Atletik Performans Ölçümlerinde Test–Tekrar Test Güvenirliği Analizleri

Süleyman ULUPINAR*

Öz

Sporcuların performansını doğru şekilde ölçebilmek, değerlendirebilmek ve gelişim süreçlerini takip edebilmek için yapılan ölçümlerin güvenilir olması en önemli ön kabullerden biridir. Güvenirlik ölçüsü olarak literatürde çoğunlukla sınıf-içi korelasyon katsayısı rapor edilse de bunun tek başına yeterli olmadığı konusundaki görüşler güncel literatüre hakim olmaya başlamıştır. Ayrıca sınıf içi korelasyon katsayısının çok çeşitli versiyonlar içermesi ve çoğunlukla rapor edilirken bu bilgilerin sunulmaması da performans araştırmalarında önemli bir eksiklik olarak değerlendirilmektedir. Son çalışmalar, bilimsel sonuçların aynı zamanda pratiğe de hitap edecek şekilde rapor edilmesinin gerekliliğine vurgu yapmaktadır. Bu çalışmanın amacı ölçümlerin standart hatası, tipik hata ve tespit edilebilir minimal değişim gibi pratik temelli hesaplamaları tanıtmak ve kullanımını yaygınlaştırmaya katkıda bulunmaktır.

Anahtar Kelimler: Performans Testleri**,** Mutlak Güvenirlik, Relatif Güvenirlik, Ölçümlerin Standart Hatası, Tipik Hata, Tespit Edilebilir Minimal Değişim

The Analyzes of Test-Retest Reliability in Athletic Performance Measurements

Abstract

A reliable measurement result is one of the most important presuppositions in investigations applied to measure and evaluate the current performance of athletes and monitor their performance developments. Although the intra-class correlation coefficient is mostly reported in the literature as a reliability measure, current opinions have come to dominate that this score alone is not sufficient. In addition, the intra-class correlation coefficient includes a wide variety of versions, an important deficiency is that the information regarding the version has been not mostly reported. Novel studies have emphasized the need to report the results of scientific studies in a way that also can be used in field practices. Therefore, the purpose of this study is to introduce practical-based calculations such as standard error of measurements, typical error, and minimal detectable change, and to contribute to common use.

Derleme Makale (Review Article)
Geliş/Received: 13.10.2020
Kabul/Accepted: 02.12.2020
DOI: https://dx.doi.org/10.17336/igusbd.809612

* Doç. Dr., Erzurum Teknik Üniversitesi, Spor Bilimleri Fakültesi, Erzurum, Türkiye.
E-posta: slymnulpnr@gmail.com ORCID https://orcid.org/0000-0002-9466-5278.

Keywords: Performance Tests, Absolute Reliability, Relative Reliability, Standard Error of Measurements, Typical Error, Minimal Detectable Change

1. Giriş

Güvenirlik, kısaca bir ölçüm ya da testin tekrarları arasındaki tutarlılığın derecesini ifade etmek için kullanılmaktadır (Bruton, Conway, ve Holgate, 2000; Hopkins, 2000; Weir, 2005). Bir testin güvenilir olması farklı zamanlarda yapılan ölçümlerde katılımcıların puanlarının veya aldıkları puanlara göre grup içerisindeki sıralamasının uyumlu olma derecesine bağlıdır (Bruton vd., 2000; Wilkinson vd., 2019). Güvenirlik, geçerlilik için vazgeçilmez bir ön şarttır, dolayısıyla bir testin geçerli olabilmesi için güvenilir olması gerekmektedir (Bruton vd., 2000; Ferreira da Silva Santos, Lopes-Silva, Loturco, ve Franchini, 2020; Wilkinson vd., 2019). Güvenilir bir cihaz ve yöntem ile yapılan bir performans ölçümünün uygun dinlenme süreleriyle arka arkaya yapılan tüm ölçümlerinde sonuçların birbirine oldukça yakın olması beklenir. Çünkü tam olarak hatadan arınık ölçümler yapabilmenin imkansız olduğu kabul edildiği için pratikte sonuçların tüm ölçümlerde aynı olmasının mümkün olmadığı varsayılır. Bu sebeple aynı kişilerde, aynı özelliğin aynı şekilde ölçülmesi durumunda bile ölçümler arasında bir miktar değişim kaçınılmazdır (Bruton vd., 2000; Ferreira da Silva Santos vd., 2020; Hopkins, 2000; Wilkinson vd., 2019).

Spor bilimlerinde bir performans ölçümünün farklı zamanlarda tekrarlanmasını içeren çalışmalarda kullanılan güvenirlik yöntemi gözlemci-içi (değerlendirici-içi) veya gözlemciler-arası (değerlendiriciler-arası) uyumdur (Balsalobre-Fernández, Glaister, ve Lockey, 2015; do Nascimento vd., 2017; Haynes, Bishop, Antrobus, ve Brazier, 2019; Özbay, Ulupınar, Çınar, ve Akbulut, 2019). Test-tekrar test güvenirliğini inceleyen çalışmalar aynı deneklerden farklı zamanlarda alınan tekrarlı ölçümler arasındaki uyumu incelemektedir. Performans ölçümlerinde güvenilirliği belirlemek için test-tekrar test analizlerini kullanan deneysel araştırma tasarımları, genellikle aynı araştırmacının aynı ekipmanı kullanarak aynı grup üzerinde farklı zamanlarda gerçekleştirdiği ölçümlerden oluşur (do Nascimento vd., 2017; Özbay ve Ulupınar, 2019; Özbay vd., 2019; Segura-Ortí ve Martínez-Olmos, 2011). Böylece ölçümler arasındaki farklar ölçüm aleti, ölçüm yöntemi ya da ölçümü yapan kişiden ziyade katılımcılar ile ilişkili hata kaynaklarına işaret etmektedir.

Güvenirlik analizlerinin yapıldığı çalışmaların tasarımları, hataya sebep olduğu düşünülen olası hata kaynağına göre değişkenlik gösterebilir (Balsalobre-Fernández vd., 2015; Ferreira da Silva Santos vd., 2020; Özbay vd., 2019; Wilkinson vd., 2019). Örneğin, güvenirliği henüz kanıtlanmamış elektronik bir baskül ile aynı kişilerin vücut ağırlıklarının bir kaç dakikalık aralıklar ile ikişer kez ölçüldüğünü varsayalım. Birkaç dakikada kişilerin vücut ağırlığı değişemeyeceği için katılımcılardan ve ölçüme doğrudan bir etkisi olmadığı için ölçümcüden (gözlemci) kaynaklanan bir hata olmadığı düşünülür. Dolayısıyla ölçümler arasındaki uyum derecesi ölçüm yapılan cihazın güvenilirliğini temsil eder. Benzer şekilde bir stadiometre ile kişilerin boy uzunluğu birkaç gün arayla ölçülürse buradaki olası hata kaynağının katılımcılar, ölçüm aleti veya ölçüm yönteminden ziyade ölçümü yapan kişi olduğu düşünülür. Son olarak ağırlık plakalarıyla koparma derecesi ölçülen haltercilerin iki farklı zamandaki ölçümleri arasındaki hatalar, sporcuların kendilerinden kaynaklanan sebeplere atfedilir. Çünkü ağırlık plakaları veya ölçümcüden kaynaklı bir hatanın olmadığı, ya da göz ardı edilebilecek kadar küçük miktarlarda olduğu varsayılır (Chiu, Wu, Chou, Yu, ve Hung, 2016; Ferreira da Silva Santos vd., 2020; Wilkinson vd., 2019). Böylece performans ölçümlerini içeren çalışmaların

amacı ve hipotezleri doğrultusunda yapılan güvenirlik analizleri varsayılan hata kaynağına ilişkin sonuçlar üretmektedir.

Güvenirlik, bir ölçümden elde edilen sonuçların içerebileceği ölçüm hatasının hangi düzeyde olduğuna dair önemli bilgiler sunar (Wilkinson vd., 2019). Güvenilir bir test, hassassiyeti yüksek sonuçlar elde edebilmeyi ve pratikte bir müdahalenin sebep olduğu gerçek bir etkiyi tespit edebilmeyi sağlar (Hopkins, 2000). Güvenirlik temelde hem mutlak hem de relatif olarak değerlendirilmektedir (Bruton vd., 2000; Haley ve Fragala-Pinkham, 2006; Weir, 2005). Mutlak güvenirlik katılımcıların ölçüm sonucundaki sayısal değerler arasındaki tutarlılığı ifade ederken, relatif güvenirlik katılımcıların grubun diğer üyelerine göre sıralamaları arasındaki tutarlılığı ifade eder (Weir, 2005). Örneğin bir spor bilimleri parkurunu 40 saniyede tamamlayan ve sıralamada 15. olan bir sporcunun ikinci ölçümü 40 saniyeye çok yakın bir sürede tamamlaması mutlak güvenirliğin iyi olduğunu gösterirken, sıralamasının 15'e yakın olması ise relatif güvenirliğin iyi olduğunu gösterir.

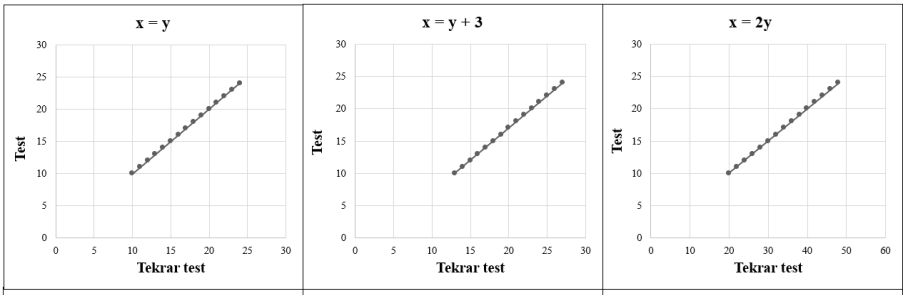
Spor bilimleri alanındaki çalışmalarda kullanılan başlıca güvenirlik analizleri denek-içi (*within-subject*) rastgele varyasyon, ortalamada sistematik değişim ve tekrar test korelasyonudur (do Nascimento vd., 2017; Özbay ve Ulupınar, 2019; Özbay vd., 2019). Eşli örneklem t-testi ve Bland-Altman analizleri mutlak güvenirlik hakkında bilgi sağlamak için kullanılırken, Pearson korelasyon analizi ölçümler arasındaki ilişkinin yönünü ve derecesi hakkında bilgiler sağlayabilmektedir (Koo ve Li, 2016; Wilkinson vd., 2019). Ancak, bu testlerin ölçümler hakkında tek taraflı bilgi sağladıkları için bir ölçümün kesinliği hakkında yargıya ulaşmak için yeterli olmadığı kabul edilmektedir (Bruton vd., 2000; Koo ve Li, 2016; Portney ve Watkins, 2009). Diğer taraftan, sınıf içi korelasyon katsayısı (ICC) hem mutlak güvenirlik hakkında hem de ölçümlerin korelasyonu hakkında daha kapsamlı bir bilgi sağladığı için araştırmalarda daha sık tercih edilen bir güvenirlik göstergesidir (do Nascimento vd., 2017; Haley ve Fragala-Pinkham, 2006; Hopkins, 2000; Koo ve Li, 2016; Özbay vd., 2019).

ICC, ölçümler arasındaki korelasyonun yanı sıra mutlak sonuçlar arasındaki uyumu da kısmen içerdiği için güvenirlik araştırmalarında Pearson korelasyon analizinden daha etkilidir (Bruton vd., 2000). Ancak ICC'nin farklı varsayımlar içeren 10 hesaplama formunun olması ve araştırmalarda çoğunlukla bu detaylara ilişkin bilgilerin rapor edilmemesi bu konudaki belirsizliği arttırmaktadır (Ateş, Öztuna, ve Genç, 2009; Koo ve Li, 2016). Ayrıca ICC'nin örneklem boyutuna ve katılımcıların heterojenliğine çok fazla duyarlı olması sebebiyle son yıllarda kullanılması önerilen bazı güvenirlik analizleri ile desteklenmesinin faydalı olacağı belirtilmektedir (Weir, 2005). Bu sebeple bu çalışmanın amacı ICC hesaplamalarına ilişkin varsayımları ve pratikte faydalı olabilecek destekleyici hesaplama yöntemlerini tanıtarak alandaki uygun kullanım sıklığını arttırmaktır.

2. Sınıf-İçi Korelasyon Katsayısı (*Intra-Class Correlation Coefficient, ICC*)

ICC, güvenirlik analizlerinin incelendiği araştırma tasarımlarında en sık tercih edilen yöntemdir (do Nascimento vd., 2017; Ferreira da Silva Santos vd., 2020; Koo ve Li, 2016; Wilkinson vd., 2019). Mevcut istatistik programları ile kolayca hesaplanan bu değerin rapor edilmesi ve uygun varsayımların seçilmesi konusunda spor bilimleri literatüründe bazı eksiklerin olduğu görülmektedir (do Nascimento vd., 2017; Özbay ve Ulupınar, 2019; Özbay vd., 2019). ICC hesaplamasına ilişkin varsayımlarda 3 farklı model (1,2 ve 3), 2 farklı hesaplama türü (tek ve ortalama) ve 2 farklı güvenirlik tanımı (mutlak ve relatif güvenirlik) olmak üzere çok sayıda seçenek kullanılmaktadır (Ateş vd., 2009; Koo ve Li, 2016). ICC hesaplamalarına ilişkin model, hesaplama türü ve güvenirlik tanımı bilgileri açıkça belirtilebildiği gibi ICC (model, tür)$_{\text{tanım}}$ şeklinde de rapor edilebilmektedir (Ateş vd., 2009; Koo ve Li, 2016). Kısaltılmış raporlamada model seçimi "1,2 ve 3"

şeklinde; hesaplama türü seçimi "1 (tek ölçüm) ve k (birden fazla ölçüm)" şeklinde; güvenirlik tanımı "A (absolute) ve C (consistency)" şeklinde belirtilebilmektedir. Örneğin model 3, ortalama hesaplama türü ve relatif güvenirlik tanımının kullanıldığı bir ICC hesaplamasında, ICC(3,k)$_C$ şeklinde bir raporlama yapılabilmektedir.



Single Measurement: One-Way Random, absolute = 1.00 Two-Way Random, absolute = 1.00 Two-Way Random, consistency = 1.00 Two-Way Mixed, absolute = 1.00 Two-Way Mixed, consistency = 1.00 **Mean Measurement:** One-Way Random, absolute = 1.00 Two-Way Random, absolute = 1.00 Two-Way Random, consistency = 1.00 Two-Way Mixed, absolute = 1.00 Two-Way Mixed, consistency = 1.00 Pearson r = 1.00	**Single Measurement:** One-Way Random, absolute = 0.798 Two-Way Random, absolute = 0.816 Two-Way Random, consistency = 1.00 Two-Way Mixed, absolute = 0.816 Two-Way Mixed, consistency = 1.00 **Mean Measurement:** One-Way Random, absolute = 0.888 Two-Way Random, absolute = 0.899 Two-Way Random, consistency = 1.00 Two-Way Mixed, absolute = 0.899 Two-Way Mixed, consistency = 1.00 Pearson r = 1.00	**Single Measurement:** One-Way Random, absolute = 0.262 Two-Way Random, absolute = 0.206 Two-Way Random, consistency = 0.800 Two-Way Mixed, absolute = 0.206 Two-Way Mixed, consistency = 0.800 **Mean Measurement:** One-Way Random, absolute = 0.709 Two-Way Random, absolute = 0.342 Two-Way Random, consistency = 0.889 Two-Way Mixed, absolute = 0.342 Two-Way Mixed, consistency = 0.889 Pearson r = 1.00

Şekil 1. Aynı veri setinde model, tür ve tanım seçimine göre hesaplanan ICC değerleri.

Şekil 1'de FSKT (*Frequency Speed of Kick Test*) olarak adlandırılan ve 10 saniyelik bir sürede, belirlenen bir alana maksimum sayıda tekme atmayı içeren bir testten hesaplanan farklı ICC sonuçları hipotetik veriler üzerinde incelenmiştir. Buna göre her iki ölçümün eşit olması durumunda tüm hesaplamalar 1.00 değerine sahiptir. Ancak ikinci ölçümde tüm değerlerin 3 birim yükselmesi, Pearson r değerini ve relatif güvenirlik değerlerini (*consistency*) değiştirmezken, mutlak güvenirliğin (*absolute*) kısmen değişmesine sebep olmaktadır. Diğer taraftan ikinci ölçümde tüm değerlerin iki katına çıkması Pearson r değerini yine değiştirmezken; relatif güvenirliği kısmen, mutlak güvenirliği ise büyük oranda değiştirebilmektedir.

2.1. Modele Göre ICC Seçimi

2.1.1. Model 1: Tek-Yönlü Rastgele-Etki Modeli (One-Way Random-Effects Model)

Bu model pratikte en az tercih edilen modeldir. Bu modelde birden fazla değerlendirici vardır ancak her denek yalnızca bir değerlendirici tarafından ölçülür (Koo ve Li, 2016). Daha büyük bir değerlendirici popülasyonundan rastgele seçilen değerlendiricilerden her biri bir alt grubunun ölçümlerini gerçekleştirir. Böylece farklı merkezlerden gelen ölçümler ortak bir havuzda toplanarak daha büyük bir örneklem sayısına ulaşılmış olur. Ancak spor bilimlerinde bazen tek bir araştırmacının deneklere birden fazla ölçüm uyguladığı bazen de birkaç araştırmacının her birinin aynı denek grubuna ölçümler uyguladığı araştırma tasarımları daha sık kullanılmaktadır (Duthie,

Pyne, Ross, Livingstone, ve Hooper, 2006; Ferreira da Silva Santos vd., 2020; Segura-Ortí ve Martínez-Olmos, 2011; Wilkinson vd., 2019). Dolayısıyla spor bilimleri alanındaki çalışmalarda güvenirlik analizleri çoğunlukla bir denekten birden fazla ölçümün alındığı tasarımları içermektedir.

Bu modelin kullanıldığı bir duruma örnek olarak, orta öğretim öğrencilerinin bel çevresinin ulusal çapta değerlendirilmeye çalışıldığını varsayalım. Bu tür geniş çaplı bir araştırmada her coğrafi bölgeden rastgele seçilen bir Beden Eğitimi Öğretmeni rastgele seçilen 100 öğrenciyi değerlendirilebilir. Böylece yedi değerlendiricinin gerçekleştirmiş olduğu toplam 700 ölçüm sonucunun güvenirliği analiz edilebilir. Bu model, değerlendiricilerin deneklere ulaşmasının mümkün olmadığı durumlar için bir avantaj sağlasa da değerlendiricilerin rastgele seçilmiş olması varsayımı ve bir denekten sadece bir kez ölçüm alınıyor olması sebebiyle spor bilimleri literatüründe nadiren kullanılmaktadır.

2.1.2. Model 2: İki-Yönlü Rastgele-Etki Modeli (Two-Way Random-Effects Model)

Bu model, benzer özelliklere sahip daha geniş bir değerlendirici popülasyonundan rastgele seçilen değerlendiriciler tarafından yapılan ölçümlerin güvenirliğini analiz etmek için kullanılır (Koo ve Li, 2016). Başka bir deyişle araştırmada yer alan değerlendiricilerden elde edilen sonuçları, benzer özelliklere sahip tüm değerlendirici popülasyonuna genellemenin amaçlandığı durumlar için iki-yönlü rastgele-etki modeli tercih edilir. Örneğin ülkemizde Sağlık Bakanlığı ve Milli Eğitim Bakanlığının işbirliği ile "Türkiye Sağlıklı Beslenme ve Hareketli Hayat Programı 2018-2023" projesi kapsamında 5-12. sınıflarda öğrenim gören tüm öğrencilerin fiziksel uygunluk ve sağlık karnesi oluşturulmaktadır. Bu karnede yer alan şınav, mekik veya esneklik ölçümlerinin rastgele seçilen dört Beden Eğitimi Öğretmeni tarafından bir öğrenci grubuna uygulandığını varsayalım. Bu durumda tüm Beden Eğitimi Öğretmenlerinin benzer özelliklere sahip olması ve bu becerileri ölçmek için aynı yöntemi kullanacak olmaları sonuçların tüm değerlendirici popülasyonuna genellenebilmesini mümkün kılmaktadır.

2.1.3. Model 3: İki-Yönlü Karma-Etki Modeli (Two-Way Mixed-Effects Model)

Bu model, spor bilimlerindeki araştırmalarda performans ölçümlerini değerlendirmede en sık tercih edilen modeldir. Bu modelde araştırma tasarımcıları tarafından belirlenen ve rastgele seçilmeyen bir değerlendirici aynı deneklere birden fazla ölçüm uygular (Koo ve Li, 2016). Bu modelde amaç, elde edilen güvenirlik sonuçlarını benzer özelliklere sahip olsalar bile tüm değerlendirici popülasyonuna genellemek değildir, dolayısıyla sonuçların yalnızca araştırmada yer alan değerlendirici için geçerli olduğu kabul edilir. Bu modelde ölçümlerdeki değerlendirici hatasının sabit olduğu varsayılır ve göz ardı edilir(Ateş vd., 2009) Örneğin tekvando sporcuları için geliştirilen ve FSKT olarak bilinen bir testte sporcular belirtilen bir hedefe 10 saniye sürede maksimum sayıda tekme atmaya çalışır. Genellikle araştırmacılardan biri olarak seçilen bir değerlendirici aynı sporcuları birkaç gün arayla iki kez ölçebilir. Bu durumda sonuçlardaki olası değerlendirici hataları göz ardı edilir ve iki ölçüm arasındaki farkların katılımcıların kendilerinden kaynaklanan faktörler ile ilgili olduğu kabul edilir.

2.2. Hesaplama Türüne Göre ICC Seçimi

Hesaplama türüne bağlı olarak ICC seçimi tek (*single*) ve ortalama (*mean*) olarak yapılabilmektedir. Tek bir ölçüm sonucuna göre ICC'yi hesaplamak birden fazla ölçümün

ortalamasına göre daha düşük sonuçlar ürettiği için daha az tercih edilmektedir (Ateş vd., 2009). Örneğin 3 değerlendiricinin (veya 3 ölçümün) ölçüm sonuçlarının ortalamasının nihai değerlendirmenin temeli olarak kullanıldığı durumlar için hesaplama türü "ortalama" olarak tercih edilir ve "k" şeklinde rapor edilir. Diğer taraftan tek bir değerlendiriciden (veya tek bir ölçümden) elde edilen sonuçların gerçek ölçüm değerleri olarak kullanıldığı durumlar için hesaplama türü "tek" olarak tercih edilir ve "1" şeklinde rapor edilir. Spor bilimleri alanında performans testlerinin uygulandığı araştırmalarda çoğunlukla ortalama hesaplama türü kullanılmaktadır. Örneğin, bel çevresinin 3 farklı araştırmacı tarafından birer kez ölçülmesi, ya da bir araştırmacı tarafından üç kez ölçülmesi durumunda ölçüm değerlerinin ortalamasını almak daha güvenilir sonuçlar vereceği için hesaplama türü "ortalama" olarak tercih edilir. Ancak bu ölçümlerden birisi gerçek ölçüm değeri olarak kullanılacaksa birden fazla ölçüm yapılmış olsa bile hesaplama türü "tek" olarak tercih edilir.

Diğer taraftan hesaplama türünün "tek" seçilmesi bazı tasarımlar için daha uygun olabilmektedir. Örneğin voleybol sporcularının sıçrama yüksekliğinin altın standart olarak bilinen bir yöntem ve yeni geliştirilmiş bir başka yöntem ile iki kez ölçüldüğünü varsayalım. Bu durumda ölçüm sonuçlarının ortalamasını almak uygun değildir. Çünkü ölçüm yöntemi ile ilgili olası hataların kaynağının yeni geliştirilen yöntem olduğu düşünülür. Benzer şekilde, bir ölçüm konusunda otorite olarak kabul edilen bir uzman ve görece daha tecrübesiz bir araştırmacının gerçekleştirdiği ölçüm sonuçları için hesaplama türünü "ortalama" olarak seçmek uygun olmayacaktır. Çünkü bu tür bir tasarımda uzman olan değerlendiriciden kaynaklanan olası hatalar göz ardı edilerek, değerlendiriciye bağlı ölçüm hataları görece daha tecrübesiz olan araştırmacıya atfedilir.

2.3. Güvenirlik Tanımına Göre ICC Seçimi

ICC hem mutlak hem de relatif güvenirlik olarak hesaplanabilmektedir (Ateş vd., 2009; Koo ve Li, 2016; Segura-Ortí ve Martínez-Olmos, 2011; Wilkinson vd., 2019). Güvenirlik tanımına ilişkin ICC seçimi, ölçümler arasındaki sayısal değerlerin mutlak uyumuna göre mi yoksa grup içerisindeki sıralamaların tutarlılığına göre mi yapılacağını içerir. Burada mutlak uyum güvenirliği, farklı ölçümlerde aynı deneklere aynı puanların verilmesini daha önemli kabul ederken, relatif güvenirlik ise bir deneğe ait ölçüm sonuçlarının aynı deneğin grup içerisindeki sıralamasını koruyabilmesini önemli kabul etmektedir. Örneğin Şekil 1'de gösterildiği gibi iki ölçümde de aynı değerlerin elde edilmesi (x = y) hem mutlak hem de relatif güvenirliği sağlarken, ikinci ölçümde tüm değerlerin 3 birim yükselmesi (x = y + 3) veya iki katına çıkması (x = 2y) durumunda relatif güvenirlik, mutlak güvenirliğe göre daha yüksektir. Çünkü tüm denekler için puanların artması deneklerin sıralamasını değil, ölçüm ortalamalarını etkilemektedir. Ancak ICC formülleri hem ölçümlerdeki mutlak uyumu hem de sıralamalardaki tutarlılığı içerdiği için Şekil 1'de görüldüğü gibi ikinci ölçümde değerlerin 2 katına çıkması durumunda grup içindeki sıralamanın değişmemesine rağmen relatif güvenirlik kısmen düşmüştür. Güvenirlik analizlerinde ICC'nin Pearson r değerine göre daha uygun olduğunun düşünülmesi temelde bu sebebe dayanmaktadır.

Bununla birlikte spor bilimleri alanındaki araştırmalarda ICC hesaplaması için çoğunlukla relatif güvenirlik tanımına göre yapılmaktadır. Çünkü relatif güvenirliği belirlemek için en uygun analiz ICC olarak kabul edilmektedir. Ölçümler arasındaki ortalamaların mutlak uyumu eşli örneklem t-testi ve Bland-Altman testleri ile veya ölçümlerin standart hatası (*standart error of measurement, SEM*) ve ölçümlerin tipik hatası (*typical error of measurement, TEM*) gibi analizler ile kolaylıkla test edilebildiği için çoğunlukla relatif güvenirlik ICC ile değerlendirilirken, mutlak güvenirlik ise bahsi geçen

diğer analizler ile değerlendirilmektedir (do Nascimento vd., 2017; Özbay vd., 2019; Weir, 2005; Wilkinson vd., 2019).

3. Ölçümlerin Standart Hatası (*Standart Error of Measurement, SEM*)

Spor bilimlerinde mutlak güvenirlik, performanstaki gerçek bir değişikliği bireysel varyasyonlar ve ölçüm hatalarından ayırt edebilmek için önemlidir (Segura-Ortí ve Martínez-Olmos, 2011). Son yıllarda, ICC'ye ek olarak pratikte sonuçların yorumlanmasını kolaylaştıracak bazı hesaplamaların da raporlanması gerektiği belirtilmektedir (Ferreira da Silva Santos vd., 2020; Özbay vd., 2019; Wilkinson vd., 2019). Ölçümler arasındaki uyumun bir göstergesi olarak kullanılması önerilen SEM, ICC ve standart sapma değerleri kullanılarak hesaplanır (Weir, 2005). SEM aynı zamanda pratikte daha kolay ve anlaşılabilir bir yoruma ulaşmayı sağlayan "minimal tespit edilebilir değişim (*minimal detectable change, MDC*)" miktarını hesaplamak için de kullanılmaktadır (Haley ve Fragala-Pinkham, 2006).

Formül 1' de gösterildiği gibi SEM hesaplanırken hem denekler-arası standart sapma, hem iki ölçümden elde edilen standart sapmaların ortalaması hem de güvenirlik skoru (ICC) kullanıldığı için pratikte çok yönlü bir yorum yapma imkanı sunmaktadır (Segura-Ortí ve Martínez-Olmos, 2011; Weir, 2005; Wilkinson vd., 2019). Benzer araştırma tasarımına sahip iki ölçümden daha düşük SEM değerine sahip olan sonuçların hatalardan daha arınık olduğu, dolayısıyla daha güvenilir bir çıkarım yapmayı mümkün kıldığı düşünülmektedir (Chiu vd., 2016; Segura-Ortí ve Martínez-Olmos, 2011). SEM ayrıca varyasyon katsayısına benzer bir şekilde ölçümlerdeki değişkenlik ve ortalama oranına ilişkin bir değeri hesaplamak için de kullanılmaktadır. Bilindiği gibi varyasyon katsayısı (*variation coefficient, CV*) denekler arası standart sapmanın ortalamaya bölünmesi ve 100 ile çarpılması ile elde edilir (CV = Ss / Ortalama × 100). SEM değerinin varyasyon katsayısı olarak kullanıldığı durumlar için ise standart sapma değeri yerine SEM değeri kullanılmaktadır (Formül 2). Bu hesaplama olası hata miktarının ortalamaya oranını sunar ve pratikte oldukça faydalı olduğu kabul edilir (Charter, 1996; do Nascimento vd., 2017; Weir, 2005).

Formül 1: $SEM = Ss_{ortalama} \times \sqrt{1 - r}$
Formül 2: $CV_{SEM} = SEM / Ortalama \times 100$

Formüllerde yer alan "$Ss_{ortalama}$" ölçümlerdeki denekler arası standart sapmanın ortalamasını; "r" ICC değerini; CV_{SEM} ise SEM değerinin varyasyon katsayısının hesaplanmasında kullanıldığı durumu ifade etmektedir. Formülden de anlaşılacağı gibi SEM denekler-arası farklılıklara ve güvenirlik katsayısına duyarlıdır. Örneğin tüm deneklerin aynı skoru alması durumunda ($0 \times \sqrt{1 - r} = 0$) veya güvenirliğin maksimum (1'e eşit olması) durumunda ($Ss \times \sqrt{1 - 1} = 0$) SEM sıfıra eşit olacaktır.

4. Ölçümlerin Tipik Hatası *(Typical Error of Measurement, TEM)*

Tipik hata ölçümden ölçüme sonuçlardaki değişimi (ölçümler arası varyasyonu) tahmin etmek için kullanılan bir hesaplamadır (do Nascimento vd., 2017; Weir, 2005). SEM ile aynı amaçla kullanılır, ancak TEM sadece ölçümler arasındaki farkların standart sapmasını içerir (Formül 3). Bu sebeple SEM, ölçümlerin kesinliğine ilişkin daha genel bir bilgi sağlarken, TEM ise ölçümden ölçüme farklılaşmaya daha duyarlıdır. Örneğin bir araştırmacı tarafından yeni geliştirilen bir mobil uygulama ile 15 kişinin 100 m sprint zamanlarının iki kez ölçüldüğünü varsayalım. Bu 15 kişinin sprint sürelerinin birbirine

yakın olması (grubun homojen olması) denekler arası standart sapmayı düşüreceği için SEM değerinin daha küçük olmasına sebep olacaktır (Formül 1). Ancak katılımcıların iki ölçümü arasındaki fark ne kadar düşükse TEM değeri buna bağlı olarak küçülecektir. Özetle, denekler arasındaki standart sapma daha çok SEM değerini etkilerken, ölçümler arasındaki farkların standart sapması TEM değerini etkilemektedir (Formül 3).

Formül 3: $TEM = Ss_{farklar} / \sqrt{2}$

Formülde yer alan "$Ss_{farklar}$" ölçümler arasındaki farkların standart sapmasını ifade etmektedir. Formülden de anlaşılacağı gibi TEM ölçümler arasındaki sabit hatalara duyarlı değildir. Örneğin Şekil 1'de gösterildiği gibi ikinci ölçümdeki değerlerinin birinci ölçüme göre 3 birim yükselmesi ($x = y + 3$) durumunda TEM sıfıra eşit olacaktır ($0 / \sqrt{2} = 0$). Dolayısıyla ikinci ölçümde değerlerin 4 birim yükseldiği ya da 2 birim düştüğü her iki durum içinde TEM sıfır olacaktır. Bu nedenle TEM ölçümler arasındaki sistematik ve tesadüfi hataları belirlemek için daha uygundur.

5. Tespit Edilebilir Minimal Değişim *(Minimal Detectable Change, MDC)*

MDC bir değişimin ölçüm hatalarından kaynaklanmayacak derecede güvenilir olduğu sonucuna varmak için performanstaki en küçük farkı ifade eder (Chiu vd., 2016; Segura-Ortí ve Martínez-Olmos, 2011; Wilkinson vd., 2019). MDC değerinden daha yüksek bir performans gelişimi, bu farkın hatalardan arınık gerçek bir değişim olduğunu yorumlamada kullanılan pratik bir göstergedir (Haley ve Fragala-Pinkham, 2006). Başka bir deyişle MDC bir ölçümde beklenen hata aralığının dışında olan en küçük eşik değerini tanımladığı için farkın hataya atfedilemeyeceği işlevsel bir bilgi sağlar (Charter, 1996; Chiu vd., 2016; Wilkinson vd., 2019). Örneğin bir sporcu grubunun skuat testinde 1 tekrarda kaldırabildikleri maksimum ağırlıkları ortalamasının 90 kg ve MDC değerinin 5,5 kg olduğunu varsayalım. Bu sporcu grubuna uygulanacak bir müdahaleden sonra grup ortalamasının 95,5 kg'a yükselmesi pratikte hata sınırlarının dışında gerçek bir değişimin gerçekleştiği çıkarımında bulunabilmeyi sağlar.

Formül 4: $MDC = SEM \times 1.96 \times \sqrt{2}$

Formülde yer alan 1,96 değeri %95 güven aralığı sınırları ile ilişkili z skorudur. Dolayısıyla formülde yer alan 1,96 değeri çalışmalarda en sık tercih edilen % 95 güven aralığı sınırları içindir, ancak araştırmacının tercihine göre bu değer değişebilir.

6. Sonuç

Sporcuların performanslarını takip etmek için yapılan testlerin ve yöntemlerin güvenilir sonuçlar vermesi kritik bir öneme sahiptir. Spor bilimleri alanındaki çalışmalarda yapılan performans ölçümlerin güvenilir olduğuna dair çıkarımlar genellikle ICC değerlerine dayandırılarak yapılmaktadır. Ancak ICC değerleri deneklerin homojen/heterojen yapısına ve örneklem büyüklüğüne aşırı duyarlı olduğu için güvenilirlik analizlerinin pratikte faydalı olabilecek diğer hesaplamalar ile desteklenmesi gerektiği düşünülmektedir. Bunun yanında ICC hesaplamalarında 10 farklı versiyonun bulunması ve çoğunlukla bu hesaplamalara ilişkin bilgilerin rapor edilmemesi bu konudaki belirsizliği arttırmaktadır. Diğer taraftan SEM ve TEM gibi hesaplamalar olası ölçüm hatalarının miktarına dair önemli bilgiler sunarken, MDC gibi hesaplamalar da hatalardan arınık gerçek bir performans değişiminin miktarı hakkında kritik bilgilere ulaşmayı kolaylaştırmaktadır. Sonuç olarak bu çalışmada akademik araştırmalar ve saha uygulamaları arasındaki işlevsel köprünün birçok disipline göre daha önemli olduğu spor

bilimlerinde bilimsel raporlamaların aynı zamanda pratikte de kullanılabilecek şekilde sunulmasının sağlayacağı yararlar vurgulanmaya çalışılmıştır.

KAYNAKLAR

ATEŞ, C., ÖZTUNA, D., ve GENÇ, Y. (2009). Sağlık araştırmalarında sınıf içi korelasyon katsayısının kullanımı. *Türkiye Klinikleri Biyoistatistik Dergisi, 1*(2), 59-64.

BALSALOBRE-FERNÁNDEZ, C., GLAISTER, M., ve LOCKEY, R. A. (2015). The validity and reliability of an iPhone app for measuring vertical jump performance. *Journal of sports sciences, 33*(15), 1574-1579.

BRUTON, A., CONWAY, J. H., ve HOLGATE, S. T. (2000). Reliability: what is it, and how is it measured? *Physiotherapy, 86*(2), 94-99.

CHARTER, R. A. (1996). Revisiting the standard errors of measurement, estimate, and prediction and their application to test scores. *Perceptual and Motor Skills, 82*(3_suppl), 1139-1144.

CHIU, E.-C., WU, W.-C., CHOU, C.-X., YU, M.-Y., ve HUNG, J.-W. (2016). Test-retest reliability and minimal detectable change of the Test of Visual Perceptual Skills-in patients with stroke. *Archives of Physical Medicine and Rehabilitation, 97*(11), 1917-1923.

DO NASCIMENTO, M. A., RIBEIRO, A. S., DE SOUZA PADILHA, C., DA SILVA, D. R. P., MAYHEW, J. L., DO AMARAL CAMPOS FILHO, M. G., ve CYRINO, E. S. (2017). Reliability and smallest worthwhile difference in 1RM tests according to previous resistance training experience in young women. *Biology of Sport, 34*(3), 279-285.

DUTHIE, G. M., PYNE, D. B., ROSS, A. A., LIVINGSTONE, S. G., ve HOOPER, S. L. (2006). The reliability of ten-meter sprint time using different starting techniques. *The Journal of Strength ve Conditioning Research, 20*(2), 251.

FERREIRA DA SILVA SANTOS, J., LOPES-SILVA, J. P., LOTURCO, I., ve FRANCHINI, E. (2020). Test-retest reliability, sensibility and construct validity of the frequency speed of kick test in male black-belt taekwondo athletes. *Ido Movement for Culture. Journal of Martial Arts Anthropology, 20*(3), 38-46.

HALEY, S. M., VE FRAGALA-PINKHAM, M. A. (2006). Interpreting change scores of tests and measures used in physical therapy. *Physical therapy, 86*(5), 735-743.

HAYNES, T., BISHOP, C., ANTROBUS, M., ve BRAZIER, J. (2019). The validity and reliability of the my jump 2 app for measuring the reactive strength index and drop jump performance. *The Journal of sports medicine and physical fitness*.

HOPKINS, W. G. (2000). Measures of reliability in sports medicine and science. *Sports medicine, 30*(1), 1-15.

KOO, T. K., ve LI, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of chiropractic medicine, 15*(2), 155-163.

ÖZBAY, S., ve ULUPINAR, S. (2019). Reliability of 1RM, 5RM and 10RM Tests in Upper Body Resistance Exercises. *The Journal of Turkish Sport Sciences for Health, 2*(1), 1-7.

ÖZBAY, S., ULUPINAR, S., ÇINAR, V., ve AKBULUT, T. (2019). Reliability of Easily Applicable Non-Laboratory Methods Used for Determination of the Upper Body Strength. *Turkiye Klinikleri Journal of Sports Sciences, 11*(2).

PORTNEY, L. G., ve WATKINS, M. P. (2009). *Foundations of clinical research: applications to practice* (Vol. 892): Pearson/Prentice Hall Upper Saddle River, NJ.

SEGURA-ORTÍ, E., ve MARTÍNEZ-OLMOS, F. J. (2011). Test-retest reliability and minimal detectable change scores for sit-to-stand-to-sit tests, the six-minute walk test,

the one-leg heel-rise test, and handgrip strength in people undergoing hemodialysis. *Physical Therapy, 91*(8), 1244-1252.

WEIR, J. P. (2005). Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. *The Journal of Strength ve Conditioning Research, 19*(1), 231-240.

WILKINSON, T. J., XENOPHONTOS, S., GOULD, D. W., VOGT, B. P., VIANA, J. L., SMITH, A. C., ve WATSON, E. L. (2019). Test–retest reliability, validation, and "minimal detectable change" scores for frequently reported tests of objective physical function in patients with non-dialysis chronic kidney disease. *Physiotherapy theory and practice, 35*(6), 565-576.

Summary

Reliability is defined as the reproducibility of a measurement results or the consistency between the repeated trials. Reliability is evaluated in absolute agreement and relative definitions. Absolute reliability refers to the degree of agreement between repeated measurements while relative reliability refers to the consistency of the ranking of group members. The intra-class correlation coefficient (ICC) is the most common method for determining relative reliability. However, it is emphasized that there are some uncertainties in the selection and reporting of the appropriate calculation in the current literature due to the fact that there are 10 different versions of ICC. In addition, since ICC coefficient is excessively sensitive to the distribution and number of the sample, it is recommended to report supporting methods in studies including reliability analysis. In the sports science literature, the reporting of calculations such as standard error of the measurements (SEM), typical error (TE) and minimal detectable change (MDC) is necessary to easily inference the reliability analyzes in practice. SEM is a calculation method used to distinguish a true variation from measurement errors and individual variations. Similarly, TE represents the variation from trial to trial and is used to assess the absolute agreement between measurements. MDC refers to the smallest threshold value outside the error limits and is used as an easy-to-interpret practical metric that allows to determine an actual performance change. Functionality between academic investigations and field practices has a more critical importance in sports sciences compared to other disciplines. Therefore, it is clear that both reporting ICC properly and supporting it with other analyzes that can be used in practice can provide important benefits. This study aimed to express the assumptions about the calculation of different ICC and the supportive reliability analysis suggested to be used in the field of sports sciences and to provide more common of their reporting.