Will the numbers really love you back? Re-examining Magnitude-based Inference

Michael L. Butson[ab]
[a] *College of Sport and Exercise Science, Victoria University, Melbourne, Australia*
[b] *Institute of Sport Exercise and Active Living, Victoria University, Melbourne, Australia*
ORCiD Identifier: 0000-0002-1166-5345

Corresponding author:
Dr Michael L. Butson
Victoria University
PO Box 14428
Melbourne, Victoria 8001
Australia
+61 3 9919 5552
Michael.Butson@vu.edu.au

Abstract
Many sports medicine and sports science researchers use Null Hypothesis Significance Testing despite it being criticized for being an amalgam of two irreconcilable methodologies. Hopkins and Batterham proposed Magnitude-based Inference as an alternative to Null Hypothesis Significance Testing. However, its validity and utility have also been questioned. Recently, it was suggested that the critics of Magnitude-based Inference lacked vision and that their objections should be ignored. However, a re-examination of Hopkins and Batterham's claims about their method indicates that they use profoundly different approaches in ways that are at odds with their theoretical foundations and intended purposes. If Hopkins and Batterham were to provide a full and explicit account of how Magnitude-based Inference is implemented, it could be comprehensively assessed. Until then, sports medicine and sports science researchers should use other theoretically valid methods that have had their utility confirmed. **Key Words:** MEDICINE, SCIENCE, SPORTS, STATISTICS, METHODS

Introduction

The credibility of scientific research has been questioned because an inadequate understanding of statistics has led many researchers to use flawed methods [1, 2]. Hopkins and colleagues [3-5] have expressed particular concern about the use of Null Hypothesis Significance Testing (NHST) in sports medicine and sports science research and have proposed MBI as an alternative. However, several statisticians have criticized the method for being theoretically unjustifiable and lacking real utility [6-8]. Hopkins and Batterham's [9, 10] response to the criticism lacked convincing evidence to support their claims. Nevertheless, prominent sports scientist Martin Buchheit [11] has argued that researchers should ignore the critics of MBI because they are overly constrained by their training and lack imagination when thinking about statistical inference. Buchheit's plea for researchers to embrace MBI invites a re-examination of Hopkins and colleagues' claims about their method.

MBI

MBI [3-5] is predicated on the belief that the size of an effect is what researchers are really interested in when analysing data. It uses confidence limits to define the range of values that would be likely to include the true value of a parameter. Then the proportion of the interval that overlaps specified effect sizes is converted into probability statements about the effect sizes. Qualitative descriptors are used to indicate the importance of the effect. Batterham and Hopkins [9] claim that MBI is probably the perfect synthesis of frequentist and Bayesian methods of statistical inference.

Frequentist Statistics

Hopkins and Batterham [4] describe Fisher's [12] and Neyman and Pearson's [13] methodologies as frequentist. However, there are appreciable differences between the two approaches. Fisher's methodology is informed by fiducial probability. From this perspective, a parameter is a random variable that has a fiducial distribution, which is a measure of the faith that can be placed in different values of the parameter [14]. Fisher's methodology is inductive, drawing inferences from the particular to the general. On the other hand, the Neyman-Pearson methodology is based on Bernoulli's Theorem, which states that if a random process is infinitely repeated, over the long run the obtained values and the predicted values will converge [15]. From the Neyman-Pearson perspective, parameters are unknown quantities that have fixed values. The approach is deductive, drawing inferences from the general to the specific [16]. Consequently, the Neyman-Pearson methodology really defines the frequentist approach to statistical inference.

NHST is a hybrid approach to inference that combines Fisher's [12] and Neyman and Pearson's [13] methodologies. Typically, researchers use Neyman-Pearson procedures but provide a Fisherian interpretation [17]. However, it has been argued that the two methodologies are fundamentally incompatible [16]. Hopkins and Batterham [3, 4] ignore this important issue. Instead, they focus on their dissatisfaction with Fisher's significance test and Neyman's hypothesis test.

Fisher's Significance Test

Fisher's significance test assesses theories by comparing observed values against a null hypothesis. The test does not consider alternative hypotheses and while its properties stem from a hypothetical infinite population, it only applies to data in hand [17]. The metric used in Fisher's hypothesis test is the *p*-value. It is the conditional probability of obtaining a value at least as extreme as the one obtained, assuming that the complete statistical model is true and that chance was operating independently when the probability was calculated [18]. *P*-values equal to or smaller than the threshold value that was set by the researcher would be surprising if the statistical model were true, whereas *p*-values larger than the threshold value would be unsurprising. Following on from this, a *p*-value equal to or smaller than the threshold value is interpreted as providing evidence against the null hypothesis [19].

Batterham and Hopkins [3] correctly state that the use of the *p*-value threshold as a decision rule for rejecting or accepting the null hypothesis in NHST, is not

compatible with Fisher's approach to inference. A *p*-value neither proves nor disproves the null; it simply indicates how compatible the data are with the complete statistical model [18]. Batterham and Hopkins also say that regardless of how the *p*-value is interpreted, the null hypothesis must always be false because there are no zero effects in nature. This statement is misleading as Fisher's significance test can legitimately be used to assess whether an effect is zero, some other specified value, or values within specified limits [18].

Assessing the Magnitude of an Effect

Batterham and Hopkins [3] dismiss the *p*-value as it does not provide information about the magnitude of an effect. In doing so, they neglect to acknowledge that Fisher recommended that in addition to reporting a *p*-value, researchers should report an estimate of the magnitude of an effect [20]. MBI uses the standardized mean difference, which is commonly known as Cohen's *d*, for that purpose. Although standardised effect size measures allow effects that have been measured on different scales to be compared [21], they still require interpretation [22]. Notably, Cohen [23] was reluctant to provide guidelines for interpreting Cohen's *d*. Instead, he encouraged researchers to use relevant theory and to look at the empirical evidence from the appropriate research domain.

When theory and the empirical evidence do not provide adequate guidance, Cohen [23] proposed using default effect size conventions. However, they were not intended to be universal; Cohen developed the conventions to aid calculating statistical power in the behavioural sciences when the effect size distribution is unknown. He acknowledged that selection of the values was somewhat arbitrary and recent findings indicate that they are not consistent with known effect size distributions [24].

Hopkins and Colleagues [3-5] disregard Cohen's [23] caveats on using the conventions. Instead, they propose their own elaboration as a general means of determining whether an effect is meaningful. However, when they increased the number of effect size categories, Hopkins and Colleagues were tacitly acknowledging that the conventions are arbitrary points on a continuum. Similarly, when they adopted different conventions for clinical and non-clinical research, they were accepting the need to consider the research area. So relying on effect size conventions remains a poor substitute for using theory and empirical evidence that is relevant to the study in question.

Furthermore, even if relevant theory and empirical evidence are used to interpret Cohen's *d,* it only provides an accurate estimate of the magnitude of the effect if data are normally distributed, there is homogeneity of variance, groups have equal numbers, sample base rates[1] are equal across groups, and there is acceptable measurements reliability [21, 25]. Additionally, Cohen's *d* is sensitive to outliers, range restriction, and it is positively biased when the sample is small [21, 25]. Although corrections are available for some of the problems, they are not widely

[1] The base rate is the proportion of a population that has a particular characteristic.

used. Notably, Hopkins does provide a formula for correcting the positive bias affect on his website (http://www.sportsci.org/resource/stats/ssmean.html). Notwithstanding the corrections that are available, it has been argued that simple effect sizes such as the raw difference between mean values are a more robust alternative to standardized effect sizes [21].

Neyman's Hypothesis Test

Neyman's [13] hypothesis test is a confidence procedure. That is, it is a way of limiting errors over the long run when deciding whether to reject a null hypothesis in favour of an alternative hypothesis [16, 26].When conducting a hypothesis test, statistical power and alpha are set by the researcher before data are collected. Statistical power is the capacity of a test to identify a true alternative hypothesis. More specifically, it is the long run probability that the null hypothesis will be rejected [18]. Alpha is the critical value that indicates the acceptable long-run error rate for mistakenly rejecting the null hypothesis, when the null and all of the assumptions associated with it are true [18]. After data are collected, if the obtained test value is within the critical region defined by alpha, the risk of rejecting the null hypothesis when it is true would be acceptable over the long run [16, 17]. Hence, alpha can be used as a decision rule.

Neyman's [27] confidence interval (CI) is another component of the hypothesis test. It assesses a confidence procedure's ability to exclude false parameter values [18]. Generally, shorter intervals are better than longer ones because they exclude false values more often [26]. CIs are based on the premise that if an infinite number of samples were taken and CIs were calculated for them, the proportion that contains the true parameter value would match the specified confidence level [26]. However, which particular confidence intervals contain the true parameter value cannot be determined. Furthermore, either the parameter is within a particular CI or it is not, chance plays no role [18, 26].

Despite criticising the hypothesis test, Hopkins and Batterham [3, 4] appropriate CIs for their own purposes. They say that in MBI a CI indicates the level of confidence a researcher should have that the true magnitude of an effect is within the range defined by the upper and lower bounds of the interval. Hopkins and Batterham acknowledge that this is at odds with the theoretical foundations and intended purpose of the CI, but justify their position by claiming to use an (intuitive) Bayesian approach to inference.

Bayesian Statistics

Whereas the frequentist approach to inference attempts to define the probability of getting the data given that a particular statistical hypothesis is true, the Bayesian approach attempts to determine the probability that a statistical hypothesis is true given the data [22]. The Bayesian approach requires researchers to state what they know about a statistical hypothesis before examining data [28]. As parameter values are unknown, knowledge about them is modelled as random [29]. This enables probability statements to be made about parameter values. Prior knowledge can then be modelled with a probability distribution. The probability distribution for prior knowledge (the prior) models the uncertainty about parameter values before examining data [28]. Once prior knowledge has been specified and data collected, the posterior probability distribution (the posterior) is calculated using Bayes Theorem.

The posterior models the researcher's uncertainty about the parameters after examining data [28].

Different types of probability distributions can be assigned to the prior. Consequently, critics of Bayesian analysis argue that the process of selecting a prior is subjective and therefore unscientific. Batterham and Hopkins [3, 4] claim that meta-analysis is a more objective way of utilizing prior knowledge. Notwithstanding that meta-analyses can be conducted using a Bayesian approach [30], Batterham and Hopkins' argument is flawed because subjective decisions are an inescapable part of research [31]. Meta-analyses are not immune to the subjective decisions that were made about what to include and exclude in the published reports that are used in a meta-analysis [32]. Furthermore, it is difficult to reconcile Batterham and Hopkins' argument in favour of conventional meta-analysis with their claim about using an (intuitive) Bayesian approach to inference.

Objective Bayes and the Reference Prior Method

There are different approaches to assigning priors. Subjective Bayes, which draws on expert knowledge to obtain an informative prior, maximizes what can be achieved with Bayesian analysis. However, it can be challenging when there is little existing knowledge [33]. Objective Bayes has been proposed as a less onerous approach that looks for structural rules to select a minimally informative prior [34]. This is the approach that Hopkins and Batterham [3, 4] say they use. Specifically, they claim to use a reference prior that is uniform (i.e. all possible parameter values have the same probability of occurring) [9], which results in the posterior having the same shape as the likelihood function[2]. Therefore, in MBI a CI is equivalent to a Bayesian credible interval that has a specified probability of containing the true parameter value. This results from the Bernstein-von Mises theorem. However, the theorem only holds when the sample is large, the statistical model can be specified with a finite number of parameters, the data share the same probability distribution, and the data are mutually independent (e.g. a standard normal distribution) [35]. Consequently, in MBI a CI will only be equivalent to a Bayesian credible interval when these conditions are met.

The use of a uniform prior is problematic because the parameter space is generally infinite. This means that the prior usually does not integrate to one and therefore it is not a probability distribution. Accordingly, the posterior may not integrate to one and so not be a probability distribution. Additionally, uniform priors are not invariant when parameters are transformed [34]. Hopkins and Batterham [4] argue that these issues are of little practical concern provided the sample size is fairly large. Whilst this claim may be justified if all of the conditions are met for the Bernstein-von Mises theorem to hold, a search of Google Scholar for studies that have used MBI was revealing.

The search found twenty-five studies (Appendix 1) with sample sizes ranging from eight to 99 participants. Five studies recruited 10 or fewer participants. Twelve studies had participants withdraw from data collection. Data were transformed in 13 studies. All studies used a repeated measures design for data analysis. Collectively,

[2] The likelihood function assesses how well a hypothesized parameter value predicts data.

this shows that the use of a uniform prior in MBI is actually of considerable practical concern.

There are also theoretical concerns about Hopkins and Batterham's [3, 4] approach, as it is not consistent with the reference prior method. The method aims to overcome the problems associated with a uniform prior by using information theory to identify a prior that is minimally informative and is appropriate given the proposed statistical model and the inference problem. The reference prior provides a baseline to which other priors can be compared in order to gauge the sensitivity of the posterior to changes in the prior [36]. As this is not the process that Hopkins and Batterham use, their claim about using a reference prior is not justified.

Best Practice

One of the studies [37] identified in the Google Scholar search was re-analysed using a genuine Bayesian approach [7]. The study examined the effect of a live-high-train-low training protocol versus an intermittent hypoxic exposure protocol on blood characteristics and running performance. Although the results were broadly similar to the MBI analysis, the Bayesian analysis was more conclusive and it afforded a direct probabilistic interpretation of effects, whereas the MBI analysis did not. In keeping with best practice [38, 39], the published report of the Bayesian analysis provided supplementary files containing data and statistical software code and used standard mathematical notation to give a full and explicit account of how the statistical models were implemented. In contrast, Hopkins and colleagues are yet to do this.

Hopkins and Batterham [4] generally use Microsoft Excel spreadsheet software [5, 40] to implement MBI. Various spreadsheets are available on Hopkins website (http://sportsci.org/). The use of Excel is questionable because it lacks the precision and stability necessary to accurately compute even the most basic statistics. Moreover, Microsoft does not provide information about the algorithms used by Excel's functions. Aspects of data analysis that are adversely affected include simulation, statistical distributions, and parameter estimation [41-43]. It is unclear why MBI is implemented using Excel when more accurate, stable, and widely used data science software is readily available for free.

Conclusion

Hopkins and colleagues [3-5] have restricted publishing their exposition of MBI to sports medicine and sports physiology journals. For the method to be properly vetted, a detailed account that includes data, statistical software code, and standard mathematical notation, needs to be published in a recognised statistics journal. Nevertheless, the available evidence shows that MBI uses aspects of fundamentally different statistical approaches [12, 13, 23, 44] in ways that are inconsistent with their theoretical foundations and the purpose(s) for which they were intended. Consequently, researchers should ignore Buchheit's [11] plea for them to use MBI and instead use theoretically justified methods that have established utility. All other things being equal, this will make it more likely that the researchers' findings will be credible.

References

1. Gelman A, Loken E. The statistical crisis in science. *American Scientist* 2014; **102**: 460-465.
2. Ioannidis JPA. Why most published research findings are false. *PLoS Medicine* 2005; **2**: e124.
3. Batterham AM, Hopkins WG. Making meaningful inferences about magnitudes. *International Journal of Sports Physiology and Performance* 2006; **1**: 50-57.
4. Hopkins WG, Batterham AM. Error rates, decisive outcomes and publication bias with several inferential methods. *Sports Medicine* 2016: DOI 10.1007/s40279-40016-40517-x.
5. Hopkins WG, Marshall SW, Batterham AM, Hanin J. Progressive statistics for studies in sports medicine and exercise science. *Medicine and Science in Sports and Exercise* 2009; **41**: 3-12.
6. Barker RJ, Schofield MR. Inference about magnitude of effects. *International Journal of Sports Physiology and Performance* 2008; **3**: 547-557.
7. Mengersen KL, Drovandi CC, Robert CP, Pyne DB, Gore CJ. Bayesian estimation of small effects in exercise and sports science. *PLoS ONE* 2016; **11**: e0147311.
8. Welsh AH, Knight EJ. "Magnitude-based Inference": A statistical review. *Medicine and Science in Sports and Exercise* 2015; **47**: 874-884.
9. Batterham AM, Hopkins WG. The case for magnitude-based inference. *Medicine and Science in Sports and Exercise* 2015: 855.
10. Hopkins WG, Batterham AM. An imaginary Bayesian monster. *International Journal of Sports Physiology and Performance* 2008; **3**: 411-412.
11. Buchheit M. The numbers will love you back in return-I promise. *International Journal of Sports Physiology and Performance* 2016; **11**: 551-554.
12. Fisher RA. *Statistical Methods for Research Workers.* (6th edn). Oliver & Boyd: Edinburgh, UK, 1936.
13. Neyman J, Pearson ES. "On the problem of the most efficient tests of statistical hypotheses". *Philosophical Transactions of the Royal Society of London, Series A, Containing Papers of a Mathematical or Physical Character* 1933; **231**: 289-337.
14. Seidenfeld T. R. A. Fisher's fiducial argument and Bayes' Theorem. *Statistical Science* 1992; **7**: 358-368.
15. Christensen R. Testing Fisher, Neyman, Pearson, and Bayes. *The American Statistician* 2005; **59**: 121-126.
16. Hubbard R, Bayarri MJ, Berk KN, Carlton MA. Confusion over measures of evidence(p's) versus errors (a's) in classical statistical testing. *The American Statistician* 2003; **57**: 171-182.
17. Perezgonzalez JD. Fisher, Neyman-Pearson, or NHST? A tutorial for teaching data testing. *Frontiers of Psychology* 2015; **6**: doi: 10.3389/fpsyg.2015.00223.
18. Greenland S, Senn SJ, Rothman KJ, Carlin JB, Poole C, Goodman SN, Altman DG. Statistical tests, p values, confidence intervals, and power: a guide to misinterpretations. *European Journal of Epidemiology* 2016; **31**: 337-350.

19. Nuzzo R. Statistical errors. *Nature* 2014; **506**: 150-152.
20. Kirk RE. Effect magnitude: A different focus. *Journal of statistical planning and inference* 2007; **137**: 1634-1646.
21. Baguley T. Standardized or simple effect size: What should be reported? *British Journal of Psychology* 2009; **100**: 603-617.
22. Kline RB. *Beyond significance testing.* American Psychological Association: Washington, DC, 2004.
23. Cohen J. *Statistical Power Analysis for the Behavioral Sciences.* (2nd edn). Academic Press: New York, NY, 1988.
24. Quintana D. Statistical considerations for reporting and planning heart rate variability case-control studies. *Psychophysiology* 2016; **10.1111/psyp.12798**.
25. Peng C-Y, J., Chen L-T. Beyond Cohen's *d*: Alternative effect size measures for between-subject designs. *The Journal of Experimental Education* 2014; **82**: 22-50.
26. Morey RD, Hoekstra R, Rouder JN, Lee MD, Wagenmakers E-J. The fallacy of placing confidence in confidence intervals. *Psychonomic Bulletin Review* 2015: DOI 10.3758/s13423-13015-10947-13428.
27. Neyman J. Outline of a theory of statistical estimation based on classical theory of probability. *Philosophical Transactions of the Royal Society A: Mathematical, Physical, and Engineering Sciences* 1937; **236**: 333-380.
28. Kruschke JK. What to believe: Bayesian methods for data analysis. *Trends in Cognitive Science* 2010; **14**: 293-300.
29. Gelman A, Robert CP. "Not only defended but also applied": The perceived absurdity of Bayesian inference. In "Not only defended but also applied": The perceived absurdity of Bayesian inference, Editor (ed)^(eds). arXiv.org: City, 2012.
30. Cochrane Collaboration. Bayesian and hierarchical approaches to meta-analysis In Bayesian and hierarchical approaches to meta-analysis Higgins J, Green S (eds). The Cochrane Collaboration, 2011.
31. Fischhoff B. *Judgment and Decision Making.* Earthscan: Milton Park, UK, 2012.
32. Forstmeier W, Wagenmakers E-J, Parker TH. Detecting and avoiding likely false-positive findings - a practical guide. *Biological Reviews* 2016 1-28.
33. Berger J. The case for objective Bayesian analysis. *Bayesian Analysis* 2006; **1**: 385-402.
34. Kass RE, Wasserman L. The selection of prior distributions by formal rules. *Journal of the American Statistical Association* 1996; **91**: 1343-1370.
35. Freedman D. On the Bernstein-von Mises theorem with infinite-dimensional parameters. *The Annals of Statistics* 1999; **27**: 1119-1140.
36. Irony TZ, Singapurwalla ND. Non-informative priors do not exist A dialogue with Jose M. Bernado. *Journal of statistical planning and inference* 1997; **65**: 159-189.
37. Humberstone-Gough C, Saunders PU, Bonetti DL, Stephens S, Bullock N, Anson JM, Gore CJ. Comparison of live high: train low altitude and intermittent hypoxic exposure. *Journal of Sports Science & Medicine* 2013; **12**: 394-401.
38. Kass RE, Caffo BS, Davidian M, Meng X-L, Yu B, Reid N. Ten simple rules for effective statistical practice. *PLoS Computational Biology* 2016; **12**: e1004961.
39. Leek J. *The elements of data analytic style: A guide for people who want to analyze data.* Leanpub: Victoria, BC, 2015.

40. Hopkins WG. A spreadsheet for deriving a confidence interval, mechanist inference and clinical inference from a p value. *Sportscience* **2007; 11**:16-20.
41. Almiron MG, Lopes B, Oliveira ALC, Medeiros AC, Frery AC. On the numerical accuracy of spreadsheets. *Journal of Statistical Software* 2010; **34**: 1-29.
42. McCullough BD, Heiser DA. On the accuracy of statistical procedures in Microsoft Excel 2007. *Computational Statistics and Data Analysis* 2008; **52**: 4570-4578.
43. Mélard G. On the accuracy of statistical procedures in Microsoft Excel 2010. *Computational Statistics* 2014; **29**: 1095-1128.
44. Berger J, Bernardo JM, Sun D. The formal definition of reference priors. *The Annals of Statistics* 2009; **37**: 905-938.

Appendix 1
MBI Studies 2008-2016

Study	Sample	Participant Attrition	Incomplete Data	Groups	Analysis Design	Transform
Antonio J, Ciccone V. The effects of pre versus post workout supplementation of creatine monohydrate on body composition and strength. Journal of the International Society of Sports Nutrition. 2013;10(1):1.	22	3		2	Repeated measures	
Brocherie F, Girard O, Faiss R, Millet GP. High-intensity intermittent training in hypoxia: a double-blinded, placebo-controlled field study in youth football players. The Journal of Strength & Conditioning Research. 2015;29(1):226-37.	16			2	Repeated measures	
Buchheit M, Haydar B, Ahmaidi S. Repeated sprints with directional changes: do angles matter? Journal of sports sciences. 2012;30(6):555-62.	12			1	Repeated measures	Log
Buchheit M, Racinais S, Bilsborough J, Hocking J, Mendez-Villanueva A, Bourdon P et al. Adding heat to the live-high train-low altitude model: a practical insight from professional football. British Journal of Sports Medicine. 2013;47(Suppl 1):i59-i69.	19	2		2	Repeated measures	Log

Chesterton P, Weston M, Butler M. The effect of mobilising the lumbar 4/5 zygapophyseal joint on hamstring extensibility in elite soccer players. International Journal of Physiotherapy and Rehabilitation. 2016;April:1-13.	25			2	Repeated measures	Log
Cockburn E, Stevenson E, Hayes PR, Robson-Ansley P, Howatson G. Effect of milk-based carbohydrate-protein supplement timing on the attenuation of exercise-induced muscle damage. Applied Physiology, Nutrition and Metabolism. 2010;35(10):270-7.	32			4	Repeated measures	Log
Gonzalez AM, Hoffman JR, Rogowski JP, Burgos w, Manalo e, Weise K et al. Performance changes in NBA basketball players vary in starters vs. nonstarters over a competitive season. Journal of Strength & Conditioning Research. 2013;27(3):611-5.	12	5		2	Repeated measures	
Humberstone-Gough CE, Saunders PU, Bonetti DL, Stephens S, Bullock N, Anson JM et al. Comparison of live high: train low altitude and intermittent hypoxic exposure. Journal of Sports Science & Medicine. 2013;12:394-401.	24	1	4 sets	3	Repeated measures	Log

Harris N, Cronin JB, Hopkins WG, Hansen KT. Squat jump training at maximal power loads vs. heavy loads: Effect on sprint ability. Journal of Strength & Conditioning Research. 2008;22(6):1742-9.	18			2	Repeated measures	
Iaia FM, Fiorenza M, Perri E, Alberti G, Millet GP, Bangsbo J. The effect of two-speed endurance training regimes on performance of soccer players. PloS one. 2015;10(9):e0138096.	18	5		2	Repeated measures	
Inoue A, Impellizzeri FM, Pires FO, Pompeu FA, Deslandes AC, Santos TM. Effects of Sprint versus High-Intensity Aerobic Interval Training on Cross-Country Mountain Biking Performance: A Randomized Controlled Trial. PloS one. 2016;11(1):e0145298.	16	4		2	Repeated measures	
Kakanis MW, Peake J, Brenu EW, Simmonds M, Gray B, Marshall-Gradisnik SM. T helper cell cytokine profiles after endurance exercise. Journal of Interferon & Cytokine Research. 2014;34(9):699-706.	10				Repeated measures	Log
Loturco I, Pereira LA, Kobal R, Kitamura K, Ramírez-Campillo R, Zanetti V et al. Muscle Contraction Velocity: A Suitable Approach to Analyze the Functional Adaptations in Elite Soccer Players. Journal of Sports	22			1	Repeated measures	

Science & Medicine. 2016;15(3):483.						
Moore DR, Areta J, Coffey VG, Stellingwerff T, Phillips SM, Burke LM et al. Daytime pattern of post-exercise protein intake affects whole-body protein turnover in resistance-trained males. Nutrition & Metabolism. 2012;9(91):1-5.	24	1		3	Repeated measures	Log
Mendiguchia J, Martinez-Ruiz E, Morin J, Samozino P, Edouard P, Alcaraz P et al. Effects of hamstring-emphasized neuromuscular training on strength and sprinting mechanics in football players. Scandinavian Journal of Medicine & Science in Sports. 2015;25(6):e621-e9.	60	9		2	Repeated measures	
Pearcey GEP, Bradbury-Squires DJ, Kawamoto J-E, Drinkwater EJ, Behm DG, Button DC. Foam rolling for delayed-onset muscle soreness and recovery of dynamic performance measures. Journal of Athletic Training. 2015;50(1):5-13.	8			1	Repeated measures	Log
Pearson J, Rowlands D, Highet R. Autologous blood injection to treat Achilles tendinopathy? A random controlled	33	10		2	Repeated measures	

trial. Journal of Sports Rehabilitation. 2012;21:218-24.						
Robertson EY, Saunders PU, Pyne DB, Gore CJ, Anson JM. Effectiveness of intermittent training in hypoxia combined with live high/train low. European Journal of Applied Physiology. 2010;110(2):379-87.	17	4		2	Repeated measures	Log
Rowlands DS, Rossler K, Thorp RM, Graham DF, Timmons BW, Stannard SR et al. Effect of dietary protein content during recovery from high-intensity cycling on subsequent performance and markers of stress, inflammation, and muscle damage in well-trained men. Applied Physiology, Nutrition, and Metabolism. 2008;33:39-51.	12			2	Repeated measures	Log
Scanlan AT, Dascombe BJ, Reaburn PRJ, Osborne M. The effect of wearing lower-body pressure garments during endurance cycling. International Journal of Sports Physiology and Performance. 2008;3:424-38.	12			2	Repeated measures	
Schubert MM, Astorino TA, Azevedo JL. The effects of caffeinated “energy shots” on time trial performance. Nutrients. 2013;5(6):2062-75.	9	3		1	Repeated measures	

Smith AE, Fukuda DH, Ryan ED, Kendall KL, Cramer JT, Stout J. Ergolytic/ergogenic effects of creatine on aerobic power. International Journal of Sports Medicine. 2011;32.	55			2	Repeated measures	Log
Tofari PJ, Cormack SJ, Ebert TR, Gardner AS, Kemp JG. Comparison of ergometer and track-based testing in junior track-sprint cyclists. Implications for talent identification and development. Journal of Sports Sciences 2016 doi:10.1080/02640414.2016.1243795.	10			1	Repeated measures	Log
West NP, Pyne DB, Cripps AW, Hopkins WG, Eskesen DC, Jairath A et al. Lactobacillus fermentum (PCC) supplementation and gastrointestinal and respiratory-tract illness symptoms a randomised control trial in athletes. Nutrition Journal. 2011;10(30):1-11.	99	12		2	Repeated measures	Log
White AC, Salgado RM, Astorino TA, Loeppky JA, Schneider SM, McCormick JJ et al. The effect of 10 days of heat acclimation on exercise performance in acute hypobaric hypoxia (4350 m). Temperature. 2016;3(1):176-85.	8			1	Repeated measures	