Accepted manuscript

Lindberg, K. A., Solberg, P. A., Bjørnsen, T., Helland, C., Rønnestad, B., Frank, M., Haugen, T., Østerås, S., Kristoffersen, M., Midttun, M., Sæland, F., Eythorsdottir, I. E. T. & Paulsen, G. (2022). Strength and Power Testing of Athletes: A Multicenter Study of Test-Retest Reliability. International Journal of Sports Physiology and Performance (IJSPP), 17 (7), 1103-1110. https://doi.org/10.1123/ijspp.2021-0558

Published in: International Journal of Sports Physiology and Performance (IJSPP)

DOI: https://doi.org/10.1123/ijspp.2021-0558

AURA: https://hdl.handle.net/11250/3057304

Lindberg, K. A., Solberg, P. A., Bjørnsen, T., Helland, C., Rønnestad, B., Frank, M., Haugen, T., Østerås, S., Kristoffersen, M., Midttun, M., Sæland, F., Eythorsdottir, I. E. T., Paulsen, G. (2022). Strength and Power Testing of Athletes: A Multicenter Study of Test-Retest Reliability. *International Journal of Sports Physiology and Performance (IJSPP), 17*(7), 1103-1110. http://dx.doi.org/10.1123/ijspp.2021-0558

Title: Strength and Power Testing of Athletes: A Multicenter Study of Test-Retest Reliability

Submission type: Original investigation

Kolbjørn Lindberg[1,2*], Paul Solberg[2], Thomas Bjørnsen[1,2], Christian Helland[2], Bent Rønnestad[2,3], Martin Thorsen Frank[1], Thomas Haugen[2,4], Sindre Østerås[2,6], Morten Kristoffersen[2,5], Magnus Midttun[2], Fredrik Sæland[2], Ingrid Eythorsdottir[7], Gøran Paulsen[2,7]

[1]Department of Sport Science and Physical Education, Faculty of Health and Sport Sciences, University of Agder, Kristiansand, Norway, [2]Norwegian Olympic and Paralympic Committee and Confederation of Sports, Oslo, Norway, [3]Department of Health and Exercise Physiology, Faculty of Social Sciences, Inland Norway University of Applied Sciences, Lillehammer, Norway, [4]School of Health Sciences, Kristiania University College, Oslo, Norway, [5]Department of Neuromedicine and Movement Science, Faculty of Medicine and Health Sciences, Center for Elite Sports Research, Norwegian University of Science and Technology, Trondheim, Norway, [6]Department of Sport and Education, Bergen University College, Bergen, Norway, [7]Department of Physical Performance, Norwegian School of Sport Sciences, Oslo, Norway

[*] **Correspondences**
Kolbjørn Lindberg, Department of Sport Science and Physical Education, Faculty of Health and Sport Sciences, University of Agder, Kristiansand, Norway
kolbjorn.a.lindberg@uia.no, +4790870067

Abstract
Purpose: This study examined the test-retest reliability of common assessments for measuring strength and power of the lower body, in high-performing athletes. **Methods:** A total of 100 participants, including both male (n=83) and female (n=17) athletes (21 [4] y, 182 [9] cm, 78 [12] kg), were recruited for this study, using a multicenter approach. The participants underwent physical testing 4 times. The first 2 sessions (1 and 2) were separated by ~1 week, followed by a period of 2 to 6 months, whereas the last 2 sessions (3 and 4) were again separated by ~1 week. The test protocol consisted of squat jumps, countermovement jumps, jump & reach, 30-m sprint, one-repetition maximum squat, sprint cycling, and a leg press test. **Results:** The typical error (TE%) ranged from 1.3% to 8.5% for all assessments. The change in means ranged from −1.5% to 2.5% for all assessments, whereas the ICC ranged from 0.85 to 0.97. The smallest worthwhile change (0.2 of baseline SD) ranged from 1.2% to 5.0%. The ratio between the TE% and the smallest worthwhile change in % ranged from 0.5 to 1.2. When observing the reliability across testing centers, considerable differences in reliability were observed (TE% ratio: 0.44 – 1.44). **Conclusions:** Most of the included assessments can be used with confidence by researchers and coaches to measure strength and power in athletes. Our results highlight the importance of controlling testing reliability at each testing center and not relying on data from others, despite having applied the same protocol.

Keywords: squat jump, countermovement jump, jump and reach, sprint running, sprint cycling, leg press

Introduction

Strength and power are fundamental physical qualities for human movement, particularly relevant for coaches and researchers working with athletes.[1-5] Consequently, these capabilities

are often assessed to monitor and evaluate acute and chronic training responses[5,6] and to classify strengths and weaknesses of athletes.[1,3] However, the usefulness of these assessments is highly dependent on the test-retest reliability[2,4,5,7-10], which highlights a concern in the field of practice with high-performing athletes where the frequent use of testing is burdened by a myriad of methods and protocols with unknown reliability.[11]

Test-retest reliability refers to the consistency in the measurements when they are repeated.[7,8] Especially when considering repeated testing over time, i.e., interday (between days) testing, several factors could affect the consistency of the measurements, such as biological and technical (equipment) variations, and test procedures.[7] Lack of test consistency, makes the test results potentially misleading and counterproductive.[7] Particularly when evaluating individual data of high-performing athletes, the requirements of interday test-retest reliability is arguably of greater importance – as subtle changes in performance are expected.[2,4,7,8]

Test-retest reliability can be investigated with different designs and statistical approaches. Within-subject variation is arguably the most important measure of test-retest reliability[7], as lower within-subject variation increases the likelihood of observing true changes in performance.[7,8] The typical error expressed in absolute (TE) or relative terms (TE% [sometimes referred to as the coefficient of variation]), is most commonly used to quantify the magnitude of the within-subject variation.[7] The TE% should preferably be as low as possible, or at least low compared to the magnitude of the true changes in performance.[7,8]

Knowledge regarding measurement errors of performance assessments appears to be increasingly recognized as important if we consider the growing number of research papers within this topic.[2,4,9,10,12-16] Indeed, several papers have reported TE% from common strength and power assessments. Jump height, measured from the squat jump (SJ), countermovement jump (CMJ), and the jump & reach test, have been reported with TE´s ranging from 1.5% to 8.6%; where the higher TE´s have typically been observed for SJ.[8,9,12,13] Furthermore, TE´s of 1.1% to 3.3% have been reported for sprint running variables such as sprint time, where the TE´s are commonly reduced with the length of the distance covered.[14,17,18] For the one-repetition maximum (1RM) squat, the TE% has been reported to range from 0.3% to 12.1%, regardless of training status or familiarisation.[10,15] Similarly, previous studies have found TE´s ranging from 1.5% to 4.6% in sprint cycling, for variables such as peak and mean power outputs.[8,19]

When interpreting repeated performance tests, it is crucial to be able to decide if changes in the test results are of important or worthwhile sizes – in addition to the test-retest reliability.[7,20] For generic performance tests, such as the SJ and CMJ, performed by a mix of athletes (as relevant at Olympic training facilities), the smallest worthwhile change (SWC) can be estimated by multiplying the baseline standard deviation (SD) of the sample by 0.2 (a small effect size).[7] From this, the SWC in relation to the TE (SWC:TE) represents the signal-to-noise ratio, which should preferably be greater than 1.[20] This approach has only recently been included in test-retest studies. [5,7,21,22]

Previous reliability studies on strength and power test for the lower body are limited by low number of participants[2,9,12,13,19,21], moderate/low level athletes[5,9,12,22] (not elite level), assessing only a single test[2,13,15,21-23], and/or calculating the reliability from merely two test sessions separated by typically one or two weeks.[5,9,12,13,15,19,21] This makes it challenging to compare the usefulness of different tests and which to include in a test-battery which is generally repeated over longer periods of time. On the contrary, in this study, several tests were used to assess strength and power of the lower body (e.g., jumping, sprinting, cycling, and lifting), in a high number of high-performing athletes over several test sessions – as more sessions and increased time frame between test sessions improve the ecological validity of the data.

Finally, most reliability studies including strength and power assessments are conducted in a single research center, whereas multicenter approaches are less frequently used.[16] The benefit of using a multicenter approach is mainly improved generalizability of the results, particularly for high-performing athletes.[16] For example, athletes training under the Norwegian Olympic Federation are typically assessed at different locations (due to domicile, travelling, and training camps), making results from a single testing center less representative for this population. Additionally, a multicenter approach allows a larger number of participants to be included, thereby gathering more data that will benefit researchers, coaches, and athletes utilizing the respective tests.[16]

The aim of the present study was therefore to examine the test-retest reliability across seven common assessments for measuring strength and power in the lower body, using a large sample of high-performing athletes, obtained with a multicenter approach.

Methods

Overview and design

The present study is part of a project investigating common assessments of physical performance in athletes. Results regarding the reliability and agreement for measures of force-velocity profiles are published elsewhere,[4] whilst the association among common assessments to evaluate strength and power is published in a twin paper.

To assess the test-retest reliability of common strength and power assessments, the participants underwent testing 4 times, at 6 different centers belonging to the Norwegian Olympic Federation. Regarding the study design in the present study, we refer to the flow chart presented in Lindberg et al.[4] The first two sessions (1 and 2) were separated by ~1 week, followed by a self-administered training period of 2 to 6 months, whereas the last two sessions (3 and 4) were again separated by ~1 week. The intent of having 2 to 6 months between testing sessions was to assess the longitudinal associations across the included measurements; these results are published in a twin paper.

The multicenter approach was chosen to reach a large sample of athletes.[16] Moreover, the multicenter approach has high ecological validity, as many athletes train and are tested at different facilities (centers).[16] A typical example is that athletes are tested at one center on a regular basis (near their home) but at a different center when training with the national team, as testing is often part of a training camp.

Since not all centers possessed the same test equipment, the sample size differed across measurement methods. In this paper, results are based on an aggregated analysis including all athletes with varying sample sizes across methods. Test leaders and equipment differed across centers but were kept constant for each athlete. Sample size for all tests is presented in the "Results" section.

Written informed consent was obtained from all participants prior to participation. The study was reviewed by the ethical committee of Inland Norway University of Applied Sciences, approved by the Norwegian Center for Research Data (reference ID: MR25102017), and performed in agreement with the Declaration of Helsinki. This study was not prospectively registered.

Participants

A total of 100 participants, including both male (n=83; 21 [4] y, 184 [8] cm, 81 [11] kg) and female (n=17; 21 [2] y, 171 [7] cm, 64 [7] kg) athletes, were recruited for this study. Most of the participants were team sports players in handball (n=31), ice hockey (n=22), soccer (n=15), and volleyball (n=15), while the remaining participants (n=20) competed in Nordic combined, ski jumping, weightlifting, athletics, badminton, and speed skating. The

competition level ranged from world-class (n=6) to club level (n=12), with the majority at a national and elite level (n=82) in their respective sport.

Test Procedures

The participants were instructed to prepare for the test days as they would for regular competition in terms of nutrition, hydration, and sleep, as well as refraining from strenuous exercise 48 h prior to testing. All testing was performed indoors, and the participants were instructed to wear identical footwear and clothing on each test day. Bodyweight was measured wearing training clothes and shoes. The different tests were separated by 5–10 min to ensure proper recovery.[4]

First, all participants performed a standardized ~10 min warm-up, explained in detail previously.[4] The test protocol consisted of SJs, CMJs, jump & reach, 30-m sprint, 1RM squat, sprint cycling, and a seated leg press test. The order of the tests varied across the multiple testing centers but was held constant within each center. All tests were separated by 5–10 min to ensure proper recovery.

SJ and CMJ

The SJs and CMJs were initially performed with bodyweight, accompanied by an incremental loading protocol.[4] For the purpose of this study, only jumps performed with bodyweight were analysed, as the test-retest results for the incremental loaded protocol are presented in a separate paper.[4]

The SJ and CMJ were performed on force platforms. Some testing centers applied a portable force platform with a sampling frequency of 200 Hz (Musclelab; Ergotest AS, Porsgrunn, Norway), while other centers used an in-ground force platform with a sampling frequency of 2000 Hz (AMTI; Advanced Mechanical Technology, Inc, Waltham Street, Watertown, USA).

Regarding the jump procedures in the present study, we refer to the paper by Lindberg et al.[4] Briefly, all jumps were performed with hands on the hips. Two valid trials were registered for each jump where the best of these was retained for further analysis. The recovery after each attempt was 2–3 min. Jump height (cm) was calculated through the impulse-momentum theorem, and registered with a minimum of 1 decimal – e.g., 0.1 cm.[4] Power was calculated as time-average (mean) instantaneous power (product of force and velocity) from the entire push-off phase for each respective jump, i.e., from peak force, obtained at the deepest position, until take-off. The power was obtained as watts (W) without any decimal points – e.g., 0W.

Jump & Reach

The jump & reach test was performed under a custom-made frame as previously described.[24] The participants marked the jump height with chalk on their fingertips. Each jump was proceeded with a 2–4 step run-up, where take-off was performed on both legs. The participants were instructed to jump as high as possible. Three jumps were performed, unless the final jump was the highest, then the participants performed an additional jump. Thirty to sixty seconds of rest was allowed between jumps. This procedure was performed twice, with a 2 min rest period. The highest measured jump (measured in cm, with a minimum of 1 decimal – e.g., 0.1 cm) was retained for further analysis.

30-m Sprint

The 30-m sprint was preceded by 2–3 strides/submaximal sprints, as a specific warm-up. During the test, 2–4 maximal sprints were performed, separated by 3–5 min recovery. Time was measured with a contact mat and wireless timing gates (Musclelab, Ergotest innovation AS, Langesund, Norway) placed with 5-m intervals and measured in seconds with a minimum

of 3 decimal points – e.g., 0.001 s. Timing was initiated when the front foot left the ground. The trial with the best 30-m time, and associated 10- and 20-m splits, were retained for analysis. Additionally, peak velocity was obtained as the highest average velocity within a 5-m time interval.

1RM Squat
The participants completed a brief warm-up consisting of submaximal squats with 2–4 repetitions at 50% and 60% of 1RM, and one repetition at 80%, 90%, and 95% of 1RM (self-estimated at the first time-point). The participants were given 2–3 trials at the 1RM loads with a rest period after each attempt of 2–3 min. After a successful 1RM attempt, the load was increased with a minimum of 2.5 kg, until no further weights could be lifted. Squat depth was standardized as top of the knee higher than the top thigh, proximal at the hip joint, and was visually controlled by the respective test leaders. The standardized squat depth was kept constant across all testing time points. The heaviest load (in kg) successfully lifted with the standardized depth was recorded as the participant´s 1RM.

Sprint Cycling
The sprint cycling test was conducted on an air-braked bicycle ergometer, where the protocol was based on the inbuilt 6 s power test feature (WattBike, WattBike Ltd, Nottingham, UK), combined with years of practical experience at the national Olympic training center. The seat of the bicycle was adjusted for each participant so that the leg was extended straight in the lower pedal position. The bicycle handlebars were adjusted to the same height as the seat. The resistance was adjusted for each participant to result in a frequency of 120–140 revolutions per minute. The participants were instructed to bicycle with maximal effort for 6 s. Three attempts were given with 3 min rest periods. If the participant was close to or outside the given revolutions per minute, the resistance was adjusted between attempts. The entire test was performed in a sitting position, without a rolling start. The highest peak and mean power outputs were used for further analysis, obtained as watts (W) from the inbuilt software without any decimal points – e.g., 0W.

Seated Leg Press
For the seated leg press test, a Keiser Air300 horizontal pneumatic device with an A420 software was used (Keiser Sport, Fresno, CA). The position of the seat was adjusted for each participant to result in approximately 80–90° of knee flexion when feet were placed on the foot pedals (180° = full extension).[25] The protocol consisted of a 10-repetition protocol, from which a force-velocity profile was calculated, as explained in Lindberg et al.[4] Participants were instructed to extend both legs with maximal effort, i.e., as fast as possible, for each repetition during the protocol.[25] The theoretical maximum power from the force-velocity profile[4] was retained for further analysis, obtained as watts (W) without any decimal points – e.g., 0W.

Statistical Analysis
All data are presented as mean (SD) unless stated otherwise. To increase the readability of the manuscript, the reliability for each respective test, is presented as the combined reliability from the double test-retest sessions – i.e., test sessions 1 and 3 (test) are presented as "Mean test", and sessions 2 and 4 (retest) as "Mean retest". Analysis for tests 1–2 and 3–4 is also provided. Additionally, the reliability between tests 1–2 and 3–4 was compared based on overlapping confidence intervals (CI) of the typical error (TE), to determine whether a learning effect was present.[7] To assess reliability across testing time points, the typical error

in absolute (TE) and relative terms (TE%), interclass correlation coefficient (ICC; 3,1)[26], and change in mean presented in relative and standardized terms (percent change; %Δ, and standardized mean difference; SMD), were calculated. The TE was calculated as the SD of the change score divided by the square root of 2.[26]Acceptable reliability was considered as ICC ≥ 0.80 and TE ≤ 10%, while good reliability was considered as ICC ≥ 0.90 and TE ≤ 5%, respectively.[27] The SWC was calculated as 0.2 of the between-athlete SD and is presented in percentage of the mean.[4] The signal to noise ratio was calculated as SWC/TE. To compare the reliability of the included performance variables across testing centers the ratio of the TE´s from the different centers was obtained according to the recommendation of Garcia-Ramos and Janicijevic.[16] The threshold to interpret a meaningful difference in the TE was 1.15.[16] CI´s for all analyses were set at 95%. All statistical analyses were performed using a customized Microsoft Excel spreadsheet.[26]

Results

The TE% ranged from 1.3% to 8.5%, for all the included assessments (Table 1). The change in mean ranged from −1.5% to 2.5% for all methods, whereas the ICCs ranged from 0.85 to 0.97, respectively (Table 1). The SWC ranged from 1.2% to 5.0%. The ratio between the typical error expressed as TE%, and the SWC in % ranged from 0.5 to 1.2 (Table 1 and Figure 1). The maximum range of the TE% ratio between testing centers was 0.44 – 1.44 (Table 2).

*** Table 1, 2 and 3 about here ***

*** Figure 1 about here ***

Discussion

The main finding of the present study was that all strength and power-related tests showed moderate to strong test-retest reliability (Table 1), which was true across all 4 test sessions (Table 1 and Table 3), and the 6 different test centers (Table 2). The CMJ height, jump & reach, sprint running, sprint cycling, and the leg press variables demonstrated good reliability (Table 1). Of all the assessments, only the sprint cycling, and the leg press test showed a signal-to-noise ratio greater than 1 (Table 1). Additionally, of the included measurements, only the SJ showed improved reliability from tests 1–2 compared with tests 3–4 (Table 3), indicating that the SJ test would potentially need more familiarization than the other tests.

Indeed, for the jump measures, there was generally better test-retest reliability for the CMJ compared with the SJ (Table 1 and Figure 1). Superior reliability for the former could simply be because initiating the jump with a countermovement is a more natural way to jump than the SJ[28], arguably leading to more stable measures. In fact, in our findings, the SJ was the only test showing improved reliability between tests 1–2 and 3–4 (see Table 3), indicating a learning effect for this particular movement. Thus, it seems as though the SJ needs to be included carefully in a test-battery, in conjunction with the familiarization of this movement in the respective group of athletes. Additionally, there might be small errors associated with SJ calculations from force platforms (using forward integration), as it is more challenging to achieve a steady stance in the bottom of the squat (SJ), compared to a steady stance in an upright position (CMJ).[4,29] Moreover, the test-retest reliability was better for jump height compared with the power output for both the SJ and CMJ, respectively (Table 1 and Figure 1). Superior reliability for jump height could be explained from previous observations that

have shown how the same jump height can be obtained with varying jump strategies within the same person[23], whereas this is not necessarily true for other variables such as power. To exemplify, McBride et al. showed how changing the depth of the CMJ caused ~5% variation in jump height and ~10% variation in power.[30] With regards to the jump & reach test, the TE% and signal-to-noise ratio were comparable to that of the SJ and CMJ heights (Table 1 and Figure 1), indicating that these tests provide similar reliability for jump height measures. Considering the availability and low cost of the associated equipment needed to measure jump height from the jump & reach test, compared with jump height measured from force platforms, these are noteworthy findings.

The power output measured in the sprint cycling and leg press test were the only variables showing a signal-to-noise ratio greater than 1. The superior reliability in these variables compared with the other measures in this study is probably caused by mainly two factors. First, in the sprint cycling and leg press test, the push-off distances and joint configurations were fixed, leading to superior standardization, which has been shown to affect power measures.[30] Secondly, the power output measured in the leg press test was calculated from a 10-repetition protocol, thus including several attempts[25], and the sprint cycling test included several maximal efforts within the 6 s. When performing multiple repetitions, the random errors inherent in each repetition tend to cancel each other out, causing beneficial effects with regards to reliability outcomes.[8] Indeed, the sprint running also included several maximal contractions, where a longer sprint resulted in a better signal-to-noise ratio (10-m = 0.7 vs 30-m = 1.0, Table 1), which is in accordance with previous findings.[14,22] For comparisons, the jump tests included less standardization in terms of joint configurations and only consisted of a few maximal attempts, whereas the latter was also true for 1RM squat. Consequently, the jump tests and 1RM squat had the worst signal-to-noise ratio of all the included tests (Table 1 and Figure 1).

The sprint running test had the lowest TE% of all the performance tests. This could possibly be caused by calculating percentages from time, rather than mechanical power or impulse, which would be closer to the actual biological variation in performance.[31] Indeed, this could explain the larger TE% for peak velocity compared with time-related variables for the sprint running test (Table 1 and Figure 1). Such considerations emphasize the necessity of presenting the SWC, as well as considering what factors underly the between-subject, as well as the between-session, variation in the measures used for performance assessments in athletes.

The different reliability outcomes across centers presented in Table 2 should be considered and are probably caused by several factors. For example, at center 1, the test-retest reliability was better than average for all the included tests, which is the center with the longest experience with athlete assessments. Other factors such as the experience of the athletes with performance assessments, and how attentive they are with regards to instructions given of e.g., not performing strenuous exercises prior to testing, might have influenced the results. Moreover, different equipment at the different testing centers could have impacted the results. On that note, only the force platforms varied across testing centers whereas the rest of the tests were performed with the same equipment. Even so, we acknowledge that the same type of equipment could yield different results across centers, due to e.g., calibration procedures. With regards to the leg press device, we have observed in a previous investigation that variables, such as power, can be used interchangeably across different testing centers within a range of ± 5%[25], which is within the range of this study (Table 1). To our knowledge, if the same type of force platforms or bicycles located at different testing centers could be used interchangeably is not known. There are no reasons to believe that this would yield meaningful errors with regards to the results in the present study, however, we do acknowledge that this potentially poses limitations to studies conducted with a multicenter

approach. Importantly, the current study design does not allow for discerning whether the reliability scores are caused by operator, equipment, and/or biological variation, and must therefore be interpreted accordingly.

Practical Applications

The present study provides novel information with regards to differences in test-retest reliability across several test sessions and training and testing centers, even when test procedures are identical. This highlights the importance of not only knowing the average TE from published research but also being aware of one's measurement reliability. However, some considerations should be made when interpreting the results from the present study.

Collecting data with a multicenter approach indicates assessing a larger range of participants than typically used in single center studies. Even though this strengthens the generalizability of the results[16], it is associated with some drawbacks from a statistical perspective. Assessing a larger range of participants usually increases the between-subject variability, which in turn could exaggerate the reliability interpreted from correlation measures such as the ICC.[16] Thus, collecting data with a multicenter approach could cause higher correlations without necessarily indicating better test-retest reliability[7,16], which is in accordance with some of our findings (Table 1). Therefore, as the TE are unaffected by between-subject variability, these should be preferred over correlation measures, when assessing test-retest reliability with a multicenter approach.[16]

The present study shows that commonly applied strength and power tests can be used to accurately quantify performance in high-performing athletes, which was also true across several test sessions and different testing centers. Practitioners and researchers can utilize the findings from this study to better interpret changes in strength and power-related variables. Note that the definition of power is test dependent[32], and has been defined differently between the included tests in this study. Hence, care must be taken when interpreting power outputs from the different tests in this study. Importantly, the results from the present study should not be interpreted as objective universal test-retest reliability measures. Rather they should be used to better inform decisions regarding which tests to use or gain an overview of typical measurement variations in commonly used strength and power measures, for high-performing athletes.

Conclusions

We advocate that most of the included lower body strength and power tests can be used with confidence by researchers and coaches to assess high-performing athletes. However, the SJ, CMJ (power), and the 1RM-squat tests should be used with caution. Indeed, the SJ variables and the 1RM squat test displayed the largest test-retest variations. CMJ height, jump & reach, 30-m sprint, sprint cycling, and the leg press test, all displayed good reliability with a TE < 5%. Of all the included tests, only the power output from the sprint cycling and the leg press test showed a signal-to-noise ratio greater than 1. Additionally, meaningful differences in reliability across testing centers were observed, while using identical test protocols. This highlights the importance of controlling testing reliability at each test center, and not relying on data from others despite having applied the same protocol.

Acknowledgments

The authors would like to thank the participating athletes for their time and effort. The authors declare no conflict of interest. No external funding was received for the present study. The

data that support the findings of this study are available from the corresponding author upon reasonable request.

References

1. Haugen T, Paulsen G, Seiler S, Sandbakk Ø. New Records in Human Power. *Int J Sports Physiol Perform.* 2018;13(6):678-686.
2. Redden J, Stokes K, Williams S. Establishing the Reliability and Limits of Meaningful Change of Lower Limb Strength and Power Measures during Seated Leg Press in Elite Soccer Players. *J Sports Sci Med.* 2018;17(4):539-546.

3. Haugen T, Hopkins W, Breitschädel F, Paulsen G, Solberg P. Fitness Tests and Match Performance in a Male Ice Hockey National League. *Int J Sports Physiol Perform.* 2021:1-8.
4. Lindberg K, Solberg P, Bjørnsen T, et al. Force-velocity profiling in athletes: Reliability and agreement across methods. *PLoS One.* 2021;16(2).
5. Sawczuk T, Jones B, Scantlebury S, et al. Between-Day Reliability and Usefulness of a Fitness Testing Battery in Youth Sport Athletes: Reference Data for Practitioners. *Measurement in Physical Education & Exercise Science.* 2018;22(1):11-18.
6. Ishida A, Travis SK, Stone MH. Short-Term Periodized Programming May Improve Strength, Power, Jump Kinetics, and Sprint Efficiency in Soccer. *J Funct Morphol Kinesiol.* 2021;6(2).
7. Hopkins W. Measures of reliability in sports medicine and science. *Sports Med.* 2000;30(1):1-15.
8. Hopkins WG, Schabort EJ, Hawley JA. Reliability of power in physical performance tests. *Sports medicine (Auckland, NZ).* 2001;31(3):211-234.
9. Thomas C, Dos'Santos T, Comfort P, Jones PA. Between-Session Reliability of Common Strength- and Power-Related Measures in Adolescent Athletes. *Sports (Basel).* 2017;5(1).
10. Grgic J, Lazinica B, Schoenfeld BJ, Pedisic Z. Test-Retest Reliability of the One-Repetition Maximum (1RM) Strength Assessment: a Systematic Review. *Sports Med Open.* 2020;6(1):31.
11. McMaster D, Gill N, Cronin J, McGuigan M. A brief review of strength and ballistic assessment methodologies in sport. *Sports Med.* 2014;44(5):603-623.
12. Thomas C, Jones PA, Comfort P. Reliability of the Dynamic Strength Index in college athletes. *Int J Sports Physiol Perform.* 2015;10(5):542-545.
13. Heishman AD, Daub BD, Miller RM, Freitas EDS, Frantz BA, Bemben MG. Countermovement Jump Reliability Performed With and Without an Arm Swing in NCAA Division 1 Intercollegiate Basketball Players. *J Strength Cond Res.* 2020;34(2):546-558.
14. Altmann S, Ringhof S, Neumann R, Woll A, Rumpf MC. Validity and reliability of speed tests used in soccer: A systematic review. *PLoS One.* 2019;14(8):e0220982.
15. Seo DI, Kim E, Fahs CA, et al. Reliability of the one-repetition maximum test based on muscle group and gender. *J Sports Sci Med.* 2012;11(2):221-225.
16. Garcia-Ramos A, Janicijevic D. Potential benefits of multicenter reliability studies in sports science: A practical guide for its implementation. *Isokinetics and Exercise Science.* 2020:1-6.
17. López-Segovia M, Pareja-Blanco F, Jiménez-Reyes P, González-Badillo JJ. Determinant factors of repeat sprint sequences in young soccer players. *Int J Sports Med.* 2015;36(2):130-136.
18. Haugen TA, Tønnessen E, Seiler S. Speed and countermovement-jump characteristics of elite female soccer players, 1995-2010. *Int J Sports Physiol Perform.* 2012;7(4):340-349.
19. Mendez-Villanueva A, Bishop D, Hamer P. Reproducibility of a 6-s maximal cycling sprint test. *J Sci Med Sport.* 2007;10(5):323-326.
20. Hopkins W. How to interpret changes in an athletic performance test. *Sportscience.* 2004;8:1-7.
21. Cormack JS, Newton UR, McGuigan RM, Doyle LT. Reliability of Measures Obtained During Single and Repeated Countermovement Jumps. *International Journal of Sports Physiology and Performance.* 2008;3:131-144.

22. Darrall-Jones DJ, Jones B, Roe G, Till K. RELIABILITY AND USEFULNESS OF LINEAR SPRINT TESTING IN ADOLESCENT RUGBY UNION AND LEAGUE PLAYERS. *Journal of strength and conditioning research.* 2015;30(5):1359-1364.
23. Moir GL, Garcia A, Dwyer GB. Intersession reliability of kinematic and kinetic variables during vertical jumps in men and women. *Int J Sports Physiol Perform.* 2009;4(3):317-330.
24. Helland C, Bojsen-Møller J, Raastad T, et al. Mechanical properties of the patellar tendon in elite volleyball players with and without patellar tendinopathy. *Br J Sports Med.* 2013;47(13):862-868.
25. Lindberg K, Eythorsdottir I, Solberg P, et al. Validity of Force-Velocity Profiling Assessed With a Pneumatic Leg Press Device. *Int J Sports Physiol Perform.* 2021:1-9.
26. Hopkins WG. Spreadsheets for Analysis of Validity and Reliability. *Sportscience.* 2017;21.
27. Koo KD, Li YM. A Guidline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of Chiropractic Medicine.* 2016;15:155-163.
28. Bobbert MF, Gerritsen KG, Litjens MC, Van Soest AJ. Why is countermovement jump height greater than squat jump height? *Med Sci Sports Exerc.* 1996;28(11):1402-1412.
29. Pérez-Castilla A, Rojas FJ, García-Ramos A. Assessment of unloaded and loaded squat jump performance with a force platform: Which jump starting threshold provides more reliable outcomes? *J Biomech.* 2019;92:19-28.
30. McBride JM, Kirby TJ, Haines TL, Skinner J. Relationship between relative net vertical impulse and jump height in jump squats performed to various squat depths and with various loads. *Int J Sports Physiol Perform.* 2010;5(4):484-496.
31. Hopkins WG, Hawley JA, Burke LM. Design and analysis of research on sport performance enhancement. *Med Sci Sports Exerc.* 1999;31(3):472-485.
32. Knudson VD. Correcting the Use of the Term «Power» in the Strength and Conditioning Literature. *Journal of strength and conditioning research.* 2009;23(6):1902-1902.

Figure captions

Figure 1 TE% and the SWC% for the CMJ, SJ, jump & reach, 30-m sprint, 1RM squat, sprint cycling test, and the leg press test, respectively

Abbreviations: TE%, typical error in percentage; SWC, smallest worthwhile change; CMJ, countermovement jump; SJ, squat jump; 1RM, one-repetition maximum. cm, centimeters; W, Watts; kg, kilograms; s, seconds; $m \cdot s^{-1}$, meters per second.

Tables

Table 1 Measures of reliability for variables obtained from the CMJ, SJ, jump & reach, 30-m sprint, 1RM squat, sprint cycling test, and the leg press test, respectively

Abbreviations: CMJ, countermovement jump; SJ, squat jump; 1RM, one-repetition maximum; Δ, change; TE, typical error; SWC, smallest worthwhile change; TE%, typical error in percentage; ICC, intraclass correlation coefficient; SD, standard deviation; CI, confidence interval; cm, centimeters; W, Watts; kg, kilograms; s, seconds; $m \cdot s^{-1}$, meters per second.

Table 2 Ratio for the TE% for testing centers 1 to 6, as well as the number of athletes, tested for each test at each center, respectively

Abbreviations: TE%, typical error in percentage; CMJ, countermovement jump; SJ, squat jump; 1RM, one-repetition maximum; cm, centimeters; W, Watts; kg, kilograms; s, seconds; $m \cdot s^{-1}$, meters per second.

Table 3 Measures of reliability for variables obtained from the CMJ, SJ, jump & reach, 30-m sprint, 1RM squat, sprint cycling test, and the leg press test, for testing timepoints 1 (test) and 2 (retest), and 3 (test) and 4 (retest)

Abbreviations: CMJ, countermovement jump; SJ, squat jump; 1RM, one-repetition maximum; Δ, change; TE, typical error; SWC, smallest worthwhile change; SMD, Standardized mean difference; ICC, intraclass correlation coefficient; SD, standard deviation; CI, confidence interval; cm, centimeters; W, Watts; kg, kilograms; s, seconds; $m \cdot s^{-1}$, meters per second.