Journal of Strength and Conditioning Research™

Statistical Tests for Sports Science Practitioners: Identifying Performance Gains in Individual Athletes

John R. Harry,[1] Jacob Hurwitz,[2] Connor Agnew,[3] and Chris Bishop[4]

[1]Department of Kinesiology & Sport Management, Texas Tech University, Lubbock, Texas [2]Department of Kinesiology, Mississippi State University, Starkville, Mississippi [3]Department of Athletics, Appalachian State University, Boone, North Carolina [4]Faculty of Science and Technology, London Sport Institute, Middlesex University, London, United Kingdom

Abstract

Harry, JR, Hurwitz, J, Agnew, C, and Bishop, C. Statistical tests for sports science practitioners: identifying performance gains in individual athletes. *J Strength Cond Res* 38(5): e264–e272, 2024—There is an ongoing surge of sports science professionals within sports organizations. However, when seeking to determine training-related adaptations, sports scientists have demonstrated continued reliance on group-style statistical analyses that are held to critical assumptions not achievable in smaller-sample team settings. There is justification that these team settings are better suited for replicated single-subject analyses, but there is a dearth of literature to guide sports science professionals seeking methods appropriate for their teams. In this report, we summarize 4 methods' ability to detect performance adaptations at the replicated single-subject level and provide our assessment for the ideal methods. These methods included the model statistic, smallest worthwhile change, coefficient of variation (CV), and standard error of measurement (SEM), which were discussed alongside step-by-step guides for how to conduct each test. To contextualize the methods' use in practice, real countermovement vertical jump (CMJ) test data were used from 4 (2 females and 2 males) athletes who complete 5 biweekly CMJ test sessions. Each athlete was competing in basketball at the NCAA Division 1 level. We concluded that the combined application of the model statistic and CV methods should be preferred when seeking to objectively detect meaningful training adaptations in individual athletes. This combined approach ensures that the differences between the tests are (a) not random and (b) reflect a worthwhile change. Ultimately, the use of simple and effective methods that are not restricted by group-based statistical assumptions can aid practitioners when conducting performance tests to determine athlete adaptations.

Key Words: data analysis, sports science, statistics

Introduction

Within sports organizations, there has been a concerted effort to form collaborative teams capable of evaluating the effectiveness of sports performance interventions using objective, scientific analysis methods. Typically, the main objective of this effort is to provide empirical assessments of a sports program's effectiveness by combining objective data with coaches' expertise to help guide actionable decisions. This has led to the development of the field known as "sports science" (16). Despite the benefits realized through this effort and the development of the larger field of sports science, there has been little advancement with respect to practices centering on individual athlete evaluations in place of team (i.e., group) average evaluations. Ultimately, practitioners and team-based sport scientists will continue to face challenges when seeking to determine whether athletes are adapting to training or related interventions. At face value, these challenges are due to the available literature to which practitioners can compare their athletes' physical performances because studies are largely limited to short-term, repeated-measures, and cross-sectional protocols that provide only mechanistic explanations of physical human performance, rely on samples of nonathlete populations, or both (1,28,32,44,59). In reality, multiple factors are to blame for the lack of ability to make actionable decisions from collected data during performance tests.

Our experiences indicate that the limited potential to make actionable decisions is predominantly due to 3 realities within the strength and conditioning community. First, in the United States, there have been very few vacancies within sports organizations that are appealing to qualified sports scientists, which forces untrained or inexperienced practitioners and scientists into such roles. Second, strength and conditioning coaches are often assigned to, or volunteer for, sports scientist roles despite a lack of training in areas concerned with research methodology, data management and analysis, and statistics. Third, formally trained scientists who assume roles in sport are predominantly exposed to statistical tests aimed at generalizing a sample's average result to the sample's larger population (26). Obviously, we cannot control the current state of the literature, personnel decisions, or the availability of appealing vacancies for adequately trained candidates in sports organizations. However, we can provide evidence-based and experience-based methodologies we have used or continue to use in real-world settings to help scientists and practitioners conduct appropriate analyses and obtain actionable test results.

Conventional statistical analyses limit practitioners to the assessment of a group's "average" response or adaptation, and this

Address correspondence to John R. Harry, john.harry@ttu.edu.

Supplemental digital content is available for this article. Direct URL citations appear in the printed text and are provided in the HTML and PDF versions of this article on the journal's Web site (http://journals.lww.com/nsca-jscr).

Journal of Strength and Conditioning Research 38(5)/e264–e272

practice is echoed in the sports science literature (4,17,18,34,50). However, replicated single-subject approaches have been used by some to detect changes, and this approach may provide the most value to sports scientists and practitioners (11,29,31,55), particularly high value for those working with smaller squads of athletes (e.g., basketball, soccer, volleyball, etc). An appropriate foundation of single-subject analyses in athletes is forming (11,29,31), helping to create the impetus needed to move away from group-average assessments, when necessary, without reliance on subjective visual inspection or trend analyses. Unfortunately, anecdotal evidence indicates that many practitioners currently employing replicated single-subject assessments rely on arbitrary thresholds for change (i.e., 5% or 10% improvement) as an indicator for real change. The next logical step to strengthen the foundation is to demonstrate methods that can be used to conduct high-quality objective assessments and provide empirical evidence for individual athletes' performance adaptations. The ultimate objective of the current article was intended to align with the National Strength and Conditioning Association's *Essentials of Sport Science* textbook, which highlights the importance of guiding necessary stakeholders in ways to quantitatively determine the effectiveness of their training or related interventions and subsequently make data-driven changes (14).

The purpose of this report was to summarize methods for practitioners to explore for their own purposes related to individual athlete assessments. We provide explanations for specific methods' selection compared with other options in addition to step-by-step guides to conduct each test for further exploration. In addition, we compare each method and provide our opinion for which method or combination of methods is most appropriate or ecologically relevant for practitioners working with athlete populations. To contextualize how each method can work in practice, we applied them to a small subset ($n = 4$) of real longitudinal data obtained from athletes competing in men's or women's basketball at the NCAA Division 1 level. Finally, we provide an editable Microsoft Excel worksheet that includes the calculations to assist practitioners in applying our recommended approach in practice.

Methods

Summary of the Exemplar Athletes, Performance Task, and Metrics Used

To demonstrate the similarities or differences among methods in detecting performance changes in athletes, we used data obtained across 5 biweekly test sessions from 2 female and 2 male NCAA Division 1 Basketball players. All were healthy, uninjured, and active members of an NCAA Division 1 Basketball program throughout the data collection period. Subjects provided written informed consent, and data were collected in accordance with the Declaration of Helsinki as approved by the Texas Tech University local Institutional Review Board. Although this report does not involve formal research methodology nor is it an "original research" article, it is a series of case examples using real data. Because of this, we felt compelled to acknowledge the ethical considerations for using real data in this report.

The countermovement vertical jump (CMJ) was selected as the test activity, with ground reaction force data obtained during testing. The CMJ was used for this report because it is commonly used in research when seeking to understand physical ability among athlete populations (3,20,28,44,49,58). The CMJ was also selected because it (and related jumps) is performed frequently during competitive play in basketball (27,48) and strongly associated with sport-specific qualities, such as speed, strength, and agility (4,45,49,51). In addition, CMJ tests are routinely performed in laboratory and practitioner settings where multiple trials are collected for each athlete, thereby satisfying the requirements of each method (see below). The modified reactive strength index (RSI_{MOD}) and vertical jump height were included as primary and secondary CMJ performance metrics, calculated as center of mass flight height and the ratio of vertical jump height and time to takeoff, respectively (30). These metrics were selected according to a recent framework produced to guide practitioners in the selection of useful metrics to examine CMJ abilities (12). This is because RSI_{MOD} is a valid and reliable surrogate for athletic explosiveness (43), and RSI_{MOD} seems to be influenced primarily by jump height and not time to takeoff (32).

Comparison Among Methods at Detecting "Change"

Performance changes from session to session were determined using the model statistic, SWC, CV, and SEM methods described previously. Table 2 provides a summary of the cumulative increases of performance detected by each method (i.e., 4 possible changes per method of analysis, per athlete, resulting in 16 possible changes). When detecting increases in RSI_{MOD}, the most sensitive methods were the SWC and CV methods, with both detecting an increase in performance between the test sessions for 38% (6 out of 16) of the total possible comparisons (Table 2). The model statistic and SEM methods were more conservative, detecting increases in performance between the test sessions for 25% (4 out of 16) and 19% (3 out of 16) of the total possible comparisons, respectively (Table 2). For jump height, the most sensitive method was SWC, with increases detected during 50% (8 out of 16) of the total number of comparisons (Table 2). The next most sensitive method was the CV method, detecting increases in jump height for 44% (7 out of 16) of the total number of comparisons (Table 2). The model statistic method detected increases in jump height for 31% (5 out of 16) of the number of comparisons, whereas the SEM method detected increases for only 6% (1 out of 16) of the total number of comparisons (Table 2).

Importantly, the 4 methods were largely inconsistent with respect to detecting performance increases for the same comparisons, as the methods were consistent during only ~13% (4 out of 32) of the total possible comparisons across the 4 athletes (Figures 1–4). The reason the SWC detected a much greater number of performance gains versus all other methods is that the equation uses only a portion of the athlete's variation (20% for the 0.2 constant; 60% for the 0.6 constant). This creates a scenario in which the SWC will inherently detect a greater number of performance changes than the other methods, and ultimately, the risk of false "gains" is greatest. As such, the methods should not be used interchangeably, nor should reports of change be compared between or among assessments with different methods for detecting change.

Motivation for Quantitatively Assessing Individual Athletes

It is beyond the scope of this report to discuss in-depth the limitations of group-level statistical testing (e.g., *t* test, analysis of variance [ANOVA], etc.) in team-based settings because this has been done elsewhere (6,8,9,22,33). The key point for practitioners is that group-level testing requires a normal distribution where the

Table 1

Critical values to determine the critical difference for the model statistic technique.*

Trial size	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.01$
3	1.3733	1.6533	2.2133
4	1.2643	1.5058	1.9867
5	1.1597	1.3662	1.7788
6	1.0629	1.2408	1.6044
7	0.9751	1.1306	1.4623
8	0.8960	1.0351	1.3473
9	0.8270	0.9536	1.2542
10	0.7673	0.8857	1.1776
11	0.7172	0.8307	1.1129
12	0.6757	0.7867	1.0581
13	0.6415	0.7516	1.0117
14	0.6132	0.7234	0.9720
15	0.5896	0.7001	0.9375
16	0.5695	0.6798	0.9070
17	0.5522	0.6618	0.8796
18	0.5371	0.6458	0.8548
19	0.5237	0.6311	0.8318
20	0.5114	0.6175	0.8102
25	0.4592	0.5572	0.7145
30	0.4194	0.5097	0.6437
35	0.3896	0.4729	0.5949
40	0.3673	0.4442	0.5626
45	0.3500	0.4207	0.5414
50	0.3352	0.4000	0.5256

*Trial size: the number of trials used to calculate the performance means for the test sessions; α: the alpha level (i.e., statistical probability criterion) or the probability that a difference between the sessions is random.

data from the sample (i.e., the athletes) are generalized before analysis. Because athlete samples are typically quite small, it is unlikely that a normal distribution will occur, leading to issues related to the sample *SD* and *SE* (13). Moreover, group-level approaches with smaller samples are limited to the interpretations based on the sample's "average" response or adaptation, which is a well-known misrepresentation of the individuals from which the group average is obtained (8). Replicated single-subject approaches (11,29,31,55) are ideal for smaller-sample settings where the individuals are not represented by the group average. However, further examples are needed to help practitioners move away from the isolated use of group-average assessments, when necessary, without reliance on subjective visual inspection or trend analyses for individual assessments. Consequently, the next logical step is to demonstrate methods that can be used to conduct high-quality assessments and provide empirical evidence for individual athletes' performance adaptations.

Importantly, there are multiple methods at one's disposal to select from, which can be used to quantitatively explore individual athlete responses. However, some of those methods might not be ideal, although they can be used at the individual level. Some of these methods, albeit less common in the contemporary literature, include nonparametric assessments such as the Mann-Whitney *U* Test (9), bootstrapping (25), and multiple regression (24). Other methods more commonly observed in the contemporary literature include assessing the percentage of nonoverlapping data points (38), counting the number of data points above a specific threshold (41), confidence intervals or effect sizes (52,53), and statistical process control (56). An advantage to these latter approaches is that they provide adequate scientific rigor because means and standard deviations of increasing or decreasing sequential data are treated in consideration for their slope. This is important when comparisons are made between very different mean values (e.g., 10 vs. 100) with similar amounts of normalized variation (i.e., 10%). The methods have additional value when seeking to determine whether an athletes' current test result is different from a previous series of tests or whether the results of one training period are different from those of a subsequent training period. However, there are consequences of these methods. First, they can involve between-subject metrics, notably the between-subject *SD*, which minimizes the fluctuations or variability among individuals within the team or group (8). Second, calculating rolling averages across several test sessions minimized the movement variations (i.e., strategies) that dictate an athlete's performance outcomes during a test session when the test sessions are pooled together. As described by Bates (6), all measurement outcomes are dependent on the state of the organism interacting with the environment at a specified moment in time. It is our opinion that individual athlete assessments must seek to account for the uniqueness of individual athlete strategy, as the available number of strategies changes over time because of ongoing changes in biomechanical or mechanical, morphological, and environmental constraints (6,35). We feel that this is best accomplished by evaluating athletes' performance changes between consecutive test sessions. Those who may be interested in procedures that involve between-subject metrics or rolling averages are referred to the previous literature (36,57,64).

The quantitative approaches we feel have the most potential to reveal test-to-test performance changes include single-subject "significance" testing [i.e., the model statistic] and comparing the magnitude of change against the smallest worthwhile change (SWC), the standard error of measurement (SEM), or the coefficient of variation (CV) (10,29,31,61,65,66). The usefulness of each of these methods is that all were designed for, or can be relatively simple to apply in, single-subject analyses. In addition, all are reasonably simple to understand, perform, and interpret by practitioners, regardless of prior statistical training. Most importantly, each method requires multiple testing efforts (i.e., trials) to be included for each comparison, which discourages the common practice of including only the "best" effort or only very few trials (42) from the test sessions. Including multiple trials is a critical component of the single-subject methodology (6,33) to reduce the variation in an athlete's data, thereby increasing statistical power (7) and provide stable performance data, which better reflects an athlete's true performance capabilities (40). Therefore, the resulting outcome for the presence of a change is obtained with consideration for both the absolute change of performance and the

Table 2

Performance increases detected by each method across athletes.*

	Model statistic		SWC		CV%		SEM	
Metric	Increases	% Total	Increases	% Total	Increases	% Total	Increases	% Total
RSI_{MOD}	4	25	8	50	6	38	8	50
Jump height	5	31	10	63	6	38	10	63

*RSI_{MOD} = modified reactive strength index; increases = number of increases detected across all 4 athletes (decreases excluded); % total = percentage of increases detected relative to the total number of comparisons; SWC = smallest worthwhile change method; CV% = coefficient of variation method, SEM = standard error of measurement method.

Table 3

Performance improvement assessments across 5 test sessions and 4 athletes.*

Athlete	Test 1		Improvement			Test 2		Improvement			Test 3		Improvement			Test 4		Improvement			Test 5		Improvement		
	Mean	*SD*	MS	CV	Both	Mean	*SD*	MS	CV	Both	Mean	*SD*	MS	CV	Both	Mean	*SD*	MS	CV	Both	Mean	*SD*	MS	CV	Both
													RSI_{MOD}												
F1	0.393	0.016	N/A	N/A	N/A	0.547	0.043	TRUE	TRUE	**TRUE**	0.389	0.010	FALSE	FALSE	FALSE	0.560	0.030	TRUE	TRUE	**TRUE**	0.504	0.113	FALSE	FALSE	FALSE
F2	0.337	0.007	N/A	N/A	N/A	0.367	0.051	FALSE	TRUE	FALSE	0.346	0.031	FALSE	FALSE	FALSE	0.327	0.010	FALSE	FALSE	FALSE	0.342	0.005	TRUE	TRUE	**TRUE**
M1	0.504	0.028	N/A	N/A	N/A	0.553	0.043	FALSE	TRUE	FALSE	0.688	0.058	TRUE	TRUE	**TRUE**	0.595	0.021	FALSE	FALSE	FALSE	0.532	0.160	FALSE	FALSE	FALSE
M2	0.777	0.126	N/A	N/A	N/A	0.536	0.012	FALSE	FALSE	FALSE	0.524	0.093	FALSE	FALSE	FALSE	0.526	0.076	FALSE	FALSE	FALSE	0.576	0.058	FALSE	FALSE	FALSE
Group	0.503	0.065	N/A	N/A	N/A	0.501	0.040	FALSE	FALSE	FALSE	0.487	0.057	FALSE	FALSE	FALSE	0.502	0.042	FALSE	FALSE	FALSE	0.489	0.102	FALSE	FALSE	FALSE
													Jump height												
F1	0.248	0.001	N/A	N/A	N/A	0.306	0.012	TRUE	TRUE	**TRUE**	0.260	0.012	FALSE	FALSE	FALSE	0.301	0.012	TRUE	TRUE	**TRUE**	0.301	0.008	FALSE	FALSE	FALSE
F2	0.254	0.010	N/A	N/A	N/A	0.275	0.007	TRUE	TRUE	**TRUE**	0.263	0.013	FALSE	FALSE	FALSE	0.277	0.003	FALSE	FALSE	FALSE	0.265	0.012	FALSE	FALSE	FALSE
M1	0.319	0.010	N/A	N/A	N/A	0.339	0.009	TRUE	TRUE	**TRUE**	0.374	0.019	TRUE	TRUE	**TRUE**	0.355	0.003	FALSE	FALSE	FALSE	0.378	0.030	FALSE	TRUE	FALSE
M2	0.452	0.012	N/A	N/A	N/A	0.424	0.027	FALSE	FALSE	FALSE	0.405	0.029	FALSE	FALSE	FALSE	0.412	0.047	FALSE	FALSE	FALSE	0.438	0.008	FALSE	FALSE	FALSE
Group	0.318	0.009	N/A	N/A	N/A	0.336	0.016	FALSE	TRUE	FALSE	0.326	0.020	FALSE	FALSE	FALSE	0.336	0.024	FALSE	FALSE	FALSE	0.346	0.017	FALSE	FALSE	FALSE

*RSI_{MOD} = modified reactive strength index; F1 = female athlete 1; F2 = female athlete 2; M1 = male athlete 1; M2 = male athlete 2; MS = model statistic method; CV = coefficient of variation method; Both = combined approach using model statistic and CV methods; TRUE = change detected by associated method; FALSE = change not detected by associated method; When TRUE is contained in bold, both methods detected change.

potential variations among trials within the test (i.e., consistency of each individual's result), which can mask changes when not accounted for (8). Importantly, accounting for variation means the performance result for each athlete's test session(s) considers their biomechanical constraints and unique response patterns to environmental feedback, which ultimately determine bodily movement and the amount of performance variability (5,35).

Overview of Session Versus Session Analysis Methods

Model Statistic

The model statistic technique is a critical difference method that can be loosely considered a single-subject dependent *t*-test, whereby the observed difference between the sessions is compared with a probabilistic critical difference (7,9). It was designed in the early 1990s by Bates et al. (7) and has been used to demonstrate the value of single-subject comparisons between 2 conditions relative to the group-level equivalent (31). Critical values were generated (7) for selected trial sizes (i.e., the number of trials used to calculate the test session average) and statistical probabilities (i.e., alpha levels; α), which are provided in Table 1. Therefore, the final decision from the test indicates the probability for whether the difference between test sessions was due to random chance, using the user's a priori choice among 10, 5, or 1% probability levels. A unique feature of the model statistic is that it does not calculate interval limits as 1 SD away from the mean and instead incorporates the weighted mean *SD* (7), which can also be described as the variation in the collective number of trials or observations used in the comparison (7,21). Ultimately, the critical values and ultimately the critical difference score are analogous to a $1.96 \times SD$ interval where the critical difference is a cutoff from which 95% of scores are between $\pm$ the critical difference. When test sessions with different trials sizes are compared, we recommended using the critical value associated with the smallest trial size because the test is more conservative when smaller trial sizes are used (i.e., more difficult to return a difference that is not due to chance). The main limitation to the model statistic is that the critical value table was created using vertical ground reaction force data obtained during the support phase of running. It may be that other types of data or other tasks might yield different critical values.

Traditional paired-samples *t*-test and effect sizes could be used in a similar way, but we favor the model statistic for a few reasons. First, *t*-tests are inappropriate for repeated test assessments (e.g., more than 2 comparative tests) without corrections for familywise error and require a minimum number of samples (i.e., trials in this case) for adequate statistical power (7), and the observations that make up the comparative means are subject assumptions related to frequency distributions and homogeneity of variance (63). Satisfying these assumptions and related criteria is not typically achievable in athlete testing environments. For effect sizes, clinical data suggest that a similar rate of "differences" will occur in comparison to the model statistic (23). However, there are lingering problems related to effect sizes. First, there is no similar evidence suggesting a similar rate of difference among athlete populations compared with the model statistic. Second, effect sizes are limited by the need to select a magnitude threshold for interpretation based on subjective scales (19,37,54), created mostly from recreationally active samples, or based on athlete training level and not primary sport or training stimulus (54). As reducing subjectivity and guesswork is critical for the applied sports scientist, we feel that effect sizes would not be appropriate in this type of scenario. Comparisons between the tests for an

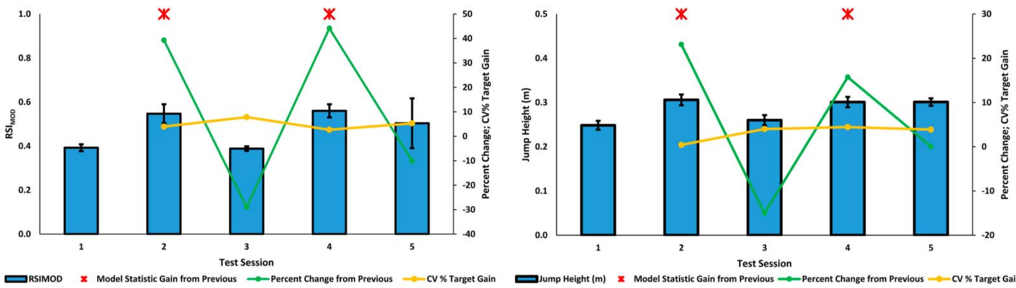

Figure 1. RSI_{MOD} (top row) and jump height (bottom row) performance results and differences detected by the model statistic and CV methods for female athlete 1. Data are presented as mean ± 1 *SD* across 3 trials for each test session; model statistic change from previous: nonrandom ($p < 0.05$) change between the adjacent test sessions; CV% target gain: threshold that must be exceeded to indicate change between the adjacent test sessions. RSI_{MOD} = modified reactive strength index; CV = coefficient of variation.

individual athlete can be performed with the model statistic technique using the following procedural steps:

1. Calculate the absolute mean difference between pretest (X_1) and posttest (X_2) sessions, ensuring a minimum of 3 trials for each test
   Mean difference = $|X_1 - X_2|$
2. Calculate the mean SD from the pretest and posttest values (SD_1 and SD_2)

$$SD_{Mean} = \left[\left(SD_1^2 + SD_2^2\right)/2\right]^{1/2}$$

3. Multiply desired critical value (i.e., test statistic) from Table 1 by the mean SD calculated in step 2, which produces a critical difference
   Critical difference = Table 1 value × SD_{Mean}
4. Compare the mean difference between the tests with the critical difference
   a. Significant difference: mean difference > critical difference
   b. Nonsignificant difference: mean difference ≤ critical difference

Smallest Worthwhile Change

The SWC approach is a form of magnitude-based inference (64) that was pioneered as an attempt to provide a more realistic assessment of performance adaptations when compared with traditional statistical tests (37). The SWC has greater specificity to sports science data sets compared with conventional Fisher-based probability tests, which, as mentioned, tend to include small sample (or trial) sizes, and have relatively higher variations among measurements; sports scientists may be less concerned with whether a change is or is not due to unexplainable chance (15,47). A constant of 0.2 is used to establish the SWC threshold (see below) for trained populations or athletes (60). This is because it aligns with the commonly accepted, albeit subjective, definition for a "small" magnitude difference or effect size (19). A constant of 0.6 can be used in the SWC calculation for untrained populations or youth athletes because large adaptations can be realized in those populations following initial or short-term periods of training (47). Although the selection of the SWC constant could be objectively calculated (46) for a specific sample or athlete, we elected to use what we determined from the sports science literature to be the most common SWC approach. Although the SWC approach is typically used at the group level, it can be easily applied at the replicated single-subject level. Comparisons between the tests for an individual athlete can be performed with the SWC approach using the following procedural steps, which are slightly modified for the single-subject assessment:

1. Calculate the athlete's mean performance display for the pretest (X_1) and posttest (X_2) sessions
2. Calculate the change from X_1 to X_2
3. Calculate the SD across all test sessions (pretest and posttest sessions; referred to here as SD_{Global})
4. Calculate the SWC as follows:
   SWC = 0.2 × SD_{Global}
5. Compare the athlete mean difference to the SWC
   a. True difference: $X_2 > X_1$ + SWC
   b. Trivial difference: $X_2 \leq X_1$ + SWC

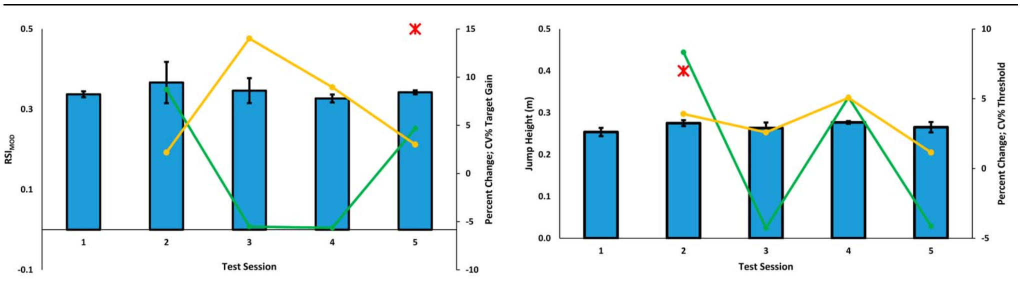

Figure 2. RSI_{MOD} (top row) and jump height (bottom row) performance results and differences detected by the model statistic and CV methods for female athlete 2. Data are presented as mean ± 1 *SD* across 3 trials for each test session. Model statistic change from previous: Nonrandom ($p < 0.05$) change between the adjacent test sessions; CV% minimum increase: threshold that must be exceeded to indicate change between the adjacent test sessions. RSI_{MOD} = modified reactive strength index; CV = coefficient of variation.

Coefficient of Variation

When comparing the percentage change between the test sessions with the CV, the objective is to determine whether performance changes or differences between the conditions are greater than the variation in the test (10). Therefore, this technique is not a probabilistic test and instead reveals whether the observed difference exceeds the "noise" (as represented by the CV) inherent in the results. Rather than using the mean difference between the pretest and posttest, the percentage difference is typically used and compared with the CV. This may be particularly beneficial because the CV is a standardized metric that is a measure of reliability, which enables the formation of a target or, if desired, a boundary used to detect positive and negative changes (60,62). A benefit of this is that it may be more feasible to explain to athletes, coaches, or other stakeholders "how" differences are determined for each comparison. In addition, the CV approach can be obtained such that the unit of measure for the test is retained and still tell the same story as when converted to a percentage value, which can simplify interpretation for some practitioners, stakeholders, or both. Although the procedure outlined here uses 1 SD to calculate the CV, 1.5 or 2 SDs could also be used to expand the "range of scores," which could be useful when seeking to modify the sensitivity of the test and account for what is quantified by each metric's interval. This should not be confused with similar processes to compare data from different numeric scales (64). Rather, it is a way to control the outcome sensitivity for performance tests known to have greater or lesser movement variability. Comparisons between the tests for an individual athlete can be performed with the CV technique using the following procedural steps:

1. Calculate the athlete's mean performance display for the pretest (X_1) and posttest (X_2) sessions
2. Calculate the athlete's SD for the pretest (SD_1) session
3. Calculate the percentage change from X_1 to X_2
   $\% \text{ Change} = 100 \times (X_2 - X_1)/X_1$
4. Calculate the CV for the pretest (CV_1) session
   $CV_1 = 100\% \times SD_1/X_1$
5. Compare the percentage change between pretest and posttest sessions to CV_1
   a. True difference: % change $> CV_1$
   b. Trivial difference: % change $\leq CV_1$

Standard Error of the Measurement

For the SEM method, the objective is to compare the change in performance to the "noise," like the CV method, with data inherently retained in the original unit of measurement. This means that the absolute mean difference (i.e., change of performance) is used in conjunction with the precision of the test data. The SEM has been described as the intraindividual version of the *SD* (13), but there seems to be lesser practical application of SEM because assessments of *individual* performance changes have historically incorporated the *SD* (31). Although there are 2 common formulae used to calculate the SEM (2), the inability to obtain an intraclass correlation coefficient (ICC) at the single-subject level requires that the SEM is calculated as the square root of the mean square error (MSE), with some modifications to align with the single-subject data set. This process may be more complicated to some than using the ICC. However, it could be beneficial because it avoids the uncertainties connected to the ICC and allows for more consistency (65) from test to test. Procedural steps for using the SEM method are as follows:

1. Calculate the athlete's mean performance display for the current test (X_1) and the mean of the 2 test sessions being compared (X_{Total})
2. Calculate the sum of squares (SS)
   $SS = (X_1 - X_{Total})^2$
3. Calculate the MSE using the SS and degrees of freedom (*df*; number of trials − 1), where the number of trials equals the sum of the number of trials recorded across sessions used for step 1
   $MSE = SS/df$
4. Calculate the SEM
   $SEM = \sqrt{MSE}$
5. 3Compare the athlete mean difference to the SEM
   a. True difference: $X_2 > X_1 + SEM$
   b. Trivial difference: $X_2 \leq X_1 + SEM$

Discussion

The risk of false-positive outcomes should come into play from the perspective of determining an actionable change. Therefore, it is our opinion that a test for change in team settings includes 3 key components. First, the test should be objective to eliminate subjective estimations or guesses. Second, it should it err on the side of conservativeness when determining whether the difference between 2 tests is random or likely to be legitimate to avoid erroneous interpretations. Third, the test should provide a simple way to determine whether a difference between 2 tests is meaningful to the athlete, practitioner, or related stakeholder. According to these requisite components and the benefits, limitations, or both discussed for each method, we recommend the model statistic and CV

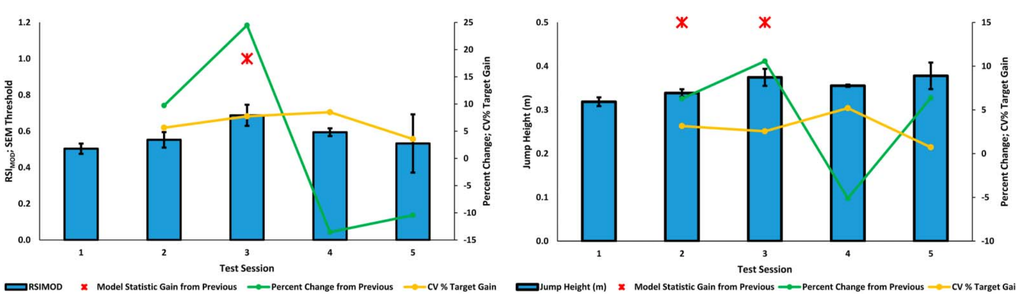

Figure 3. RSI_{MOD} (top row) and jump height (bottom row) performance results and differences detected by the model statistic and CV methods for male athlete 1. Data are presented as mean ± 1 *SD* across 3 trials for each test session. Model statistic change from previous: nonrandom ($p < 0.05$) change between the adjacent test sessions; CV% minimum increase: threshold that must be exceeded to indicate change between the adjacent test sessions. RSI_{MOD} = modified reactive strength index; CV = coefficient of variation.

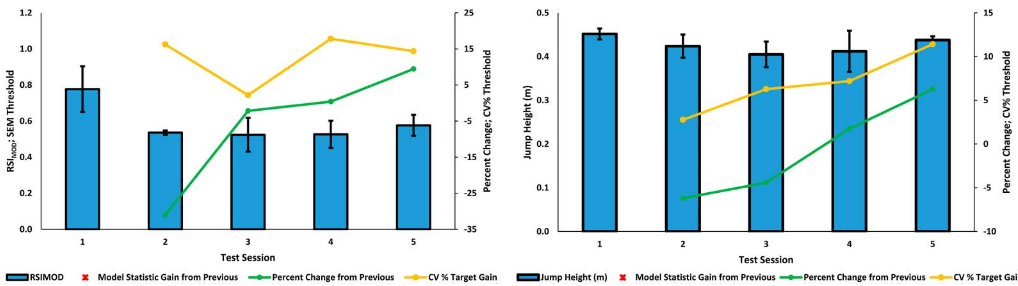

Figure 4. RSI$_{MOD}$ (top row) and jump height (bottom row) performance results and differences detected by the model statistic and CV methods for male athlete 2. Data are presented as mean ± 1 *SD* across 3 trials for each test session. Model statistic change from previous: nonrandom ($p < 0.05$) change between the adjacent test sessions; CV% minimum increase: threshold that must be exceeded to indicate change between the adjacent test sessions. RSI$_{MOD}$ = modified reactive strength index; CV = coefficient of variation.

approaches to be used in parallel. This combined use of approaches will provide practitioners with the ability to identify actionable changes in their athletes. Furthermore, it eliminates the likelihood for decisions to be made based on the presence of potentially random differences that seem valuable or nonrandom differences of little value. This "2-pronged" comparative approach should be familiar to sports science practitioners, and its value to practitioners and key stakeholders can be reviewed elsewhere (37).

Table 3 shows the cumulative increases of performance based on the model statistic, CV, and combined approach (the table is also available as an editable Microsoft Excel file, provided with the article as Table 1, Supplemental Digital Content 1, http://links.lww.com/JSCR/A462). The fact that the model statistic and CV approaches did not identify athlete-specific differences with the same pattern supports the potential use of the 2 methods in parallel. The reason for the different patterns of differences between the approaches relates to the objectives for which the tests were designed. From the sports scientist's perspective, the model statistic indicates whether the test results are *legitimately* different due to its conservativeness, whereas the CV indicates the *meaningfulness* of the difference between the test results. As such, they are complementary methods as opposed to the ones that should be compared against one another.

The recommendation for using the model statistic may be obvious because it is the only method providing whether the difference between the tests is statistically significant like convention group analyses (e.g., *t*-test, ANOVA). However, the recommendation for parallel use of the CV method may be less clear because the CV, SEM, and SWC methods incorporate the session variation or the athlete's consistency across trials performed with the session(s). Our position for omitting the SWC relates to its reliance on subjectivity when determining the threshold for importance of a change. Moreover, the SWC's use of 20% or 60% of the athlete's variation makes it excessively sensitive, as mentioned previously, which is not ideal for athlete populations with a wide range of responses to training stimuli.

The main limitation to the SEM method is that it involves modified equations to permit its use at the individual athlete level. In addition, the SEM method returned the same number of differences as the SWC method, albeit with a similar yet not identical pattern of differences (not shown in tables or figures). Thus, there may be an increased risk for erroneous conclusions using the SWC or SEM methods, in which a certain number of significant performance changes from the model statistic would coincide with an inflated number of "meaningful" changes versus the CV method. In turn, practitioners using SWC or SEM might be motivated to conclude that an athlete has demonstrated a positive adaptation when there was no adaptation, or the adaptation was trivial.

One final point to note relates to the level of influence a practitioner or sports scientist places on the results of a single or repeated performance tests because the importance of any test can vary among practitioners, athletes, or both depending on training objectives. For context, Figures 1–4 demonstrate our individual athletes' CMJ performance changes across 5 sessions (i.e., 10 weeks) from the start of full-time training until approximate start of the competition season. The nature of the training interventions for each athlete, which was somewhat unique to each individual, was centered on progressive changes in training volume and intensity for exercises to continue improving "explosive strength" (67). As mentioned, we operationally defined "increased explosive strength" as an increase of RSI$_{MOD}$, with jump height used as a secondary performance metric to help explain changes in RSI$_{MOD}$. We use this approach because of previous work demonstrating that RSI$_{MOD}$ is a valid and reliable surrogate for athletic explosiveness (43) in addition to our work suggesting that increased RSI$_{MOD}$ is influenced primarily by jump height and not its other component part, time to takeoff (32). As such, if jump height was not an adequate explanatory metric for RSI$_{MOD}$ changes, it would mean that the change was primarily due to altered times to takeoff and that would be considered in any recommendations provided. Thus, all athletes were expected to realize CMJ performance gains at each test session. If a change was not observed on a given test day, we would reflect on the results and other contributing factors to overall workload, such as physical training, on-court work, test-day fatigue, or athlete readiness. Those data will not be discussed here but can be explored in part elsewhere (39). According to those reflections, recommendations were given to the strength and conditioning staff to decide whether training or related changes were needed.

For female athlete 1 (Figure 1), the session-to-session changes in RSI$_{MOD}$ were similar to their changes in jump height. For the purposes of this report, we will focus on the way in which we use CMJ results to make data-driven training changes. In particular, the large decreases in RSI$_{MOD}$ and jump height from test session 2 to 3 were concerning. A member of the sports science team provided the strength and conditioning staff with recommendations, and they decided whether training or related workload modifications were appropriate. The changes they implemented were shown to be successful according to the athlete's CMJ results at test session 4, which were both statistically significant (model statistic) and meaningful (CV). For female athlete 2, the CMJ test results were not concerning enough to recommend specific changes or considerations between test sessions 1 and 5, although there was a statistically significant (model statistic) and meaningful (CV) increase in RSI_{MOD} at test session 5.

It is important to note that our training, recommendation-based changes, or both, do not always return positive changes, as shown between test sessions 4 and 5 for male athlete 1 (Figure 3). For instance, this athlete demonstrated somewhat positive results from test session 1 to test session 2 because RSI_{MOD} was shown to meaningfully increase (CV), whereas jump height was shown to increase significantly (model statistic) and meaningfully (CV). This performance further improved from test session 2 to test session 3 because RSI_{MOD} and jump height increased significantly (model statistic) and meaningfully (CV). However, there was an alarming drop in performance, as evidenced by decreases in both RSI_{MOD} and jump height. Although data-driven recommendations were presented and training or related modifications were prescribed, the athlete did not display statistically significant (model statistic) or meaningful (CV) improvements in RSI_{MOD} at test session 5 because of substantial variation across the session 5 trials. Further to this, male athlete 2 displayed concerning results from test session 1 to test session 2, where a ~30% decrease in RSI_{MOD} occurred. As the RSI_{MOD} decrease appeared to be primarily driven by time to takeoff because of the <5% decrease in jump height, specific recommendations and training or related modifications were prescribed specific to the athlete's display. Interestingly, the athlete's subsequent results at test session 3 did not stimulate targeted change in RSI_{MOD} or jump height. This trend continued through test session 5, suggesting that the athlete was, for some reason, more resistant to the changes we were recommending or the strength and conditioning staff were prescribing. Although detailed exploration into those reasons is beyond the scope of this report, it did provide insight into ways test results could be used to consider or better adjust training.

A final point for this section relates to our objectives during the 10-week test period presented because this approach was specific to the groups of athletes and the goals established for those athletes. Although the overarching objective for this subset of athletes and most athletes in the basketball programs is to increase "explosive strength," the way in which that is stimulated varies across seasons, training blocks, and athletes. This is why the specific training programs, recommendations, and complimentary data are not discussed because it would be of little relevance to the reader. The objective for this practical application section was to show *when* test data reveal performance changes that are objective in nature and should stimulate some consideration as to whether a modification is needed for training or some related aspect of the overall program.

Practical Applications

We summarized 4 methods, specifically the model statistic, SWC, CV, and standard error of the measurement (SEM), with respect to detecting individual athletes' change during performance tests, using the CMJ as the example test. We provide support for our recommendation that the combined use of the model statistic and CV should be preferred when seeking to objectively detect *real* and *important* training adaptations in individual athletes. We applied our recommended approach to a small subset of real data obtained in 4 different athletes, competing in men's or women's basketball at the NCAA Division 1 level, to contextualize how these methods can work in practice, highlighting when we would or would not recommend or prescribe a training-related modification.

References

1. Adams K, O'Shea JP, O'Shea KL, Climstein M. The effect of six weeks of squat, plyometric and squat-plyometric training on power production. *J Strength Cond Res* 6: 36–41, 1992.
2. Atkinson G, Nevill AM. Statistical methods for assessing measurement error (reliability) in variables relevant to sports medicine. *Sports Med* 26: 217–238, 1998.
3. Barker LA, Harry JR, Mercer JA. Relationships between countermovement jump ground reaction forces and jump height, reactive strength index, and jump time. *J Strength Cond Res* 32: 248–254, 2018.
4. Barnes JL, Schilling BK, Falvo MJ, et al. Relationship of jumping and agility performance in female volleyball athletes. *J Strength Cond Res* 21: 1192–1196, 2007.
5. Bates BT. Scientific basis of human movement. *J Phys Educ Recreat* 48: 68–75, 1977.
6. Bates BT. Single-subject methodology: An alternative approach. *Med Sci Sports Exerc* 28: 631–638, 1996.
7. Bates BT, Dufek JS, Davis HP. The effect of trial size on statistical power. *Med Sci Sports Exerc* 24: 1059–1065, 1992.
8. Bates BT, Dufek JS, James CR, Harry JR, Eggleston JD. The influence of experimental design on the detection of performance differences. *Meas Phys Educ Exerc Sci* 20: 200–207, 2016.
9. Bates BT, James CR, Dufek JS. Single subject analysis. In: *Innovative Analyses of Human Movement*. Stergiou N, ed. Champaign, IL: Human Kinetics, 2004. pp. 3–28.
10. Bishop C, Abbott W, Brashill C, Read P. *Effects of Strength Training on Bilateral and Unilateral Jump Performance, and the Bilateral Deficit in Premier League Academy Soccer Players*. UK Strength and Conditioning Association, 2021.
11. Bishop C, Lake J, Loturco I, et al. Interlimb asymmetries: The need for an individual approach to data analysis. *J Strength Cond Res* 35: 695–701, 2021.
12. Bishop C, Turner A, Jordan M, et al. A framework to guide practitioners for selecting metrics during the countermovement and drop jump tests. *Strength Cond J* 44: 95–103, 2021.
13. Bland JM, Altman DG. Measurement error. *BMJ* 312: 1654, 1996.
14. Brewer C. Performance interventions and operationalizing data. In: *NSCA's Essentials of Sport Science*. French DN, Torres Ronda L, eds. Champaign, IL: Human Kinetics, 2022. pp. 339–351.
15. Buchheit M. Chasing the 0.2. *Int J Sports Physiol Perform* 11: 417–418, 2016.
16. Burwitz L, Moore PM, Wilkinson DM. Future directions for performance-related sports science research: An interdisciplinary approach. *J Sports Sci* 12: 93–109, 1994.
17. Castagna C, Castellini E. Vertical jump performance in Italian male and female national team soccer players. *J Strength Cond Res* 27: 1156–1161, 2013.
18. Cohen DD, Restrepo A, Richter C, et al. Detraining of specific neuromuscular qualities in elite footballers during COVID-19 quarantine. *Sci Med Footb* 5(Suppl 1): 26–31, 2020.
19. Cohen J. A power primer. *Psychol Bull* 112: 155–159, 1992.
20. Drinkwater EJ, Pyne DB, McKenna MJ. Design and interpretation of anthropometric and fitness testing of basketball players. *Sports Med* 38: 565–578, 2008.
21. Dufek JS, Bates BT, Davis HP, Malone LA. Dynamic performance assessment of selected sport shoes on impact forces. *Med Sci Sports Exerc* 23: 1062–1067, 1991.
22. Dufek JS, Bates BT, Stergiou N, James CR. Interactive effects between group and single-subject response patterns. *Hum Mov Sci* 14: 301–323, 1995.
23. Dufek JS, Harry JR, Eggleston JD, Hickman RA. Walking mechanics and movement pattern variability in monozygotic twins with autism Spectrum disorder. *J Dev Phys Disabil* 30: 793–805, 2018.
24. Dufek JS, Zhang S. Landing models for volleyball players: A longitudinal evaluation. *J Sports Med Phys Fitness* 36: 35–42, 1996.
25. Edgington ES. Randomized single-subject experiments and statistical tests. *J Counsel Psychol* 34: 437–442, 1987.
26. Fisher RA. On the mathematical foundations of the theory of statistics. In: *Theory of Statistical Estimation (Proceedings of the Cambridge Philosophical Society)*. England: Cambridge Philosophical Society, 1925. pp. 700–725.
27. Hammond KG, Harry JR, Hays CW, Schilling BK. Time-motion analysis of men's professional beach volleyball. *J Sport Hum Perform* 8, 2020.
28. Harry JR, Barker LA, James R, Dufek JS. Performance differences among skilled soccer players of different playing positions during vertical jumping and landing. *J Strength Cond Res* 32: 304–312, 2018.

29. Harry JR, Barker LA, Tinsley GM, et al. Relationships among countermovement vertical jump performance metrics, strategy variables, and inter-limb asymmetry in females. *Sports Biomech* 1–19, 2021.
30. Harry JR, Blinch J, Barker LA, Krzyszkowski J, Chowning L. Low pass filter effects on metrics of countermovement vertical jump performance. *J Strength Cond Res* 36: 1459–1467, 2022.
31. Harry JR, Eggleston JD, Dufek JS, James CR. Single-subject analyses reveal altered performance and muscle activation during vertical jumping. *Biomechanics* 1: 15–28, 2020.
32. Harry JR, Paquette MR, Schilling BK, et al. Kinetic and electromyographic sub-phase characteristics with relation to countermovement vertical jump performance. *J Appl Biomech* 34: 291–297, 2018.
33. Hecksteden A, Kraushaar J, Scharhag-Rosenberger F, et al. Individual response to exercise training-a statistical perspective. *J Appl Physiol* 118: 1450–1459, 2015.
34. Heishman A, Daub B, Miller R, et al. Countermovement jump inter-limb asymmetries in collegiate basketball players. *Sports* 7: 103, 2019.
35. Higgins S. Movement as an emergent form: Its structural limits. *Hum Mov Sci* 4: 119–148, 1985.
36. Hopkins WG, Batterham AM. Error rates, decisive outcomes and publication bias with several inferential methods. *Sports Med* 46: 1563–1573, 2016.
37. Hopkins WG, Marshall SW, Batterham AM, Hanin J. Progressive statistics for studies in sports medicine and exercise science. *Med Sci Sports Exerc* 41: 3–13, 2009.
38. Huberty CJ, Lowman LL. Group overlap as a basis for effect size. *Educ Psychol Meas* 60: 543–563, 2000.
39. Hurwitz J, Rich D, Agnew C, Harry J. Workload management strategies to optimize athlete performance in collegiate men's and women's basketball. *Int J Strength Cond* 2, 2022.
40. James CR, Herman JA, Dufek JS, Bates BT. Number of trials necessary to achieve performance stability of selected ground reaction force variables during landing. *J Sports Sci Med* 6: 126–134, 2007.
41. Kazdin AE, Tuma AH. *Single-case Research Designs*, 1982.
42. Kennedy RA, Drake D. Improving the signal-to-noise ratio when monitoring countermovement jump performance. *J Strength Cond Res* 35: 85–90, 2021.
43. Kipp K, Kiely MT, Geiser CF. Reactive strength index modified is a valid measure of explosiveness in collegiate female volleyball players. *J Strength Cond Res* 30: 1341–1347, 2016.
44. Krzyszkowski J, Chowning LD, Harry JR. Phase-specific predictors of countermovement jump performance that distinguish good from poor jumpers. *J Strength Cond Res* 36: 1257–1263, 2022.
45. Loturco I, Pereira LA, Cal Abad CC, et al. Vertical and horizontal jump tests are strongly associated with competitive performance in 100-m dash events. *J Strength Cond Res* 29: 1966–1971, 2015.
46. Mann JB, Ivey PA, Mayhew JL, Schumacher RM, Brechue WF. Relationship between agility tests and short sprints: Reliability and smallest worthwhile difference in National Collegiate Athletic Association Division-I football players. *J Strength Cond Res* 30: 893–900, 2016.
47. Marocolo M, Simim MAM, Bernardino A, et al. Ischemic preconditioning and exercise performance: Shedding light through smallest worthwhile change. *Eur J Appl Physiol* 119: 2123–2149, 2019.
48. McInnes S, Carlson J, Jones C, McKenna MJ. The physiological load imposed on basketball players during competition. *J Sports Sci* 13: 387–397, 1995.
49. McMahon J, Lake JP, Ripley N, Comfort P. Vertical jump testing in rugby league: A rationale for calculating take-off momentum. *J Appl Biomech* 36: 370–374, 2020.
50. McMahon JJ, Murphy S, Rej SJ, Comfort P. Countermovement-jump-phase characteristics of senior and academy rugby league players. *Int J Sports Physiol Perform* 12: 803–811, 2017.
51. Nuzzo JL, McBride JM, Cormie P, McCaulley GO. Relationship between countermovement jump performance and multijoint isometric and dynamic tests of strength. *J Strength Cond Res* 22: 699–707, 2008.
52. Odgaard EC, Fowler RL. Confidence intervals for effect sizes: Compliance and clinical significance in the journal of consulting and clinical psychology. *J Consult Clin Psychol* 78: 287–297, 2010.
53. Parker RI, Hagan-Burke S. Useful effect size interpretations for single case research. *Behav Ther* 38: 95–105, 2007.
54. Rhea MR. Determining the magnitude of treatment effects in strength training research through the use of the effect size. *J Strength Cond Res* 18: 918–920, 2004.
55. Robertson S, Bartlett JD, Gastin PB. Red, amber, or green? Athlete monitoring in team sport: The need for decision-support systems. *Int J Sports Physiol Perform* 12: S273–S279, 2017.
56. Sands W, Cardinale M, McNeal J, et al. Recommendations for measurement and management of an elite athlete. *Sports* 7: 105, 2019.
57. Sands WA, Kavanaugh AA, Murray SR, McNeal JR, Jemni M. Modern techniques and technologies applied to training and performance monitoring. *Int J Sports Physiol Perform* 12: S263–S272, 2017.
58. Sporis G, Jukic I, Ostojic SM, Milanovic D. Fitness profiling in soccer: Physical and physiologic characteristics of elite players. *J Strength Cond Res* 23: 1947–1953, 2009.
59. Stone MH, Sands WA, Stone ME. The downfall of sports science in the United States. *Strength Cond J* 26: 72–75, 2004.
60. Turner A, Brazier J, Bishop C, et al. Data analysis for strength and conditioning coaches: Using excel to analyze reliability, differences, and relationships. *Strength Cond J* 37: 76–83, 2015.
61. Turner AN, Parmar N, Jovanovski A, Hearne G. Assessing group-based changes in high-performance sport. Part 1: Null hypothesis significance testing and the utility of p values. *Strength Cond J* 43: 112–116, 2021.
62. Turner AN, Parmar N, Jovanovski A, Hearne G. Assessing group-based changes in high-performance sport. Part 2: Effect sizes and embracing uncertainty through confidence intervals. *Strength Cond J* 43: 68–77, 2021.
63. Vincent W, Weir J. *Statistics in Kinesiology*. Champaign, IL: Human Kinetics, 2012.
64. Ward P, Coutts AJ, Pruna R, McCall A. Putting the "I" back in team. *Int J Sports Physiol Perform* 13: 1107–1111, 2018.
65. Weir JP. Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. *J Strength Cond Res* 19: 231–240, 2005.
66. Williams CA, Oliver JL, Faulkner J. Seasonal monitoring of sprint and jump performance in a soccer youth academy. *Int J Sports Physiol Perform* 6: 264–275, 2011.
67. Zatsiorsky VM, Kraemer WJ, Fry AC. Task-specific strength. In: *Science and Practice of Strength Training*. Champaign, IL: Human Kinetics, 2021. p. 25.