Trivial effects are *clearly* important

Martin Buchheit [1]

[1] Paris Saint Germain, Saint-Germain-En-Laye, France.

Expert Opinion | Statistics | MBI | Clear effects

Headline

No later than the 19th of December, I was very privileged to attend a pretty interesting PhD defense at the French Institute of Sports (INSEP, Paris, France). It was about the role of post-training protein supplementation on body composition and muscle strength adaptations, and more importantly the effects of the type of those proteins (i.e., slow vs. fast assimilation). This topic is super trendy at the moment (1), and highly relevant for us practitioners working in the field with athletes. During 50 min, the candidate gave a nice and entertaining presentation of the key findings of her researches, which are partly summarized below. After her talk, no one could stop congratulating her for the hard work she put in while running such a longitudinal experimentation. However, there were intense discussions between the referees (all very renowned scientists), the candidate and her supervisors about the statistical treatment of the results and the study design (detailed below too). Being a simple attendee it would have been inappropriate to try to take part into the debate, so I listened and stayed quiet. But you'll guess that I wasn't really satisfied with was finally agreed from that debate. I am now taking the opportunity to react in a very open and transparent manner.

Discussion

The type of protein doesn't matter (?). Here is the abstract of the study I will refer to (2):

> *While effects of the two classes of proteins found in milk (i.e. soluble proteins, including whey, and casein) on muscle protein synthesis have been well investigated after a single bout of resistance exercise (RE), the combined effects of these two proteins on the muscle responses to resistance training (RT) have not yet been investigated. Therefore, the aim of this study was to examine the effects of protein supplementation varying by the ratio between milk soluble proteins (fast-digested protein) and casein (slow-digested protein) on the muscle to a 9-week RT program. In a double-blind protocol, 31 resistance-trained men, were assigned to 3 groups receiving a drink containing 20g of protein comprising either 100% of fast protein (FP(100), n=10), 50% of fast and 50% of slow proteins (FP(50), n=11) or 20% of fast protein and 80% of casein (FP(20), n=10) at the end of training bouts. Body composition (DXA), and maximal strength in dynamic and isometric were analyzed before and after RT. Moreover, blood plasma aminoacidemia kinetic after RE was measured. The results showed a higher leucine bioavailability after ingestion of FP(100) and FP(50) drinks, when compared with FP(20) ($p<0.05$). However, the RT-induced changes in lean body mass ($p<0.01$), dynamic ($p<0.01$), and isometric muscle strength ($p<0.05$) increased similarly in all experimental groups. To conclude, compared to the FP(20) group, the higher rise in plasma amino acids following the ingestion of FP(100) and FP(50) did not lead to higher muscle long-term adaptations.* **Key words:** *Milk soluble protein and Casein; Lean Body Mass; Isometric and dynamic muscle strength*

1. **Key findings of the published paper (2):** While there was a significant main training effect ($P<0.01$ = fat-free mass improved in all groups when analyzed together), all time x supplement interaction effects were non-significant ($P>0.05$, Figure 1) = there was no difference between the treatments.
   → In real words for practitioners it means that the type of protein ingested post-training may not matter if you want to build muscle or strength. However, whether the overall improvement seen in the three groups was actually meaningful is unknown. "It's significant, but does it matter?" In fact, this could not be answered with the present statistical approach (3).
2. **Key findings as presented during the talk:** (*in response to one of the referee's comments during the pre-defense review, she thankfully added some magnitude-based inferences (MBI) stats into her presentation, Figure 1*): 1) Fat-free mass increased by a ***possibly*** **small** amount in all the three groups (within-group changes, Figure 1 and 2: the actual improvements in fat-free mass (≈+2.5% or ≈+0.25 in standardized unit following Cohen'd approach) were greater than the so-called smallest worthwhile change (SWC, 0.2 in standardized unit = 0.2 x between-athlete SD, ≈2.3%), however 2) the differences between the effects of the 3 groups (between-group differences in the changes) were all ***clearly*** **trivial** (Figure 3).
   While the mention of the effects magnitude were partly noted by some of the referees during the talk, they were mainly seen as superficial analyses that complicated the presentation of the results ("why add this stuff?"), rather than improving the overall message of the study.
   → In real words for practitioners it means that the present strength program and associated supplementation options are all ***possibly*** to lead to worthwhile improvement in fat-free mass, although the increase is possibly to be of a **small** magnitude only. However, it is ***unlikely*** that the actual composition of post-exercise drinks affects the changes in fat-free mass following such a strength program.

Let's re-do, but better (?). This is the main conclusion from the discussion between the experts and the candidate: with the 3 groups, leading to not-so-large samples size (n = 10, 11 and 10), it was very likely that the analysis lacked power; it is therefore likely that the lack of between-group differences (no significant interaction ^^) was related to a type 2 error (when there may be a difference but the lack of power doesn't allow you to get tiny P values, and then you fail to show that difference). They all agreed that this lack of power was a problem and concluded that if they had to redo the study, they would split the subjects into two groups only instead of three (which in effect increases power). Everyone congratulated each other once more and then we all had a drink (or two).

Seeking for the best shaker composition, in the end. In fact, those people wouldn't need to complain about the poor samples size if they had had the insight to consider 1) the actual **magnitude** of the effects that the candidate still kindly added in her talk on the D day (Figure 1) and 2) the likelihood for these effects to be true (Figure 2). I invite readers to read

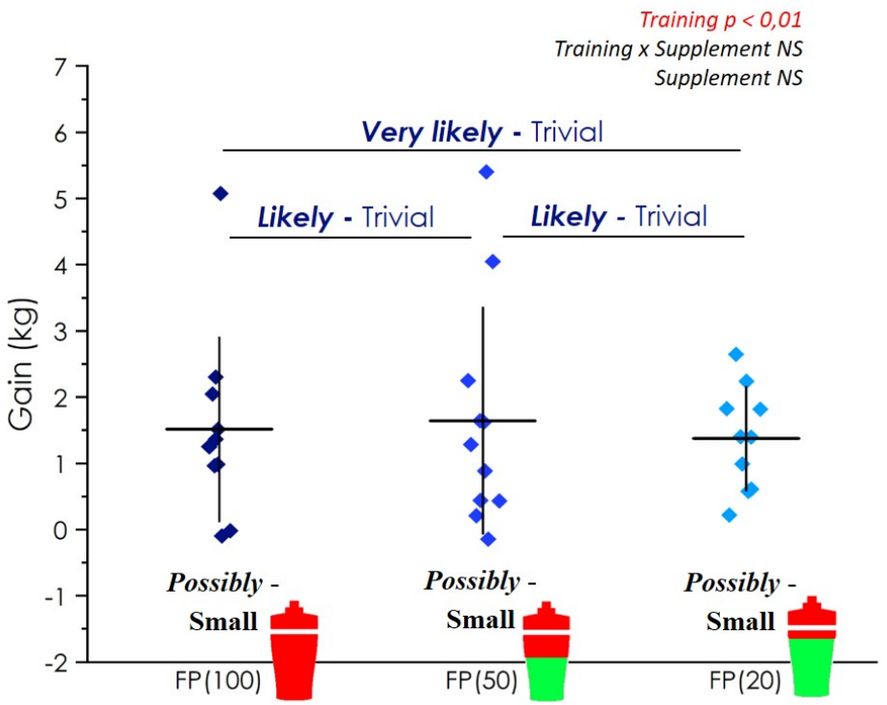

Fig. 1. Figure 1. Changes in free-fat mass following the training intervention on the three groups receiving a drink containing 20g of protein comprising either 100% of fast protein (FP(100), n=10), 50% of fast and 50% of slow proteins (FP(50), n=11) or 20% of fast protein and 80% of casein (FP(20), n=10) at the end of training bouts. Reproduced with permission from a slide presented during the PhD defense (2).

Log-transformed Data		One predictor	
Custom effects in standardized units		Custom effect, not adjusted	Custom effect, not adjusted
Confidence level (%)		90	90
Degrees of freedom		10	10
Standardized custom effect		**0.23**	**0.23**
Confidence limits	lower	0.10	0.10
	upper	0.36	0.36
	"±"	0.13	0.13
Thresholds for inferences	+ive or harmful	0.20	0.20
	-ive or beneficial	-0.20	-0.20
Chances (% and qualitative) that the true value of the statistic is non-clinically or clinically (practically)...	substantially positive (+ive) or harmful	66 possibly	66 possibly
	trivial	34 possibly	34 possibly
	substantially negative (-ive) or beneficial	0 most unlikely	0 most unlikely
Non-clinical inference, based on threshold chances of 5% for substantial magnitudes		possibly +ive	possibly +ive
Clinical inference, based on threshold chances of harm and benefit of 0.5% and 25%		possibly harmful	possibly harmful
Odds ratio for benefit/harm		0	0
Clinical inference as above, but declaring beneficial when odds ratio of benefit/harm is >66		possibly harmful	possibly harmful

Log-transformed Data		One predictor	
Difference in custom effects in standardized units		Custom effect, adj. for X1	Custom effect, not adjusted
Confidence level (%)		90	90
Degrees of freedom		17	19
Standardized difference in custom effects		**0.01**	**0.01**
Confidence limits	lower	-0.16	-0.15
	upper	0.18	0.17
	"±"	0.17	0.16
Thresholds for inferences	+ive or harmful	0.20	0.20
	-ive or beneficial	-0.20	-0.20
Chances (% and qualitative) that the true value of the statistic is non-clinically or clinically (practically)...	substantially positive (+ive) or harmful	4 very unlikely	3 very unlikely
	trivial	94 likely	95 very likely
	substantially negative (-ive) or beneficial	3 very unlikely	2 very unlikely
Non-clinical inference, based on threshold chances of 5% for substantial magnitudes		likely trivial	very likely trivial
Clinical inference, based on threshold chances of harm and benefit of 0.5% and 25%		likely trivial	very likely trivial
Odds ratio for benefit/harm		1	1
Clinical inference as above, but declaring beneficial when odds ratio of benefit/harm is >66		likely trivial	very likely trivial

Fig. 2. **Left**: Example of the results obtained using the Post-only Crossover Group trial spreadsheet,(4) to examine the changes following training using the FP(50) drink. There were 66% chances (***possibly***) that the standardized change (Cohen's d: 0.23) may be **small** (i.e., greater than the smallest worthwhile change, SWC, Cohen's d: 0.2 - considering that a Cohen's d from 0.2 to 0.6 is considered as small (5)). **Right**. Example of the results obtained using the Pre-Post Parallel Group trial spreadsheet (6) to compare the difference in the changes following training using the FP(50) vs. FP(100) drinks. It appears that the difference in the change (Cohen's d: 0.01) is ***likely*** to be **trivial** (94% of chances to be lower than the SWC), with the chances for the changes with FP(100) to be greater and lower than that with FP(50) being ***very unlikely*** (with 4 and 3% to be greater than the SWC, respectively). All effects were adjusted for pre-training fat-free mass (results highlighted in yellow – note that the adjustment didn't affect much the results in comparison with the non-adjusted output). Note that both spreadsheets are available for download, with a short explanatory note on how to fill some of those cells.

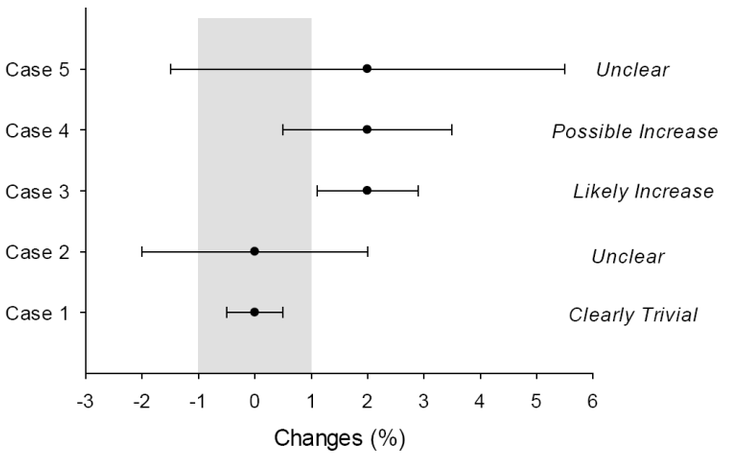

Fig. 3. Example of possible decisions when interpreting changes using magnitude-based inferences. Note the clear vs. unclear cases (based on confidence limits, in relation to the shaded trivial area), which i) is one of the extreme beauty of magnitude-based inferences and ii) provide no insight through null hypothesis significance testing. Note also how, for clear effects, the likelihood of changes increases as the confidence limits shrink. Adapted from (3).

more on the topic (3), but keep in mind that one of the beauty of MBI is to deal with the concept of ***clear vs. unclear*** effects (Figure 3), which can't be touched using null hypothesis significance testing. Case 1, 3 and 4: example when we have enough confidence to say that the effects are of the observed magnitude, since the confidence intervals (CI) stay within (**trivial**, Case 1) or don't touch much (Case 3 and 4) the smallest worthwhile change (SWC); it's ***clearly***... of whatever **magnitude**. Case 2 and 5: the CI are too large and overlap the SWC, so that we can't make a decision; it's ***unclear***. In that case, AND IN THAT CASE ONLY, increasing sample size is important since it will decrease the CI and allow to make the decision about where the actual effects sit.

When this reasoning is applied to our present study:

1. With more subjects, the **small** improvements seen after the three conditions would likely remain **small**, but it will just become more certain that those improvements were actually of that magnitude (e.g., from ***possibly*** **small** to ***almost certainly*** **small**). So why would we need more subjects? Note that we anticipate no change in the main effect magnitude with increased sample size, unless those 'new' subjects were to respond completely differently to the same protocol (which is another possible problem due to individual responses and sample variations, but that is something out of the scope of that paper that would also affect the decisions based on P values). What is expected in general is that if the response recorded is robust enough for a sub-sample, every other subject (of similar characteristic) using the same drink and training the same should respond more or less in the same manner. This is what stats are all about, making inferences from sub-samples responses into what should happen for a larger population. So in this case, since we already had a ***clear*** decision on that small effect, why would we need more subjects!? Please keep the three groups (which adds a lot to the study to me)!!
2. The last point is the one that pushed me to write the present note. "Lack of difference between the group effects, and the fear of the type 2 error". The candidate reported on her slides (Figure 1) some ***likely*** **trivial** differences between the effects of the three different drinks. This means that the magnitude of the differences in the effects was too small to be important (<SWC, i.e. **trivial**), but more importantly, that the confidence level was already good enough to make the decision (***clearly*** **trivial**)! So if it was already ***clearly*** **trivial**, no matters how many more subjects we would add, the difference would remain trivial!! (excluding again the possible sample variation issue of course). The consideration of these ***likely*** **trivial** differences shows that the recommendation of the experts to increase sample was in fact unfounded, and more importantly, unnecessary!

Conclusion and practical applications. When we all left the amphitheater after the drinks, I believe that if I had interviewed the audience, the main take-home message (based on the experts discussion) would have been that while the candidate had produced amazing efforts and did well with her talk, we were still waiting for a definitive answer to prepare the best post-training drinks, and that only re-doing another study with larger sample sizes will allow finding the truth. Not sure what those people and the PhD referees will put in their next shakers, but as far as I am concerned, I don't need to wait for that new study at all. I won't pay much attention to the whey content of my shakers and will likely focus on other aspects of training and supplementation to make a difference!

Twitter: Follow Martin Buchheit @mart1buch.

References

1. Phillips, S.M., A Brief Review of Critical Processes in Exercise-Induced Muscular Hypertrophy. Sports Med, 2014. 44(Suppl 1): p. S71–S77.
2. Fabre, M., et al., Effects of Post-Exercise Protein Intake on Muscle Mass and Strength During Resistance Training: is There an Optimal Ratio Between Fast and Slow Proteins. Int J Sport Nutr, 2017: p. In Press.
3. Buchheit, M., The Numbers Will Love You Back in Return-I Promise. Int J Sports Physiol Perform, 2016. 11(4): p. 551-4.
4. Hopkins, W.G. Spreadsheets for analysis of controlled trials, crossovers and time series. http://www.sportsci.org/resource/stats/xPostOnlyCrossover.xlsx 2017. 21, 1-4 DOI: http://newstats.org/xcl.xls.
5. Hopkins, W.G., et al., Progressive statistics for studies in sports medicine and exercise science. Med Sci Sports Exerc, 2009. 41(1): p. 3-13.

6. Hopkins, W.G. Spreadsheets for analysis of controlled trials, parallel groups and time series. http://www.sportsci.org/resource/stats/xParallelGroupsTrial.xlsx Sportscience, 2017. 21, 1-4 DOI: http://newstats.org/xcl.xls.