Assessing group-based changes in high-performance sport.
Part 2: Effect sizes and embracing uncertainty through confidence intervals

Anthony N Turner, Nimai Parmar, Alex Jovanoski, Gary Hearne
London Sport Institute, Middlesex University

Abstract

Today's strength and conditioning coach must extend their skill set to include data analysis, understating the validity and utility of *p* values, effect sizes, confidence intervals, and terms such as the smallest worthwhile change, and minimal difference. The aim of part two of this two-part review is to now build on our discussion of null hypothesis significance testing (covered in part one), and introduce effect sizes, measures of variability, and confidence intervals, culminating in recommendations as to which may be the most viable options within the context of performance-based sport, and thus potential methods to report group-based changes. This paper has a series of worked examples to aid the reader.

Introduction

With null hypothesis significance testing (NHST), we are analysing the probability of observing our data or data more extreme, under the null hypothesis, and assuming all model assumptions are true; this is written as $P(D|H_0)$. In performance-based sport, however, what we often want to know is the *practical significance* of our data. That is, how big a difference is there between our groups as a consequence of our intervention for example, and importantly, what are the range of values that are compatible with our data, as well as the precision of our estimate. This latter question enables us to make inferences from our sample of athletes to the wider population from which it was drawn. Achieving this outcome is the aim of part 2 of this 2-part review, where we will discuss the definition and utility of effect sizes and confidence intervals, such that applied practitioners can incorporate them within their practice. Importantly, these can be computed through Microsoft excel and thus are available to most coaches, with this paper including a series of worked examples to facilitate this.

Effect sizes

Effect sizes (ES) are an important outcome of empirical studies (Lakens, 2013), given they enable researchers to report the magnitude of an effect (e.g., some novel training stimulus) and

thus provide applied practitioners with some measure of "*practical significance*". A common ES is a standardised mean difference, calculated by dividing the mean difference by some standardiser such as the standard deviation (SD) (Lakens, 2013). With a standardised ES, the numerator can be considered to represent the signal, and the denominator some estimate of *noise*, or the natural (unexplainable) variation we would see across repeated trials or across different groups. This ratio is important because it helps us detect true changes and again, not be fooled by random variation and error. Cohen's *d* is probably the most commonly used estimate of the standardized effect when reporting change, although it is known to be upwardly biased (leading to overestimation) with small samples ($n < 20$); see **Equation 1**. An alternative is Hedges *g*, which provides a *corrected effect size* for small samples and is calculated as per **Equation 2**. Both of these can be interpreted as a percentage of the SD, such that a value of $d = 0.5$ means a difference (or effect) of half an SD (but we need to look at the context to determine what exactly this is a percentage of). Again, thresholds to quantify the magnitude of change have been provided to assist researchers in interpreting the data, when no better basis for estimating the ES index is available; for example, $d = 0.2$ is a "*small*" effect, $d = 0.5$ is a "*moderate*" effect, and $d = 0.8$ is a "l*arge*" effect (Cohen, 1988). But as described in part 1, conventional thresholds often fail to capture context, uncertainty, and provoke cynicism in the data and any interpretation of it – something we will therefore aim to tackle in the latter half this paper. The calculations for Cohen's d_s and Hedges g_s are shown below (Lakens, 2013), and are also presented as they would be typed in to Microsoft Excel. The $_s$ signifies we are undertaking an independent samples analysis.

Equation 1. Cohen's d_s

$$d = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{(n_1 - 1)SD_1^2 + (n_2 - 1)SD_2^2}{n_1 + n_2 - 2}}}$$

In Excel, type:

= (X1- X2)/SQRT(((n1-1)*SD1^2+(n2-1)* SD2^2)/(n1+n2-2)))

where: X_i are the group means; SD_i are the group sample standard deviations; n_i are the respective sample sizes. These values can be typed in directly, but it is better to use cell references.

Equation 2. Hedges g_s

$$g = \text{Cohen's } d \times \left(1 - \frac{3}{4(n_1 + n_2 - 9)}\right)$$

In Excel type =d*(1-(3/(4*(n1+n2-9)))

where: *d* is the cell reference for equation 1 above.

For reference, one example of an ES calculation for within – subjects designs is referred to as Cohen's d_{av} and calculated as per **Equation 3**, whereby the *av* denotes that the average SD of both repeated measures is the standardizer (Lakens, 2013). The numerator in this equation is the difference between the mean (*M*) of the difference scores. We should also point out that there are many different effect size calculations available (for between – and within – subjects designs), with an in-depth coverage available via the work of Lakens (2013) and Caldwell and Vigotsky (2020).

Equation 3. Cohen's d_z

$$\text{Cohen's } d_{av} \times \left(1 - \frac{M_{diff}}{\frac{SD_1 + SD_2}{2}}\right)$$

A worked example

The following data set and worked example highlights the benefits of assessing the magnitude of change (or difference) through ES, compared to basing decisions solely on measuring the compatibility of the data against the null (i.e., the *p* value), and certainly when compared to *yes – no* decisions driven by a critical threshold such as $\alpha = 0.05$. **Table 1** identifies hypothetical jump height data collected from three elite teams of soccer, basketball, and hockey athletes. While we have chosen to example our analysis with this metric, it should be viewed as a proxy measure of some fundamental performance variable that the reader better relates to. The practical question here is, which group of athletes (soccer *vs.* basketball *vs.* hockey) jumps highest?

Table 1. hypothetical jump height data collected from three teams

	Soccer	Basketball	Hockey
	43.7	46.8	47.4
	36.0	37.6	35.7
	42.8	44.7	42.5
	46.2	54.5	50.8
	37.0	38.7	36.7
	45.1	47.1	44.8
	41.2	47.2	44.9
	39.6	42.4	39.3
	44.9	46.9	44.6
	40.1	41.9	39.8
	45.3	47.3	45.0
	47.7	49.8	47.4
	31.5	38.0	33.3
	36.7	56.1	49.3
	46.7	48.8	46.4
	43.7	45.7	43.4
Mean	**41.8**	**45.9**	**43.2**
SD	**4.6**	**5.3**	**5.0**

Here we will analyse the data from **Table 1** (using IBM SPSS Statistics, version 25) using a one-way ANOVA as our statistical model to calculate the relevant test-statistic (i.e., an *F*-statistic), and compare groups to determine if there is a "*statistically significant*" difference between them. The null hypothesis is that there isn't, with the alternative being that at least one is different from the others. We need to take care here, since ANOVA assumes the group variances are equal (known as homoscedasticity), and the individual means can be affected by skewness and extreme values that can either mask or accentuate any differences. **Figure 1** below shows that this is not an issue with this data set, although it is particularly problematic when dealing with such small sample sizes; we will return to the why this matters later in this paper. The one-way ANOVA produces a *p* value of 0.072, which is greater than the conventional α of 0.05. Thus, we fail to reject the null hypothesis and state that given our test-statistic, sample size, statistical power, and *chosen* false-positive error rate, there is no *statistically significant* difference between our three groups – as such, some researchers would choose to end the analysis here. However, as we noted in part 1, the *p* value is a continuous variable and so what we should say is that, assuming the null hypothesis were true and all the assumptions made by the underlying model, the probability of obtaining such a result or more extreme, is 7.2%.

Equally in part 1, we suggested that having some critical threshold (i.e., the alpha level) maybe a good idea as it ensures consistent decision making that's grounded in long run probabilities, ensuring we operate within a boundary of acceptable error. However, according to Lakens et al., (2018), we should set that false-positive error rate based on the context at hand and on the implications of false-negatives (Type II errors). Again, as we discussed in part 1, we could probably justify raising the alpha level to 0.1 (i.e., 10%), given that high-performance athletes are close to their genetic ceiling and sporting success is based on the smallest of margins, therefore it is often preferred to risk an increase in false-positives (Type I errors) than tightly guard against false-negatives. This line of thinking seems further justified when we consider that our statistical tests are often underpowered (on account of low sample sizes) and the fact that our training interventions rarely risk injury – of course the alpha level should be lowered if operating under a false-positive could incur financial harm and/or risk injury, health, and potentially be fatal. Now, operating under this contextualised threshold of 0.1, there does seem to be a difference, which is supported by examining the simple effect sizes noted when comparing means. In summary, for us to *act* as though any *p* value does indeed represent a difference, it should be less than our pre-defined alpha level, which should be based on the context at hand, and our desire to control the false-positive error rate. In this example, it is probably ok to have an alpha level to 0.1 and be content in *acting* as if there is a difference, if the probability value falls below 10%, and thus us only being wrong in the long run, 10% of the time. Of note, if the null is true, all *p* values are equally likely.

We now need to look at calculating an important and practically significant statistic, the standardised ES, noting that some may choose to not even engage in NHST in the first instance. Using **Equation 2** (given we have a sample size below 20), we can note that there is a "large" difference between the soccer team and the basketball team, and a "small" difference between the soccer team and the hockey team. Also, there is a "moderate" difference between the basketball team and the hockey team. **Figure 1** illustrates these changes, with the values of Hedges *g* identified along with the associated descriptors according to Cohen. Therefore, in this example we can see that dichotomised thinking using conventional error rates (i.e., 5%) can potentially mask differences that are practically meaningful, at least within a sport setting where high-performance teams may look to aggregate a series of "marginal gains" to improve the chance of sporting success, and where they are relatively certain that higher error rates (above 5%) do not endanger athletes, support staff, or the financial integrity of the club. Now, the remaining issue to contend with here, is that we have just used Cohen's standard thresholds

to determine the magnitude of effect we noted through our ES analysis. In the next sections, therefore, we will aim to determine our own thresholds by appreciating the context at hand, as well as our precision to actually determine true changes; these will provide justification for our choices.

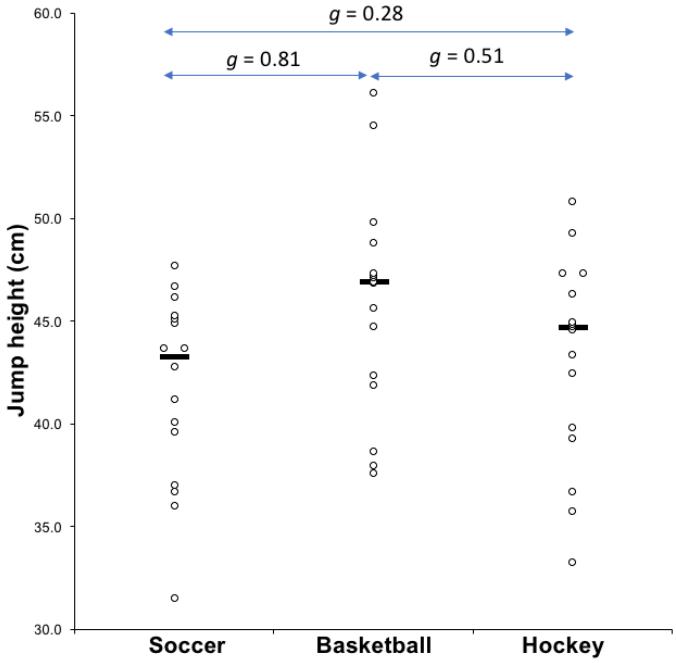

Figure 1. Using effect sizes to calculate the magnitude of difference between three teams. The plot shows that the group ranges are similar, as are the distributions. There is some suggestion of negative skew in all three, but no obvious outliers.

The Smallest Effect Size of Interest

Given effect sizes tell us the magnitude of change between interventions or between groups, they can therefore be reverse engineered to define target scores. This approach utilises what is known as the smallest worthwhile change (SWC), first described by Hopkins (2004). This statistic aims to establish an athlete's new target score by multiplying the between subject standard deviation by 0.2 (which represents a small change according to Cohen) and then adding or subtracting it from the athlete's current score. Clearly for sprints, where we want times to be as low as possible, we subtract, whereas for jumps, where we want scores as high as possible, we add. So, using the soccer team as an example, we could define the SWC as 0.2 × 4.6 cm = 0.9 cm. The soccer team would therefore have to jump (as a group average) 42.7 cm (i.e., 41.8 cm + 0.9 cm) to define the first meaningful improvement in score.

Now, it is important to acknowledge that the definition of "*meaningful*" in the above example is open to interpretation. Let's say that we planned on running this jump height assessment each year, and next time around, the soccer team were determined to be crowned the best athletes at jumping high. In this scenario, a SWC of 0.9 cm would not be meaningful, as it would still leave the soccer team in last place (assuming no change in the other teams). For it to be meaningful (given our context) it would have to be ≥ 1.4 cm assuming they were happy with a second-place finish, or ≥ 4.1 cm if nothing but a first-place finish would suffice. If we assume the former, then the ES required to achieve this (i.e., 2nd place) would be 0.3, and calculated as follows: ES × 4.6 = 1.4, therefore the ES = 1.4 ÷ 4.6 = 0.3. For the latter, the ES required to achieve this (i.e., first place) would be 0.9, and calculated as follows: ES × 4.6 = 4.1, therefore the ES = 4.1 ÷ 4.6 = 0.9. So, in determining the utility of the SWC within this context (where we have an idea of the change required to make meaningful improvements), we must again be critical of simply abiding by arbitrary values (e.g., 0.2). Instead we should adjust these to represent meaningful ones based on the context at hand. This concept is described as the *smallest effect size of interest* (SESOI), and can also be applied in power calculations (discussed in part 1) to define the ES a researcher does not want to miss (Caldwell & Cheuvront, 2019). An alternate method to calculate the SESOI is to base its value on measurement error, that is, what is the lowest ES you can measure, and be sure that it falls outside the noise generated by random variation in the data; this concept is discussed in the next section.

Using Noise to Set the Target

Firstly, there are several methods to determine target differences, and we refer readers to the work of Hislop et al., (2014) for a detailed review of the pros, cons, and utility of each. Here however, we will focus only on distribution based methods, which typically involve using a value which is larger than the inherent imprecision in the measurement, and therefore likely to represent a minimal level for a meaningful difference (Hislop, et al., 2014). As such, for us to establish meaningful changes, we will consider the natural variation and error (or noise) of our tests and ensure our target score falls outside of this. In this regard, we can view our error value as defining the boundaries within which our test cannot reliably determine change. Weir (2005) describes one such method, referred to as minimal difference (MD), which is calculated as: MD = SEM × 1.96 × √2. The SEM is defined as the standard error of the measurement and acts as a measure of precision of scores, and is calculated as the SD of difference scores (SD*d*) divided by √2 (the SEM is referred to as the typical error by some (Hopkins, 2004)). The 1.96

then defines the *z*-score associated with 95% confidence intervals (which is described more fully in the following section), and the √2 is in place to provide an estimation of the difference we might get between two random measurements on the same subject. Consider that this difference has a variance equal to the sum of the two variances, that is $SD^2 + SD^2$, which equals $2 \times SD^2$. Given that the SD is calculated as the square root of the variance, this would then equal $\sqrt{2} \times SD$ (of note, the calculation uses variances rather than SDs as they are additive) – in the context of calculating the MD, the SEM represents the SD. A worked example of the SEM is provided below (Table 2) using the soccer team's data, whereby each athlete had two attempts to jump as high as possible. In completing multiple trials, the error (or noise) of a test can be defined, which enables practitioners to gauge its sensitivity to (real) change. Therefore, conducting multiple trials should be seen as good practice by S&C coaches when fitness testing, without which, data could be deemed unusable.

Table 2. To determine the reliability of a test, the athlete conducts two trials from which the precision of the measurement can be calculated. Here reliability is calculated as the standard error of the measurement (SEM) by first calculating the standard deviation (SD) of the difference scores (= 2.3), and then dividing this value by √2 (= 1.6). The minimal difference (= 4.4) is then calculated by multiplying the SEM by 1.96 (to establish 95% confidence intervals) and then by √2 to account for the estimated error in the post intervention follow up test. In any follow-up test, the athlete's score would have to change by the MD for it to be deemed a real change.

Athlete	CMJ Trial 1	CMJ Trial 2	Trial 2 - Trial 1
A	43.7	42.8	-0.9
B	34.8	36.0	1.2
C	42.8	40.2	-2.6
D	46.2	45.8	-0.4
E	37.0	36.2	-0.8
F	45.1	44.0	-1.1
G	36.0	41.2	5.2
H	38.5	39.6	1.1
I	41.9	44.9	3.0
L	37.5	40.1	2.6
M	42.7	45.3	2.6
N	47.7	45.8	-1.9
O	31.5	30.1	-1.4
P	33.8	36.7	2.9
Q	46.7	45.1	-1.6
R	43.7	42.7	-1.0
SD of change scores			**2.3**
SEM (cm)			**1.6**

Minimal difference (cm)	4.4

As an additional method to detect real changes based on error, we could calculate the coefficient of variation (CV), which will provide another measure reliability, enabling us to form a boundary to enable the detection of real changes (Turner, et al., 2015). The CV scales the SD relative to the mean, again making it a standardized metric. Furthermore, it can be easily computed in Excel with more than two trials of a test, and each athlete can have their own measure of precision. Significant to this latter point, athletes that generate low variability in their test movement will have a lower CV than their less consistent colleagues, and thus their personal test scores will be more sensitive to change. When the reliability score is pooled, some athletes will be advantaged (who had a high CV), while others disadvantaged (who had a low CV). The same data from **Table 2** is now used to identify the CV values for each athlete, the pooled CV value, and changes required by each athlete individually, or taken collectively as a group (Table 3). Here we should note that much like the MD method discussed above, CV based change scores, both individually and pooled, can be multiplied by √2 and then by 1.645 (for 90 % confidence intervals) or 1.96 (for 95% confidence intervals) depending on preference, to represent a maximal difference. These values have also been calculated in the table for each athlete (see CVMD column) and as a team (see final row). When this is done, there is not much difference between the two methods, noting that had the MD method been calculated with 90% confidence intervals rather than 95%, the value would = 3.7 cm. Therefore, the choice of which to use may just come down to preference, which can be most readily calculated, or if values for each athlete are required.

Table 3. To determine the reliability of a test, the athlete conducts at least two trials from which the precision of the measurement can be calculated. Here reliability is calculated as the coefficient of variation (CV) by dividing the standard deviation by the mean for each athlete. This generates a reliability (or consistency) coefficient for each athlete (see athlete CV column). These scores are then averaged to generate the group (pooled) CV (= 3.3%). This pooled CV can be used to generate the change in score required in any follow-up test, to class it as real (= 1.4 %). This can be done on an individual level too (see CV based change column; CVBC) and where practitioners feel this is too lenient, then as per the minimal difference method, the CV can be multiplied by √2 and then 1.645 (the latter corresponding with 90% confidence intervals), to calculate a CV based MD (see final column).

Athlete	CMJ Trial 1	CMJ Trial 2	Best effort	Athlete CV (%)	CVBC (cm)	CVMD (cm)
A	43.7	42.8	43.7	1.5	0.6	1.5
B	34.8	36.0	36.0	2.4	0.9	2.0

C	42.8	40.2	42.8	4.4	1.9	4.4
D	46.2	45.8	46.2	0.6	0.3	0.7
E	37.0	36.2	37.0	1.5	0.6	1.3
F	45.1	44.0	45.1	1.7	0.8	1.8
G	36.0	41.2	41.2	9.5	3.9	9.1
H	38.5	39.6	39.6	2.0	0.8	1.8
I	41.9	44.9	44.9	4.9	2.2	5.1
L	37.5	40.1	40.1	4.7	1.9	4.4
M	42.7	45.3	45.3	4.2	1.9	4.4
N	47.7	45.8	47.7	2.9	1.4	3.2
O	31.5	30.1	31.5	3.2	1.0	2.4
P	33.8	36.7	36.7	5.8	2.1	5.0
Q	46.7	45.1	46.7	2.5	1.2	2.7
R	43.7	42.7	43.7	1.6	0.7	1.7

Average of best effort	**41.8**
Pooled CV (%)	**3.3**
CV based change (pooled) (cm)	**1.4**
CV based Minimal Difference (pooled) (cm)	**3.3**

In concluding this section, we can identify some examples of contextualised ES thresholds based on the error of our data; these represent the lowest values we can choose, given we have attempted to separate the signal from the noise. Using the MD ($z = 1.96$) to calculate the SESOI for the soccer team would result in a ES of 0.96, calculated as: ES × 4.6 = 4.4, therefore ES = 4.4 ÷ 4.6 = 0.96. If the MD is instead based on 90% CIs ($z = 1.645$), then it would be: ES × 4.6 = 3.7, therefore ES = 3.7 ÷ 4.6 = 0.80. Alternatively, if we choose to use the CV (single) method: ES × 4.6 = 1.4, therefore ES = 1.4 ÷ 4.6 = 0.30, and finally, if the CV method is also multiplied by 1.645 and √2 (to account for error in the follow-up test): = ES × 4.6 = 3.3, therefore ES = 3.3 ÷ 4.6 = 0.72.

In keeping with our narrative, we would prefer to avoid a Type II error, so we will define our SESOI using the single CV method. We will also continue to use thresholds as often coaches and players better appreciate qualitative descriptors, rather than being left to interpret the magnitude of difference as inferred by the SD; we will however, endeavour to not use arbitrary ones, although acknowledging that this nonetheless promotes dichotomisation. So, let's say values up to 0.3 are just noise and thus represent "*trivial*" changes, and anything over 0.9 represents a "*large*" change as that gives us a first-place finish. Let's also continue to graduate

our changes by taking 0.31 to 0.6 as "*small*" changes and 0.61 to 0.9 as "*moderate*" changes, with changes in both these regions associated with a second-place finish, but the second region signifying we are edging ever closer to wining the competition. Interestingly, in going with the CV (single) method, and thus the lowest threshold of ES we can detect is 0.3, this means that based on our distribution method of detecting change, the ES we noted in Figure 1 between soccer and basketball athletes is not "real" and thus attributed to random variation in the data. Therefore, under this assumption, if these two teams were to come back the following day and repeat the test for example, we should not be surprised if the team's positions, as determined by differences in means (simple effect), swapped, and the soccer team now jumped highest.

Back to Frequentist Statistics, but this time using Confidence Intervals

The group mean values for each team as well as the ES values we have generated above, essentially provide us with a *point estimate*, that is, one value that represents the best estimate of a statistic of interest, e.g., the mean of our squad of athletes. This is problematic as our data may be influenced by outliers for example, that is, an athlete (or several) that scores much higher or lower than the rest of the squad that are otherwise quite tightly clustered together. Secondly, it affects our ability to use our data such that we can make inferences about the wider population from which our sample was drawn. And finally, it does not provide direct insight into the precision of our estimate. With respect to the latter, and as per *p* values discussed in part one, we need to appreciate that even between two perfect replication studies, the chances of achieving the same ES is low, and thus we must learn to embrace uncertainty and be humble in our interpretations (Amrhein, Greenland, & McShane, 2019). We can however, largely overcome these issues, if we use a confidence interval (CI), but noting that these would also be dissimilar between studies as their calculation is based on the sample size and SD, and as ever, subject to random variation, which can have a profound effect (Amrhein, Greenland, & McShane, 2019).

Confidence intervals are considered an alternate inferential statistic to NHST and were first proposed by Neyman (1937), and have since been heavily supported by others (Cumming & Finch, 2001; Cumming G. , 2014; Amrhein, Greenland, & McShane, 2019; Gardner & Altman, 1986). Confidence intervals can be applied to means, correlations, and effect sizes for example, to define a range of scores for which you can *act* as though the true population value lies within. Again, we must use the term "*act*" given that CIs are part of the frequentist framework and thus actually provide a statement about the performance of the procedure of drawing such

intervals in repeated use (Greenland, et al., 2016). More specifically then, our model suggests that were we to calculate 95% CIs repeatedly, across similar samples, then 95% of them would include the true (population) ES. So, in summary, a CI enables us to present sample statistics as estimates of results that would be obtained if the whole population from which the sample was drawn, was tested (Gardner & Altman, 1986). A CI also indicates the precision of our estimate (via the width of intervals), which is an important point to highlight in sport, given we often have small samples and thus the degree of variability in the factor being investigated can be high – as such, CI's can also be used as an alternative to power calculations. Finally, a CI facilitates us moving away from using point estimates to describe our entire data set and encourages us to embrace uncertainty and describe the practical implications of all values inside the interval, especially the point estimate and the limits (Amrhein, Greenland, & McShane, 2019). Amrhein et al., (2019) remind researchers that it is nonsensical to focus on one value, given that all values between the interval's limits are reasonably compatible with the data, given the statistical assumptions used to compute the interval. In fact, many statisticians are now advocating CI's be renamed as *compatibility intervals* for this reason. Such a name change should serve to promote humility and thus temper confidence when interpreting their meaning and extrapolating the findings, as well as encouraging us to consider and embrace all values inside the interval (Amrhein, Greenland, & McShane, 2019).

Some additional points to note regarding confidence intervals (or rather, compatibility intervals), are that they can also be part of the NHST framework by considering if the interval contains the null value or not. Let's assume a null value of zero and an alpha level set at 0.05 for this next example, and suppose we were using ES analysis to quantify change from baseline testing in two groups. If group one revealed 95% CI's ranging from 0.1 to 0.6, while group two had 95 % CI's ranging -0.1 to 0.4, then, given this data, we could say that group one is statistically significantly different from zero, whereas group two is not. We should also point out that this design is not suitable to determine if the groups are statistically significantly different from each other, as this needs to be done via direct comparison of means between groups with the use of baseline as a covariate, rather than by separate analysis of changes from baseline in each parallel group (Bland & Altman, 2015). So, assuming an assessment of independent means, then if an exact *p* value was required, an ANCOVA would be used. However, we could again use our CI's within the NHST framework here, with Cumming and Finch (2005) providing a useful "inference by eye" guide, which would suit a quick assessment by practitioners or those examining only the figures. In this scenario, if the 95% CI's overlap

by no more than about half the average margin of error (i.e., half the length of one CI arm), then *p* can be considered to equal < 0.5. If the error bars do not overlap at all, then *p* can be considered to equal < 0.1. These relationships are sufficiently accurate when both sample sizes are at least 10 and the margins of error do not differ by more than a factor of 2.

Finally, much like *p* values, the correct interpretation of CI's is not well understood by all those who use them (Hoekstra, Morey, Rouder, & Wagenmakers, 2014), which again is unfortunate given that along with *p* values, they constitute the main tools by which sport scientists draw conclusions from data. To reiterate their meaning, using the definition as provided by Hoekstra et al., (2014), a CI is a numerical interval constructed around the estimate of a parameter (noting that a statistic describes a sample, while a parameter describes a population). As is typical for the frequentist technique, the interval does not directly indicate a property of the parameter, but rather a property of the procedure, with conclusions regarding the data based on a procedures average performance for a hypothetical infinite repetition of experiments (i.e., the sample space). Specifically, it tells us that this procedure, when used repeatedly across a series of hypothetical data sets, produces intervals containing the true parameter value in 95 % of the cases. We can therefore note that it is incorrect to interpret a CI as the probability that the true value is within the interval, however, we may choose to *act* as though this is the case.

So, if we wanted to use our sample data from Table 1 (in which our model is based only on raw scores and includes no covariates such as age, height, weight, or factors such as gender) to infer on what the actual average jump height would be from the populations from which our soccer, basketball, and hockey athletes was drawn, we could calculate CIs, as per **Equation 4**. In general, when the sample size is ≥30, we use *z*-scores to build our intervals around our point estimate. However, if our sample is <30, as is often the case in sport (and in our example), we use *t*-statistics. As we have already discussed above, the *z*-scores corresponding with a 95% CI = 1.96, and 90% = 1.645 (by convention, researchers typically calculate a 95% CI). However, we can calculate this for any interval using Excel's NORM.S.INV function as shown in **Equation 5**. For *t*-statistics, we use Excel's T.INV.2T as shown in **Equation 6**, basing the value not only on the size of the CI (e.g., 95% or 90%), but also the sample size (or rather the *degrees of freedom*). In our example, where we have a sample size of 16, we would use t = 2.131 for a 95% CI and t = 1.753 for a 90% CI. So, using **Equation 4**, we can calculate the CI associated with each of the teams, with the data illustrated in **Table 4** and **Figure 2**. We would also note that in practice, it may be more convenient to stick to using the *t* distribution, since it

is most suitable for smaller samples, but the results converge towards those of the normal distribution as the sample size increases anyway. Finally, a CI can also be built around our calculated ES for the differences between teams, but for this we need to use **Equation 7** (Nakagawa & Cuthill, 2007), with data illustrated in **Table 5**.

Equation 4. Calculating confidence intervals for the mean, using small groups (<30)

$$= X \pm t * \frac{SD}{\sqrt{n}}$$

Where X = mean value (e.g., group average or mean improvement); SD = standard deviation; n = sample size; and t is based on the chosen width of CI (e.g., 95% or 90%) as well as the sample size (See equation 5 for how to calculate this). When the sample size is >30, substitute t for z.

Equation 5. Using Excel's NORM.S.INV function to return the inverse of the standard normal cumulative distribution. The distribution has a mean of zero and a standard deviation of one.

In excel type =NORM.S.INV(probability/2)

The "probability" is the chosen alpha level, typically 0.05 but can be as high as the research team deem appropriate – here we suggest using 0.1. Choosing 0.5 would result in generating the value associated with 95% confidence intervals (i.e., 1 – alpha), whereas choosing 0.1 would result in generating the value associated with 90% confidence intervals. Here we divide by two as our hypothesis is two-tailed, with values either side of the mean likely.

Equation 6. Using Excel's T.INV.2T function to calculate the inverse of a two-tailed *t*-distribution.

In excel type =T.INV.2T(probability,deg_freedom)

The "probability" is the chosen alpha level, typically 0.05 but can be as high as 0.1. The "deg_freedom" is calculated as the number of athletes minus the number of groups. So, if we are just testing one squad of 16 athletes, this would be 16 – 1 = 15. Therefore, we would type in 15.

Table 4. Confidence intervals (CI) for the mean of each team

Team	Sample size (*n*)	Mean jump height (cm); point estimate	Standard deviation (cm)	95% CI	90%CI
Soccer	16	41.8	4.6	39.3 – 44.2	39.8 – 43.8
Basketball	16	45.9	5.3	43.0 – 48.7	43.5 – 48.2
Hockey	16	43.2	5.0	40.5 – 45.9	41.0 – 45.4

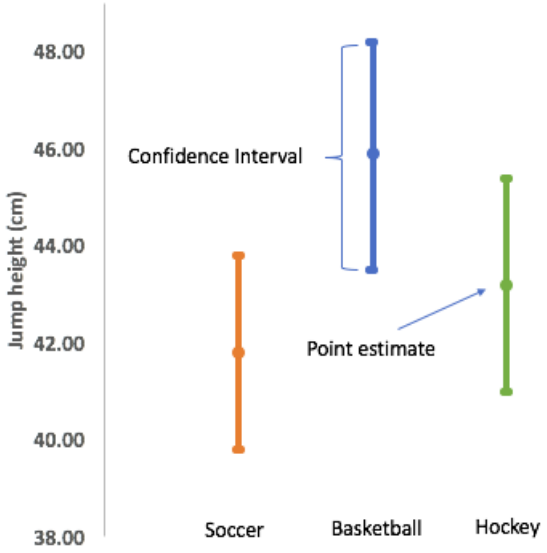

Figure 2. 90% Confidence intervals for the mean of each team

Equation 7. Calculating 90% confidence intervals for effect sizes, for small groups (<30), with the equation written as it would be entered into Excel. These intervals are then constructed either side of the point estimate. This equation provides the standard error (SE) for a bias-corrected standardised mean difference (Hedges *g*)

90% CI = 1.64*SE

$$SE = \sqrt{\frac{n_1 + n_2}{n_1 n_2} + \frac{d^2}{2(n_1 + n_2 - 2)}}$$

In excel type =SQRT((*n*1+ *n*2)/(*n*1* *n*2)+ES^2/(2*(*n*1+ *n*2-2)))

Where SE = standard error; *n* = sample size; ES = effect size

Table 5. Confidence intervals (CI) for the effect size (ES) that describes the difference in jump height between teams

Team comparison	Sample size (*n*)	ES (*g*); point estimate	95% CI	90%CI
Basketball vs. Soccer	16	0.81	0.08 – 1.53	0.20 – 1.41
Basketball *vs.* Hockey	16	0.51	-0.19 – 1.22	-0.08 – 1.10
Hockey vs. Soccer	16	0.28	-0.41 - 0.98	-0.30 - 0.87

Let's now re-examine our ES data, ensuring to not just focus our analysis on the point estimate, but on the extreme values of each interval, which are also compatible with our data given the statistical assumptions used to compute the interval. We will also use the thresholds and justification of them as presented in the previous section (i.e., 0-0.3 = trivial, 0.31-06 = small, 0.61-0.9 = moderate, and >0.9 = large). So, we could say something along the lines of: we found basketball players demonstrated trivial to larger jump heights compared to soccer players (g = 0.81, 90% CI, 0.20 – 1.41) and hockey players (g = 0.51, 90% CI, -0.08 – 1.10), and while the difference between soccer and hockey players was trivial (g = 0.28, 90% CI, -0.30 – 0.87), hockey players demonstrating greater jump heights by a small to moderate distance is also reasonably compatible with our data, given our assumptions.

Conclusion

Effect sizes may be the most important part of the results, providing us with the practical significance of our data and an indication of the magnitude of difference observed. Practitioners should consider the SESOI when defining the threshold for their ES, and that the lowest this value can be should be determined by quantifying the error of a test. Here we have presented two examples for this, using the MD and CV, and the choice of which to use may, as ever, come down to whether practitioners prefer to risk a Type I or Type II error.

Confidence intervals should envelop our ES, allowing us to make inferences from our data to the wider population from which our sample was drawn. They should also serve to steer us away from fixating on point estimates, encouraging us to embrace the uncertainty in our data as well as describing the practical implications of all values inside the interval, especially the point estimate and the limits. Finally, we must remember that a CI defines a range of scores for which we can *act* as though the true population value lies within, given they (like *p* values) are part of the frequentist framework and thus actually provide a statement about the performance of the procedure in the long run. More specifically, they inform us that were we to calculate

90% CIs repeatedly, across similar samples, then 90% of them would include the true population value.

In closing, whenever possible, collaborate with a statistician!

References

1. Amrhein, V., Greenland, S., & McShane, B. (2019). Scientists rise up against statistical significance. *Nature, 567*, 305.
2. Bland, J., & Altman, D. (2015). Best (but oft forgotten) practices: testing for treatment effects in randomized trials by separate analyses of changes from baseline in each group is a misleading approach. *The American journal of clinical nutrition, 102*(5), 991-994.
3. Caldwell, A., & Cheuvront, S. (2019). Basic statistical considerations for physiology: The journal Temperature toolbox. *Temperature, 6*, 181-210.
4. Caldwell, A., & Vigotsky, A. (2020). Does One Effect Size Fit All? the Case Against Default Effect Sizes for Sport and Exercise Science. *SportRχiv. https://doi.org/10.31236/osf.io/tfx95*.
5. Cumming, G. (2014). The new statistics: Why and how. *Psychological science, 25*(1), 7-29.
6. Cumming, G., & Finch, S. (2001). A primer on the understand- ing, use, and calculation of confidence intervals that are based on central and noncentral distributions. *Educational and Psychological Measurement, 61*, 532-574.
7. Cumming, G., & Finch, S. (2005). Inference by eye: confidence intervals and how to read pictures of data. *American psychologist, 60*(2), 170-180.
8. Gardner, M., & Altman, D. (1986). Confidence intervals rather than P values: estimation rather than hypothesis testing. *Br Med J, 292*, 746-750.
9. Greenland, S., Senn, S., Rothman, K., Carlin, J., Poole, C., Goodman, S., & Altman, D. (2016). Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. *European journal of epidemiology, 31*(4), 337-350.
10. Hislop, J., Adewuyi, T., Vale, L., Harrild, K., Fraser, C., Gurung, T., & Norrie, J. (2014). Methods for specifying the target difference in a randomised controlled trial: the Difference ELicitation in TriAls (DELTA) systematic review. *PLoS Med, 11*(5), e1001645.
11. Hoekstra, R., Morey, R., Rouder, J., & Wagenmakers, E. (2014). Robust misinterpretation of confidence intervals. *Psychonomic bulletin & review, 21*(5), 1157-1164.

12. Hopkins, W. (2004). How to interpret changes in an athletic performance test. *Sportscience , 8*, 1-7.
13. Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. *Frontiers in psychology, 4*, 1-12.
14. Lakens, D., Adolfi, F., Albers, C., Anvari, F., Apps, M., Argamon, S., & Buchanan, E. (2018). Justify your alpha. . *Nature Human Behaviour, 2*, 168-171.
15. Nakagawa, S., & Cuthill, I. (2007). Effect size, confidence interval and statistical significance: a practical guide for biologists. *Biological reviews, 82*, 591-605.
16. Neyman, J. (1937). Outline of a Theory of Statistical Estimation Based on the Classical Theory of Probability. *Phil. Trans. R. Soc. Lond. A, 236*, 333-380.
17. Turner, A., Brazier, J., Bishop, C., Chavda, S., Cree, J., & Read, P. (2015). Data Analysis for Strength and Conditioning Coaches. *Strength And Conditioning Journal, 37*, 76-83.
18. Weir, J. (2005). Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. *J. Strength Cond. Res, 19*(1), 231–240.