Sample size: How many repeats?

E-mail distributed on 18-05-2026

Dear all,

A common problem in experimental design is deciding how often to repeat the outcome measurement of interest. Such "repeats" can arise in a variety of different ways, for example: **(a)** items measuring the same construct in a questionnaire, **(b)** trials in a behavioral task, **(c)** spatial sampling in multiple locations, **(d)** longitudinal sampling at multiple time points, or **(e)** uninterrupted time series sampling. While many of these have an intrinsic temporal aspect (e.g., trials, longitudinal data, time series), this need not strictly be the case (e.g., questionnaire items, spatial sampling). Likewise, while repeated measures are more often encountered in within-subject designs, they can also be employed in between-subject designs. Typically, the repeated measures are aggregated to obtain a final outcome measure, such as a mean or median outcome value.

But what is the impact of repeats on your statistical analysis, particularly its power, and how can this inform sample size planning? In this month's newsletter, I address some of these issues.

1. Sample size

For outcomes averaged from repeated measures, the gain in statistical power can be substantial. Vickers (2003) estimated the gains between 30% and 70% power, potentially, though critically the actual amount depends on the strength of repeated measures correlation. That is, when repeated measures correlation is high, outcomes will be very similar, and hence little reduction in error variance is achieved by averaging, leading to similar power as under a one-shot measurement design. Conversely, when repeated measures correlation is low, averaging can substantially reduce error variance, and hence more power is achieved with fewer subjects.

Figure 1 shows a simulation of a traditional difference of independent means (*t*-test, equal variance assumed), for a moderate effect size of Cohen's $d = 0.5$, and statistical significance level $\alpha = 0.05$. Four design parameters were varied, the total sample size N (40, 60, 100; N/2 per group), the number of "trials" over which the outcome is averaged (1 to 15), and the strength of within-subject correlation ρ (0.0, 0.3, 0.5, 0.8). Each combination of design parameters was re-run 500 items, and the proportion of significant *p*-values counted.

As shown in the upper left panel, for the lowest repeated measures correlation, power increases rapidly with only 3 or 4 trials, even when the total sample size is small. In fact, for $N = 40$, having 4 trials per subject more than doubles the observed power over having just 1 trial. These gains

diminish as repeated measures correlation becomes stronger. For ρ = 0.8, there is little advantage in trial-averaging, with stronger power gains being made by increasing the number of subjects instead.

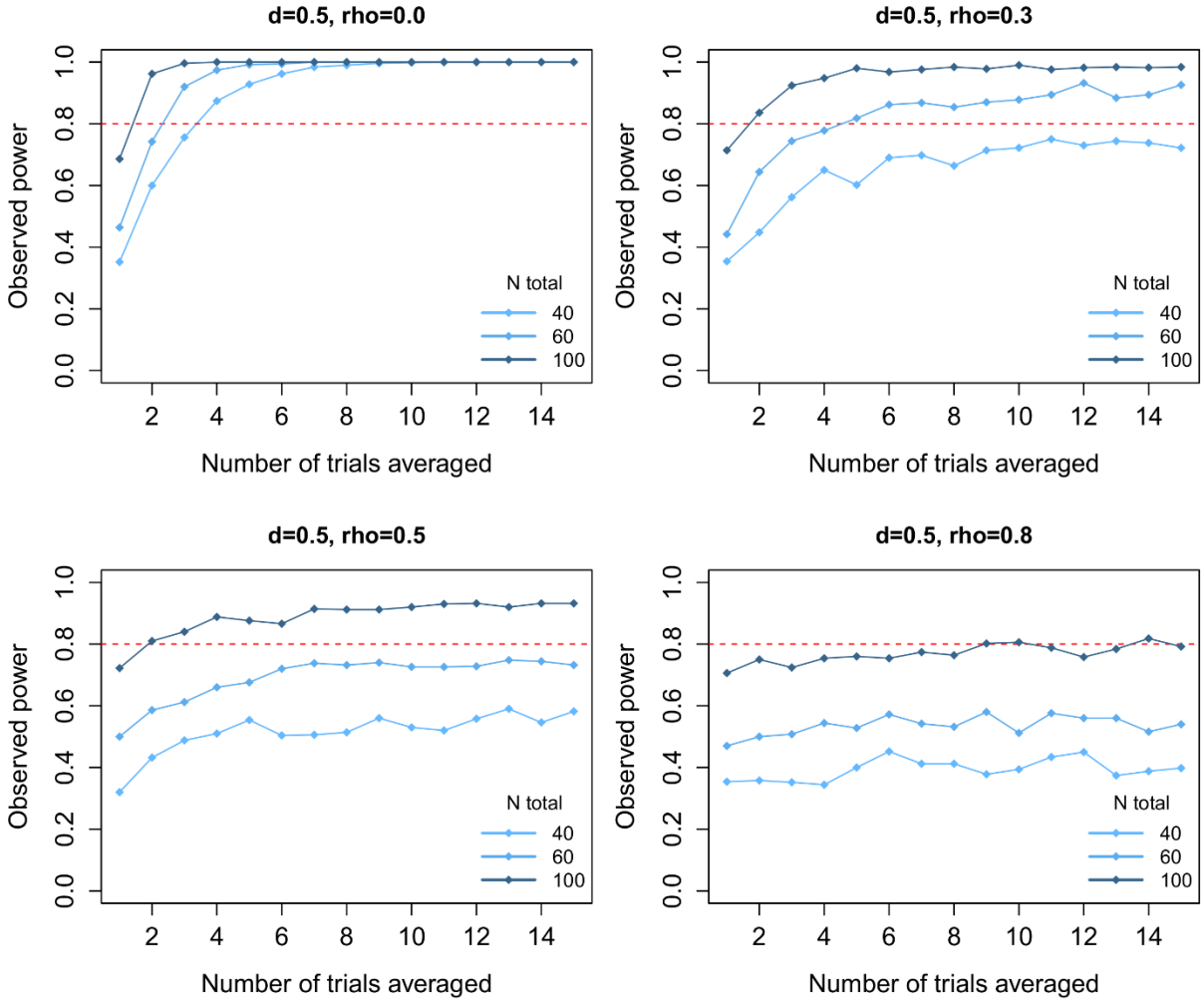

Figure 1. Simulated power for a comparison of independent means (*t*-test), for varying sample size (*N*), trial number, and repeated measures correlation (ρ). Effect size was set to a standardized difference of Cohen's d = 0.5, with significance level α = 0.05, and desired power level 80% (red dashed line).

As in Vickers (2003), the largest power gains are made for the first few repeats, before stagnating around the 6th or 7th trial. Similar results were reported by Fitzmaurice, Laird and Ware (2005; Table 15.1) in the context of longitudinal designs. In general, it appears that having repeated measures is almost always a good design practice.

Unfortunately, popular software for power and sample size calculation (e.g., G*Power, rpwrss) does not include fields to specify the number of repeats in the design, except for the traditional case of within-subjects designs, where each "repeat" coincides with a single measurement of each condition. In fact, for the simulated 2-group difference, G*Power suggests that a total sample size of 128 (64 per group) is required for this effect size. This is likely an overestimate for designs that include trial-averaging. If you find yourself designing a multiple-trial experiment, you should therefore consider

simulating the design directly, rather than relying on G*Power, and may be required to convince reviewers (e.g., funding applications, registered reports) of sample sizes that seem unrealistic or subject to debatable power parameters. While G*Power has become a gold standard for power calculations and its calculations technically correct, I have discussed earlier some of its more dubious handling of within-subject designs.

2. Reliability

An aspect closely linked to statistical power and reducing error variance is that of "reliability", or "precision". Figure 2 shows an estimated 2-group difference and its standard error (SE), as a function of the number of trials it was averaged over. Using an average of 50 trials, the SE is reduced to about 20% of the size it was at 4 trials, substantially increasing the precision of the point estimate, and bringing it closer to its final value, when averaged over 120 trials (dashed line).

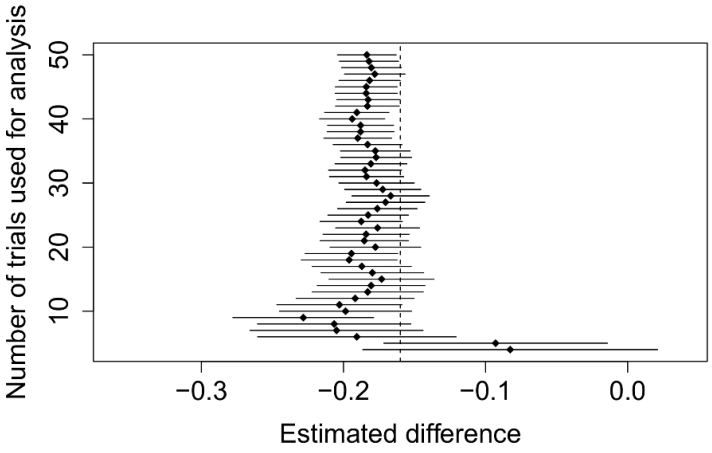

Figure 2. Estimated 2-group difference (paired t-test) and its standard error (SE), as a function of the number of task trials used for estimation. Dashed line reflects the value of the estimated difference when all trials of the experiment are used (120; not depicted).

However, one should bear in mind that high reliability does not *automatically* imply increases in power. As shown in Figure 1, when there is low repeated measures correlation, an average of trials will be more reliable/precise than the value of a single trial, without implying that the average is objectively reliable (or good at measuring at what it purports to measure). Conversely, a measure could be highly reliable in the sense that it has strong repeated measures correlation (i.e., the same values are consistently observed) but, as also shown in Figure 1, without much gain in statistical power. In fact, somewhat paradoxically, the most reliable measure would have the least tradeoff in power by having repeats. This scenario is often encountered in psychometric scales, where the items should be strongly intercorrelated, when assumed to measure the same underlying construct.

3. Many repeats versus many participants

A question that is often asked is whether it is better for a sample to have many participants or many repeats per participant. It is sometimes falsely believed that increasing the number of repeats can generally compensate for a lack of participants. The simulation in Figure 1 should go some way in

clarifying under which conditions this is true, and that the tradeoff—if any—is limited to the first couple of trials. In general, it is still better to focus more on collecting many participants, especially for the purpose of **generalizability**. That is, large samples not only guarantee statistical power, reliability, and normality, they increase the likelihood that the sampled participants are representative of the population of interest.[1]

Despite the limitations of many repeats with respect to power, researchers may have other reasons to include a large number of repeats. In multilevel models, repeats are typically not aggregated and analyzed directly at the trial-level. This enables the analyst to explicitly separate effects at the population level from effects at the subject-level, by the use of random effects (in particular random slopes). For the purpose of estimating reliable random slopes, 7 repeats should not be considered sufficient. If within-subject modelling is of interest, it would be advised to use similar sample size criteria as for any between-subjects effect.

4. When repeats are detrimental

Finally, there are cases when having repeats, and/or averaging over repeats, is detrimental to the design of the study or the insights of its data analysis. For example, in affective science, stimuli intended to evoke emotional responses may quickly lose their potency when presented too often. The best measure of the emotion may simply correspond to the first trial, and averaging over subsequent repeats would dilute its value. One should always inspect how the outcome values are distributed over trials, and choose the right statistic of aggregation accordingly. When strongly skewed or multimodal, an average may not be representative. A similar scenario occurs when repeats are taken not in time (i.e., over consecutive trials) but in spatially linked locations, for example electrode locations in electro-encephalogram (EEG). Typically, the effect of interest is localized to a specific electrode or set of electrodes. Averaging over all locations would again dilute the effect of interest.

Finally, there is the practical concern that having many trials can exhaust participants, and makes later outcome measurements in a session not comparable to earlier ones.

References

Fitzmaurice, G. M., Laird, N. M., & Ware, J. H. (2004). *Applied Longitudinal Analysis*. Hoboken, NJ: John Wiley & Sons.

Vickers, A. J. (2003). How many repeated measures in repeated measures designs? Statistical issues for comparative trials. *BMC medical Research Methodology*, *3*(1), 22.

--

Ben Meuleman, Ph.D.
Statistician
Swiss Center for Affective Sciences
University of Geneva | Campus Biotech
Chemin des Mines 9 | CH-1202 Genève
ben.meuleman@unige.ch | +41 (0)22 379 09 79

[1] For an earlier discussion of these concerns, see the newsletter on sample size.