkin	0,23	0,40	Muhtelif	0,25	0,34
Saptamak	0,45	0,45	Sav	0,33	0,44	Erk	0,22	0,36	Yatkın	0,49	0,52
Atamak	0,68	0,51	Divan	0,61	0,44	Edim	0,21	0,31	Tutucu	0,26	0,35
Adamak	0,63	0,46	Vade	0,52	0,63	Demeç	0,19	0,33	Baki	0,32	0,48
Bulamak	0,70	0,50	Protokol	0,62	0,57	Cevher	0,72	0,59	Beşerî	0,37	0,50
Algılamak	0,69	0,30	Haz	0,76	0,63	İştirak	0,44	0,37	Yetkin	0,48	0,41
Öfke	0,92	0,46	Şerit	0,80	0,68	Rant	0,32	0,41	Kuytu	0,63	0,52
Kıyı	0,93	0,46	Yargıç	0,83	0,57	Fail	0,38	0,41	Harici	0,30	0,34
Yargı	0,86	0,49	Seyir	0,35	0,43	Yalı	0,75	0,63	Rutin	0,66	0,56
Taban	0,92	0,47	Erdem	0,77	0,56	Skandal	0,59	0,59	Bezemek	0,38	0,45
Bütçe	0,93	0,44	Evrim	0,79	0,56	Görkemli	0,68	0,59	Zikretmek	0,39	0,54
Sal	0,89	0,45	Tahsil	0,38	0,54	Engin	0,42	0,46	Hırpalamak	0,69	0,57
Kanıt	0,78	0,38	Söyleşi	0,72	0,60	Soylu	0,69	0,65	Sarfetmek	0,31	0,33
Zirve	0,89	0,46	Taktik	0,55	0,47	Ender	0,58	0,54	Yıpratmak	0,40	0,37
Tasarım	0,77	0,36	Dehşet	0,85	0,52	Kronik	0,27	0,32	Körüklemek	0,50	0,51
Sivil	0,88	0,49	Finans	0,81	0,54	Saygın	0,60	0,62			

Madde ayırtedicilik indeksinin 0.40 ve üzerinde olması maddelerin yüksek işlevselliğe iye olduğunu göstermektedir. Bununla birlikte, 0.30-0.39 aralığında yer alan maddeler de ölçme aracına doğrudan alınabilecek niteliktedir. Ayırtedicilik indeksi 0.20-0.29 arasında olan maddeler ise kuramsal

gerekçelerle ya da kapsam geçerliliği açısından gerekli görülmesi durumunda ölçme aracında tutulabilir; ancak bu maddelerin gözden geçirilmesi ve gerekirse düzeltilmesi önerilmektedir. Yeterince uzun ölçme araçlarında ayırtedicilik indeksinin 0.20'nin altına düşmemesi gerektiği vurgulanmaktadır. Ayrıca ayırt etme gücü kabul edilebilir düzeyde olan bir maddenin güçlük oranının en uygun olarak .50 dolayında olması ve araçta her güçlük düzeyinden maddelere yer verilmesi önerilmektedir (Baykul, 2015; Büyüköztürk, 2020; DeVellis, 2017; Tekin, 2004; Turgut, 1997; Yıldırım, 1983).

Bu ölçütler doğrultusunda Çizelge 2 incelendiğinde, araçta her güçlük düzeyinden sözcüğün yer aldığı görülmektedir. Özellikle ilk 2.000'lik dilimde yer alan bazı sözcüklerin güçlük düzeylerinin düşük olduğu göze çarpmaktadır. Ancak bu aşamada sözcüklerin ölçme aracından çıkarılıp çıkarılmamasına karar verilirken yalnızca istatistiksel ölçütler değil, kuramsal gerekçeler ve sözcüklerin ilişkin oldukları sıklık dilimleri de göz önünde bulundurulmuştur. Nitekim 2×1000 dilimi, günlük yaşamda sözlü ve yazılı ortamlarda karşılaşılan sözcüklerin yaklaşık %80'ini oluşturmakta ve en sık kullanılan sözcükleri kapsamaktadır (Ülper ve Kiraz, 2020; Şahin ve Ülper, 2021; Dağhan ve Ülper, 2022). Bu nedenle bu dilimde yer alan sözcüklerin bilinme olasılığı yüksektir ve görece "kolay" değerler üretmeleri beklenmektedir.

İlerleyen aşamada yapılacak MTK analizlerinde madde özelliklerinin (parametrelerinin) daha ayrıntılı biçimde incelenecek olması nedeniyle, bu aşamada yalnızca güçlük düzeyi düşük olduğu gerekçesiyle sözcük atılmasının gerekli olmadığına karar verilmiştir. Bu koşullarda ölçme aracının betimsel görünümü aşağıdaki gibidir:

Tablo 3
Ölçme Aracının Genel Görünümü

N	Min	Max	$\bar{X}$	SS	KR20
549	14	188	119,5	40,82	,982

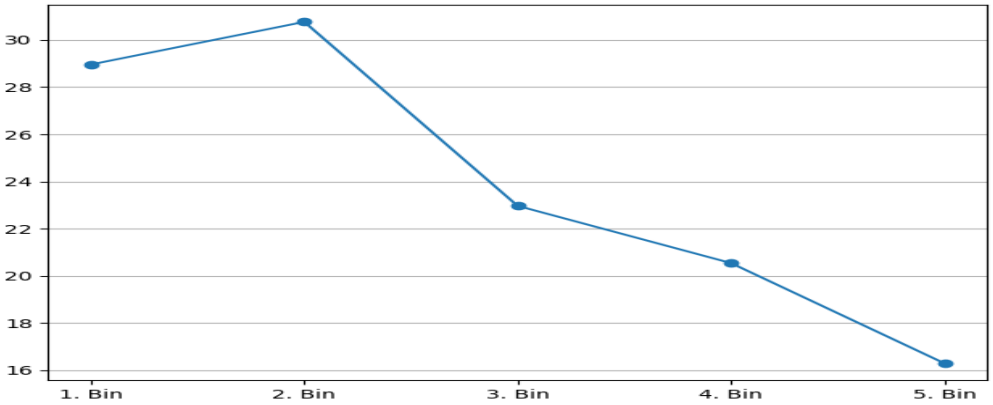
Çizelgeden de anlaşılacağı gibi, toplam 549 geçerli katılımcının ölçme aracına verdiği yanıtlar veri olarak kullanılmıştır. Bu puanların genel dağılımı 14-188 doğru arasında değişmektedir. Ortalaması 119,5 ve standart sapması 40,82 olarak hesaplanmıştır. Bu görünüm araçtan genel olarak orta düzeyde başarı elde edildiğini ve puanların önemli ölçüde ortalama çevresinde dağıldığını göstermektedir. Ancak kimi katılımcıların ortalamanın altında, kimi katılımcıların ise üstünde puan aldığını da belirtmek yerinde olacaktır.

Söz konusu araç ikili puanlanan maddelerden oluştuğu için iç tutarlılığa dayalı güvenirlik durumunu saptamak amacıyla Kuder-Richardson 20 (KR20) katsayısına bakılmış ve bu değer ,982 olarak bulunmuştur. Alanyazındaki (DeVellis, 2017; George ve Mallery, 2003) bilgilere göre ,90 ve üzerindeki değerler çok yüksek olarak anlamlandırıldığı için SASÖ'nün çok yüksek derecede iç tutarlılığa iye olduğu, ölçülmek istenen özelliği tutarlı bir biçimde ölçtüğü ve güvenirliğinin çok yüksek olduğu belirtilebilir. Ayrıca tetrakorik korelasyon matrisi üzerinden McDonald's omega katsayısı hesaplanmıştır. Omega toplam (ω_t) değeri .99 olarak elde edilmiştir. Omega sonuçları da ölçeğin yüksek güvenirliğe iye olduğunu desteklemektedir.

Yapı Geçerliliği Bulguları

Alanyazında yapı geçerliliği hakkında bilgi elde etmek için kuramsal olarak ortaya konmuş olan bir yapının ölçme aracından elde edilen puanlar aracılığıyla sınanabileceği belirtilmektedir (Cronbach ve Meehl, 1955). Bu bağlamda sözcük bilme bakımından ilgili araştırmalara dayanarak sıklık dilimleri ve sınıf düzeyleri arasında var olduğu belirtilen ayrımların bu ölçme aracı puanlarıyla da doğrulanması yapı geçerliliği açısından önemli bir kanıt olarak görülebilir. Sık kullanılan sözcüklerle daha çok karşılaşıldığı için kişilerin bu sözcükleri daha az karşılaşılan sözcüklere göre daha iyi bilmeleri dolayısıyla sıklık düzeyi yüksek sözcüklerin daha kolay, düşük sözcüklerin ise daha zor olması gerekir, biçiminde bir sav ileri sürmek usa yatkın görünmektedir. Bu sav düşük sıklık düzeylerindeki sözcükler için daha az geçerli gibi görünse de araştırmalarla da desteklenmektedir (Beglar, 2010; McLean, Kramer ve Beglar, 2015; Milton, 2009; Nation, 1990). Bu bağlamda sıklık dilimlerinin karşılaştırılması sonucu elde edilecek olan verilerin yapı geçerliği açısından güçlü bir kanıt oluşturacağı ve dolayısıyla maddelerin ve ölçme

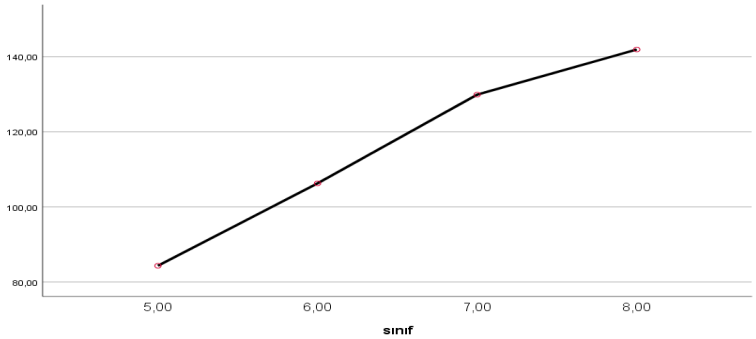
aracının uygunluğuna gönderimde bulunacağı açıktır. Şekil 1'de binlik dilimlere göre ortalama doğru puanları ve bu puanların sıklık dilimlerine göre akışı yer almaktadır.



Şekil 1. Sözcük Sıklığına Göre Ortalama Doğru Sayılarının Akışı

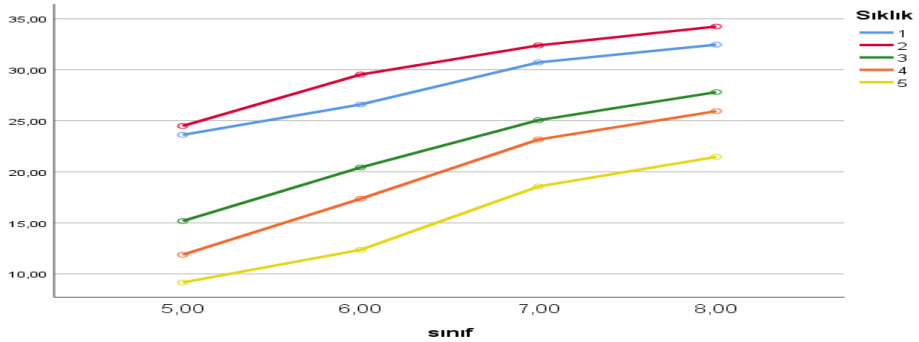
Birinci binlik dilimdeki sözcüklerin doğru yanıtlanma ortalaması 28.96; ikinci binlik 30.77; üçüncü binlik 22,96; dördüncü binlik 20.53 ve beşinci binlik 16.28'dir. Bu görünüme göre genellikle daha sık karşılaşılan sözcüklerin doğru yanıtlanma ortalamaları daha yüksektir ve alanyazınla uyumlu bir biçimde bu yönde ikinci binlik dilim dışında doğrusal bir ilerleyiş vardır. Bu aşamada ortalama puanlar arasındaki bu ayrımın anlamlı olup olmadığına belirlemek amacıyla "tekrarlı ölçümler için tek yönlü ANOVA" yapılmıştır. Anova sürecinde Mauchly's Küresellik Testi anlamlı ($\chi^2(9) = 321{,}041$, $p <$,001) çıktığı için Greenhouse-Geisser testi sonucuna bakılmıştır. Buna göre binlik dilimler arasında anlamlı bir ayrım [$F(3{,}02, 1655{,}41) = 994{,}80$, $p <$,001, $\eta^2_p = {,}645$] olduğu ve etki büyüklüğünün çok yüksek olduğu görülmüştür. Bu ayrımın hangi dilimler arasında olduğunu belirlemek için Bonferroni düzeltmesiyle ikili karşılaştırmalar yapılmıştır. Bunun sonucunda tüm binlik dilimlerin birbirinden anlamlı ($p <$,001) biçimde ayrı olduğu bulgulanmıştır.

Yaşla ve sınıf düzeyinin artmasıyla birlikte karşılaşılan sözcük sayısının artmasına bağlı olarak ileri sınıftakilerin düşük sınıftakilere göre daha çok sözcük bilmeleri olası bir durumdur. Bu durum kimi araştırmalarla (bkz. Keuleers vd., 2015) da ortaya konmuştur. Bu bağlamda sınıf düzeyleri arasında her bir binlik dilim bağlamında bir ayrım olup olmadığının da belirlenmesinin verilerin yapı geçerliği açısından güçlü bir kanıt oluşturacağı ve dolayısıyla maddelerin uygunluğuna gönderimde bulunacağı biçiminde yorumlanmıştır.



Şekil 2. Sınıf Düzeylerine Göre Genel Ortalama Doğru Sayılarının Akışı

Şekil 2'den anlaşılacağı gibi, sınıf düzeyleri 5'ten 8'e doğru gittikçe öğrencilerin bildiği sözcük sayıları ortalaması da düzenli (84.29; 106.25; 129.88; 141.90) bir biçimde artmaktadır. Buna göre sınıf düzeyleri bilinen sözcük sayıları için önemli bir belirleyici konumundadır.



Şekil 3. Sınıf Düzeylerine Göre Sıklık Bağlamında Ortalama Doğru Sayılarının Akışı

Bunun yanında Şekil 3'ten de anlaşılacağı gibi, her binlik dilimde sınıflara göre de belirgin bir artış göze çarpmaktadır. Birinci binlik dilimde 5. sınıftan 8. sınıfa doğru ortalamalar 23,61-32.45 arasında; ikinci binlik dilimde 24,48-34,25 arasında; üçüncü binlik dilimde 15.17-27.80 arasında; dördüncü binlik dilimde 11.86-25.94 arasında; beşinci binlik dilimde 9.14-21.46 arasında değişmektedir. Bu ayrımların istatistiksel olarak anlamlı olup olmadığını belirlemek için "Tek Yönlü ANOVA" yapılmıştır. Ancak ANOVA işlemleri sürecinde varyansların eşit olmadığı LEVENE testi ile saptandığı için grup ortalamaları arasındaki ayrımların anlamlılığı Welch ANOVA ile analiz edilmiştir. Elde edilen sonuçlara göre anlamlı [Welch F(3, 277,39) = 72,46, p < ,001] bir ayrım bulunmuştur. Aynı biçimde bu ayrım 5x1000'lik dilimlerin tümünde de gözlemlenmiştir (p < ,001). Ancak henüz hangileri arasında bir ayrım olduğu belli değildir. Bunu ortaya koymak amacıyla yapılan Games-Howell çoklu karşılaştırma testine göre genel toplamda tüm sınıf düzeyleri arasında anlamlı bir ayrım vardır (p < ,05). Yine 5x1000'lik dilimlerin tümünde de sınıf düzeyleri arasında anlamlı ayrım olduğu görülmektedir. Burada tek istisna olarak karşımıza 1000 ve 2000 düzeyinde 7 ve 8. sınıflar arasındaki ayrımın anlamsız oluşu çıkmaktadır.

Yine taslak ölçme aracının alfa değerlerinin yüksek (,982) çıkması ve bu değerlerin madde silinse de aynı kalıyor görünmesi veri niteliğinin iyi olduğunu yani maddeler arasında tutarlılık olduğunu, madde çıkartmaya gerek olmadığını ve ölçme aracının gelişigüzel yanıtlanmadığını göstermektedir. MTK öncesinde KTK'ye dayanarak ortaya konan bu geçerlik ve güvenirlikle ilgili bilgiler ön denetim açısından yerinde görünmektedir. Tüm bu durumlar şu aşamaya dek SASÖ'nün yapı geçerliğinin uygun olduğuna ve herhangi bir madde atmaya gerek olmadığına ilişkin önemli bir gerekçe oluşturmaktadır. Bu aşamadan sonra yapı geçerliliğinin bir diğer yönü olarak dışsal yönüne de bakmak yerinde olacaktır.

Yapı Geçerliliğinin Dışsal Yönü

Yapı geçerliğinin dışsal yönü için Messick (1989; 1994) ölçme aracının diğer araçlarla olan ilişkisinin önemine gönderimde bulunur. Ancak Türkçe eğitimi alanında anadili olarak doğrudan alıcı sözcük dağarcığını ölçmek amacıyla geliştirilmiş bir ölçme aracı olmadığı için dışsal görünümü belirlemek amacıyla Çetinkaya vd. (2023) tarafından geliştirilen alıcı eşdizimlik testi ile olan ilişkisine bakılmıştır. İki ölçme aracının da alıcı becerileri ölçmesi ve eşdizimlik testinin sözcük boyutu yanında belirli ölçüde anlam boyutuyla da ilgili olması nedeniyle dışsal görünüm için şu an kullanılabilecek en uygun araç durumunda olduğu açıktır. Yine elde edilen alıcı sözcük dağarcığı puanlarının işlev açısından ne anlama geldiğini yorumlayabilmek için öğrencilerin okuma becerileri ile olan ilişkisine de bakılmıştır. Schmitt, Nation ve Kremmel (2020) sözcük dağarcığı ölçme aracı performansının dinleme, konuşma, okuma veya yazma gibi bir tür dil kullanımıyla ilişkili olması gerektiğini ileri sürer. Bu bağlamda alıcı sözcük dağarcığını ölçmeyi amaçlayan SASÖ'nün okuma becerisiyle ilişkilendirilmesi anlamlı görünmektedir. Alanyazındaki ilgili çalışmalar (Perfetti, 2007; Qian, 2002) da okuma ve sözcük dağarcığı arasındaki ilişkiyi açıkça ortaya koymaktadır. Bunun için Ülper vd. (2017) tarafından geliştirilen okuma anlama testi kullanılmıştır.

Yapılan Pearson momentler çarpımı korelasyon analizi sonucunda, eşdizimlik ölçümü toplam puanları ile sözcük dağarcığı ölçümü toplam puanları arasında orta düzeyde ve pozitif yönlü, istatistiksel olarak anlamlı bir ilişki olduğu belirlenmiştir (r = .47, p < .001, N = 344); benzer biçimde okuma anlama ölçümü toplam puanları ile sözcük dağarcığı ölçümü toplam puanları arasında da orta düzeyde, pozitif yönlü ve anlamlı bir ilişki saptanmıştır (r = .46, p < .001, N = 344). Bu bulgular hem eşdizimlik düzeyi hem de okuma başarısı arttıkça sözcük dağarcığı düzeyinin de artma eğiliminde olduğunu göstermektedir. Bu veriler dışsal geçerlilik açısından önemli kanıtlar olarak değerlendirilmiştir.

Katılımcılarla Görüşme

Bazı durumlarda öğrencilerin, bilgi düzeylerinden bağımsız olarak, bazı soruları doğru yanıtlama olasılığı bulunmaktadır. Bu durumda aracın ürettiği puanların ne denli gerçekçi olduğunu anlayabilmek için kestirimde bulunarak verilen doğru yanıtların ve dikkatsizlik sonucu verilen yanlış yanıtların belirlenmesi gerekir. Schmitt (1999) TOEFL ile ilgili çalışmasında görüşmeler yaparak katılımcıların sözcük dağarcığı ölçme araçlarında gerçekte ne bildiklerinin belirlenmesinin maddelerin geçerliği açısından doğrudan kanıtlar sunması bakımından önemli olduğunu vurgular. Bu durumların saptanması için öğrencilerle görüşmeler yapılmıştır. Ayrı sınıf ve sosyo-ekonomik düzeylerdeki toplam 50 katılımcı görüşmelerde yer almıştır. Görüşmeler uygulamadan bir gün sonra araştırmacı tarafından eğitilen 5 kişilik bir ekip ile yapılmıştır. Her kişi on öğrenci ile görüşme yapmıştır. Görüşme sürecinde her bir öğrenciye her binlik dilimden ad, eylem ve sıfat türlerinden gelişigüzel seçilen 9 sözcük, toplamda ise 45 sözcük olacak biçimde soru sorulmuştur. Öğrencilerden verilen sözcüğün anlamını açıklamaları istenmiştir. Öğrencilerin verdikleri yanıtlar SASÖ'deki açıklamalarla birebir örtüşmese bile bir biçimde doğru açıklama yapılmış ise biliyor olarak (1 puan), yanlış açıklama yapılmış ise bilmiyor olarak (0 puan) belirlenmiş ve puanlanmıştır. Sonuçlara bakılınca yazılı uygulamada bilmese de görüşmede sözcüğün anlamını bilen bir öğrenci olmamıştır buna karşın yazılıda ve sözlüde aynı sayıda sözcüğü bilen öğrenci oranı %96'dir. Diğerleri (%4) görüşmede yazılıda bildikleri sözcükten 1 ya da 2 sözcük daha az bilmişler diğer bir deyişle 2 öğrenci yazılı uygulamada 1-2 sözcüğü kestirimde bulunarak ya da şans eseri doğru yanıtlamış denebilir. SASÖ'deki toplam madde sayısına uyarlayınca 4-8 arasında sözcüğün bu %4'lük kesim tarafından kesin olarak bilinmediği ve kestirimde bulunulmuş olabileceği düşünülebilir. Bu oran SASÖ'deki sözcüklerin %5'inin altındadır ve kabul edilebilir bir düzeydedir. Bu durum SASÖ'nün işlevine uygun bir biçimde katılımcıların sözcük dağarcığına ilişkin kabuledilebilir bir ölçüm yaptığını göstermektedir. Katılımcılara yanıtlarken kestirimde bulunup bulunmadıkları sorulduğunda çok büyük bir çoğunluğu kestirimde bulunmadıklarını, bilmedikleri sözcükle karşılaştıklarında boş bıraktıklarını belirtmişlerdir. SASÖ'nün uygulanmasından hemen önce öğrencilere yapılan açıklamalarda bu yönde bir uyarı yapılmış olması da bunda önemli bir etken olarak düşünülebilir.

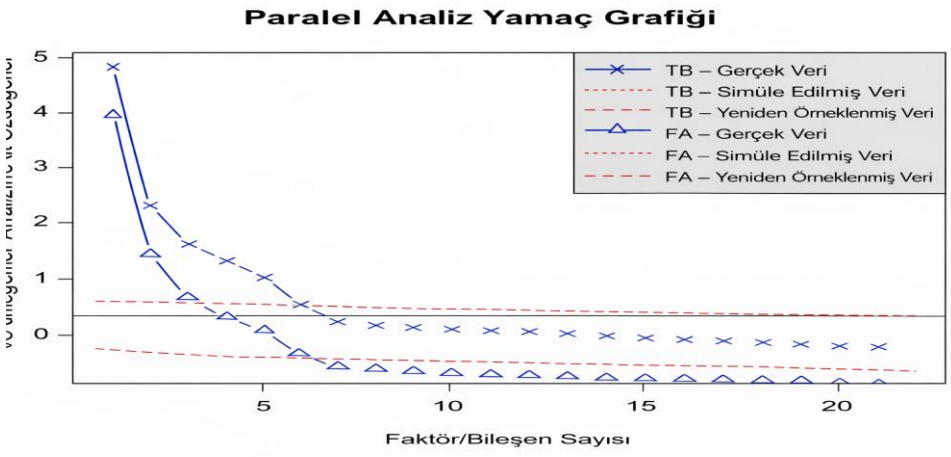
Madde Tepki Kuramı Varsayımlarına İlişkin Bulgular

Alıcı sözcük dağarcığını ölçmek amacıyla tasarlanan SASÖ, 1-0 olarak puanlanan bir araçtır. Yukarıda da ayrıntılı biçimde belirtildiği gibi, KTK bağlamında yapılan madde ayırtedicilik analizleri maddelerin bu bakımdan birbirinden ayrımlaştığını göstermektedir. Bu nedenle tüm maddelerin eşit ayırtedicilik değerine iye olduğu varsayımına dayanan Rasch modeli (DeMars, 2016) yeğlenmemiş bunun yerine madde güçlüğü ve ayırtediciliğinin ayrı ayrı kestirilmesine olanak sağlayan (DeMars, 2016) bir model olduğu için iki parametreli lojistik model (2PL) yeğlenmiştir.

Açımlayıcı Faktör Analizi

Ölçme aracının yapısal açıdan nasıl boyutlandığını belirleyerek 2PL'nin uygun olup olmadığına karar verebilmek için gerekli adımlardan birini atmak amacıyla açımlayıcı faktör (AFA) analizi yapılmıştır (Hambleton vd., 1991). Geliştirilmekte olan ölçme aracı öğrencilerin alıcı sözcük bilgisini ölçmeye dönük bir araçtır. Bu bakımdan alıcı sözcük bilgisi gibi adlandırılabilecek tek bir yapıdan oluşması gerektiği varsayılmaktadır. Bu bağlamda Webb vd. (2017) ölçme aracının tek boyutluluğu yakalamasını istenen bir durum olarak belirtir. Çünkü bu durum amaçlanan yapının yansıtılmasını sağlar.

Alanyazında MTK'nin tek boyutlu bir örtük özelliğe dayandığı kabul edilmektedir (Embretson ve Reise, 2000; Hambleton vd., 1991). Bu doğrultuda tek boyutluluğun görgül (ampirik) olarak sınanmasında ortak varyansa dayalı faktör modelleri sıklıkla kullanılmaktadır. Ölçme aracı maddeleri ikili (0 = yanlış, 1 = doğru) olarak puanlandığından analizler tetrakorik korelasyon matrisi üzerinden yürütülmüştür. Çünkü dikotom verilerde Pearson korelasyon katsayısı sürekli ve normal dağılım varsayımına dayandığı için ilişkileri kestirme sürecinde sapmalar olabilmektedir (Flora ve Curran, 2004; Holgado-Tello vd., 2010). Faktör sayısının belirlenmesinde paralel analiz yöntemi uygulanmış (Hayton, Allen ve Scarpello, 2004; Horn, 1965) ve analiz sonuçlarına göre 34 faktör saptanmıştır. Bununla birlikte özdeğer dağılımı incelendiğinde birinci faktörün özdeğerinin ($\lambda_1 = 69.51$) diğer faktörlere kıyasla belirgin biçimde baskın olduğu ve ikinci faktörden itibaren keskin bir düşüş gözlendiği görülmüştür. Bu görüntü baskın bir genel faktörü imlemektedir. Güçlü bir birinci faktörün varlığı ve sonraki faktörlerin sınırlı katkı sağlaması tek boyutlu yapının göstergeleri arasında görülmektedir (Reise, 2012; Reise vd., 2000). Ayrıca birinci faktörün toplam varyansın %35.34'ünü açıklaması da genel faktörün baskınlığını destekleyen bir bulgu olarak yorumlanmıştır. Yamaç-birikinti çizgesi (scree plot grafiği) de bu durumu desteklemektedir.



Şekil 4. Paralel Analiz Yamaç Grafiği

Yamaç-birikinti çizgesi incelenince birinci faktörün özdeğerinin simülasyon verisinden belirgin biçimde ayrışmakta olduğu, sonraki faktörlerde bu ayrışmanın büyük ölçüde azalmakta olduğu göze çarpmaktadır. Bu görüntü Cattell'in (1966) tanımladığı kırılma noktası ile tutarlıdır. Özdeğerlerdeki keskin düşüş ve varyans açıklama oranları; birinci faktörün özdeğerinin diğer faktörlere göre baskın olması ve %20'den büyük olması yine baskın faktörün diğer faktöre göre yaklaşık 8 kat büyüklüğünde olması (Reckase, 1979; Hattie, 1985) birlikte değerlendirildiğinde ölçme aracı tek boyutlu olarak yorumlanmıştır. Bu koşullar altında MTK analizlerine geçmek için önemli bir adım atılmıştır. Buna karşın Hambleton, Swaminathan ve Rogers'e (1991) göre baskın bir tek boyutun olması MTK modellerinin uygulamasına geçebilmek için gerekli ama yeterli değildir. Bu nedenle yerel bağımsızlık ve model-veri uyumuna da ayrıca bakılmalıdır.

Yerel Bağımsızlık

Yerel bağımsızlığı incelemek için yaygın olarak kullanılan Q3 istatistiğine göre madde çiftlerinin büyük bir kesiminin olması gereken değer olarak ,20'nin altında olduğu görülmüştür. Bununla birlikte bu değer toplam madde çiftlerinin %1.1'inde Q3 > 0.20; %0,3'ünde ise Q3 > 0.30 olarak görülmüştür. Bu oranlar 18915 olan tüm madde çiftleri düşünülünce çok küçük oranlardır. Bu derece küçük oranlar özellikle madde sayılarının 100'ün üzerinde olması durumunda alanyazında (Chen ve Thissen, 1997; De Ayala, 2009; Yen, 1984) da belirtildiği gibi, yerel bağımsızlığı bozacak nitelikte değildir.

Model Uyumu ve Madde Düzeyinde Uyum

Yukarıda belirtilen yerel bağımsızlık ile ilgili bulgular model uyumu açısından önemli göstergelerden biri olarak değerlendirilebilir. Buna karşın tek başına yeterli görülmemektedir. Bu bakımdan model uyumunu incelemek için diğer yollara da başvurulmuştur. Aşağıda bunlara ilişkin bilgiler yer almaktadır.

Geliştirilen SASÖ 1-0 olarak puanlanan ve çok sayıda maddeden oluşan bir ölçme aracıdır. Bu tür verilerde DFA kökenli CFI, TLI ve klasik X^2 gibi uyum indekslerinin MTK çerçevesinde kuram ve yöntem açısından uygun olmadığı bu nedenle de model uyumunun ayrı yollarla değerlendirilmesi gerektiği alanyazında vurgulanmaktadır (de Ayala, 2009; Embretson ve Reise, 2000). Bu durum bir sorun değil MTK'nin doğasından kaynaklanan bir sınırlılık olarak ele alınmaktadır. Dolayısıyla model uyumunu belirlemek için diğer seçeneklere yönelinmiştir. İlk aşamada böyle bir ölçme aracı için olası modellerden biri olan 1PL ile seçilen model olarak 2PL modeli karşılaştırılmıştır. Tablo 4'te buna ilişkin sonuçlar yer almaktadır.

Tablo 4
1PLve 2PL Karşılaştırması

Model	Deviance (−2LL)	Parametre Sayısı	AIC	BIC	$\Delta\chi^2$	sd
1PL (Rasch)	98,257.71	196	98,649.71	99,494.10	1612.91	194
2PL	96,644.80	390	97,424.80	99,104.95	–	–

Çizelgede açıkça görüldüğü üzere 2PL modeli 1PL modeline göre -2LL, AIC ve BIC değerleri bakımından daha düşük değerlere iyedir. Buradaki ayrımın anlamlılığına ilişkin yapılan karşılaştırmanın sonucu ($\Delta\chi^2 = 1612.91$, sd = 194, $p < .001$) da 2PL modelinin veriyle anlamlı düzeyde daha iyi uyum sağladığını göstermektedir.

Kuramsal açıdan bakılınca 2PL modeli hem güçlük (b) hem de ayırtedicilik (a) parametrelerinin ayrımlaşmasına olanak tanımaktadır. Buna karşın 1PL modeli ise tüm maddelerin ayırt ediciliklerinin eşit olduğu varsayımına dayanmaktadır (DeMars, 2016). Bu bağlamda sözcük bilgisinin bilişsel ve anlamsal yükünün eşit olmadığı gerçeği göz önünde bulundurulunca madde ayırtediciliğinin eşit olmayacağı açıktır. Bu bakımdan da 1PL modeli uygun görülmemiştir. Bu açılardan 2PL modelinin kullanılmasının daha uygun olduğu düşünülmüştür.

Bunların yanında 2PL modelinin uyumunu sınamak için artık temelli göstergelere de bakılmıştır. Bu bağlamda SRMR=.036, SRMSR=.048 ve 100xMADCOV=.678 değerleri elde edilmiştir. Bu değerlerden SRMR'nin .08'in altında olması kabuledilebilir düzeyde bir uyumu, .05'in altında olması ise çok iyi uyumu imlemektedir (Hu ve Bentler, 1999). İkili veriler için önerilen rezidüel temelli uyum ölçütlerine göre SRMSR ve 100xMADCOV göstergeleri için elde edilen bu değerler, artık yapının sınırlı olduğunu ve genel uyumun güçlü olduğunu göstermektedir (Maydeu-Olivares, 2013).

Karşılaştırılması gereken bir diğer model olarak 3PL modeli de değerlendirilmiştir. Ancak bu model yanıtlanma sürecinde kestirimde bulunmanın da etkisi olduğunu varsaymakta ve bu ek parametrenin güvenilir biçimde kestirilebilmesi için görece daha büyük bir örneklem kümesini gerekli kılmaktadır (DeMars, 2016). Buna karşın geliştirmekte olduğumuz ölçme aracı kestirim olasılığı en aza indirgenecek biçimde yapılandırılmıştır. Bu nedenle kestirim parametresinin modele eklenerek yorumlanabilirliğin karmaşık duruma getirilmesi uygun görülmemiştir. Yine örneklem sayısı da 3PL modeli için düşük düzeydedir. Tüm bu kuramsal ve görgül bulgular birlikte değerlendirilince 2PL modeli öne çıkmaktadır.

Maddelerin modele uyumunu saptamak amacıyla kullanılan S-X^2 istatistiği (Orlando ve Thissen, 2000, 2003) sonucunda, 195 maddeden yalnızca 19'unun (%9,7) iki parametreli lojistik (2PL) modelle uyumsuz olduğu, diğer maddelerin ise modele uyum gösterdiği belirlenmiştir. Bununla birlikte, uyumsuz olarak işaretlenen maddelere ilişkin S-X^2, serbestlik derecesi (sd), RMSEA ve p değerleri birlikte incelendiğinde, istatistiksel olarak anlamlı görünen uyumsuzlukların çoğunlukla oldukça düşük RMSEA değerleriyle (0,022–0,035 arası) birlikte ortaya çıktığı göze çarpmaktadır. Alanyazında RMSEA değerlerinin 0,05'in altında olmasının iyi uyumu imlediği kabul edilmektedir (Browne ve Cudeck, 1993;

Kline, 2016). Bu bağlamda, söz konusu maddelerde gözlenen uyumsuzlukların mutlak bir model ihlali olarak değerlendirilmemesi gerektiği söylenebilir. Model uyumuna ilişkin yapılan tüm analizler, kuramsal çerçeve ve yerel bağımsızlıkla ilgili bulgular birlikte ele alındığında, verilerin 2PL modeliyle daha iyi örtüştüğü sonucuna ulaşılmaktadır.

2PL Modeli Analiz Bulguları

Madde Parametreleri, Madde Karakteristik Eğrileri ve Test Bilgi Fonksiyonunun Bütüncül Değerlendirilmesi

İki parametreli lojistik model (2PL) bağlamında ölçme aracının maddelerinin işlevi, ayrı yetenek düzeylerindeki katılımcıları ayırt edebilme gücünü gösteren ayırtedicilik *(a)* ve maddenin göreli zorluk durumunu gösteren güçlük *(b)* parametreleri üzerinden değerlendirilir. *a* parametresi *b* parametresiyle birlikte maddelerin hangi yetenek aralığında ne ölçüde etkili olduğunu ortaya koyar ve bu da ölçme aracının bilgi fonksiyonu yapısını belirler (Embretson ve Reise, 2000; deMars, 2016). Bu bağlamda ölçme aracından elde edilen *a* ve *b* parametrelerine ilişkin değerler alanyazında (deMars, 2016; Baker ve Kim, 2004) yaygın olarak kullanılan düzeylendirmeye uygun olarak aşağıda Çizelge 5'te sunulmaktadır. Ayrıca ekler bölümünde güçlük ve ayırtedicilik bakımından en üst ve en altta yer alan 10'ar maddeye ilişkin çizelge sunulmuştur.

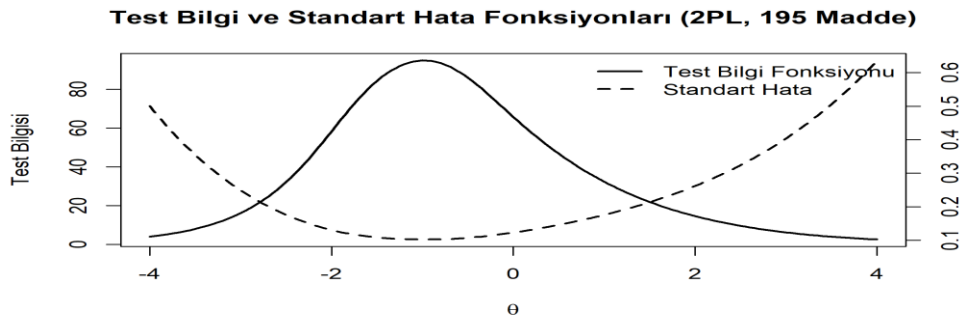
Tablo 5

Madde Ayırtedicilik (a) ve Madde Güçlük (b) Parametrelerinin Dağılımı

Parametre	Aralık	Anlamı	Madde sayısı (n)	Oran (%)
Ayırtedicilik (a)	$a < 0.65$	Düşük	0	0.0
	$0.65 \leq a < 1.35$	Orta	83	42.6
	$1.35 \leq a < 2.00$	Yüksek	75	38.5
	$a \geq 2.00$	Çok yüksek	37	19.0
Güçlük (b)	$b \leq -2.0$	Çok kolay	7	3.6
	$-2.0 < b \leq -1.0$	Kolay	55	28.2
	$-1.0 < b \leq 0$	Orta kolaylıkta	73	37.4
	$0 < b \leq 1.0$	Orta zorlukta	40	20.5
	$b > 1.0$	Zor	20	10.3

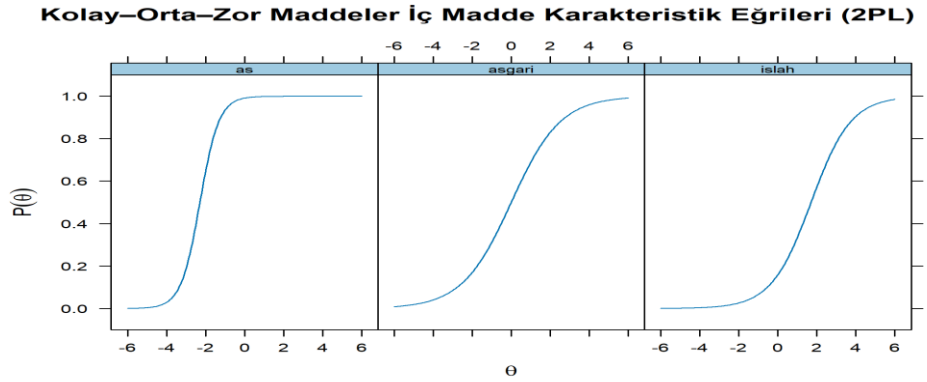
Not. Oranlar toplam madde sayısı (N = 195) üzerinden hesaplanmıştır.

Bu çizelgeye bakılınca maddelerin ayırtedicilik ve güçlük parametreleri bakımından dengeli bir dağılıma iye olduğu göze çarpmaktadır. Düşük ayırtedicilik düzeyinde herhangi bir madde yokken maddelerin büyük bir oranı orta ve yüksek düzey ayırt ediciliğe iyedir. Bu görünüm maddelerin bireyler arasındaki yeterlik ayrımlarını etkili biçimde yapabileceği biçiminde yorumlanabilir. Güçlük parametreleri bakımından bir değerlendirme yapılınca maddelerin genellikle orta kolay ve kolay düzeylerinde yer aldığı görülmektedir. Bu durum SASÖ'nün genellikle orta ve düşük yeterlik düzeyindeki bireylerin ölçümünde daha duyarlı olduğunu imlemektedir. Test bilgi çizgesi de benzer bir duruma gönderimde bulunmaktadır.



Şekil 5. Test Bilgi ve Standart Hata Fonksiyonları

Test bilgi çizgesine bakınca -1.5 ile -1 yetenek aralığında test bilgisinin en yüksek düzeye ulaştığı göze çarpmaktadır. Bununla birlikte -2 ile 0 aralığında genel olarak test bilgisinin yüksek olduğu ve bu aralıkta ölçümün en güvenilir düzeyde olduğu belirtilebilir. Buna karşın θ < −3 ve θ > 2' den sonra bilgi belirgin bir biçimde düşmekte ve buna bağlı olarak da ölçme duyarlığı azalmaktadır. Bu durum ölçme aracının uç yetenek düzeylerinde daha az duyarlı olduğuna gönderimde bulunmaktadır. Ancak bu ölçme aracının uç yetenek düzeylerinde tamamen işlevsiz olduğu anlamına gelmemektedir. Maddelerin genel dağılımına bakılınca orta ve zor maddelerin de araçta yer alması SASÖ'nün aynı süremde üst yetenek düzeylerinde de ölçüm yapabildiğini göstermektedir. Bu durum maddelerin güçlük dağılımlarıyla ilişkilendirilince uyumlu bir görüntü ortaya çıkmaktadır. Madde karakteristik eğrileri de bu açıdan benzer nitelikte bilgiler sunmaktadır.



Şekil 6. Kolay-Orta-Zor Maddeler İç Madde Karakteristik Eğrileri

Yukarıdaki çizgede kolay, orta ve zor maddelere ilişkin birer örnek yer almaktadır. Buna göre kolay maddeye ilişkin eğri yetenek düzleminde sola yani düşük yetenek kesimine daha yakın görünmektedir. Bu durum düşük yetenek düzeylerinde doğru yanıtlanma olasılığının arttığını göstermektedir. Ortadaki eğri ortalama yeterlik düzeyindeki öğrenciler açısından, sağdaki eğri ise daha çok yüksek düzeydeki öğrenciler açısından doğru yanıt olasılığının %50'ye ulaştığını göstermektedir. Dolayısıyla SASÖ ayrı yetenek aralıklarında bilgi üretebilecek niteliktedir. Eğrilerin çok yatık olmaması ve benzer derecede dik olması ayırtedicilik parametresi bakımından birbirine yakın olduklarını göstermektedir. Maddelerin güçlük düzeyleri ayrı olsa da benzer ölçme duyarlılığına iye oldukları görülmektedir. Bu durum yukarıdaki açıklamalarla da örtüşmektedir. Dolayısıyla SASÖ bütüncül olarak değerlendirildiğinde ayırtedicilik bakımından güçlü, ölçüm duyarlılığı bakımından daha çok düşük ve orta yetenek düzeyine yoğunlaşan ancak her iki uç yetenek düzeyinde görece daha sınırlı bilgi sunsa da ölçüm yapabilen bir yapıya iye olduğu söylenebilir.

Yapı Geçerliğinin Genellenebilirliği

SASÖ'den elde edilen puanların genellenebilirliği, kişi (p), madde (i) ve hata (p×i + e) varyans bileşenlerinin kestirilmesine dayalı olarak Genellenebilirlik Kuramı çerçevesinde incelenmiştir. Bu bağlamda kişi varyansı bireyler arasındaki gerçek yetenek ayrımlarını, madde varyansı maddelerin güçlük düzeylerindeki ayrımlıkları, hata varyansı ise kişi-madde etkileşimi ile ölçme sürecine ilişkin rastlantısal hataları temsil etmektedir (Brennan, 2001; Shavelson ve Webb, 1991). Analiz sonuçlarına göre kişi varyansı 0.043, madde varyansı 0.043 ve hata varyansı 0.151 olarak hesaplanmıştır. Hata varyansının görece yüksek olması, bireylerin madde düzeyindeki tepkilerinde belirli bir düzensizlik bulunduğunu göstermektedir ve bu durum ölçme sonuçlarının kararlılığını sınırlayan temel ögelerden biri olarak değerlendirilmektedir (Shavelson ve Webb, 1991). Bununla birlikte, toplam puan düzeyinde hata bileşeni madde sayısına bölünerek hesaba katıldığından, çok sayıda maddeden oluşan araçlarda bu hata önemli ölçüde azalmakta ve ölçümlerin kararlılığı artmaktadır (Brennan, 2001). Bu varyans bileşenlerine dayalı olarak hesaplanan genellenebilirlik katsayısı G = 0.98'dir. Bu değer, SASÖ'den elde edilen toplam puanların son derece kararlı ve yüksek düzeyde genellenebilir olduğunu göstermektedir.

Alanyazında da yüksek madde sayısına iye ölçme araçlarında hata varyansının toplam puan üzerindeki etkisinin azaldığı ve genellenebilirlik katsayılarının yükseldiği vurgulanmaktadır (Brennan, 2001). Bu bulgular, test bilgi fonksiyonu ve standart hata değerleriyle birlikte değerlendirildiğinde, geliştirilen ölçme aracı yüksek ve tutarlı bir ölçme gücüne iye görünmektedir. Sonuç olarak hata varyansı madde düzeyinde görece yüksek görünse de SASÖ geniş madde yapısından dolayı bu hata toplam puan düzeyinde büyük ölçüde ödünlenmekte, elde edilen puanların son derece güvenilir ve genellenebilir olduğu anlaşılmaktadır.

Tartışma ve Sonuç

Bu çalışmada Türkçeyi anadili olarak öğrenen öğrenciler için alıcı sözcük bilgisini ölçmeyi amaçlayan SASÖ'nün psikometrik özellikleri, KTK, MTK ve GK çerçevesinde çok yönlü olarak incelenmiştir. Ölçme aracının 5-8. sınıf öğrencilerine uygulanması sonucunda geçerli 549 veriden elde edilen bulgular, geliştirilen ölçme aracının hem kuramsal hem de istatistiksel açıdan güçlü geçerlik ve güvenirlik kanıtları sunduğunu göstermektedir. Ölçme araçlarının geçerliğinin tek bir kanıt türüyle değil, ayrı kaynaklardan elde edilen bulguların bütüncül biçimde değerlendirilmesiyle temellendirilmesi gerektiği yönündeki çağdaş yaklaşım (Kane, 2013; Messick, 1989) göz önüne alındığında, SASÖ'nün çok katmanlı bir geçerlik yapısına iye olduğu söylenebilir.

Yapı geçerliğinin içerik boyutu, sıklık görünümlerine göre hazırlanan 5000 sözcüklük evreni örnekseyecek biçimde yapılandırılmasıyla güvence altına alınmıştır. Bu evreni örnekseyecek biçimde 195 sözcüğün seçilmesi, ölçülmek istenen yapının kapsamlı ve gerçekçi biçimde örneksenmesini sağlamıştır. Gyllstad, McLean ve Stewart'ın (2021) önerileriyle uyumlu biçimde güven aralıklarını daraltmak ve ölçümlerin kararlılığını artırmak için her binlik dilim başına 39 sözcük seçilmiştir. Bunun yanında ayrıca uzman görüşlerinin alınması, sözcük seçiminde izlenen ilkeler ve alanyazındaki benzer biçimde (McLean vd., 2015; Schmitt, Schmitt ve Clapham, 2001; Webb vd., 2017) çoklu eşleştirme dizgesinin kullanılması içerik uygunluğuna ilişkin güçlü kanıtlar sunmaktadır. Bu yönüyle SASÖ, Nation ve Beglar'ın (2007) geliştirdiği Sözcük Genişliği Testi (Vocabulary Size Test) gibi uluslararası ölçekte yaygın kullanılan araçlarla aynı kuramsal altyapıyı paylaşmakta ve benzer örnekseme gücü sergilemektedir.

KTK bağlamında yapılan madde analizleri sonucunda SASÖ'de yer alan maddelerin hiçbirinin ayırtedicilik bakımından alt sınırın altında olmadığı, tersine büyük çoğunluğunun kabul edilebilir ve yüksek düzeyde ayırtediciliğe iye olduğu belirlenmiştir. Maddelerin güçlük düzeylerinin geniş bir aralığa yayılması, aracın türdeş olmayan (heterojen) yeterlik düzeylerine duyarlı olduğunu göstermektedir. Özellikle 1x1000 ve 2x1000 gibi yüksek sıklıklı dilimlerdeki sözcüklerin görece daha kolay maddeler üretmesi, sözcük sıklığı ile bilinirlik arasındaki ilişkinin alanyazındaki bulgularıyla (Nation, 1990; Milton, 2009; Beglar, 2010) örtüşmektedir. SASÖ'nün KR-20 güvenirlik katsayısının ,982; McDonald's Omega katsayısının ise ,99 gibi çok yüksek bir değer alması, DeVellis (2017) ve Büyüköztürk'ün (2020) belirttiği ölçütler doğrultusunda iç tutarlılığın son derece güçlü olduğunu ve maddelerin büyük ölçüde aynı yapıyı ölçtüğünü ortaya koymaktadır. Bu değerler, alanyazında rapor edilen benzer araçların güvenirlik katsayılarıyla ($\alpha \approx .95$–.98) uyumludur (Beglar, 2010; Nation ve Beglar, 2007).

Yapı geçerliğine ilişkin bulgular, sıklık dilimleri ve sınıf düzeyleri arasında beklenen doğrultuda anlamlı ayrımlar bulunduğunu göstermiştir. Daha sık karşılaşılan sözcüklerin daha yüksek doğruluk oranlarına iye olması ve sınıf düzeyi yükseldikçe bilinen sözcük sayısının düzenli biçimde artması, sözcük gelişiminin aşamalı ve birikimli bir süreç olduğu ve hem yaş hem de eğitim düzeyiyle düzenli biçimde arttığı yönündeki kuramsal beklentilerle uyumludur (Cronbach ve Meehl, 1955; Milton, 2009; Nation, 1990). Bu sonuçlar aynı süremde ayrı dillerden elde edilen bulgularla da (Gyllstad vd., 2021; Nation ve Beglar, 2007) örtüşmektedir. Bu bağlamda SASÖ, Türkçe için geliştirilen ilk kapsamlı alıcı sözcük dağarcığı ölçme aracı olarak, evrensel sözcük edinimi örüntülerinin Türkçe bağlamında da geçerli olduğunu deneysel olarak ortaya koymaktadır.

Dışsal yapı geçerliği açısından sözcük bilgisinin kuramsal olarak ilişkili olan eşdizimlik ve okuma anlama ölçüm sonuçlarıyla olan ilişkiye bakılmış ve orta düzeyde anlamlı ilişkiler bulunmuştur. Bu ilişkiler ölçme aracının diğer dil becerileriyle uyumlu bir görünüm sergilediğini göstermesi bakımından dışsal yapı geçerliği için güçlü kanıtlar sunmaktadır (Perfetti, 2007; Schmitt vd., 2020; Qian, 2002). Bu

ilişki aynı süremde SASÖ'nün yalnızca yüzeysel bir tanıma bilgisini değil, işlevsel alıcı sözcük yeterliğini yansıttığını göstermektedir. Benzer biçimde Perfetti'nin (2007) "Sözcüksel Nitelik Varsayımı (Lexical Quality Hypothesis)" çerçevesinde ortaya koyduğu üzere, sözcük örneksemelerinin niteliği arttıkça okuma süreçleri daha akıcı ve etkili duruma gelmektedir. SASÖ bulguları, bu kuramsal savın Türkçe bağlamında da geçerli olduğunu desteklemektedir

Yazılı uygulamadan hemen sonra katılımcılarla yapılan görüşmeler, SASÖ'den elde edilen puanların kestirimden uzak bir biçimde büyük ölçüde gerçek bilgiye dayandığını göstermektedir. Yazılı ve sözlü ölçümler katılımcıların %96'lık kesiminde birebir örtüşmektedir. Bu oran, SASÖ'nün alıcı sözcük bilgisini amacına ve işlevine uygun biçimde ölçtüğünü ortaya koymaktadır. Böylece alanyazındaki (Schmitt, 2010) çoktan seçmeli sözcük bilgisi ölçme araçlarına yöneltilen kestirim boyutuna ilişkin eleştirilerinin geliştirilen dizge bağlamında büyük ölçüde geçerliliğini yitirdiğini düşündürmektedir.

MTK kapsamında yapılan analizler sonucunda maddelerin büyük çoğunluğunun orta ve yüksek düzeyde ayırtediciliğe iye olması, SASÖ'nün bireyler arasındaki yeterlik ayrımlarını etkili biçimde ortaya koyabildiğini göstermektedir (DeMars, 2016; Embretson ve Reise, 2000). Bu bulgular, Beglar (2010) ile Webb, Sasao ve Ballance'ın (2017) sözcük ölçme araçlarının çoğunlukla tek boyutlu bir yapıyı ölçtüğünü ve ayırtedicilik parametrelerinin yeterlik ayrımlarını ortaya koymada etkili olduğunu gösteren sonuçlarıyla uyumludur. SASÖ'deki maddelerin büyük bölümünün orta ve yüksek ayırtediciliğe iye olması, Türkçe bağlamında geliştirilen bu aracın uluslararası ölçekte kullanılan sözcük ölçme araçlarıyla benzer psikometrik özellikler sergilediğini ortaya koymaktadır.

Test bilgi fonksiyonunun özellikle düşük ve orta yetenek düzeylerinde yoğunlaşması, Nation ve Beglar'ın (2007) SGT için rapor ettikleri bulgularla aynı yöndedir. Söz konusu çalışmada da sözcük ölçme araçlarının çoğunlukla orta düzey yeterlik aralığında daha yüksek bilgi ürettiği, uç düzeylerde ölçme duyarlılığının görece azaldığı belirtilmektedir. Bu durum, sözcük bilgisinin gelişimsel doğasıyla ilişkilidir: Öğrenenlerin büyük çoğunluğu orta bantta kümelenmekte, uç düzeylerde ise hem birey sayısı hem de ölçülebilir varyans sınırlı kalmaktadır. SASÖ için elde edilen bu bulgular, bu evrensel eğilimin Türkçe anadili konuşurları için de geçerli olduğunu göstermektedir. Dolayısıyla geliştirilen aracın hedef kitlesi olan 5-8. sınıf öğrencilerinin büyük çoğunluğu için yüksek ölçme gücüne iye olduğunu ve genellenebilir olduğunu göstermektedir. Bu bağlamda Genellenebilirlik Kuramına göre hesaplanan genellenebilirlik katsayısının G=0.98 olması, araçtan elde edilen puanların ayrı bireyler bağlamında son derece genellenebilir olduğunu ortaya koymaktadır (Brennan, 2001; Shavelson ve Webb, 1991). Bu bulgu, SASÖ gibi yüksek madde sayısına iye ölçme araçlarında hata varyansının etkisinin azaldığını vurgulayan alanyazınla tutarlıdır (Brennan, 2001). SASÖ'nün 195 maddelik yapısı, bu açıdan hem ayrıntılı ölçüm yapabilen hem de sonuçları yüksek düzeyde genellenebilir olan bir araç ortaya konmasına katkı sağlamıştır. Bu özellik, aracın yalnızca araştırma amaçlı değil, geniş ölçekli eğitimsel değerlendirme süreçlerinde de güvenle kullanılabileceğini göstermektedir.

SASÖ'den elde edilen bulgular, alanyazında ayrı diller ve öğrenici kesimleri için geliştirilen ölçme araçlarının sonuçlarıyla büyük ölçüde örtüşmektedir. Özellikle Nation ve Beglar'ın (2007) geliştirdiği "Sözcük Genişliği Testi (Vocabulary Size Test), Beglar'ın (2010) Japon EFL öğrencileriyle yürüttüğü çalışmalar ve McLean ve arkadaşlarının (2015) ayrı yaş kümeleri üzerinde uyguladıkları araçlar, sözcük sıklığı temelli yapılandırılmış ölçme araçlarının yüksek geçerlik ve güvenirlik sunduğunu göstermektedir. Bu çalışmada SASÖ için elde edilen yüksek iç tutarlılık katsayısı (KR-20 = .982) ve güçlü madde ayırtedicilik değerleri, söz konusu araçlarda rapor edilen güvenirlik düzeyleriyle (SGT için $\alpha \approx$ .95–.98) benzerlik göstermektedir. Bu benzerlik, sözcük sıklığına dayalı örnekleme yaklaşımının sözcüksel evreni örneksemede evrensel ölçekte işlevsel olduğunu ortaya koymaktadır.

KTK, MTK ve GK bulguları birlikte değerlendirildiğinde, SASÖ'nün hem klasik hem de modern ölçme kuramları çerçevesinde güçlü psikometrik özelliklere iye olduğu görülmektedir. KTK, aracın genel güvenirliğini ve madde işlevselliğini ortaya koyarken MTK, aracın hangi yetenek aralıklarında daha duyarlı ölçüm yaptığını göstermiş; GK ise elde edilen puanların genellenebilirliğini kanıtlamıştır. Bu çok katmanlı yaklaşım, Kane'in (2013) savunduğu "kanıta dayalı geçerlik savı" anlayışıyla örtüşmekte ve geliştirilen ölçme aracının bilimsel sağlamlığını önemli ölçüde güçlendirmektedir.

SASÖ'den elde edilen puanların yorumlanabilirliğini artırmak amacıyla, MTK çerçevesinde kestirilen yetenek düzeyleri (θ) ile ham puanlar arasında bir eşleme yapılmış ve öğrenciler dört başarı

düzeyinde sınıflandırılmıştır. Buna göre θ < −1 aralığında yer alan öğrenciler "Başarısız" (0-76 puan; %0–39), −1 ≤ θ < 0 aralığında yer alanlar "Gelişmekte" (77-127 puan; %40–65), 0 ≤ θ < 1 aralığında yer alanlar "Başarılı" (128–161 puan; %66–83) ve θ ≥ 1 düzeyinde yer alanlar "İleri" (162-195 puan; %84–100) olarak tanımlanmıştır. Bu sınıflama, ham puanların yalnızca nicel sonuçlar olarak değil, öğrencilerin sözcük yeterliğinin gelişimsel düzeylerine karşılık gelen anlamlı ulamlar olarak yorumlanmasını olanaklı kılmaktadır. Messick'in (1989; 1994) vurguladığı "puan yorumu" boyutu açısından değerlendirildiğinde, bu eşleme aracın yapı geçerliğine ilişkin önemli bir kanıt sunmaktadır.

Sonuç olarak SASÖ'den elde edilen bulgular, İngilizce ve diğer diller için geliştirilen çağdaş sözcük ölçme araçlarından elde edilen sonuçlarla büyük oranda örtüşmektedir. Bu durum, aracın yalnızca yerel bir ölçme aracı olmadığını aynı süremde uluslararası ölçme yaklaşımlarıyla kuramsal ve yöntemsel açıdan uyumlu bir araç olduğunu göstermektedir. SASÖ, anadili Türkçe olan öğrencilerin alıcı sözcük bilgisini geçerli, güvenilir ve kuramsal olarak temellendirilmiş biçimde ölçebilen bir ölçme aracıdır. Araç, özellikle 5-8. sınıf düzeyindeki öğrenciler için yüksek ölçme gücü sunmakta; sözcük öğretimi, dil gelişimi araştırmaları ve eğitimsel değerlendirme süreçlerinde kullanılabilecek güçlü bir araç niteliği taşımaktadır. Bununla birlikte, uç yetenek düzeylerinde ölçme duyarlılığını artırmaya yönelik ek maddeler geliştirilmesi ve ayrı örneklemler üzerinde yeniden sınanması, kapsamını ve genellenebilirliğini daha da güçlendirecektir. Bu yönleriyle SASÖ, Türkçe için alıcı sözcük bilgisini hem toplam düzeyde hem de binlik sıklık dilimleri bağlamında ölçmede alanyazındaki önemli bir boşluğu doldurmakta; Türkçe bağlamında dil gelişimi araştırmaları için karşılaştırılabilir, güvenilir ve geçerli bir ölçme altyapısı sunmaktadır. Aynı süremde yalnızca bireyler arasındaki ayrımları ortaya koyan bir araç değil, öğrencilerin sözcük gelişim düzeylerine ilişkin pedagojik olarak kullanılabilir bilgiler üreten, tanılayıcı ve biçimlendirici değerlendirme, norm belirleme, okuma güçlüklerini tanılama ve hedef belirleme gibi amaçlar için işlevsel bir ölçme aracı olarak değerlendirilebilir.

Araştırma ve Yayın Etiği

Bu çalışmada "Yükseköğretim Kurumları Bilimsel Araştırma ve Yayın Etiği Yönergesi" kapsamında uyulması belirtilen tüm kurallara uyulmuştur. Yönergenin ikinci bölümü olan "Bilimsel Araştırma ve Yayın Etiğine Aykırı Eylemler" başlığı altında belirtilen eylemlerden hiçbiri gerçekleştirilmemiştir.

Etik Kurul İzni

Kurul adı=Mehmet Akif Ersoy Üniversitesi Girişimsel Olmayan Klinik Araştırmalar Etik Kurulu
Karar tarihi= 18.06.2025 Çarşamba
Belge sayı numarası= GO 2025/1722

Yazarların Katkı Oranı

Makale tek yazar tarafından hazırlamıştır.

Yapay Zekâ Kullanım Beyanı

Makale hazırlık sürecinde Madde Tepki Kuramına göre analizleri yapabilmek için R programında kullanılacak olan kodların yazımında yapay zekâ araçlarından yararlanılmıştır.

Çıkar Çatışması

Makale tek yazarlıdır, kişi ya da kurumlarla ilgili herhangi bir çıkar çatışması durumu yoktur.

Kaynaklar

Aksan, Y., Aksan, M., Mersinli, Ü., & Demirhan, U. U. (2017). *A frequency dictionary of Turkish: Core vocabulary for learners*. Routledge.

Anderson, R. C., & Nagy, W. E. (1993). *The vocabulary conundrum* (Technical Report No. 570). Center for the Study of Reading, University of Illinois at Urbana-Champaign.

Baker, F. B., & Kim, S. H. (2004). *Item response theory: Parameter estimation techniques* (2nd ed.). Marcel Dekker.

Baykul, Y. (2015). *Eğitimde ve psikolojide ölçme: Klasik test kuramı ve uygulaması*. Pegem Akademi.

Beglar, D. (2010). A Rasch-based validation of the Vocabulary Size Test. *Language Testing, 27*(1), 101–118. https://doi.org/10.1177/0265532209340194

Brennan, R. L. (2001). *Generalizability theory*. Springer.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136-162). Sage.

Büyüköztürk, Ş. (2020). *Sosyal bilimler için veri analizi el kitabı* (27. bs.). Pegem Akademi.

Cameron, L. (2002). Measuring vocabulary size in English as an additional language. *Language Teaching Research, 6*(2), 145-173. https://doi.org/10.1191/1362168802lr103oa

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research, 1*(2), 245-276.

Chen, W. H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics, 22*(3), 265-289. https://doi.org/10.3102/10769986022003265

Christensen, L. B., Johnson, R. B., & Turner, L. A. (2015). *Araştırma yöntemleri: Desen ve analiz* (A. Alpay, Çev.). Anı.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281-302. https://doi.org/10.1037/h0040957

Çetinkaya, G., Kesici, S., & Polat, B. (2023). Alıcı ve üretici eşdizimlilik bilgisini değerlendirme aracının geliştirilmesi. *Dil Dergisi, 174*(2), 23-44.

Çetinkaya, G., & Uzun, L. (2018). Türkçe ders kitaplarındaki metinlerin okunabilirlik özellikleri. H. Ülper (Ed.) *Türkçe ders kitabı çözümlemeleri* içinde (s. 141-155). Pegem Akademi.

D'Anna, C. A., Zechmeister, E. B., & Hall, J. W. (1991). Toward a meaningful definition of vocabulary size. *Journal of Literacy Research, 23*(1), 109-122. https://doi.org/10.1080/10862969109547729

Dağhan, O., & Ülper, H. (2022). Türkçe ders kitaplarını anlamak için bilinmesi gereken sözcük sayılarının belirlenmesi. *International Journal of Language Academy, 43*, 293-316.

Daller, H., Milton, J., & Treffers-Daller, J. (Eds.). (2007). *Modelling and assessing vocabulary knowledge*. Cambridge University.

Dang, T. N. Y., & Webb, S. (2025). Applications of word lists in second language learning and teaching. *Language Teaching, 58*(3), 291-311. https://doi.org/10.1017/S0261444825000059

De Ayala, R. J. (2009). *The theory and practice of item response theory*. Guilford Press.

DeMars, C. (2016). *Madde tepki kuramı* (H. Kelecioğlu, Çev. Ed.). Nobel Akademi.

DeVellis, R. F. (2017). *Ölçek geliştirme: Kuram ve uygulama* (T. Totan, Çev.). Nobel Akademi.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Lawrence Erlbaum.

Fan, M. (2000). How big is the gap and how to narrow it? An investigation into the active and passive vocabulary knowledge of L2 learners. *RELC Journal, 31*(2), 105-119. https://doi.org/10.1177/003368820003100205

Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods, 9*(4), 466-491. https://doi.org/10.1037/1082-989X.9.4.466

George, D., & Mallery, P. (2003). *SPSS for Windows step by step: A simple guide and reference* (4th ed.). Allyn & Bacon.

Goulden, R., Nation, P., & Read, J. (1990). How large can a receptive vocabulary be? *Applied Linguistics, 11*(4), 341-363. https://doi.org/10.1093/applin/11.4.341

Gyllstad, H., McLean, S., & Stewart, J. (2021). Using confidence intervals to determine adequate item sample sizes for vocabulary tests: An essential but overlooked practice. *Language Testing, 38*(4), 558-579. https://doi.org/10.1177/0265532220979562

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Sage.

Hayton, J. C., Allen, D. G., & Scarpello, V. (2004). Factor retention decisions in exploratory factor analysis: A tutorial on parallel analysis. *Organizational Research Methods, 7*(2), 191-205.

Hattie, J. (1985). Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement, 9*(2), 139-164. https://doi.org/10.1177/014662168500900204

Henriksen, B., Albrechtsen, D., & Haastrup, K. (2004). The relationship between vocabulary size and reading comprehension in the L2. *Angles on the English-Speaking World, 4*, 129-140.

Holgado-Tello, F. P., Chacón-Moscoso, S., Barbero-García, I., & Vila-Abad, E. (2010). Polychoric versus Pearson correlations in exploratory and confirmatory factor analysis of ordinal variables. *Quality & Quantity, 44*(1), 153-166. https://doi.org/10.1007/s11135-008-9190-y

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika, 30*(2), 179–185.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6*(1), 1-55. https://doi.org/10.1080/10705519909540118

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

Kelecioğlu, H., & Göçer Şahin, S. (2014). Geçmişten günümüze geçerlik. *Journal of Measurement and Evaluation in Education and Psychology, 5*(2), 1–11. https://doi.org/10.21031/epod.41706

Keuleers, E., Stevens, M., Mandera, P., & Brysbaert, M. (2015). Word knowledge in the crowd: Measuring vocabulary size and word prevalence in a massive online experiment. *Quarterly Journal of Experimental Psychology, 68*(8), 1665-1692. https://doi.org/10.1080/17470218.2015.1022560

Kline, R. B. (2016). *Principles and practice of structural equation modeling* (4th ed.). Guilford.

Laufer, B. (2005). Focus on form in second language vocabulary learning. *EUROSLA Yearbook, 5*, 223-250. https://doi.org/10.1075/eurosla.5.11lau

Laufer, B., & Goldstein, Z. (2004). Testing vocabulary knowledge: Size, strength and computer adaptiveness. *Language Learning, 54*(3), 399–436. https://doi.org/10.1111/j.0023-8333.2004.00260.x

Laufer, B., & McLean, S. (2016). Loanwords and Vocabulary Size Test scores: A case of different estimates for different L1 learners. *Language Assessment Quarterly, 13*(3), 202-217. https://doi.org/10.1080/15434303.2016.1202041

Maydeu-Olivares, A. (2013). Goodness-of-fit assessment of item response theory models. *Measurement: Interdisciplinary Research and Perspectives, 11*(3), 71-101. https://doi.org/10.1080/15366367.2013.831680

McLean, S., Kramer, B., & Beglar, D. (2015). The creation and validation of a listening vocabulary levels test. *Language Teaching Research, 19*(6), 741-760. https://doi.org/10.1177/1362168814567889

Meara, P. (1990). A note on passive vocabulary. *Second Language Research, 6*(2), 150-54. https://doi.org/10.1177/026765839000600204

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.). Macmillan.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*(2), 13-23. https://doi.org/10.3102/0013189X023002013

Milton, J. (2009). *Measuring second language vocabulary acquisition*. Multilingual Matters.

Milton, J., & Treffers-Daller, J. (2013). Vocabulary size revisited: The link between vocabulary size and academic achievement. *Applied Linguistics Review, 4*, 151-172. https://doi.org/10.1515/applirev-2013-0007

Nation, I. S. P. (1990). *Teaching and learning vocabulary*. Newbury House.

Nation, I. S. P. (2001). *Learning vocabulary in another language*. Cambridge University.

Nation, I. S. P. (2006). How large a vocabulary is needed for reading and listening? *The Canadian Modern Language Review, 63*(1), 59-82. https://doi.org/10.3138/cmlr.63.1.59

Nation, I.S.P. & Beglar, D. (2007) A vocabulary size test. *The Language Teacher*, *31*(7), 9-13.

Orlando, M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement, 24*(1), 50–64. https://doi.org/10.1177/01466216000241003

Orlando, M., & Thissen, D. (2003). Further investigation of the performance of S-X^2: An item fit index for use with dichotomous item response theory models. *Applied Psychological Measurement, 27*(4), 289–298. https://doi.org/10.1177/0146621603027004004

Perfetti, C. A. (2007). Reading ability: Lexical quality to comprehension. *Scientific Studies of Reading, 11*(4), 357-383. https://doi.org/10.1080/10888430701530730

Qi, S., Teng, M. F., & Fu, A. (2024). LexCH: A quick and reliable receptive vocabulary size test for Chinese learners. *Applied Linguistics Review, 15*(2), 643–670. https://doi.org/10.1515/applirev-2022-0006

Qian, D. D. (2002). Investigating the relationship between vocabulary knowledge and academic reading performance: An assessment perspective. *Language Learning, 52*(3), 513-536. https://doi.org/10.1111/1467-9922.00193

Read, J. (2000). *Assessing vocabulary*. Cambridge University.

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics, 4*(3), 207-230. https://doi.org/10.2307/1164671

Reise, S. P. (2012). The rediscovery of bifactor measurement models. *Multivariate Behavioral Research, 47*(5), 667-696.

Reise, S. P., Waller, N. G., & Comrey, A. L. (2000). Factor analysis and scale revision. *Psychological Assessment, 12*(3), 287-297.

Schmitt, N. (1999). Relationship between TOEFL vocabulary items and meaning, association, collocation, and word-class knowledge. *Language Testing, 16*(2), 189-216. https://doi.org/10.1177/026553229901600204

Schmitt, N. (2010). *Researching vocabulary: A vocabulary research manual*. Palgrave Macmillan.

Schmitt, N. (2014). Size and depth of vocabulary knowledge: What the research shows. *Language Learning, 64*(4), 913-951. https://doi.org/10.1111/lang.12077

Schmitt, N., Nation, I. S. P., & Kremmel, B. (2020). Moving the field of vocabulary assessment forward: The need for more rigorous test development and validation. *Language Teaching*, *53*(1), 109-120. https://doi.org/10.1017/S0261444819000326

Schmitt, N., Schmitt, D., & Clapham, C. (2001). Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test. *Language Testing, 18*(1), 55-88. https://doi.org/10.1177/026553220101800103

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Sage.

Şahin, M., & Ülper, H. (2021). Yabancı dil olarak Türkçe öğrenen öğrencilerin günlük yaşamda bilmeleri gereken sözcük sayılarının belirlenmesi. *International Journal of Language Academy, 9*(3), 258-279.

Tekin, H. (2004). *Eğitimde ölçme ve değerlendirme*. Yargı.

Turgut, M. F. (1997). *Eğitimde ölçme ve değerlendirme metotları*. Yargıcı Matbaası.

Ülper, H. (2023). *Sözcük ve öğretimi*. Pegem Akademi.

Ülper, H., & Kiraz, E. (2020). Türkçe öğrenen yabancı öğrenciler açısından gazeteleri okuyabilmek için gereksinim duyulan sözcük sayısının belirlenmesi. *Erzincan Üniversitesi Eğitim Fakültesi Dergisi, 22*(3), 708-722.

Ülper, H., Çetinkaya, G., & Bayat, N. (2017). Okuduğunu anlama testinin geliştirilmesi. *Ahi Evran Üniversitesi Kırşehir Eğitim Fakültesi Dergisi, 18*(1), 175-187.

van Zeeland, H., & Schmitt, N. (2013). Lexical coverage in L1 and L2 listening comprehension: The same or different from reading comprehension? *Applied Linguistics, 34*(4), 457-479. https://doi.org/10.1093/applin/ams074

Webb, S. (2005). Receptive and productive vocabulary learning: The effects of reading and writing on word knowledge. *Studies in Second Language Acquisition, 27*(1), 33-52. https://doi.org/10.1017/S0272263105050023

Webb, S., Sasao, Y., & Ballance, O. (2017). The updated Vocabulary Levels Test: Developing and validating two new forms of the VLT. *ITL-International Journal of Applied Linguistics*, *168*(1), 33-69. https://doi.org/10.1075/itl.168.1.02web

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement, 8*(2), 125-145. https://doi.org/10.1177/014662168400800201

Yıldırım, C. (1983). *Eğitimde ölçme ve değerlendirme (öğretmenler için el kitabı)*. ÖSYM Eğitim.

Zhang, L. J., & Anual, S. B. (2008). The role of vocabulary in reading comprehension: The case of secondary school students learning English in Singapore. *RELC Journal,* 39(1), 51–76. https://doi.org/10.1177/0033688208091140

Extended Abstract

This study introduces a contemporary assessment instrument, the Frequency-Based Receptive Vocabulary Test (FRVT), developed to measure receptive vocabulary knowledge among native speakers of Turkish, and presents empirical evidence of its validity and reliability. The primary aim of the study is to profile vocabulary knowledge among middle school students (Grades 5–8) within frequency-based thousand-word bands and to develop a tool that reliably estimates how many words students know at each frequency level. The absence of instruments comparable to the Vocabulary Levels Test (VLT)—which is widely used in English and grounded in frequency-based principles—for Turkish, particularly among native speakers, makes this study both significant and original in the literature.

Vocabulary knowledge is a multidimensional construct comprising form, meaning, and use (Nation, 2001). However, not all language skills require mastery of all these dimensions. Whereas form–meaning connections are generally sufficient for receptive skills such as reading and listening, productive skills such as writing and speaking additionally require knowledge of use. Empirical research consistently demonstrates that receptive vocabulary is substantially larger than productive vocabulary, with only 16–50% of receptively known words being available for productive use (Laufer, 2005; Fan, 2000). Accordingly, any vocabulary assessment instrument must clearly specify the type of lexical knowledge it targets. FRVT is explicitly designed to assess receptive vocabulary knowledge and is conceptually aligned with reading ability.

The literature documents robust associations between receptive vocabulary size and reading achievement (Perfetti, 2007; Qian, 2002). Successful reading comprehension presupposes knowledge of a large proportion of the words in a text. It is estimated that compelling reading requires approximately 8,000–9,000 words in English and around 14,000 words in Turkish (Nation, 2006; Ülper and Kiraz, 2020). Determining the extent to which learners approach these thresholds is feasible only with reliable measurement instruments. FRVT addresses this need by functioning both diagnostically and developmentally, revealing students' lexical knowledge across successive frequency bands.

Test items were selected from the most up-to-date Turkish frequency dictionary (Aksan et al., 2017). Five-thousand-word bands (1×1000–5×1000) were established, and a total of 195 target words—39 from each band—were included. This number exceeded the minimum recommended for the stable representation of frequency bands (Gyllstad, McLean and Stewart, 2021). Words were sampled proportionally across grammatical categories (nouns, verbs, adjectives); proper nouns, technical terms, slang, multiword expressions, and multiple derivatives from the same root were excluded. Variables known to affect lexical difficulty, such as syllable length and loanword status, were also considered.

FRVT employs a multiple-matching format. Each cluster consists of six target words and three definitions, and students are required to match each definition with the appropriate word. Definitions were constructed primarily using words from the most frequent 2,000-word band, thereby ensuring that performance reflected knowledge of the target items rather than comprehension of the prompts. The test was administered in a paper-and-pencil format and completed within a single class period.

The sample comprised 549 students attending public schools in the central district of Burdur, selected from Grades 5–8 and representing diverse socioeconomic backgrounds. The mean age of the participants was 12.6 years. Data were analyzed using both Classical Test Theory (CTT) and Item Response Theory (IRT). Responses were dichotomously scored (1 = correct, 0 = incorrect).

CTT analyses indicated that items displayed a wide range of difficulty and satisfactory discrimination. Internal consistency was extremely high (KR-20 = .982), indicating that FRVT measured receptive vocabulary with high reliability and coherence. Total scores ranged from 14 to 188, with a mean of 119.5 (SD = 40.82), suggesting adequate variability and strong discriminatory power across proficiency levels.

Construct validity was examined through multiple procedures. First, mean accuracy rates were analyzed across frequency bands. High-frequency words were recognized significantly more accurately than low-frequency words. Repeated-measures ANOVA revealed significant differences among all thousand-word bands, with a huge effect size, thereby confirming the theoretical relationship between lexical frequency and learnability.

Second, grade-level comparisons were conducted. Vocabulary knowledge increased systematically from Grade 5 to Grade 8. Significant differences emerged across grades, both in overall scores and within each frequency band. These findings indicated that the test was developmentally sensitive and could capture age-related growth in vocabulary knowledge.

External aspects of construct validity were examined by correlating FRVT scores with a receptive synonymy test and a reading comprehension test. The resulting correlations were moderate, positive, and statistically significant ($r \approx .46–.47$), indicating that receptive vocabulary knowledge was closely associated with lexical relations and reading achievement and that FRVT scores possessed clear functional meaning.

Prior to IRT analyses, the assumption of unidimensionality was evaluated. Exploratory factor analysis revealed a dominant general factor, with the first eigenvalue approximately five times larger than the second and the majority of items exhibiting factor loadings above .40. Local independence diagnostics further supported the suitability of the data for IRT modelling. Consequently, a two-parameter logistic (2PL) model was applied. Model comparisons demonstrated that the 2PL model fit the data significantly better than the Rasch model. Global fit indices (SRMR, SRMSR, 100×MADCOV) indicated excellent model–data fit. The vast majority of items conformed well to the model, with only a small subset displaying limited misfit.

Qualitative interviews further indicated students engaged in minimal guessing and written test performance closely corresponded to actual lexical knowledge. Agreement between written and oral assessments reached 96%, supporting the interpretation that FRVT scores accurately reflected receptive vocabulary knowledge.

In sum, FRVT constituted a contemporary, frequency-based instrument for assessing receptive vocabulary in Turkish, characterized by strong validity and reliability. It enabled the profiling of students' lexical knowledge both globally and across frequency bands and demonstrated meaningful associations with reading ability. The test has substantial potential to establish national norms, guide vocabulary instruction, diagnose reading difficulties, and monitor lexical development. As such, it offers a robust methodological contribution to research on vocabulary assessment in both first-language education and Turkish-as-a-foreign-language contexts.

Ekler

Ek 1. Güçlük ve Ayırtedicilik Bakımından En Üst ve En Alt 10'ar Madde

Grup	Madde	a (Ayırt edicilik)	b (Güçlük)
En kolay	Aş	2.06	-2.29
	yaymak	1.18	-2.21
	tepki	1.62	-2.20
	yatırım	1.36	-2.18
	boyut	1.26	-2.08
	yanıt	1.67	-2.07
	kişilik	1.88	-2.01
	yat	2.26	-1.95
	yorum	1.37	-1.90
	bütçe	2.63	-1.81
En zor	ıslah	0.98	1.70
	edim	0.93	1.67
	demeç	1.07	1.64
	ekol	1.01	1.59
	us	1.01	1.57
	meşru	0.97	1.45
	sentez	0.70	1.44
	tabu	0.92	1.44
	erk	1.10	1.39
	kronik	0.80	1.38