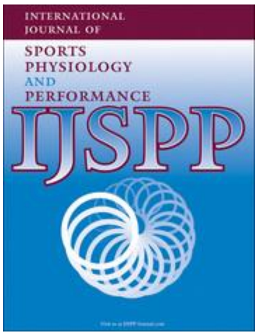


2016, 11, 551 – 554

Title: **The numbers will love you back in return – I promise**

Submission type: **Invited Commentary**

Authors: *M. Buchheit.*[1]

[1] Performance Department, Paris Saint-Germain Football Club, Saint-Germain-en-Laye, France

Running Head: Magnitude-based inferences

Contact details:
Martin Buchheit
Performance Department, Paris Saint-Germain Football Club,
4a avenue du président Kennedy
78100 Saint-Germain-en-Laye, France
Tel.: +33 1 61 07 10 77
E-mail: mbuchheit@psg.fr

Abstract word count: **248**

Text-only word count: **1727**

Number of Tables: **0**

Number of Figures: **3**

Disclosures: nothing to disclose

1. Abstract

The first sport science-oriented and comprehensive paper on magnitude-based inferences (MBI) was published 10 years ago in the first issue of this journal. While debate continues, MBI is today well-established in sports science and in other fields, particularly clinical medicine where practical/clinical significance often takes priority over statistical significance. In this commentary, some reasons why both academics and sport scientists should abandon null hypothesis significance testing (NHST) and embrace MBI are reviewed. Apparent limitations and future areas of research are also discussed. The following arguments are presented: P values and in turn, study conclusions, are sample-size dependent, irrespective of the size of the effect; significance doesn't inform on magnitude of effects, yet magnitude is what matters the most; MBI allows authors to be honest with their sample size and better acknowledge trivial effects; the examination of magnitudes *per se* helps provide better research questions; MBI can be applied to assess changes in individuals; MBI improves data visualisation; and lastly, MBI is supported by spreadsheets freely available on the internet. Finally, recommendations to define the smallest important effect and improve the presentation of standardized effects are presented.

Keywords: magnitude-based inferences; null hypothesis significance testing; sample size; trivial effect; smallest important effect.

2. Introduction

I discovered magnitude-based inferences (MBI) in 2008 while reading Impellizzeri et al.'s repeated-sprint testing paper in professional soccer.[1] I found the first figure of their paper to be simply fascinating. First, changes in repeated-sprint performance were compared in reference to a typical threshold representative of a smallest important or meaningful change (later to be termed the smallest worthwhile change, SWC[2]). Second, instead of a classical 'yes or no' type response, the authors reported both quantitatively and qualitatively the probabilities for these changes to be 'real' (Figure 1). Never had I previously read of anything more meaningful to that day. The message displayed within that figure spoke to both sport scientists and practitioners alike. In France, as in most other countries at that time, statistical lectures exclusively sang the praises of null hypothesis significance testing (NHST). The understanding of these statistics took long hours of self-driven and motivated learning. However, this new statistical approach, driven largely by Will G. Hopkins's efforts, has changed my life, both as an academic and practitioner in elite sport.[3]

Perhaps analogous to spirituality and religion, where individuals follow their own God, editors and reviewers of most journals (even those with high impact factors[4]) can find it difficult to think outside their bubble, believing only what they were taught in graduate school. Since authors driven by their H-index know that providing everything other than a P value increase dramatically their chances of seeing their paper rejected,[3] they simply stick to NHST to facilitate reviews and expedite publication. Fortunately, things have progressively moved on in some sports science journals.[5, 6] While some may view such occurrences as a coincidence, the first sport science-oriented and comprehensive paper on MBI was published 10 years ago in the first issue of our journal,[7] and remains one of the most cited papers on the topic, together with the 2009 update in another journal.[8] Somewhat unexpectedly last year, some authors claimed that MBI had questionable theoretical foundations and suffered from apparently high rates of type I errors (i.e., false positives), which lead them to advise researchers against using MBI.[9] In March this year, Hopkins and Batterhamm[10] provided evidence to dismiss the critiques and to reassure researchers and practitioners that MBI is in reality superior to NHST. While the debate will likely continue, MBI is today a well-established analytical approach in sports science and in other fields, particularly clinical medicine where practical/clinical significance often takes priority over statistical significance.

While the present work is only an invited commentary, and should not be considered as journal policy, I personally wish that MBI is influential with other scientists, as it has been to me. I take this opportunity to put forth the following recommendations, limitations and future areas of research, to assist researchers and practitioners to make better decisions with our numbers.

Reasons why academics should abandon NHST and embrace MBI (using the probable effect of a new nutritional supplement on performance as an example).

1. **P values and in turn, study conclusions, are sample-size dependent (the greater the n, the lower the P), irrespective of the size of the effect.** While it can be concluded that the nutritional supplement is ineffective with a sample of 12 athletes ($P>0.05$), the same comparison may turn useful with $n = 14$ ($P<0.05$). In other words, the drop-out of a few athletes, or the lucky involvement of 2 more subjects can induce a 180 degree change in a study conclusion.[11] This sample size issue explains also a large portion of the publication bias in research,[12] where only significant results tend to be submitted (among the studies with small sample size only those showing large effects –more likely significant ($P<0.05$)– are submitted and published).[10]
2. **Significance doesn't inform on magnitude of effects, yet magnitude is what matters the most.[13]** With a large enough sample size, even very small, trivial or non-practical effects can turn significant ($P<0.05$). In practice, with 200 athletes showing a 0.01% improvement in performance, NHST would suggest that the nutritional supplement works, while the effects may in fact be negligible. In my experience, coaches and athletes (and probably most of our readers

too) are first interested in knowing what kind of performance benefits may be expected from the supplement (i.e., how much, the actual magnitude), and how likely this magnitude is of practical importance (i.e., likelihood of the effect to be greater than the SWC).

3. **MBI allows authors to be honest with their sample size and better acknowledge trivial effects.** While a P>0.05 is often interpreted as a lack of an effect/difference, it is actually impossible to be confident that this is the right interpretation of the data analysis (sample size issue, Type II error resulting from low statistical power). The beauty of MBI is that it allows deciphering between clear (confidence limits within the SWC) and unclear (CL overlapping the SWC) trivial effects (Figure 1). This can't be touched by NHST. An unclear effect/difference is not to be interpreted as a lack of an effect, but suggests the need to increase sample size to improve precision.
4. **The examination of magnitudes *per se* helps provide better research questions**. Considering that the size of an effect matters more than a simple yes or no answer (NHST), typical hypotheses that do not have clear foundations (e.g., "We hypothesized that the new supplement would be beneficial for performance") can be replaced by a simpler and more relevant statement: "Our aim was to quantify the performance benefit of that supplement, if any".
5. **MBI is supported by spreadsheets freely available on the internet (e.g., [14])**

Reasons why magnitude-based inferences are the essential statistical tool for practitioners in the field

1. **MBI can be applied to assess changes in individuals.** In essence, conventional statistics allow analysis of population-based responses, which are impractical for monitoring performance changes in individuals (Figure 2). While individual score changes can be assessed in various ways (e.g., Z-scores,[15] standard difference score[16]), MBI additionally allow us to assess the likelihood of these changes to be true for any given athlete, once the typical error of the test of interest and the SWC are known.[17, 18]
2. **MBI improves data visualisation.** MBI principles should be applied to graphical reports produced by sport scientists, where shaded trivial areas and confidence limits (or typical errors for individual data) are presented systematically to acknowledge the fact that not all changes are worthwhile and that some uncertainty always remains (Figures 1, 2 and 3).

An apparent limitation of MBI is that, in contrast to NHST, researchers have to define *a priori* both the magnitude of the smallest important effect and the thresholds used to qualify likelihoods (e.g., *very likely*).[19] My view is that instead of being a limitation, this forces researchers to adopt a conscious process when analysing their data. "NHST is easy, but misleading. MBI is hard but honest" (W.G. Hopkins, personal communication). The importance of an appropriate SWC definition is often overlooked[20, 21] and may directly impact decisions: while a larger SWC may lead to more conservative decisions, a smaller SWC increases the chance of effects/differences being substantial. In fact, the most appropriate SWC is variable-dependent and based on either theoretical or practical considerations. While for individual athlete performance, a third of the performance coefficient of variation (CV) is generally suggested, and a fifth of the between-athlete SD is often used for performance variables in team sports.[17] A limitation however of using the SD for standardization is that the SWC may be affected by group homogeneity; for that reason, performance clues may be sometimes used instead, e.g., based on empirical observations of direct performance benefits, such as a distance of 20–50 cm that one soccer player needs to be ahead of the opponent to win a ball, corresponding to a 1% improvement in 20-m sprint time.[22] For physiological data with no direct link to performance (e.g., heart rate variability), using multiples of the within-athlete SD is a relevant option. In contrast, when an association with performance can be established for a physiological variable (i.e., submaximal HR), the actual change in this variable that relates to the smallest important change in performance is often preferred.[23] There are some variables however for which the most appropriate SWC remains to be determined. For match running performance data in team sports for example, which are neither related to actual physical capacities nor match outcomes,[24] using the between-athlete SD is questionable, but using within-athlete variation is not easy either. In fact, the magnitude of within-athlete variations may depend on both tracking variables and intensity zones.[25]

How standardized changes/differences are presented is crucial for a better understanding of magnitudes. While percentages are commonly used to report changes/differences both in research and field practice, there are no clear thresholds to interpret their magnitudes, and they often bias the comparison of variables that differ in units[26] (e.g., in terms of athlete trainability, while a 3% increase in sprinting speed may be considered remarkable,[22] the same improvement in maximal oxygen update may be relatively negligible). For these reasons, using Cohen's effect size principle (d) is generally the first step toward standardization (Figure 3).[27] However, if we consider that the actual method of SWC determination may be variable-dependent (Cohen's d vs. within-athletes CV vs. performance clues), the same approach could be applied to standardize the changes in different variables. The thresholds for small, moderate, large and very large standardized changes (Cohen's d) being 0.2, 0.6, 1.2 and 2, respectively, means that any change of 1x, 3x, 6x and 10x SWC can be considered as small, moderate, large and very large, respectively (Figure 3). Reporting effects/changes as multiples of the SWC[28] is relevant for at least two reasons: i) in manuscripts, the changes in all variables can be easily aggregated into a single figure with a single shaded trivial area (Figure 3) and ii) for coaches and athletes, the message cannot be simpler than: "the effect is *x* times greater than what generally matters to you guys".

3. Conclusion.

The introduction of MBI into sports science nearly 15 ago represents one of the most important analytical progressions in our field. While there are still areas that need to be developed, there is no doubt that we should all be leaning toward a more mature and conscious process of analysing and presenting our data.[19] "The numbers are where the discussion should start, not end."[29]

Acknowledgments. I will be forever indebted to Will G. Hopkins (Victoria University, Austalia) for his overall work on applied statistics, and his very helpful and critical comments on the drafts of the present manuscript. Warm thanks too to Paul B. Laursen (High Performance Sports New Zealand) for his edits on the present manuscript, and Alberto Mendez-Villanueva (Aspire Academy, Qatar) for holding my hand in 2009 when we took the risky decision (at that time) not to ever report a P value anymore in any of our manuscripts.

References

1. Impellizzeri, F.M., et al., Validity of a Repeated-Sprint Test for Football. *Int J Sports Med*, 2008;29:899-905.
2. Hopkins, W.G., Statistical vs clinical or practical significance. *Sportscience* 2002;6:http://www.sportsci.org/jour/0201/Statistical_vs_clinical.ppt.
3. Buchheit, M. *Any Comments?* 2013; Available from: https://herearemycomments.wordpress.com/category/best-hopeless-comments.
4. Tressoldi, P.E., et al., High impact = high statistical standards? Not necessarily so. *PLoS One*, 2013;8(2):e56180.
5. Atkinson, G., A.M. Batterham, and W.G. Hopkins, Sports performance research under the spotlight. *Int J Sports Med*, 2012;33(12):949.
6. Winter, E.M., G.A. Abt, and A.M. Nevill, Metrics of meaningfulness as opposed to sleights of significance. *J Sports Sci*, 2014;32(10):901-2.
7. Batterham, A.M. and W.G. Hopkins, Making meaningful inferences about magnitudes. *Int J Sports Physiol Perform*, 2006;1(1):50-7.
8. Hopkins, W.G., et al., Progressive statistics for studies in sports medicine and exercise science. *Med Sci Sports Exerc*, 2009;41(1):3-13.
9. Welsh, A.H. and E.J. Knight, "Magnitude-based inference": a statistical review. *Med Sci Sports Exerc*, 2015;47(4):874-84.
10. Hopkins, W.G. and A.M. Batterham, Error Rates, Decisive Outcomes and Publication Bias with Several Inferential Methods. *Sports Med*, 2016.
11. McCormack, J., B. Vandermeer, and G.M. Allan, How confidence intervals become confusion intervals. *BMC Med Res Methodol*, 2013;13:134.
12. Kuhberger, A., A. Fritz, and T. Scherndl, Publication bias in psychology: a diagnosis based on the correlation between effect size and sample size. *PLoS One*, 2014;9(9):e105825.
13. Cohen, J., Things I have learned (so far). *American Psychologist*, 1994;45:1304-1312.
14. Hopkins, W.G. *A spreadsheet for deriving a confidence interval, mechanistic inference and clinical inference from a P value.* Sportscience, 2007. 11, 16-20. DOI: http://newstats.org/xcl.xls.
15. McGuigan, M.R., S.J. Cormack, and N.D. Gill, Strength and Power Profiling of Athletes: Selecting Tests and How to Use the Information for Program Design. *Strength and Conditioning Journal*, 2013;37(6):7-14.
16. Pettitt, R.W., The standard difference score: a new statistic for evaluating strength and conditioning programs. *J Strength Cond Res*, 2010;24(1):287-91.
17. Hopkins, W.G., How to Interpret Changes in an Athletic Performance Test. *Sportscience*, 2004; 8:1-7.
18. Al Haddad, H., B.M. Simpson, and M. Buchheit, Monitoring changes in jump and sprint performance: best or average values? *Int J Sports Physiol Perform*, 2015;10(7):931-4.
19. Schaik, P.V. and M. Weston, Magnitude-based inference and its application in user research. *International Journal of Human-Computer Studies*, 2016(88):38-50.
20. Atkinson, G., Does size matter for sports performance researchers? *J Sports Sci*, 2003;21(2):73-4.
21. Buchheit, M., A. Rabbani, and H.T. Beigi, Predicting changes in high-intensity intermittent running performance with acute responses to short jump rope workouts in children. *J Sports Sci Med*, 2014;13(3):476-82.
22. Haugen, T. and M. Buchheit, Sprint Running Performance Monitoring: Methodological and Practical Considerations. *Sports Med*, 2015.
23. Buchheit, M., Monitoring training status with HR measures: do all roads lead to Rome? *Front Physiol*, 2014;27(5):73.

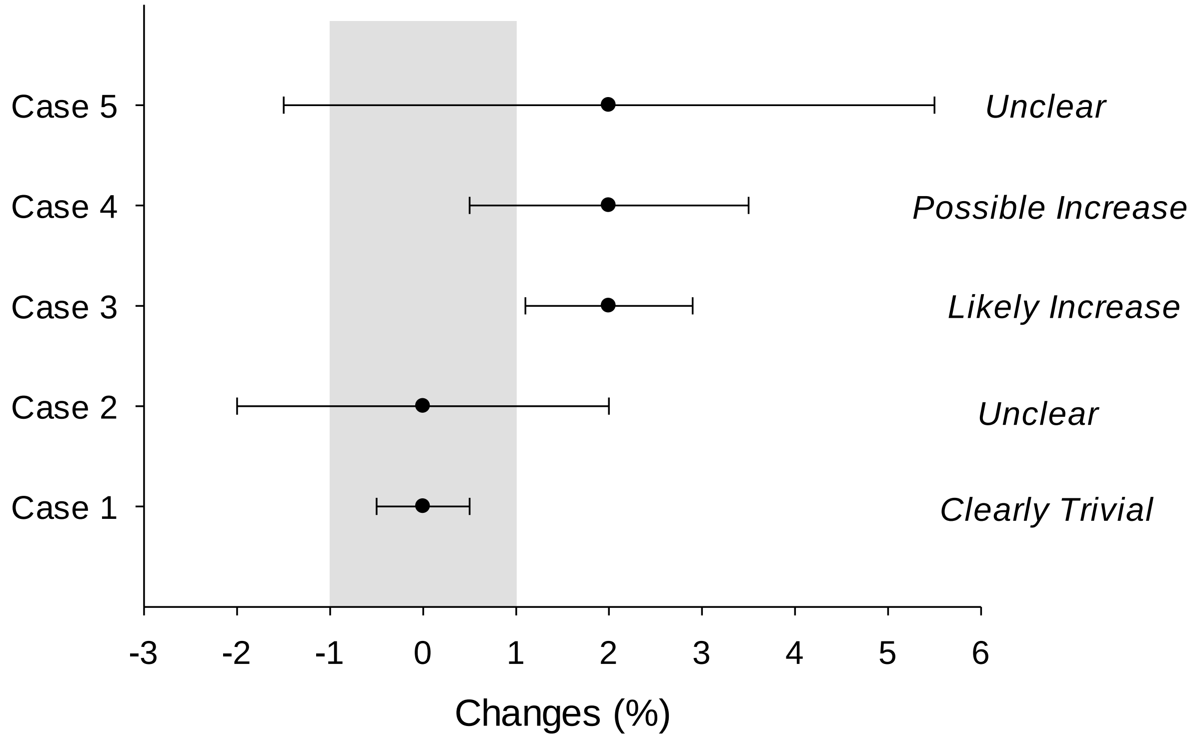
24. Mendez-Villanueva, A. and M. Buchheit, Physical capacity-match physical performance relationships in soccer: simply, more complex. *Eur J Appl Physiol*, 2011;111(9):2387-9.
25. Buchheit, M., et al., Integrating different tracking systems in football: multiple camera semi-automatic system, local position measurement and GPS technologies. *J Sports Sci*, 2014;32(20)(20):1844-1857.
26. Buchheit, M. and A. Rabbani, 30-15 Intermittent Fitness Test vs. Yo-Yo Intermittent Recovery Test Level 1: Relationship and Sensitivity to Training. *Int J Sports Physiol Perform*, 2014;9(3):522-524.
27. Cohen, J., *Statistical Power Analysis for the Behavioral Sciences*. 1988, Hillsdale: Lawrence Erlbaum. 599.
28. Buchheit, M., et al., Physiological, psychometric, and performance effects of the Christmas break in Australian football. *Int J Sports Physiol Perform*, 2015;10(1):120-3.
29. Nuzzo, R., Scientific method: statistical errors. *Nature*, 2014;506(7487):150-2.
30. Buchheit, M. and A. Mendez-Villanueva, Effects of age, maturity and body dimensions on match running performance in highly trained under-15 soccer players. *J Sports Sci*, 2014;32(13):1271-8.

Figures Legends

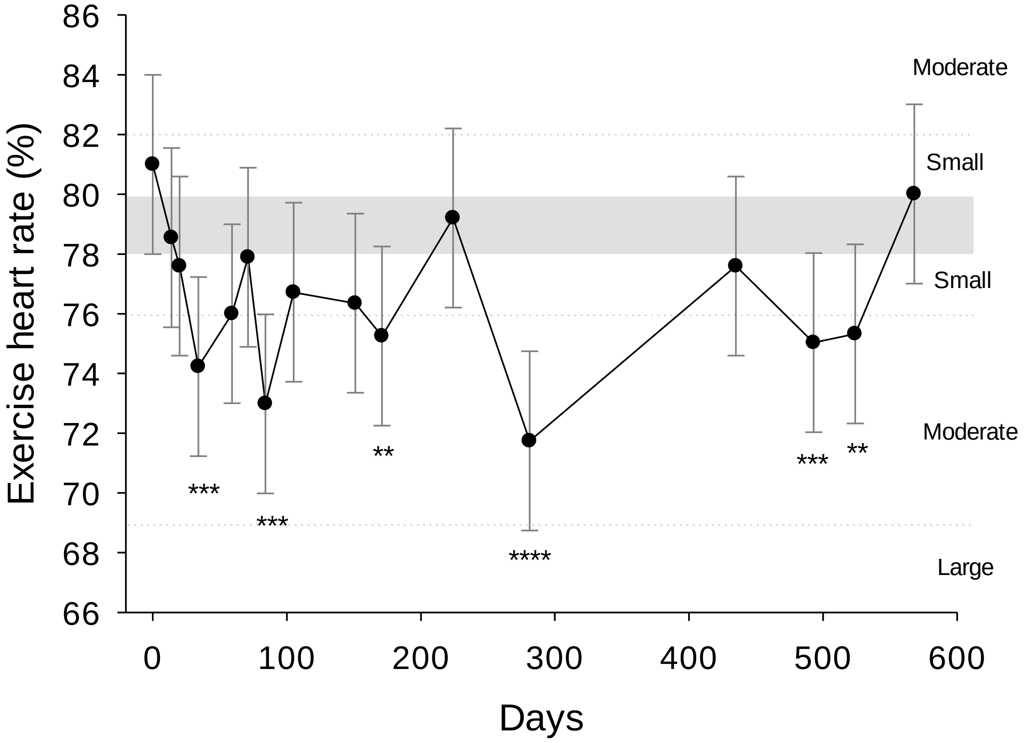
Figure 1. Example of possible decisions when interpreting changes using magnitude-based inferences. Note the clear vs. unclear cases (based on confidence limits, in relation to the shaded trivial area), which i) is one of the extreme beauty of magnitude-based inferences and ii) provide no insight through null hypothesis significance testing. Note also how, for clear effects, the likelihood of changes increases as the confidence limits shrink.

Figure 2. Individual changes in submaximal heart rate in a professional soccer player when running at 12 km/h throughout 1.5 competitive seasons (% of maximal heart rate). The shaded area represents trivial changes (1%).[23] The error bars represent the typical error of measurement (3%).[23] The numbers of * indicate the likelihood for the changes to be substantial, with 1 symbols referring to possible changes, 2 to likely, 3 to very likely and 4 to almost certain changes.

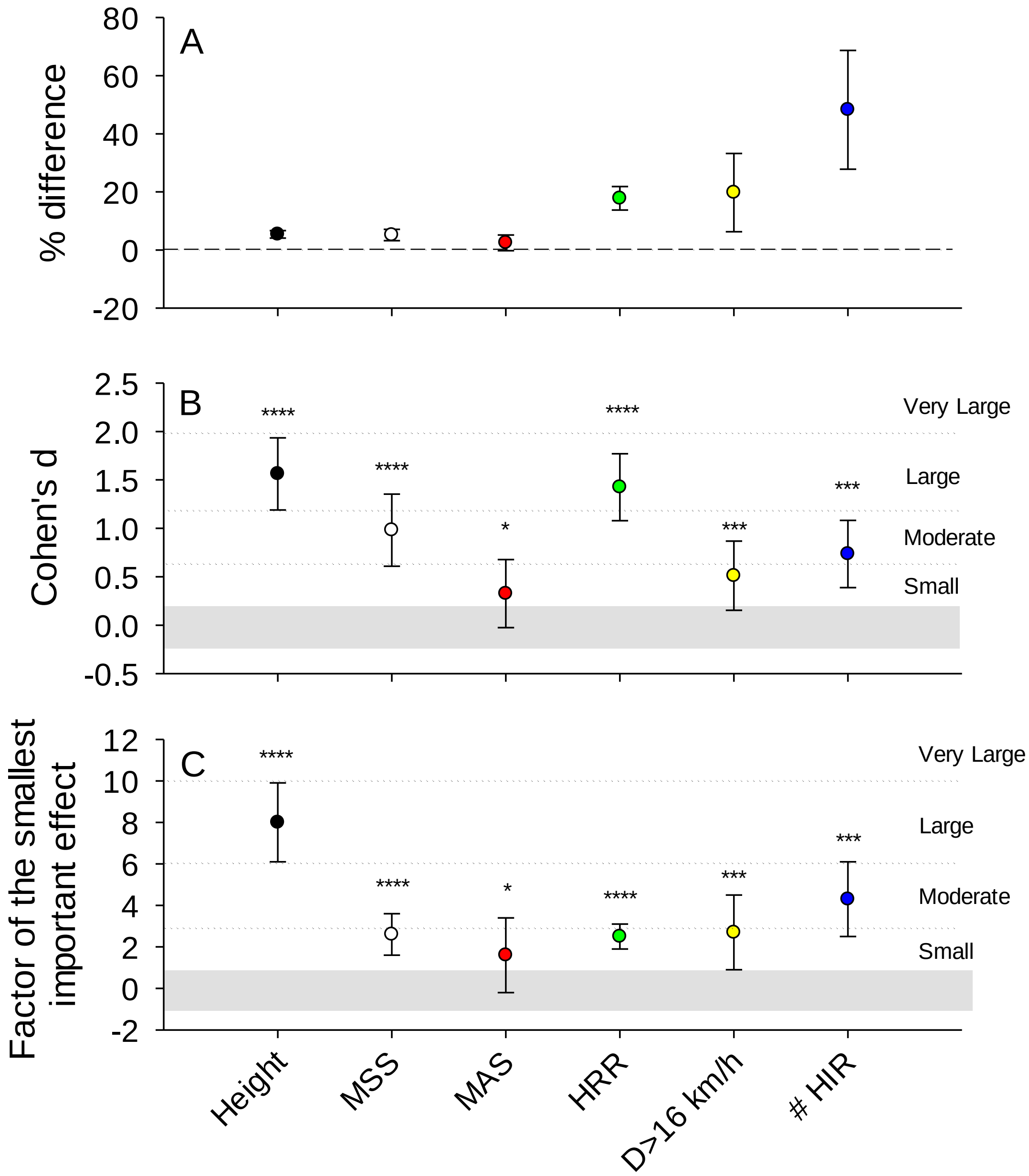
Figure 3. Differences in various anthropometric, physiological and performance measures between two groups of young soccer players differing by their maturity status (0.9 ± 0.3 vs. -0.2 ± 0.4 years from predicted peak height velocity)[30] when expressed in percentages (A), using Cohen's effect size principle (B) and as a factor of variable-specific smallest worthwhile differences (SWD) (C):[28] 0.2 x between-athletes SD for height, MAS and matches tracking data; performance-related changes for HRR and MSS (7[23] and 2[22]%, respectively). The numbers of * indicate the likelihood for the between-group differences to be substantial, with 1 symbols referring to possible difference, 2 to likely, 3 to very likely and 4 to almost certain differences. Note that that magnitude of the between-group differences and their likelihood varies between the panels. My suggestion is to use the method used in panel C (with a variable-specific SWD). MSS: maximal sprinting speed, MAS: maximal aerobic speed, HRR: heart rate recovery after submaximal exercise, D>16 km/h: distance ran above 16 km/h during matches, #HIR: number of high intensity runs during matches.



Case 5
Case 4
Case 3
Case 2
Case 1
Unclear
Possible Increase
Likely Increase
Unclear
Clearly Trivial
-3
-2
-1
0
1
2
3
4
5
6
Changes (%)



Exercise heart rate (%)
86
84
82
80
78
76
74
72
70
68
66
0
100
200
300
400
500
600
Days
Moderate
Small
Small
Moderate
Large
***
***
**
****
***
**


A
% difference
80
60
40
20
0
-20
B
Cohen's d
2.5
2.0
1.5
1.0
0.5
0.0
-0.5
Very Large
Large
Moderate
Small
C
Factor of the smallest important effect
12
10
8
6
4
2
0
-2
Very Large
Large
Moderate
Small
Height
MSS
MAS
HRR
D>16 km/h
HIR