But did my athlete improve!?

Assessing performance changes when N = 1

Author Profile

Dr Anthony Turner is an Associate Professor in Strength and Conditioning, and the Research Degrees Coordinator for Sport, at the London Sport Institute, Middlesex University. Links to Anthony's university profile, Research Gate, and YouTube channel are provided, and his twitter handle is @anthonyturneruk.

Abstract

This paper outlines how coaches can analyse the data of a single athlete so they can determine if their athlete responded favourably to their training programme and improved from test to test. The method presented is simple and can be calculated in excel or comparable programmes like google sheets. The paper also presents a discussion of Type I and II errors, as the coaches' philosophical stance around which they prefer to avoid most, will ultimately determine the threshold at which they class any changes in scores as real and meaningful.

Introduction

I've written a few papers on analysing data and how to profile a single athlete. Aside from bettering my own understanding, there were principally two motives for doing this. The first is because we're mainly taught about and read about group-level analysis, i.e., has my team, on average, improved? But when we're not involved in experimental research (and thus how the results translate to the wider population) and instead actually working with athletes (whether a squad of athletes or just one), this is not particularly helpful. In these scenarios what we really want to know is, has my athlete's score appreciably changed? From a performance perspective that is, did my (each) athlete get better at the test, or, from a monitoring perspective, are the fluctuations in my athlete's scores real and thus require me to take action? My second motive was to uncover the holy grail – that one single method that was unequivocally the right one... But unfortunately, it turns out that even in something as seemingly binary as statistics, which is underpinned by mathematical equations, that there are still several ways to answer a single question! And yes, whichever route you choose, there is a guaranteed entry into some debate you hoped to avoid by undertaking all this research in the first place. But in any case, let me discuss one such method such that you can defend its use (assuming you choose to adopt it),

while still recognising its inevitable limitations. And what's great is that it's actually a really simple method to compute (as I'll show), without any advanced statistics or software!

Now, in walking us to this single-athlete analysis, I want to start off by considering how we measure differences at a group level – as this is what we are most familiar with. So, consider that when testing a group of athletes we try to separate the signal from the noise such that we can identify the value above which we can say with some degree of confidence, that our athlete has indeed improved. For example, if the trial-to-trial and day-to-day variability in our test score is 2% or 2 cm, then crudely put, we cannot categorise any change as real unless it's greater than this value. And that variability may be driven by the athlete, physiologically and behaviourally, in terms of how they interreact with their environment and prepare for and undertake the test, or it could just be generated by the testing equipment and setup, which will never be perfect. We would typically compute that 2% and 2 cm using the coefficient of variation (CV), and the standard error of the measurement (SEM) and so here I will briefly define these methods, because again, all of this will walk us through to or single athlete analysis.

The CV and the SEM

Now, the CV represents the relative variability in test scores, given it is the ratio of the standard deviation (SD) to the mean. So a CV of 10% suggests that the SD is 10% of the mean. The higher the CV, the less consistent the data points and vice versa. The CV would probably be the best measure of reliability if you were comparing across tests with different units, i.e., when asking the question, which test is the most reliable? For example, which is the more reliable test out of a jump height system with an SD of 3 cm, or a peak force system with an SD of 100 N? It's tempting to say the jump height one, but actually we need to know what this is relative to. Across multiple trials, the mean score may be 40 cm and 2000 N respectively. So relative to these, the jump height system has a trial to trail variability of 7.5% (i.e., [3 / 45 = 0.075]*100), while the peak force system is 2.5% (i.e., [100 / 2000 = 0.05]*100). This example highlights when the CV is the standout winner for measuring reliability. The SEM alternatively, is the SD of error of measurement and again, the higher the SEM, the less precise a test is. The SEM is also calculated in the units of the test too – so for jumps, we would be able to report the error in cm for example, as opposed to the CV which presents it as a percentage.

The SD

Now what you will note, is that both measures of reliability have the SD at the very heart of the calculation; so let's explore this value a little further too. The SD represents the spread of data around the mean. A small SD represents data that is more closely clustered around the mean, enabling us to infer that the tested group performed similarly, or from a testing standpoint in a single athlete, that the scores are similar (and thus quite consistent). A larger SD suggests more variance in test scores, suggesting a wider range of abilities amongst the group, and thus, perhaps, the group are not particularly similar. Equally from a testing standpoint in a single athlete, it suggests a lack of consistency and thus accuracy of the test data (Figure 1).

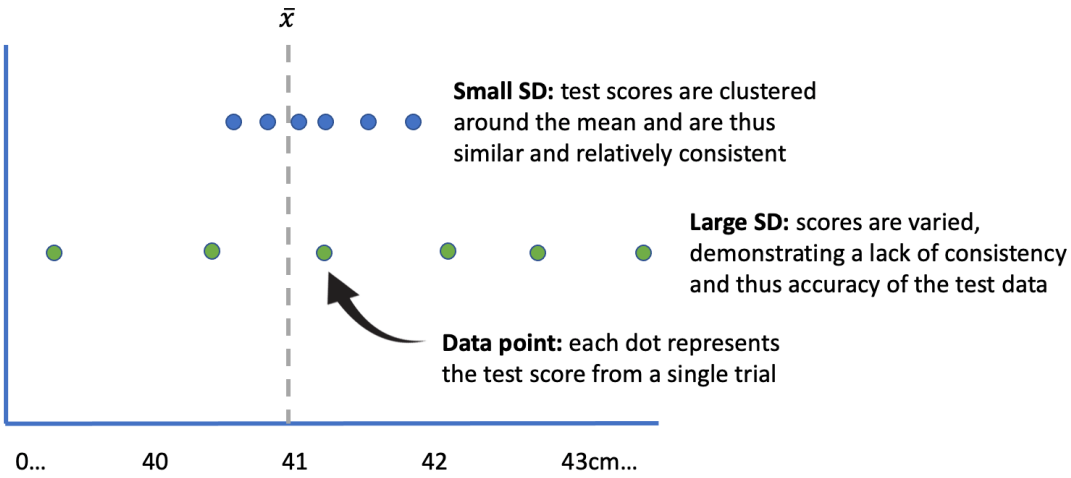

Figure 1. Trial-to-trail and day-to-day variability of test scores in a single athlete. The green dots represent trails collected from an unreliable test, while the blue dots represent trails collected from a reliable test. $\bar{x}$ = mean, SD = standard deviation

When looking at a group of athletes, the SD and indeed the mean, would represent the average of all individual athletes. While the SEM is generally calculated from group-based data only, the CV can actually be calculated for each individual athlete too; you just need to divide each athlete's SD across trials by their mean across trials, and multiple by 100 (the last bit is to turn that value into a %). For example, if jump height in trial 1 = 43.7 cm and in trial 2 = 42.8 cm, then SD = 0.64 cm, mean = 43.25 cm, and thus CV (i.e., 0.64/43.25 *100) = 1.47 %. Now, we can use this CV value to define the absolute error in the test, i.e., the change required in any follow-up test, for me to say, yes, my athlete got better. Of note, the SEM would be ideal here but it is calculated at a group level, hence me only using the CV right now, as we transition to our single-athlete analysis. So, what is 1.47% of 43.25 cm (i.e., the mean)? Well that is

calculated as follows: (43.25/100)*1.47 = 0.64 cm… Which equals the SD, as we have essentially just reversed engineered the previous equation! So, assuming we calculate the absolute change from the mean, as opposed to the best score (i.e., 43.7 cm), the value simply equals the SD. In summary then, when not looking a group-based data, there is no need to determine the threshold for meaningful change via the CV or SEM, instead, just keep your analysis simple and set your threshold for change in your single-athlete analysis based on the SD.

How much confidence should you have in your data?

Before I draw your attention to a video link which explains all this further including the calculation of it in excel, let's just explore the SD a little more and the concept of confidence in your data. So, when we build a range of scores around the mean using the SD, we account for 68% of the variability. That is, if our athlete was to do 100 trials, we would expect 68 of them to land within that range (this is therefore the same for the CV and SEM at the group level). In our example, that means that we would expect our athlete to jump between 42.6 cm (calculated as the mean – SD) and 43.9 cm (calculated as the mean + SD) 68 times out of 100 trials. If we instead build a range of scores around the mean using 2 or 3 SD's, then we would expect 95 and 99 trails out of 100 to land within them, respectively. These would naturally generate larger ranges of 42.0 cm – 44.5 cm and 41.3 cm – 45.2 cm, respectively.

But when describing our data, we typically just state the best score, or in this example, the highest jump, as a single point estimate of our athlete's physical capacity that day. There is nothing wrong with that, but just note that this score contains the true score plus all the error from the aforementioned variability, so it could actually be a slight over or underrepresentation of their physical capacity. In reality, their jump height is better described as currently being between 42.6 and 43.9 cm, with our confidence in our range of estimates growing as we increase the intervals form 1 SD to 2 and 3 SDs. But as you can clearly note form these ranges, as the intervals increase and our confidence therefore grows in us having actually captured the true score, the opposite happens in the likelihood of us being able to identify small but potentially meaningful changes (I'll elaborate on this point in just a moment). So, what intervals should we aim to build around our point estimate, such that we are confident in having captured the signal while limiting the noise, yet capable of detecting small but real improvements? Should our confidence represent 68% of the variability (1SD), 95% (~2SD) or 99% (~3SD), or how about some middle ground such as 80% (1.28SD) or 90% (1.64SD)?

Personally, I think 1SD provides a wide enough interval within the context of high-performance sport. It provides an interval indicative of the fact that we acknowledge our point estimate contains some error that has either over or underestimated the true score, but an interval not so wide has to mask small but meaningful changes in follow-up tests. My preference, however, is largely a philosophical one, based on my preference to avoid Type II errors. Let me elaborate and explore this concept now, such that you can adjust your intervals based on your own philosophy.

Type I or Type II error. That is the question.

A Type I error is a *false-positive* – where you claim a difference when there is none, and a Type II is a *false-negative* – where you claim no difference when there was one. As already stated, no matter how well you collect your data, there will always be error or rather noise, and that at the analysis stage, when you aim to limit noise, you also limit the signal (which is where the athlete's 'true' score is found). Similarly, when you aim to amplify the signal, you amplify the noise. So, it's a case of which is the lesser of two evils? Do you prefer to be conservative and wait till a change can be classed as all but certain, or do you prefer to be sensitive to smaller changes, but inevitably claiming some changes were real which were really just noise? You do have to pick a side and you need to stick with your decision throughout the season – otherwise if you adjust the intervals each testing battery, you're essentially biasing the data based on what you told the team would happen following your awesome training programme! There is no right or wrong here, so it's totally up to you and the team. Sometimes you'll be right, sometimes you'll be wrong, so you need to focus on the consequences of each scenario to help you choose.

Back to my opinion. In a separate but related 2-part paper (see Ref 1 and 2), I explained my take on whether a Type I or II error was better in sport, which went along the following lines. In sport, unlike say medicine, the consequence of making a Type I error (i.e., a false-positive) will unlikely be fatal or lead to any health complications, let alone lead to injury, so it's probably okay to increase the sensitivity of the test (by reducing the intervals). I mean, at worse we persist with something that isn't working like we think it is, but at best, we don't miss out on some new training technique that is actually working, albeit by the tiniest of margins. But tiny changes, just creeping past the threshold of trivial, are really important in high-performance sport, with success based on the smallest of differences; a statement that every Olympic Games final will prove testament to. Equally, professional athletes are often butting

up against their genetic ceiling and change is hard to come by, so if we do induce some positive change, it won't be big, but it will certainly be meaningful. As such, I think it is more important to reduce false-negatives and thus potentially avoid missing out on recognising interventions that may stimulate tiny yet positive adaptations. In summary then, higher sensitivity would be used by those for whom false-positives are relatively inconsequential, and lower sensitivity would be used by those for whom false-positives could be disastrous. Perhaps you would use the latter if you were investing thousands of pounds into a new bit of kit, in which case, you would want to be pretty certain it worked!

Determining change in a single athlete

To conclude then, to analyse a single athlete from one test to the next, just use the SD and base decisions of if they improved, stayed the same, or somehow got worse, on whether the confidence intervals overlap or not (Figure 2). And the width of your confidence intervals should be based on a desire to avoid Type I or Type II errors, which in turn is based on the consequences of being wrong relative to missing small changes.

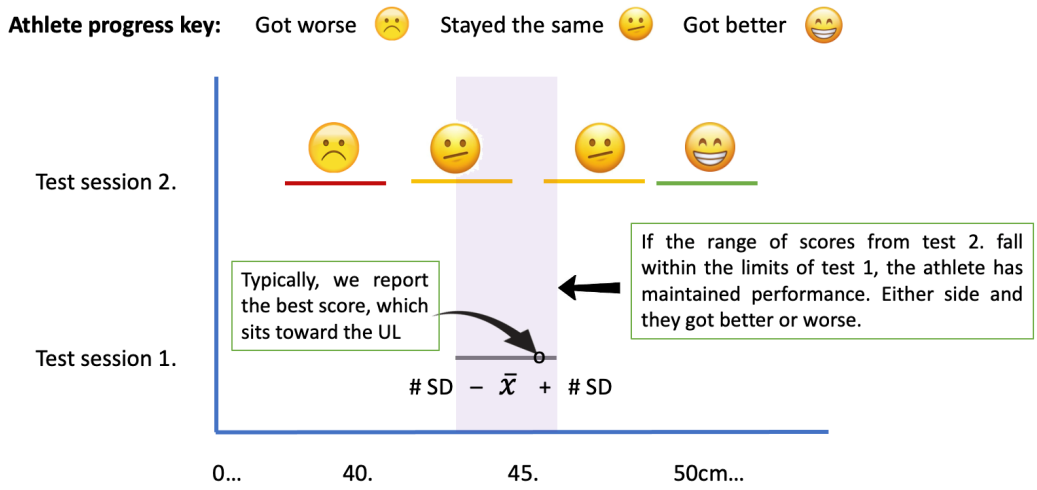

Figure 2. Determining if the athlete got better from testing session one to two. $\bar{x}$ = mean, SD = standard deviation, UL = upper limit, # = chosen multiple of the SD based on a philosophical desire to avoid Type I or II errors.

Note here that, given we have both sets of data now, i.e., jump height in testing battery one and jump height in testing battery two, we can make our estimation of improvement or not, based on the error (or intervals) generated from both testing batteries – this is key. While we can set targets for our athletes using the data generated from the first battery, i.e., next time we test them we want them to aim to jump the mean + 2*SD (we are using double our chosen SD now

to account for error in the follow-up test), the reality is that scores above or below this pre-set value may actually result in a score that would be classed as 'better', but we will never know until we make a comparison with the second testing session and its associated SD value. Interestingly it is useful to note that our 2SD target produces a 95% CI, which is what is also used to infer change in statistical significance testing… but anyway, moving swiftly on.

Is the juice worth the squeeze?

Actually, one last thing. It is important to keep an eye on each athlete's SD. If theirs is much higher than everyone else's, then they need to spend more time learning how to perform the test, as until then, the intervals will be so wide that any meaningful change will be lost in all the noise. Alternatively, if you note that everyone's SD is high (relative to their mean), then you should question the validity of the test for monitoring purposes. Just because a test has seemingly high theoretical utility, does not mean it has high practical utility. If you are witnessing these relatively high SD's amongst the team, then you have principally two options, a) invest in better equipment (assuming it exists and you have the finances) or b) stop using it! Often b is the best solution but the hardest to execute, as we all so desperately want to measure something, anything will do, and theoretical at least, it has to works… surely!?

Video link

I've provided a QR code that will take you to a video I put together for this paper, where I briefly talk through the points covered here and show you how to analyse the data in excel and identify if your athlete got better, worse, or just stayed the same. I hope you find it useful!

References

1. Turner, A., Parmar, N., Jovanoski, A., & Hearne, G. (2021). Assessing group-based changes in high-performance sport. Part 1: Null hypothesis significance testing and the utility of p values . *Strength and Conditioning Journal*.
2. Turner, A., Parmar, N., Jovanoski, A., & Hearne, G. (2021). Assessing group-based changes in high-performance sport. Part 2: Effect sizes and embracing uncertainty through confidence intervals. *Strength and Conditioning Journal*.