Doi: 10.51224/SRXIV.364 SportRχiv Preprint version 1

Adequate statistical power in strength and conditioning may be achieved through longer interventions and high frequency outcome measurement.

Paul, A. Swinton

Doi: 10.51224/SRXIV.364
SportRχiv hosted preprint version 1
28/12/2023
PREPRINT - NOT PEER REVIEWED

Contact details
Dr. Paul Swinton
School of Health Sciences, Robert Gordon University
Garthdee Road
Aberdeen, UK,
AB10 7QG
p.swinton@rgu.ac.uk, +44 (0) 1224 262 3361
Twitter Handle:
@PaulSwinton9

Please cite as: Swinton, PA. Adequate statistical power in strength and conditioning may be achieved through longer interventions and high frequency outcome measurement. 2023. Pre-print available from SportRχiv. https://doi.org/10.51224/SRXIV.364.

Doi: 10.51224/SRXIV.364 SportRχiv Preprint version 1

Abstract

Pre-post randomised controlled trials (RCT) are the most common design used to build an evidence base in strength and conditioning but are limited by small effects, small samples sizes, and concomitant low statistical power. The purpose of this study was to explore the effects of manipulating a range of factors including intervention length, frequency, and pattern of outcome measurements on the sample size required to achieve adequate statistical power. A case-based approach was used to enhance applicability.

Realistic data generating patters were considered for hypothetical RCTs investigating resistance training interventions to improve maximum strength as measured by the 1RM bench press. Improvements for the 'reference' intervention and subsequent average treatment effect for the 'testing' intervention were matched to data summarised in recent large meta-analyses. Different measurement error magnitudes were added to the high frequency RCT data to recreate the use of 1RM prediction methods that could be used during training sessions. A closed form solution linking statistical power and sample size was used to explore different strategies with simulations performed as a final check.

The results showed large improvements in statistical power could be achieved when conducting interventions over a longer period (e.g. 18 weeks), and/or performing multiple outcome measurements. Efficient reductions in required sample sizes could be achieved by performing multiple measurements at baseline and post-intervention. This strategy, however, may be limited by induced fatigue or training effects. Similar reductions in sample size could be achieved by performing high-frequency measurement throughout the intervention. This reduction in sample size was demonstrated despite acute increases in measurement error (factor of 1.5 and 2) that would occur when using prediction methods.

In conclusion, very low statistical power is likely the norm in pre-post RCTs in strength and conditioning. Simply increasing sample size is unlikely to remedy the situation given the resource constraints that are common in the discipline. The results of this study suggest that researchers should consider other strategies including longer interventions and high frequency data collection to obtain adequate statistical power with feasible resources.

Doi: 10.51224/SRXIV.364 SportRχiv Preprint version 1

Introduction

A broad range of strength and conditioning approaches and training regimes have been shown to cause readily observable improvements across many populations and outcome domains of interest. Within this context, it is common for athletes and coaches to seek interventions that maximise improvements. To assist in this regard, researchers frequently compare similar training practices using randomised controlled trials (RCTs). Previous reviews of RCTs in strength and conditioning have shown that most studies employ a pre-post design (single baseline and post-intervention measurement) with interventions lasting between eight to twelve weeks and include sample sizes of ten to twenty per group.[1,2] In a recent simulation study creating data to reflect patterns typical in strength and conditioning, I demonstrated that statistical power is likely to be very low.[3] Statistical power was shown to be influenced by a range of factors including the average treatment effect (ATE), sample size, treatment response heterogeneity, and measurement error.[3] When analysing a small ATE contextualised for strength and conditioning, statistical power tended to remain below ~0.2 for sample sizes less than 20 per group, and only reached ~0.4 for small measurement errors and a sample size of 50 per group.[3] Given there is a strong desire in strength and conditioning to identify the most effective approaches and training regimes, RCTs will frequently compare similar interventions such that the ATE should be expected to be small. With greater sample sizes a concomitant increase in statistical power will occur. However, given 90% of RCTs in strength and conditioning include sample sizes of 20 or less per group,[1] it does not appear that simply increasing the number of participants is a viable strategy and alternatives are required that fit within the typical resource constraints of the discipline.

I have suggested a move-away from pre-post designs and instead collecting high-frequency data measured daily or multiple times per week.[3] Whilst, the collection of high-frequency data is unlikely to be a solution to low statistical power in all scenarios, there may be a range for which the approach obtains adequate statistical power for realistic sample sizes. The ability to collect high-frequency data will depend on the intervention and the outcome of interest. For cases where measurements are not cost or resource-intensive and do not interact with the intervention itself, high frequency measurement may be easy to adopt. Examples include measures of body composition through body mass and bioimpedance analysis that can be measured on a daily basis.[4-6] Additionally, more laboratory-based body composition measurements including use of ultrasound or 3D body scanning could potentially be used at each participant visit if training regimes are supervised.[7,8] There is also the potential for participant reported outcomes such as pain and muscle soreness to be recorded multiple times per day with digital technologies available to robustly collect the data.[9] Similarly, technology

exists for participants to conduct their own physical tests such as the vertical jump and accurately measure a range of variables using smartphone applications.[10]

In contrast to cases where testing does not interact with training interventions, there are likely to be more cases where interactions are severe. For example, in many studies comparing resistance training interventions the primary outcome of interest is maximum strength. The frequent measurement of this outcome domain using standard tests such as 1RMs are likely to alter the training stimulus rendering the RCT invalid. It may be possible, however, to predict maximum strength or other relevant outcomes using data that can be collected during the training intervention itself. Where the intervention includes performance of repetition maximums, previously validated regression equations that predict 1RM from the load lifted and the number of repetitions performed can be used.[11,12] Alternatively, where sets are not performed to momentary failure, but repetitions are performed with the intention to move the load as fast as possible, more novel approaches including use of individualised load-velocity relationships can be used to predict 1RM.[13] Velocity of the barbell can be measured using a range of technologies including low-cost options [14] with our recent review suggesting this approach would be useful for high frequency data collection and monitoring.[13] With many of the approaches suggested measurement error may be increased compared with tests that are commonly used for pre-post designs. The increased frequency of data collection, however, may offset the acute increase in measurement error and ultimately improve ATE precision and thereby statistical power.

There are a range of statistical approaches that can be used to analyse high-frequency data collected within an RCT. It has been argued that linear mixed models (LMMs) should be the default statistical approach to analyse experimental data [15] and there have been calls for greater use in sport and exercise science beyond traditional methods such as repeated measures ANOVA.[16] The advantages of LMMs over traditional methods include greater statistical power, improved ability to handle missing data, and estimation of parameters that provide relevant insights into the data generating mechanism.[17] LMMs include fixed effects and random effects, the former represent population-level (i.e., average) trends that should be persistent across experiments, and the latter represent the extent to which these trends vary across some grouping factor such as participants.[17] LMMs are linear in their parameters but can include non-linear variables such as time from baseline allowing curvilinear changes over an intervention to be modelled.[18] A large observational analysis conducted by Steele et al.[19] of resistance training over seven years was shown to be appropriately modelled with linear-log growth such that 30-50% of improvements were obtained over the first year. In the context of most RCTs in strength and conditioning,

however, interventions are generally short, and an assumption of linear change may be appropriate, estimable, and easier to interpret. The purpose of this study, therefore, was to explore whether a simple class of LMM could be applied to high frequency data from a typical RCT conducted in strength and conditioning to achieve adequate statistical power. Exploration was conducted using a closed form solution linking statistical power and sample size. Representative parameter values were used and combined with systematic variation of measurement error, and pattern of outcome measurements to achieve the study purpose.

Methods

Approach to the problem

In this study, results from recent large scale meta-analyses in strength and conditioning were combined with research on the accuracy of prediction methods to generate realistic data patterns across a range of scenarios that could be investigated for statistical power and sample size. To enhance the interpretability of the findings, a case study approach was used. The case was of an intervention study investigating the ATE between a "reference" resistance training intervention known to be successful and a new "testing" resistance training intervention hypothesised to be superior in improving maximum strength. The primary outcome measure was the 1RM bench press, which could be predicted using individualised load-velocity relationships.[20-31] Statistical power was approximated using a closed form solution and compared: 1) traditional pre-post design; 2) multiple pre-post measurements; and 3) high-frequency data collected throughout the intervention. A range of realistic scenarios were investigated with manipulations made to: 1) the ATE magnitude; 2) the sample size in each group; 3) the frequency of measurements; 4) the magnitude of measurement errors; and 5) the strength of relationship between baseline values and change scores.

Data generating and statistical model

Typically, we conceptualise the data generating model for RCT's in discrete time with the true scores of participants denoted by Y_{ijk}, where $i = 1,2,\dots,N$ indexes participants, $j = 0,1$ indexes the reference and testing interventions, and $k = 0,1$ indexes the baseline and post-intervention measurements. Previously I outlined two data generating mechanisms referred to as the independent and constrained linear cases.[3] For the independent case we have $Y_{ij1} = Y_{ij0} + \Delta_j + \xi_{ij1}$, where Δ_j is a constant describing the mean change in group j, and $\xi_{ij1} \sim N(0, \nu_1^2)$. The ATE is equal to $\Delta_1 - \Delta_0$ and ν_1^2 quantifies individual variation in change scores. For the constrained linear case we assume that change across the intervention is influenced by the baseline value. In notation we have $Y_{ij1} = Y_{ij0} + \Delta_j + \tau Y_{ij0} + \tilde{\xi}_{ij1}$, where τ sets the slope of the linear relationship between the true change score and baseline true score which is the same for each group, and $\tilde{\xi}_{ij1} \sim N(0, \tilde{\nu}_1^2)$ describes any further variation in true change score and is independent from Y_{ij0}. The data generating mechanism is constrained by the fact that τ and $\tilde{\nu}_1^2$ are the same across the groups.

Where we can collect data during an intervention it is advantageous to consider other data generating mechanisms and associated statistical analyses. It is common to refer to RCTs with intermediate testing as $S:T$ repeated measures designs,

where S denotes the number of measurements made at baseline, and T denotes the number of measurements made after the intervention has started.[32] LMMs provide a powerful and flexible framework to conceptualise and analyse longitudinal data. A simple LMM that could be used to analyse relatively short interventions (e.g. up to 6 months) is the random intercept and slope LMM with treatment by linear time interaction.[32] The model can be expressed as

$y_{ijt} = b_{0i} + \beta_0 + \beta_1 G_{ij} + (\beta_2 + \beta_3 G_{ij} + b_{1i})t + \epsilon_{ijt}$, where b_{0i} is the participant random intercept which is distributed as $b_{0i} \sim N(0, \sigma_{b0}^2)$ and denotes how much above or below the mean baseline value (β_0) each participant is, G_{ij} is a group indicator variable equal to 0 for the standard intervention, and 1 for the test intervention, β_1 describes any group offset at baseline, β_2 is the mean rate of change of the standard group (change in outcome per week), β_3 is the difference in rate of change of the test intervention relative to the standard intervention (also referred to as the treatment by linear time interaction or comparative effect), and b_{1i} is the participant random slope which is distributed as $b_{1i} \sim N(0, \sigma_{b1}^2)$ and denotes how much each participants rate of change is above or below the mean group rate of change. b_{0i} and b_{1i} follow a multivariate normal distribution $\begin{matrix} b_{0i} \\ b_{1i} \end{matrix} \sim \mathbb{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_{b0}^2 & \rho\sigma_{b0}\sigma_{b1} \\ \rho\sigma_{b0}\sigma_{b1} & \sigma_{b1}^2 \end{bmatrix}\right)$, and $\epsilon_{ijt} \sim N(0, \delta^2)$ describes variability due to factors not included in the model and measurement error. The multivariate relationship between b_{0i} and b_{1i} enables the random effects to be related such that when $\rho = 0$ we have the independent case and pre-intervention value is not related to change, and when $\rho \neq 0$ they are related and we can match this to the constrained linear case (see the appendix for details).

An advantage of considering data to be generated by the above LMM is that simple albeit approximate closed form solutions are available to link sample size and statistical power.[33] These formulas can then be explored across various manipulations to identify for example what intervention length and measurement frequencies can obtain adequate statistical power given constraints on sample size. Using maximum likelihood estimates, Tango [32] derived the following notation and equation to link sample size and statistical power:

Y_{ijt_k}, where $k = -(S-1), -(S-2), \ldots, 0, 1, \ldots T$,

$n_{S,T} = 2\frac{(Z_{\alpha/2} + Z_\phi)^2}{\beta_3^2}\left(\sigma_{b1}^2 + \delta^2 Q(S,T)^{-1}\right)$,

where α is the probability of a Type I error (usually set to 0.05), statistical power is $1 - \phi$ (usually set to 0.8), $n_{S,T}$ is the sample size required for both groups, Z_γ denotes the upper 100γ% percentile of the standard normal distribution, β_3 is the

comparative effect size, σ_{b1}^2 describes primarily between subject variability (e.g. response variability), and δ^2 describes within subject variability due to a broad conception of measurement error.[34] We also have that

$$Q(S,T) = \sum_{k=0}^{T}(t_k - \bar{t})^2 + \bar{t}^2\,\frac{(S-1)(T+1)}{S+T},$$

where $\bar{t} = \sum_{k=0}^{T}\frac{t_k}{T+1}$, $t_0 = 0$, and our T post-baseline time of measurements are rounded to a decimal value, so that if our first two post-baseline measurements were in the middle and end of week 1 we would have $t_1 = 0.5$ and $t_2 = 1.0$.

To explore statistical power and sample size using the above formulas and make it relevant to strength and conditioning, a range of scenarios were considered with systematic modifications made to: 1) the intervention length (6,12, and 18 weeks); 2) the measurement structure (multiple baseline and post intervention measurements, or single baseline and equally spaced measurements); 3) the measurement frequency (1, 2, 4, and 7 days/week); 4) the comparative effect size (small: β_3 =0.1875 and medium: β_3 =0.375); 5) the magnitude of measurement errors (5, 7.5, 10 kg); and 6) changes to sampling to reduce σ_{b0} (σ_{b0} =15, σ_{b0} =10, and σ_{b0} =5 kg). Further discussions of parameter values and attempts to align the LMM with standard conceptions of pre-post designs are presented in the appendix. Statistical power-sample size curves were generated with statistical power set from 0.1 to 0.9. To provide a check on the values obtained, direct simulations were conducted on those scenarios where statistical power of 0.8 was obtained with sample size $\leq$50. Simulations were performed in R,[35] with the doParallel package [36] used to conduct parallel computation. R code for the study is presented in the appendix.

Results

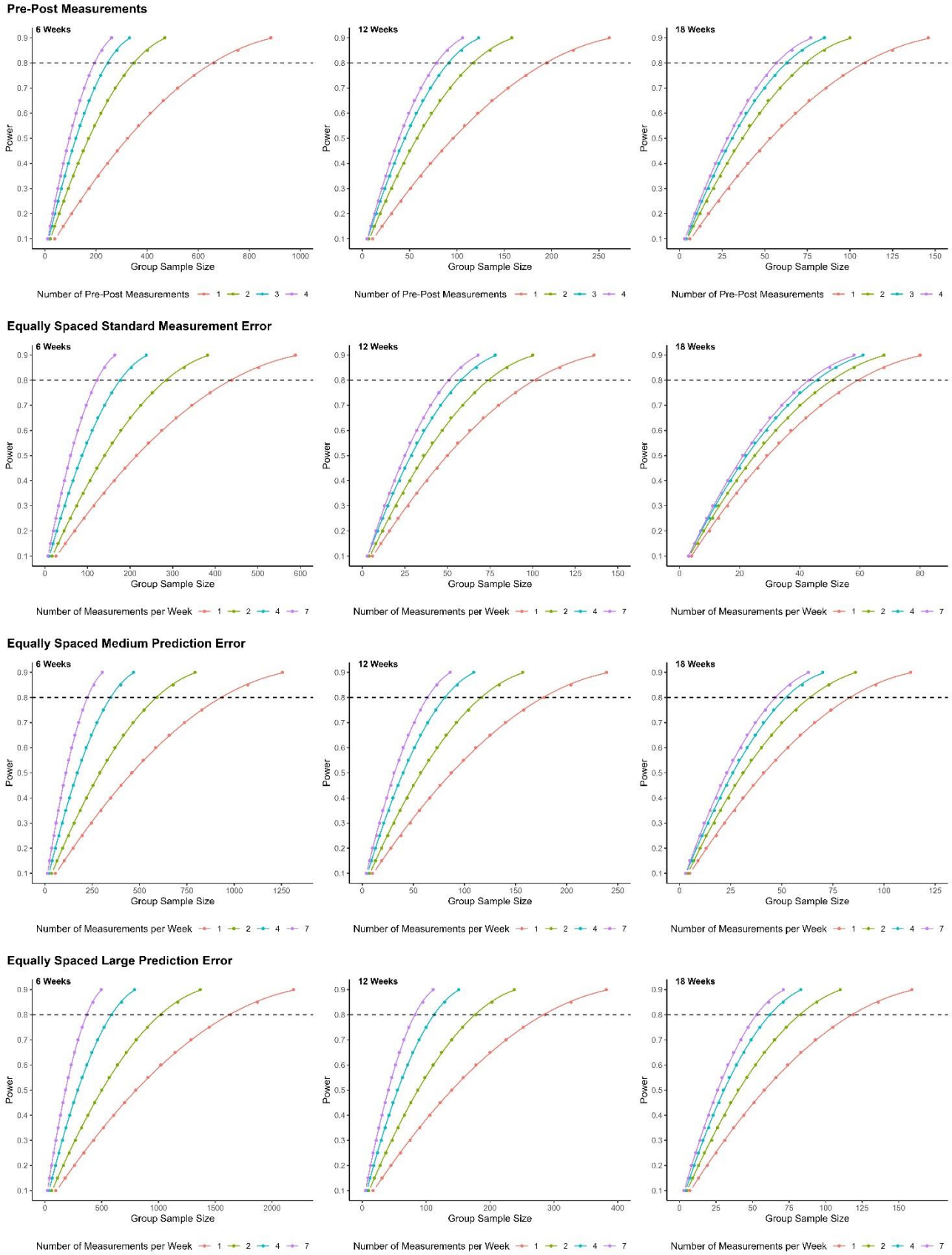
Based on the formula $n_{S,T} = 2\frac{\left(Z_{\alpha/2}+Z_{\phi}\right)^2}{\beta_3^2}\left(\sigma_{b1}^2 + \delta^2 Q(S,T)^{-1}\right)$, and once α and ϕ are set, it can be seen that researchers greatest control over the sample size required is through within person variability (δ^2) and the structure of measurements made (e.g. $Q(S,T)$). Whilst the researcher has some control over the variation in interindividual response (e.g. σ_{b1}^2), this is limited. From the formula $Q(S,T) = \sum_{k=0}^{T}(t_k - \bar{t})^2 + \bar{t}^2\frac{(S-1)(T+1)}{S+T}$ we can see that increasing the number of measurements and altering their structure can reduce the effects of within participant variability (δ^2) thereby reducing the sample size required for a specified statistical power. It can also be seen that for a given number of measurements, sample size requirements are best reduced by performing these at baseline and as long as possible following baseline (e.g. long-intervention lengths). Figures 1 and 2 illustrate this phenomenon. With the small comparative effect size, low measurement error (5 kg) and one measurement made at baseline and post-intervention, sample sizes of 660, 195, and 109 per group are required for statistical power of 0.8 for intervention lengths of 6, 12, and 18 weeks, respectively (Figure 1). In contrast, if four measurements are made at baseline and post-intervention, sample sizes required decrease to 195, 90, and 57 (Figure 1). With a medium comparative effect size, the single and quadruple set of measurements lead to required sample sizes of 165, 49, 27; and 49, 20, and 14, respectively (Figure 2).

Figures 1 and 2 also provide statistical power-sample size curves for equally spaced measurements across different intervention lengths and measurement error values. Again, large reductions in sample size are obtained with greater intervention lengths. Eventually, increasing the frequency of measurements provides limited additional benefits. For example, with the small comparative effect size, medium measurement error (10 kg) and intervention length of 18 weeks, we require sample sizes of 53 and 47 to obtain statistical power of 0.8 when measuring four times a week or seven times a week, respectively (Figure 1). For a medium comparative effect, sample sizes required are reduced to 13 and 12, respectively (Figure 12).

Simulations performed to provide a check showed that the formula tended to overestimate the sample size required to obtain statistical power of 0.8. This overestimation was greatest for the small comparative effect size, with overestimations of approximately 10 participants per group. However, overestimations for the medium comparative effect size were generally between 1 and 2 participants per group.

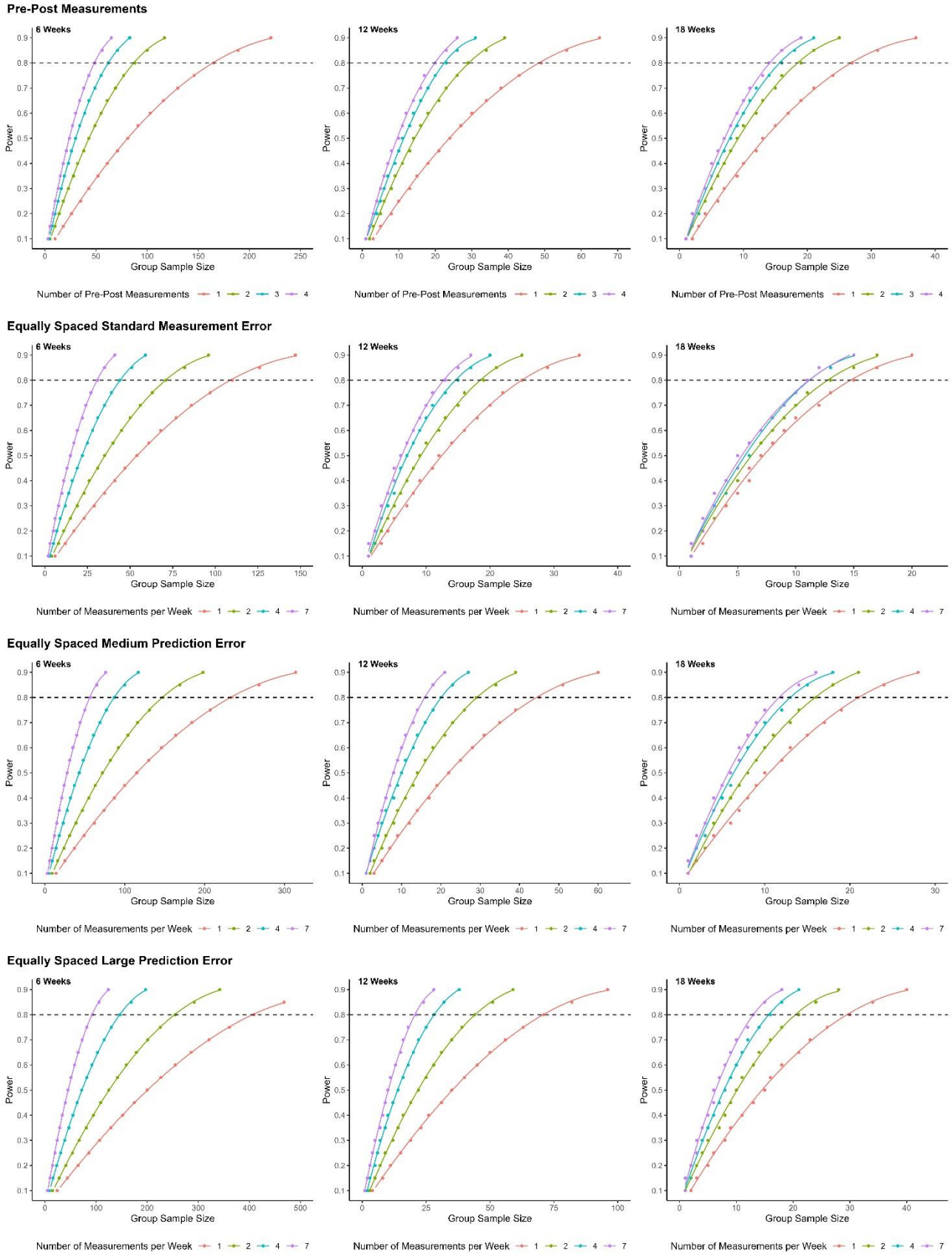
Doi: 10.51224/SRXIV.364 SportRχiv Preprint version 1

Figure 1: Statistical power-sample size curves for small comparative effect size across different intervention lengths, measurement strategies, measurement frequencies, and measurement error magnitudes.



Top-panel illustrates measurements made only at baseline and post-intervention. Other panels show equally spaced measurements comprising one at baseline and a multiple per week throughout the intervention.

Figure 2: Statistical power-sample size curves for medium comparative effect size across different intervention lengths, measurement strategies, measurement frequencies, and measurement error magnitudes.



Top-panel illustrates measurements made only at baseline and post-intervention. Other panels show equally spaced measurements comprising one at baseline and a multiple per week throughout the intervention.

Where there is a relationship between change scores and baseline values, we can see that $\sigma_{b1}^2 = \frac{\tau^2\sigma_{b0}^2+\hat{v}_1^2}{t_T^2}$ (see appendix for further details). Therefore, between participant variability σ_{b1}^2 can be reduced by recruiting a more homogenous group such that σ_{b0}^2 is reduced. We can see from the statistical power-sample size formula, that reducing σ_{b1}^2 will have a constant reduction in required sample size regardless of the number or spacing of measurements. It can also be seen that, greater reductions in required sample size for a given α and statistical power will be obtained for lower comparative effect sizes. In the present study eight conditions (2×2×2) were investigated with modifications to the comparative effect size, the amount of change score variance contributed by the relationship between σ_{b0}^2 and σ_{b1}^2, and the magnitude of reduction in σ_{b0}^2 through sampling. Derivations of the formulas to calculate the change in sample size are presented in the appendix. For the small comparative effect size, the reduction in required sample size for α=0.05 and statistical power of 0.8 ranged from 6 to 18, and for the medium comparative effect size ranged from 1 to 4.

Discussion

The results from this study suggest that research designs, measurement strategies, and more contemporary statistical practices can be combined to obtain adequate statistical power for RCTs conducted in strength and conditioning with sample sizes that are feasible given typical resource constraints of the discipline. The following sections will discuss the findings and their implications for future research.

Underpinning this study is the assumption that the response of individuals to a single training intervention can be adequately described by a linear model. Whilst long-term observational studies show that improvements to strength and conditioning interventions will be non-linear,[19] over the shorter-term which is typical of RCT's, a linear model may be appropriate. In our large meta-analysis, we identified an ordered effect with greater magnitude improvements obtained with intervention durations exceeding 10 weeks as compared to intervention durations of 6 to 10 weeks, and less than 6 weeks.[37] What is probably not as well understood in strength and conditioning research, is the large improvements in statistical power that can be achieved if interventions are conducted over longer durations (assuming the underlying data generating model presented is appropriate). As a reference calculation, the sample size required for statistical power of 0.8 with a pre-post design, small comparative effect and relatively low measurement error (e.g. 5 kg for the bench press), was 660, 195, and 109 when the post-intervention measurement was made at 6, 12, and 18 weeks, respectively. Similarly with a medium comparative effect size, the sample sizes reduced to 165, 49, and 27. These represent enormous savings in resources that can be achieved with conducting longer interventions, which are also more reflective of strength and conditioning in practice and the time course for stable adaptations.[38,39]

Another practice that is not well understood in strength and conditioning research, is the large improvements in statistical power that can be achieved if baseline and post-interventions measurements are repeated. As the statistical power-sample size formula used highlights, for a given number of measurements made, statistical power is increased the most by performing these at baseline and post-intervention.[33] Combining the long intervention period of 18 weeks with repeated baseline and post-intervention measures, the results from this study show that the sample size required for statistical power of 0.8 (small comparative effect size and relatively low measurement error) was 109, 75, 63, and 57 for 1, 2, 3, and 4 sets of measurements, respectively. For the medium comparative effect size, the required sample sizes reduced to 27, 19, 16, and 13. The combination of longer interventions and duplicate measurements at baseline and post-intervention should be strongly considered as a strategy to obtain adequate statistical power for future RCTs in strength and conditioning. There is

the possibility, however, that multiple testing at baseline and post-intervention will not be feasible as the repeat testing may cause training or fatigue effects that disrupt assessment of the interventions.

An alternative and more feasible strategy may be to conduct additional measurements during the intervention. As highlighted in the introduction, there may be cases where the same measurement process used at baseline and post-intervention is also appropriate during the intervention. There are likely to be many cases, however, where this is not possible, and outcomes are best obtained with prediction methods that add some additional measurement error. Despite the increase in measurement error with prediction methods, the results from this study highlight with sufficient frequency of measurement, this problem can be overcome, and required sample sizes lowered to that achieved when performing multiple baseline and post-intervention measurements. For example, with the small comparative effect size and 18-week intervention, sample sizes of 53 and 62 were found to provide statistical power of 0.8 when measurements were made four times per week and errors were increased by a factor of 1.5 and 2, respectively. If the measurements could be made each day in the week, the sample sizes reduced to 47 and 53. For the medium comparative effect size, the sample sizes were 13 and 16 for measurements made four times per week, and 12 and 13 for measurements made seven times per week.

Collectively, the results of the study highlight that when the comparative effect size is small, there may still be substantial resource required to obtain adequate statistical power (e.g. sample sizes >40). However, the resource is much smaller than the hundreds of participants that are required with small comparative effect sizes and short interventions using a pre-post design. The statistical-power sample size calculation used for this study highlights that there are limitations with regards to high frequency measurements. This strategy reduces the effects of within participant variation but does not alter the influence of between participant variability. Where this latter variation is large, statistical power is likely to be low unless more participants are recruited. Where there is a relationship between baseline values and change scores, between participant variation can be reduced by recruiting more homogenous participants. However, the results of this study show that the reduction may be small and restricting recruitment of participants to be more homogenous may lead to issues with generalising findings. The sample recruited should reflect the population the researcher wishes to generalise to and for which the data generating mechanism studied is appropriate.

Further research on the statistical analysis of high frequency data collection to improve RCTs is required. The present study introduced a relatively simple model and did not consider complexities that exist within actual studies. Where

high frequency measurements are possible, the structure of the data may include dependencies beyond that considered here. Further research is required to identify the structure of high frequency data in strength and conditioning, which may include for example autoregressive errors.[40] With a better understanding of the structure of the data, more accurate statistical models can be applied that may further improve statistical power with realistic sample sizes. Further research is also required to investigate the effects of issues such as missing data and drop-outs that may increase with high frequency data collection. The performance of more complex models including multivariate models is required to determine how best to accommodate RCTs with multiple outcome variables that are likely to be correlated.[3] Further research should also be conducted to explore the use of Bayesian approaches and how previous data and structured expert elicitation can be used to build informative priors that better leverage pre-existing knowledge.

References

1. Swinton PA, Burgess K, Hall A, Greig L, Psyllas J, Aspe R, Maughan P, Murphy A. Interpreting magnitude of change in strength and conditioning: Effect size selection, threshold values and Bayesian updating. Journal of sports sciences. 2022;40(18):2047-54. https://doi.org/10.1080/02640414.2022.2128548
2. Swinton, PA. Murphy, A. Comparative effect size distributions in strength and conditioning and implications for future research: A meta-analysis. 2022. Pre-print available from SportRχiv. https://doi.org/10.51224/SRXIV.202
3. Swinton, PA. Is it time to rethink pre-post randomised controlled trials in strength and conditioning? A review of statistical approaches with derivations and simulations. 2023. Pre-print available from SportRχiv. https://doi.org/10.51224/SRXIV.363
4. Choi A, Kim JY, Jo S, Jee JH, Heymsfield SB, Bhagat YA, Kim I, Cho J. Smartphone-based bioelectrical impedance analysis devices for daily obesity management. Sensors. 2015;15(9):22151-66. https://doi.org/10.3390/s150922151
5. Kooreman P, Scherpenzeel A. High frequency body mass measurement, feedback, and health behaviors. Economics & Human Biology. 2014;14:141-53. https://doi.org/10.1016/j.ehb.2013.12.003
6. Moon JR. Body composition in athletes and sports nutrition: an examination of the bioimpedance analysis technique. European journal of clinical nutrition. 2013;67(1):S54-9. https://doi.org/10.1038/ejcn.2012.165
7. Franchi MV, Longo S, Mallinson J, Quinlan JI, Taylor T, Greenhaff PL, Narici MV. Muscle thickness correlates to muscle cross-sectional area in the assessment of strength training-induced hypertrophy. Scandinavian journal of medicine & science in sports. 2018;28(3):846-53. https://doi.org/10.1111/sms.12961
8. Wong MC, Bennett JP, Leong LT, Tian IY, Liu YE, Kelly NN, McCarthy C, Wong JM, Ebbeling CB, Ludwig DS, Irving BA. Monitoring body composition change for intervention studies with advancing 3D optical imaging technology in comparison to dual-energy X-ray absorptiometry. The American Journal of Clinical Nutrition. 2023;117(4):802-13. https://doi.org/10.1016/j.ajcnut.2023.02.006
9. Pyper E, McKeown S, Hartmann-Boyce J, Powell J. Digital Health Technology for Real-World Clinical Outcome Measurement Using Patient-Generated Data: Systematic Scoping Review. Journal of Medical Internet Research. 2023;25:e46992. https://doi.org/10.2196/46992
10. Gençoğlu C, Ulupınar S, Özbay S, Turan M, Savaş BÇ, Asan S, İnce İ. Validity and reliability of "My Jump app" to assess vertical jump performance: a meta-analytic review. Scientific Reports. 2023;13(1):20137. https://doi.org/10.1038/s41598-023-46935

11. Brzycki M. Strength testing—predicting a one-rep max from reps-to-fatigue. Journal of physical education, recreation & dance. 1993;64(1):88-90. https://doi.org/10.1080/07303084.1993.10606684
12. Reynolds JM, Gordon TJ, Robergs RA. Prediction of one repetition maximum strength from multiple repetition maximum testing and anthropometry. The Journal of Strength & Conditioning Research. 2006;20(3):584-92. https://doi.org/10.1519/R-15304.1
13. Greig L, Aspe RR, Hall A, Comfort P, Cooper K, Swinton PA. The predictive validity of individualised load-velocity relationships for predicting 1RM: a systematic review and individual participant data meta-analysis. Sports medicine. 2023;53(9):1693-1708. https://doi.org/10.1007/s40279-023-01854-9
14. Balsalobre-Fernández C, Marchante D, Baz-Valle E, Alonso-Molero I, Jiménez SL, Muñóz-López M. Analysis of wearable and smartphone-based technologies for the measurement of barbell velocity in different resistance training exercises. Frontiers in physiology. 2017:649. https://doi.org/10.3389/fphys.2017.00649
15. McElreath R. Statistical rethinking: A Bayesian course with examples in R and Stan. Chapman and Hall/CRC; 2018. https://doi.org/10.1201/9781315372495
16. Newans T, Bellinger P, Drovandi C, Buxton S, Minahan C. The utility of mixed models in sport science: a call for further adoption in longitudinal data sets. International Journal of Sports Physiology and Performance. 2022;1(aop):1-7. https://doi.org/10.1123/ijspp.2021-0496
17. Brown VA. An introduction to linear mixed-effects modeling in R. Advances in Methods and Practices in Psychological Science. 2021;4(1):2515245920960351. https://doi.org/10.1177/2515245920960351
18. Mirman D. Growth curve analysis and visualization using R. CRC press; 2017.
19. Steele J, Fisher JP, Giessing J, Androulakis-Korakakis P, Wolf M, Kroeske B, Reuters R. Long-term time-course of strength adaptation to minimal dose resistance training through retrospective longitudinal growth modeling. Research Quarterly for Exercise and Sport. 2022:1-8. https://doi.org/10.1080/02701367.2022.2070592
20. Balsalobre-Fernández C, Kipp K. Use of machine-learning and load–velocity profiling to estimate 1-repetition maximums for two variations of the bench-press exercise. Sports. 2021;9(3):39. https://doi.org/10.3390/sports9030039
21. Fernandes JF, Dingley AF, Garcia-Ramos A, Perez-Castilla A, Tufano JJ, Twist C. Prediction of one repetition maximum using reference minimum velocity threshold values in young and middle-aged resistance-trained males. Behavioral Sciences. 2021;11(5):71. https://doi.org/10.3390/bs11050071

22. Pérez-Castilla A, Fernandes JF, Garcia-Ramos A. Validity of the bench press one-repetition maximum test predicted through individualized load-velocity relationship using different repetition criteria and minimal velocity thresholds. Isokinetics and Exercise Science. 2021;29(4):369-77. https://doi.org/10.3233/IES-202247
23. Macarilla CT, Sautter NM, Robinson ZP, Juber MC, Hickmott LM, Cerminaro RM, Benitez B, Carzoli JP, Bazyler CD, Zoeller RF, Whitehurst M. Accuracy of predicting one-repetition maximum from submaximal velocity in the barbell back squat and bench press. Journal of Human Kinetics. 2022;82(1):201-12. https://doi.org/10.2478/hukin-2022-0046
24. Pérez-Castilla A, Piepoli A, Garrido-Blanca G, Delgado-García G, Balsalobre-Fernández C, García-Ramos A. Precision of 7 commercially available devices for predicting bench-press 1-repetition maximum from the individual load–velocity relationship. International Journal of Sports Physiology and Performance. 2019;14(10):1442-6. https://doi.org/10.1123/ijspp.2018-0801
25. Nickerson BS, Williams TD, Snarr RL, Garza JM, Salinas G. Evaluation of load-velocity relationships and repetitions-to-failure equations in the presence of male and female spotters. The Journal of Strength & Conditioning Research. 2020;34(9):2427-33. https://doi.org/10.1519/JSC.0000000000003731
26. García-Ramos A, Haff GG, Pestaña-Melero FL, Pérez-Castilla A, Rojas FJ, Balsalobre-Fernández C, Jaric S. Feasibility of the Two-Point Method for Determining the One-Repetition. International Journal of Sports Physiology and Performance. 2018; 13(4):474-481. https://doi.org/10.1123/ijspp.2017-0374
27. Balsalobre-Fernández C, Marchante D, Muñoz-López M, Jiménez SL. Validity and reliability of a novel iPhone app for the measurement of barbell velocity and 1RM on the bench-press exercise. Journal of sports sciences. 2018;36(1):64-70. https://doi.org/10.1080/02640414.2017.1280610
28. Williams TD, Esco MR, Fedewa MV, Bishop PA. Bench press load-velocity profiles and strength after overload and taper microcyles in male powerlifters. The Journal of Strength & Conditioning Research. 2020;34(12):3338-45. https://doi.org/10.1519/JSC.0000000000003835
29. Jiménez-Alonso A, García-Ramos A, Cepero M, Miras-Moreno S, Rojas FJ, Pérez-Castilla A. Velocity performance feedback during the free-weight bench press testing procedure: an effective strategy to increase the reliability and one repetition maximum accuracy prediction. Journal of Strength and Conditioning Research. 2022;36(4):1077-83. https://doi.org/10.1519/JSC.0000000000003609
30. Caven EJ, Bryan TJ, Dingley AF, Drury B, Garcia-Ramos A, Perez-Castilla A, Arede J, Fernandes JF. Group versus individualised minimum velocity thresholds in the prediction of maximal strength in trained female athletes.

International Journal of Environmental Research and Public Health. 2020;17(21):7811. https://doi.org/10.3390/ijerph17217811

31. Janicijevic D, Jukic I, Weakley J, García-Ramos A. Bench press 1-repetition maximum estimation through the individualized load–velocity relationship: comparison of different regression models and minimal velocity thresholds. International Journal of Sports Physiology and Performance. 2021;16(8):1074-81. https://doi.org/10.1123/ijspp.2020-0312
32. Tango T. Repeated measures design with generalized linear mixed models for randomized controlled trials. CRC Press; 2017. https://doi.org/10.1201/9781315152097
33. Fitzmaurice GM, Laird NM, Ware JH. Applied longitudinal analysis. John Wiley & Sons; 2012. https://doi.org/10.1002/9781119513469
34. Swinton PA, Hemingway BS, Saunders B, Gualano B, Dolan E. A statistical framework to interpret individual response to intervention: paving the way for personalized nutrition and exercise prescription. Frontiers in nutrition. 2018;5:41. https://doi.org/10.3389/fnut.2018.00041
35. R Core Team (2023). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.
36. Weston S, Analytics R. doParallel: Foreach parallel adaptor for the parallel package. R package version. 2019;1:15.
37. Swinton, PA. Burges, K. Hall, A. Greig L. Psyllas J. Aspe R. Maughan P. Murphy A. A Bayesian approach to interpret intervention effectiveness in strength and conditioning: Part 1. A meta-analysis to derive context-specific thresholds. 2021. Pre-print available from SportRχiv. https://doi.org/10.51224/SRXIV.9.
38. Appleby B, Newton RU, Cormie P. Changes in strength over a 2-year period in professional rugby union players. Journal of Strength & Conditioning Research. 2012;26(9):2538-46. https://doi.org/10.1519/JSC.0b013e31823f8b86
39. Ahtiainen JP, Pakarinen A, Alen M, Kraemer WJ, Häkkinen K. Muscle hypertrophy, hormonal adaptations and strength development during strength training in strength-trained and untrained men. European Journal of Applied Physiology. 2003;89(6):555-63.
40. Chi EM, Reinsel GC. Models for longitudinal data with random effects and AR (1) errors. Journal of the American Statistical Association. 1989;84(406):452-9. https://doi.org/10.2307/2289929

Doi: 10.51224/SRXIV.364 SportRχiv Preprint version 1

Adequate statistical power in strength and conditioning may be achieved through longer interventions and high frequency outcome measurement.

Paul, A. Swinton

Appendices

Data generating mechanisms

In this section we introduce the notation used and explain the data generating mechanisms assumed to produce pre- and post-intervention values for traditional RCT analyses and higher frequency data to be analysed with linear mixed models.

Notation for pre- and post-intervention data

Y_{ijk} is the true score of participant i $(i = 1,2, \dots, n)$, in group j $(j = 0,1)$ at time k $(k = 0,1)$.
y_{ijk} is the observed score (true score plus measurement error). Groups $j = 0,1$ are both considered intervention groups, and times $k = 0,1$ refer to baseline and post-intervention, respectively.
Pre-intervention we draw from a normal distribution with mean μ_0 and standard deviation σ_0 to obtain $Y_{ij0} \sim N(\mu_0, \sigma_0^2)$. We assume that j is randomly assigned with equal probability and is independent of $Y_{i.0}$.
What we observe in data collection is the true value plus error given by $y_{ijk} = Y_{ijk} + \epsilon_{ij.}$, where $\epsilon_{ij.} \sim N(0, \delta^2)$ is independent of Y_{ijk}.

Data generation

The baseline and post-intervention true scores are generated from a multivariate normal distribution $\begin{matrix} Y_{ij0} \\ Y_{ij1} \end{matrix} \sim \mathbb{N}\left(\begin{bmatrix} \mu_0 \\ \mu_{j1} \end{bmatrix}, \begin{bmatrix} \sigma_0^2 & \rho_j \sigma_0 \sigma_{j1} \\ \rho_j \sigma_0 \sigma_{j1} & \sigma_{j1}^2 \end{bmatrix}\right)$, where μ_{j1} and σ_{j1} are the group mean and standard deviation in post-intervention true scores, and ρ_j is the group correlation between baseline and post-intervention true scores. Throughout we only consider the homogenous case, that is where $\rho_0 = \rho_1$ and $\sigma_{01}^2 = \sigma_{11}^2$. We consider two homogenous cases including: 1) the independent case where $\mathrm{Corr}\left(Y_{ij0}, Y_{ij1} - Y_{ij0}\right) = 0$; and 2) the constrained linear case where $\mathrm{Corr}\left(Y_{ij0}, Y_{ij1} - Y_{ij0}\right) \neq 0$.

Independent case

For the independent case we express the data generating mechanism as
$Y_{ij1} = Y_{ij0} + \Delta_j + \xi_{ij1}$,
where Δ_j is a constant describing the average change in group j, and $\xi_{ij1} \sim N(0, \nu_1^2)$ describes the variation in true change score which is the same across groups to fit with the homogenous case. From this data generating mechanism we have the following relationships:

$$\begin{aligned} \mathrm{Cov}\left(Y_{ij0}, Y_{ij1}\right) &= E\left(Y_{ij0} Y_{ij1}\right) - \mu_0 \mu_{j1} \\ &= E\left(Y_{ij0}\left(Y_{ij0} + \Delta_j + \xi_{ij1}\right)\right) - \mu_0(\mu_0 + \Delta_j) \\ &= \sigma_0^2 + \mu_0^2 + \mu_0 \Delta_j - \mu_0(\mu_0 + \Delta_j) \\ &= \sigma_0^2. \end{aligned} \qquad \text{Result 1}$$

$$\begin{aligned} \mathrm{Cov}\left(Y_{ij0}, Y_{ij1} - Y_{ij0}\right) &= E\left(Y_{ij0}\left(Y_{ij1} - Y_{ij0}\right)\right) - \mu_0(\mu_{j1} - \mu_0) \\ &= \sigma_0^2 + \mu_0 \mu_{j1} - (\sigma_0^2 + \mu_0^2) - \mu_0(\mu_{j1} - \mu_0) \\ &= 0. \end{aligned} \qquad \text{Result 2}$$

The distribution of baseline and post-intervention true scores for the independent case are thus

$$\begin{matrix} Y_{ij0} \\ Y_{ij1} \end{matrix} \sim \mathbb{N}\left(\begin{bmatrix} \mu_0 \\ \mu_0 + \Delta_j \end{bmatrix}, \begin{bmatrix} \sigma_0^2 & \sigma_0^2 \\ \sigma_0^2 & \sigma_0^2 + \nu_1^2 \end{bmatrix}\right). \qquad \text{Result 3}$$

Constrained linear case

For the constrained linear case we assume the data generating model
$Y_{ij1} = Y_{ij0} + \widetilde{\Delta}_j + \tau Y_{ij0} + \tilde{\xi}_{ij1}$,

where τ sets the slope of the linear relationship between the true change score and baseline true score which is the same for each group, and $\widetilde{\Delta}_j$ sets any group offset. $\tilde{\xi}_{ij1} \sim N(0, \tilde{v}_1^2)$ describes any further variation in true change score not caused by differences in baseline true score and is independent from Y_{ij0}. The data generating mechanism is constrained by the fact that τ and $\tilde{v}_1^2$ are the same across the groups. From this data generating mechanism we have the following relationships:

$$\begin{aligned}\mathrm{Cov}(Y_{ij0}, Y_{ij1}) &= E(Y_{ij0}Y_{ij1}) - \mu_0\mu_{j1} \\ &= E\left(Y_{ij0}(Y_{ij0} + \widetilde{\Delta}_j + \tau Y_{ij0} + \tilde{\xi}_{ij1})\right) - \mu_0(\mu_0 + \widetilde{\Delta}_j + \tau\mu_0) \\ &= \sigma_0^2 + \mu_0^2 + \mu_0\widetilde{\Delta}_j + \tau(\sigma_0^2 + \mu_0^2) - \mu_0^2 - \widetilde{\Delta}_j\mu_0 - \tau\mu_0^2 \\ &= (1+\tau)\sigma_0^2. \end{aligned}$$ Result 4

$$\begin{aligned}\mathrm{Cov}(Y_{ij0}, Y_{ij1} - Y_{ij0}) &= E\left(Y_{ij0}(Y_{ij0} + \widetilde{\Delta}_j + \tau Y_{ij0} + \tilde{\xi}_{ij1} - Y_{ij0})\right) - \mu_0(\mu_0 + \widetilde{\Delta}_j + \tau\mu_0 - \mu_0) \\ &= E\left(Y_{ij0}(\widetilde{\Delta}_j + \tau Y_{ij0} + \tilde{\xi}_{ij1})\right) - \mu_0(\widetilde{\Delta}_j + \tau\mu_0) \\ &= \widetilde{\Delta}_j\mu_0 + \tau(\sigma_0^2 + \mu_0^2) - \widetilde{\Delta}_j\mu_0 - \tau\mu_0^2 \\ &= \tau\sigma_0^2. \end{aligned}$$ Result 5

The distribution of baseline and post-intervention true scores for the constrained linear case are thus

$$\begin{matrix} Y_{ij0} \\ Y_{ij1} \end{matrix} \sim \mathbb{N}\left(\begin{bmatrix} \mu_0 \\ \mu_0(1+\tau) + \widetilde{\Delta}_j \end{bmatrix}, \begin{bmatrix} \sigma_0^2 & (1+\tau)\sigma_0^2 \\ (1+\tau)\sigma_0^2 & (1+\tau)^2\sigma_0^2 + \tilde{v}_1^2 \end{bmatrix}\right).$$ Result 6

If we want the post-intervention means to match between the independent and constrained linear cases, then we have
$\mu_0 + \Delta_j = \mu_0(1+\tau) + \widetilde{\Delta}_j \rightarrow \widetilde{\Delta}_j = \Delta_j - \tau\mu_0.$

If we want to match the change score variances between the independent and constrained linear cases, then we have
$v_1^2 = \tau^2\sigma_0^2 + \tilde{v}_1^2.$
We set $\tau^2\sigma_0^2$ to a proportion $0 < c < 1$ of the variance v_1^2, such that
$\tau^2\sigma_0^2 = cv_1^2 \rightarrow \tau = \pm\frac{\sqrt{c}v_1}{\sigma_0}$, and we select τ to be negative.
We then have $\tilde{v}_1^2 = (1-c)v_1^2$.

Notation for high-frequency data
Y_{ijt} is the true score of participant i $(i = 1,2,\dots,n)$, in group j $(j = 0,1)$ at time t, where t is expressed as a decimal with integer values representing the number of weeks from baseline, such that $t = 10$, would equal 10 weeks from pre-intervention. And $t = \frac{71}{7} \sim 10.14$ would equal 10 weeks and 1 day from pre-intervention.
y_{ijt} is the observed score (true score plus measurement error). Groups $j = 0,1$ are both considered intervention groups.
Pre-intervention we draw from a normal distribution with mean β_0 and standard deviation σ_{b0} to obtain $Y_{ij0} \sim N(\beta_0, \sigma_{b0}^2)$. We assume that j is randomly assigned with equal probability and is independent of $Y_{i.0}$.
What we observe in data collection is the true value plus error given by $y_{ijt} = Y_{ijt} + \epsilon_{ij.}$, where $\epsilon_{ij.} \sim N(0, \delta^2)$ is independent of Y_{ijt}.

Data generation
The high-frequency data are generated from a linear mixed model with random intercept and slope

$$Y_{ijt} = b_{0i} + \beta_0 + (\beta_2 + \beta_3 G_{ij} + b_{1i})t,$$

where b_{0i} is the participant random intercept which is distributed as $b_{0i} \sim N(0, \sigma_{b0}^2)$ and denotes how much above or below the mean baseline value each participant is, β_2 is the mean rate of change of the standard group (change in outcome per week), G_{ij} is a group indicator variable equal to 0 for the standard intervention, and 1 for the test intervention, β_3 is the difference in rate of change of the test intervention relative to relative to standard intervention (also referred to as the treatment by linear time interaction or comparative effect size), and b_{1i} is the participant random slope which is distributed as $b_{1i} \sim N(0, \sigma_{b1}^2)$ and denotes how each participants rate of change is above or below the mean group rate of change. b_{0i} and b_{1i} follow a multivariate normal distribution

$$\begin{matrix} b_{0i} \\ b_{1i} \end{matrix} \sim \mathbb{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_0^2 & \rho\sigma_0\sigma_{b1} \\ \rho\sigma_0\sigma_{b1} & \sigma_{b1}^2 \end{bmatrix}\right).$$

This multivariate relationship allows b_{0i} and b_{1i} to be related such that when $\rho = 0$ we have the independent case and pre-intervention value is not related to change, and when $\rho \neq 0$ they are related. When we have measurement error our data generation mechanism becomes

$$y_{ijt} = b_{0i} + \beta_0 + (\beta_2 + \beta_3 G_{ij} + b_{1i})t + \epsilon_{ij.},$$

Where $\epsilon_{ij.}$ is independent b_{0i} and b_{1i}.

From this data generating mechanism we have the following
$\mathrm{Var}(Y_{ijt}) = \sigma_{b0}^2 + \sigma_{b1}^2 t^2 + 2t\rho\sigma_{b0}\sigma_{b1}.$

$$\begin{aligned}\mathrm{Cov}(Y_{ijt}, Y_{ijt'}) &= E(Y_{ijt}, Y_{ijt'}) - \mu_0\mu_{j1} \\ &= E\left((b_{0i} + \beta_0 + (\beta_2 + \beta_3 G_{ij} + b_{1i})t)(b_{0i} + \beta_0 + (\beta_2 + \beta_3 G_{ij} + b_{1i})t')\right) - \\ &\quad (\beta_0 + \beta_2 t + \beta_3 G_{ij} t)(\beta_0 + \beta_2 t' + \beta_3 G_{ij} t') \\ &= E(b_{0i}^2 + (t + t')b_{0i}b_{1i} + b_{1i}^2 tt') \\ &= \mathrm{Var}(b_{0i}^2) + (t + t')\mathrm{Cov}(b_{0i}, b_{1i}) + tt'\mathrm{Var}(b_{1i}^2) \\ &= \sigma_{b0}^2 + (t + t')\rho\sigma_{b0}\sigma_{b1} + tt'\sigma_{b1}^2.\end{aligned}$$

If we want to match the pre-, post-intervention models and the linear mixed model, then to achieve this, we ensure that the variance at the post-intervention is the same as the variance of the final data point in the linear mixed model, and we ensure that the covariance between pre-, and post-intervention is the same as the covariance of the initial and final data points in the linear mixed model. For the independent case we set $\rho = 0, \sigma_0^2 = \sigma_{b0}^2$, and $t = T_{END}$ (the final week of the intervention) such that
$\sigma_{b0}^2 + \nu_1^2 = \sigma_{b0}^2 + \sigma_{b1}^2 T_{END}^2 \rightarrow \sigma_{b1}^2 = \frac{\nu_1^2}{T_{END}^2}.$

For the constrained linear case we have
$(1 + \tau)\sigma_{b0}^2 = \sigma_{b0}^2 + T_{END}\rho\sigma_{b0}\sigma_{b1} \rightarrow \rho = \frac{\tau\sigma_0^2}{T_{END}\sigma_{b0}\sigma_{b1}}.$
For the variances at post-intervention we have
$(1 + \tau)^2\sigma_{b0}^2 + \tilde{\nu}_1^2 = \sigma_{b0}^2 + \sigma_{b1}^2 T_{END}^2 + 2T_{END}\rho\sigma_{b0}\sigma_{b1}.$
Inserting ρ from the previous equation we have
$(1 + \tau)^2\sigma_{b0}^2 + \tilde{\nu}_1^2 = \sigma_{b0}^2 + \sigma_{b1}^2 T_{END}^2 + 2T_{END}\left(\frac{\tau\sigma_{b0}^2}{T_{END}\sigma_{b0}\sigma_{b1}}\right)\sigma_{b0}\sigma_{b1} \rightarrow \tau^2\sigma_{b0}^2 + \tilde{\nu}_1^2 = \sigma_{b1}^2 T_{END}^2 \rightarrow \sigma_{b1}^2 = \frac{\tau^2\sigma_{b0}^2 + \tilde{\nu}_1^2}{T_{END}^2}.$

In the data generating model we did not include a group difference at baseline, but it common to include and fit the model

$$y_{ijt} = b_{0i} + \beta_0 + \beta_1 G_{ij} + (\beta_2 + \beta_3 G_{ij} + b_{1i})t + \epsilon_{ij.}.$$

Parameter selection for analysis

To explore the effects of manipulations on statistical power-sample size curves, an approximation using a closed form solution for the random intercept and slope model with treatment by linear time interaction was used. The notation for the formula includes the number of measurements S made at baseline, and the number of measurements T made following baseline, expressed as
Y_{ijt_k} where $k = -(S-1), -(S-2), \ldots, 0, 1, \ldots T$. We also have $t_{-S+1} = t_{-S+2} = \cdots = t_0 = 0$.
The sample size for each group given the number of measurements S and T is denoted $n_{S,T}$, and we have [33]

$$n_{S,T} = 2\frac{(Z_{\alpha/2} + Z_\phi)^2}{\beta_3^2}(\sigma_{b1}^2 + \delta^2 Q(S,T)^{-1}),$$

where α is the probability of a Type I error (set to 0.05), statistical power is $1 - \phi$ (explored from values of 0.1 to 0.9), Z_γ denotes the upper 100γ% percentile of the standard normal distribution, β_3 is the comparative effect size, σ_{b1}^2 describes primarily between subject variability (e.g. response variability), and δ^2 describes within subject variability due to a broad conception of measurement error.[34] We also have that

$$Q(S,T) = \sum_{k=0}^{T}(t_k - \bar{t})^2 + \bar{t}^2\frac{(S-1)(T+1)}{S+T}.$$

The study was based on interventions to improve maximum strength in the bench press, and based on previous research.[20-31] I assumed a baseline mean $\mu_0 = \beta_0 = 100$ kg, a population standard deviation of $\sigma_0 = \sigma_{b0} = 15$ kg. A measurement error value for conducting a 1RM test of the bench press of $\delta = 5$ kg was assumed. For the standard intervention, an improvement of 6 kg was assumed over 12 weeks

which returned $\beta_2 = 0.5$ kg/wk. Using small and medium comparative effect sizes for strength and conditioning ,[2] the small value was set to $\beta_3 = \frac{0.15\times15}{12} = 0.1875$ kg/wk, and the medium value set to $\beta_3 = \frac{0.3\times15}{12} = 0.375$ kg/wk. σ_{b1} was set so that only 5% of those in the standard group would have a slope that was negative, such $\sigma_{b1} = 0.3$, that is qnorm(0.05,0.5,0.3)=0.

To investigate the effect of manipulating σ_{b1} through changes in sampling practices (that is reducing σ_{b0}), eight conditions were considered based on manipulating three factors (e.g. 2×2×2 factorial). The first was the proportion of the change variance accounted for by the relationship between the change score and baseline ($c = 0.25$ and 0.5). The second was the restricted value of σ_{b0} (5 and 10). The third was the comparative effect size (small and medium). To determine the subsequent change in σ_{b1}, we start with the relations $v_1 = \sigma_{b1}T_{END}$ and $\tau = -\frac{\sqrt{c}v_1}{\sigma_{b0}}$ such that $\tau = -\frac{\sqrt{c}\sigma_{b1}T_{END}}{\sigma_{b0}}$.

To identify the changed σ_{b1}^2 from the original σ_{b1}^{*2} we use $\sigma_{b1}^2 = \frac{\tau^2\sigma_{b0}^2+\tilde{v}_1^2}{T_{END}^2}$, and set $\tilde{v}_1^2$ to $(1-c)v_1 = (1-c)\sigma_{b1}^{*2}T_{END}^2$. This gives

$$\sigma_{b1}^2 = \frac{\frac{c\sigma_{b1}^{*2}T_{END}^2}{\sigma_{b0}^{*2}}\sigma_{b0}^2+(1-c)\sigma_{b1}^{*2}T_{END}^2}{T_{END}^2}$$
$$= \frac{c\sigma_{b1}^{*2}\sigma_{b0}^2}{\sigma_{b0}^{*2}} + (1-c)\sigma_{b1}^{*2}.$$

For $c = 0.25$ we have

$$\sigma_{b1}^2 = 0.75\sigma_{b1}^{*2} + 0.25\left(\frac{\sigma_{b0}^2}{\sigma_{b0}^{*2}}\right)\sigma_{b1}^{*2}$$
$$= \left(\frac{3+\sigma_{b0}^2/\sigma_{b0}^{*2}}{4}\right)\sigma_{b1}^{*2}.$$

For $c = 0.5$ we have

$$\sigma_{b1}^2 = 0.5\sigma_{b1}^{*2} + 0.5\left(\frac{\sigma_{b0}^2}{\sigma_{b0}^{*2}}\right)\sigma_{b1}^{*2}$$
$$= \left(\frac{1+\sigma_{b0}^2/\sigma_{b0}^{*2}}{2}\right)\sigma_{b1}^{*2}.$$

From the closed form solution, we can see that the sample size that is reduced from lowering σ_{b1}^2 for $c = 0.25$ is

$$2\frac{(Z_{\alpha/2}+Z_\phi)^2}{\beta_3^2}\left(\sigma_{b1}^{*2}\left(1-\left(\frac{3+\sigma_{b0}^2/\sigma_{b0}^{*2}}{4}\right)\right)\right),$$

And for $c = 0.5$ is

$$2\frac{(Z_{\alpha/2}+Z_\phi)^2}{\beta_3^2}\left(\sigma_{b1}^{*2}\left(1-\left(\frac{1+\sigma_{b0}^2/\sigma_{b0}^{*2}}{2}\right)\right)\right).$$

Simulation check

The table below provides results of simulations used to provide a check on the statistical power-sample size results with the approximate closed form solution. Checks were made on those results where statistical power of 0.8 was achieved with a sample size of 50 or less.

Effect size	Intervention length	Measurement error	Measurement frequency	Sample size formula n, for power of 0.8	Power of sample size with simulation	Sample size for power of 0.8 calculated with simulation
0.1875	18	5	4×Week	46	0.97	35
0.1875	12	5	7 ×Week	50	0.94	40
0.1875	18	5	7 ×Week	43	0.97	31
0.375	12	5	1 ×Week	25	0.81	25
0.375	18	5	1 ×Week	15	0.87	13
0.375	12	7.5	1 ×Week	45	0.81	45
0.375	18	7.5	1 ×Week	21	0.84	19
0.375	18	10	1 ×Week	30	0.83	28
0.375	12	5	2 ×Week	19	0.88	17
0.375	18	5	2 ×Week	13	0.90	11
0.375	12	7.5	2 ×Week	29	0.84	28
0.375	18	7.5	2 ×Week	16	0.88	14
0.375	12	10	2 ×Week	44	0.82	43
0.375	6	5	4 ×Week	44	0.84	42
0.375	12	5	4 ×Week	15	0.90	13
0.375	18	5	4 ×Week	11	0.78	12
0.375	12	7.5	4 ×Week	20	0.85	18
0.375	18	7.5	4 ×Week	13	0.90	12
0.375	12	10	4 ×Week	28	0.82	28
0.375	18	10	4 ×Week	16	0.89	14
0.375	6	5	7 ×Week	31	0.82	30
0.375	12	5	7 ×Week	13	0.89	11
0.375	18	5	7 ×Week	11	0.92	9
0.375	12	7.5	7 ×Week	16	0.86	15
0.375	18	7.5	7 ×Week	12	0.92	10
0.375	12	10	7 ×Week	21	0.85	19
0.375	18	10	7 ×Week	13	0.89	11

R Code

```
library(MASS)
library(ggplot2)
library(tidyr)
library(dplyr)
library(scales)
library(lmerTest)
library(foreach)
library(doParallel)
library(Hmisc)
library(cowplot)

# Statistical power-sample size function

NSTPowerRange = function(t,beta3,sigmab,sigmaE,alpha=0.05,phi=rev(seq(0.1,0.9,0.05))){
  S = sum(t==0)
  Tint = sum(t!=0)
  t0 = c(0,t[t!=0])
  tbar = sum(t0)/(Tint+1)
  QST = sum((t0-tbar)^2) + (((S-1)*(Tint+1))/(S+Tint))*tbar^2
  nst = ((2*((qnorm(1-(alpha/2)) + qnorm(1-phi))^2))/beta3^2) *
    (sigmab^2+((sigmaE^2)*(1/QST)))
  names(nst) =c("0.1","0.15","0.2","0.25","0.3","0.35","0.4","0.45","0.5",
                "0.55","0.6","0.65","0.7","0.75","0.8","0.85","0.9")
  return(nst)
}

# Function to simulate LMM for a given number of iterations

PostlmmallParams = function(n,t,sigmab0,sigmab1, beta0,beta2,beta3,rho,Iter){
  Datalmm = array(NA, c(2*n,length(t),Iter))
  G = c(rep(0,n),rep(1,n))
  for(i in 1:Iter){
    U = mvrnorm(n,c(0,0),
                matrix(c(sigmab0^2,rho*sigmab0*sigmab1,rho*sigmab0*sigmab1,sigmab1^2),ncol=2))
    for(j in 1:length(t)){
      Datalmm[,j,i] = beta0 + U[,1] + t[j]*(beta2 + G*beta3 + U[,2])}}
  return(Datalmm)}

# Function to return estimates of the LMM and p values

LMMAFunP = function(t,Data,ErrorSD){
  n = length(Data[,1])
  Timep = length(Data[1,])
  LmmResult=c(NULL)
  DataError = Data +  matrix(rnorm(n*Timep,0,ErrorSD),nrow=n)
  DataErrorDF = as.data.frame(DataError)
  DataErrorDFL = as.data.frame(pivot_longer(DataErrorDF, cols = everything()))
  DataErrorDFL$ID = rep(1:n,each=Timep)
  DataErrorDFL$Group = c(rep("Standard",(n/2)*Timep),rep("Test",(n/2)*Timep))
  DataErrorDFL$Time = rep(t,n)
  DataErrorDFL$ID = factor(DataErrorDFL$ID)
  DataErrorDFL$Group = factor(DataErrorDFL$Group)

  lmmA = summary(lmer(value~Group+Group*Time + (1+Time|ID),
                      data=DataErrorDFL))
  LmmResult= c(as.numeric(lmmA$coefficients[4,c(1,2,5)]),
               as.data.frame(lmmA$varcor)[c(1,2,4,3),5])
  return(LmmResult)}

# Function to turn results across iterations into a data frame for analysis
```

```
DataSortFun = function(DataA){
  OutP = DataA
  OutPDF = t(as.data.frame(OutP))
  rownames(OutPDF)=NULL
  OutPDF = as.data.frame(OutPDF)
  OutPDFL = pivot_longer(OutPDF, cols = everything())
  OutPDFL$Statistic =rep(c("ATERate","SE","Pvalue","Sigma0","Sigmab","Sigma","rho"),Iter)
  return(OutPDFL)}

# Function to summarise results and calculate statistical power

  OutputSum = function(OutPDFL){
    OutputSumc = OutPDFL%>%
    group_by(Statistic) %>%
    summarise(mean = round(mean(value),3),sd = round(sd(value),3),
              Q25 = round(quantile(value, 0.25),3),Median = round(median(value),3),
              Q75 = round(quantile(value, 0.75),3),
              P005 = round(100*(mean(value<0.05)),1))
  Out = list(OutputSumc,OutPDFL)
  return(Out)
}



# Study parameters
# Beta0 =100, sigma0 = b0 = 15, sigmab1 =0.3, Beta2 = 0.5, Beta3 = (0.15*15)/12 =0.1875, nu = 0.1875*T

# Examples of use of statistical power-sample size formula
PrePost6One = round(NSTPowerRange(c(0,6),0.1875,0.3,5,alpha=0.05,rev(seq(0.1,0.9,0.05))),0)
PrePost12One = round(NSTPowerRange(c(0,12),0.1875,0.3,5,alpha=0.05,rev(seq(0.1,0.9,0.05))),0)
PrePost18One = round(NSTPowerRange(c(0,18),0.1875,0.3,5,alpha=0.05,phi=rev(seq(0.1,0.9,0.05))),0)

# Double Pre and Post
PrePost6Two = round(NSTPowerRange(c(0,0,6,6),0.1875,0.3,5,alpha=0.05,rev(seq(0.1,0.9,0.05))),0)
PrePost12Two = round(NSTPowerRange(c(0,0,12,12),0.1875,0.3,5,alpha=0.05,rev(seq(0.1,0.9,0.05))),0)
PrePost18Two = round(NSTPowerRange(c(0,0,18,18),0.1875,0.3,5,alpha=0.05,phi=rev(seq(0.1,0.9,0.05))),0)

# Examples of simulation check

n.cores = parallel::detectCores() - 4
my.cluster = parallel::makeCluster(
  n.cores,
  type = "PSOCK"
)
doParallel::registerDoParallel(cl = my.cluster)

lmmS18SME4Wk46 = PostlmmallParams(n=46,t=seq(0,18,0.25),sigmab0=15,sigmab1=0.3,
beta0=100,beta2=0.5,beta3=0.1875,rho=0,Iter=10000)
lmmS18SME4Wk46A_5 = foreach(i = 1:10000,.packages=c('lmerTest','tidyr')) %dopar%
  LMMAFunP(t=seq(0,18,0.25),Data=lmmS18SME4Wk46[,,i],ErrorSD=5)
lmmS18SME4Wk46A_5DF = DataSortFun(lmmS18SME4Wk46A_5,1000)
OutputSum(lmmS18SME4Wk46A_5DF)[[1]]
```