STATISTICAL SUPPORT

SWISS CENTER FOR AFFECTIVE SCIENCES

Caution on mixed ANOVAs with G*Power

E-mail distributed on 13-05-2024

Dear all,

G*Power remains the most popular software in social sciences for calculating power and setting the sample size for a planned study (ver 3.1.9.6, as of this writing). Using a point-and-click interface, the user specifies the analysis of interest (e.g., chi-square, *t*-test, *F*-test) and sets the desired parameters (e.g., effect size, significance level α, power level β) to arrive at a sample size estimate. While this seems straightforward enough, ambiguities and even inconsistencies arise when you try to obtain agreement across different G*Power menus.

Specifically, let us consider a mixed ANOVA—such as for an intervention study—with two groups (**Group**: Control, Experiment) and two measurement times, pre- and post-intervention (**Time**: T0, T1). There are three equivalent ways to construct the Group × Time interaction to capture the intervention effect:

- Group × Time interaction in a repeated measures ANOVA
- Group × Time interaction in a repeated measures MANOVA
- Group main effect in a regression with the T0−T1 difference as the outcome

Statistically all three will yield **identical** *F*-tests and *p*-values.[1] However, when selecting these methods in G*Power, an inconsistency emerges between the rm-ANOVA and the two other approaches (Fig. 1 and 2). For the same medium effect size (Cohen's f = 0.25), the required sample size for rm-ANOVA is only N=34, versus N=128 for the other two methods! What happened?

It turns out that G*Power is not consistent in its use of Cohen's f and the assumption it makes for the same effect. The authors of the software clarified the point on StackExchange, admitting that the rm-ANOVA assumes a "double dissociation effect", where e.g., the T0−T1 difference in the Experiment group is reversed in equal size in the Control group. No such assumption is made for rm-MANOVA. In the G*Power dialog for rm-ANOVA, there is an "options" button to change this calculation, with one of the options "Cohen (1988) – Recommended" (Fig .3). When one does, the effect size changes from f to f(V) and the previous value of 0.25 doubles to 0.50, revealing the true assumed value. Resetting f(V) back to 0.25 produces N=128, consistent with the other approaches. Stunningly, the "Cohen (1988)" option not the default in G*Power!

[1] For designs with within-subjects factors that have more than 2 levels, the rm-ANOVA and rm-MANOVA will differ!

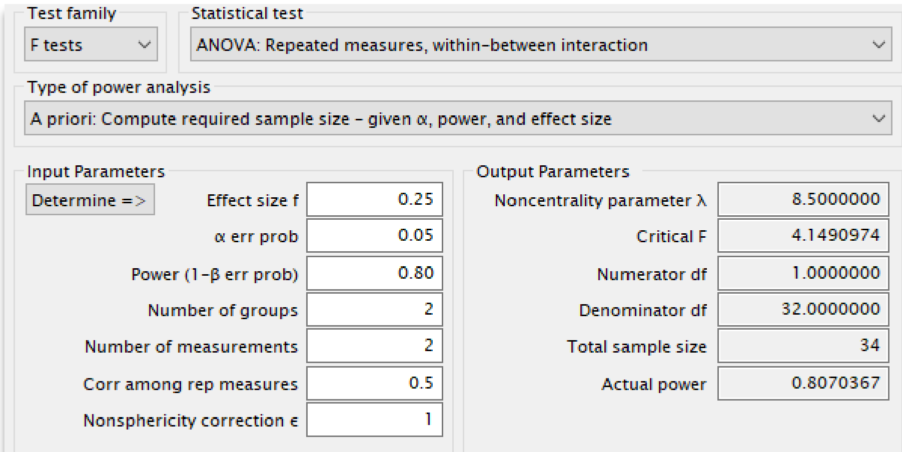

Figure 1. Sample size for repeated measures ANOVA.

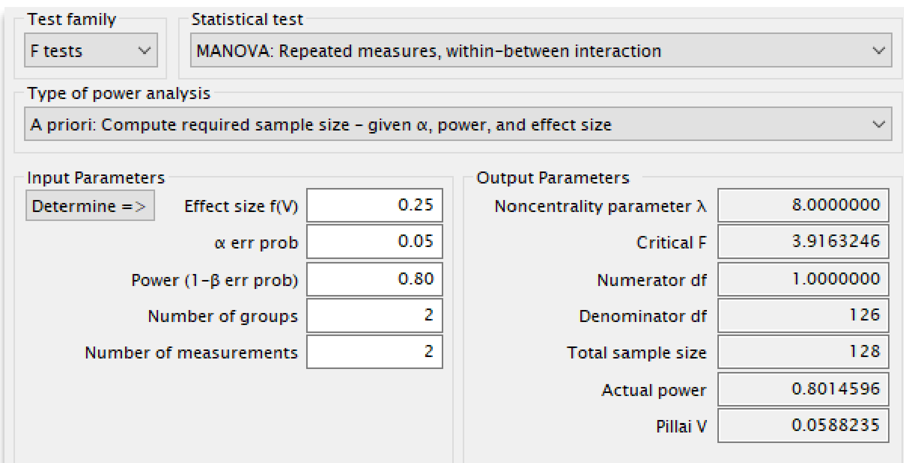

Figure 2. Sample size for repeated measures MANOVA.

As such, extreme caution is warranted when planning sample sizes for mixed designs, as the rm-ANOVA will return overly optimistic estimates. I suspect that these estimates are popular precisely because they are so low, however realistically one should prefer the rm-MANOVA, or one risks to consistently run underpowered studies. While the news that you may need 4 times as many participants will be tough to swallow, I remind that planning and justifying sample size is not just a consideration of statistical power. Moreover, G*Power does not allow to enter certain design specifications that do have a substantial impact on power:

- **Within-subject correlation:** When the correlation among within-subjects conditions is strong, fewer subjects are needed to detect the same effect.
- **Covariates:** Analyses that include covariates with a strong relationship to the outcome will boost the precision with which the experimental effect is estimated (Maxwell & Delaney, 2004).

- **Multiple trials:** Outcome variables that were aggregated from multiple trials or multiple questionnaire items will be much more precise than single-trial measurements.

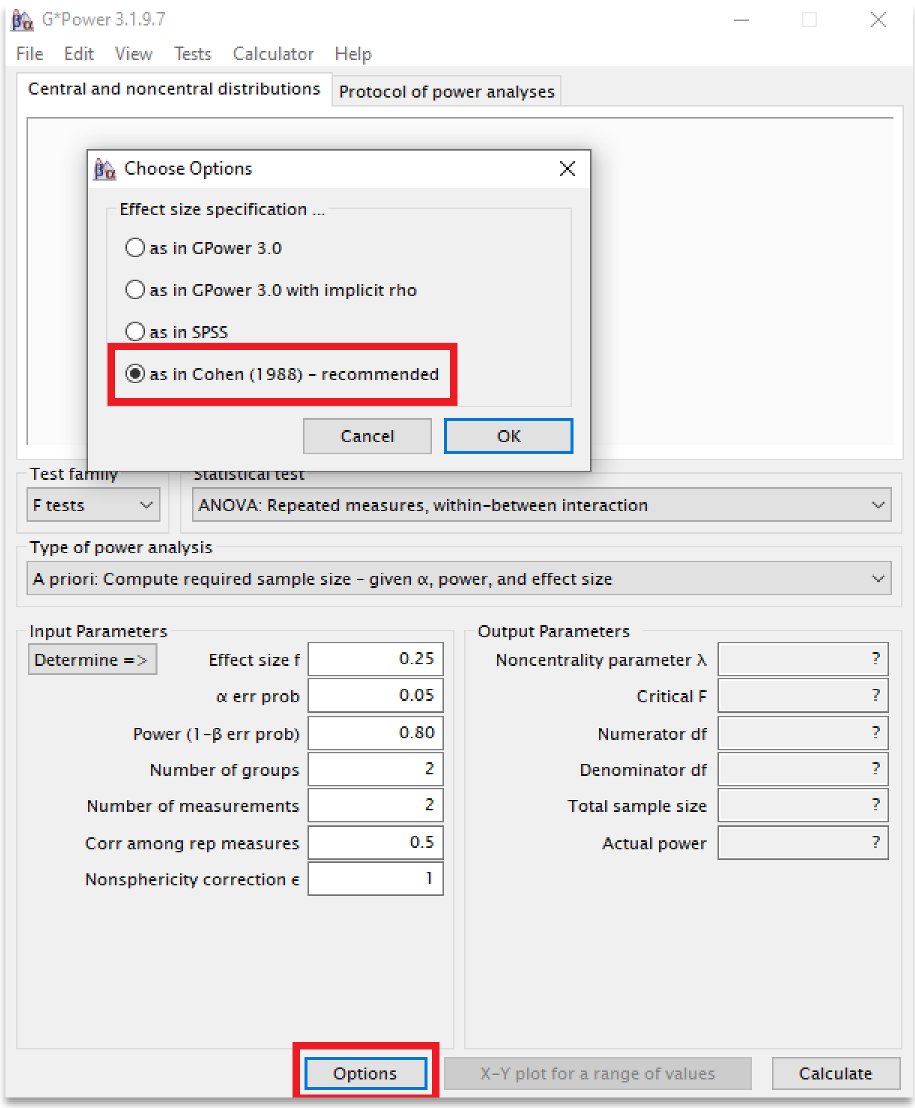

Figure 3. Cohen (1988) option for repeated measures ANOVA.

Especially in the case of multiple trials the gain in power can be substantial, which means that the estimate in the above example of N=128 may just be a conservative upper bound. In practice, your analysis may have more power than set a priori.

What about avoiding G*Power altogether? Firstly, be warned that the R package pwrss, which largely replicates the menus and options of G*Power, will produce the same inconsistency between rm-ANOVA and rm-MANOVA. Secondly, for exact specification of your design details (e.g., assignment, trials, covariates), the best solution may be a simulation. This has been the default for multilevel regression for some years, in fact, where it is the only practical approach, using packages such as longpower or simr. For other types of designs this may not be straightforward, unless pilot data is available to set reasonable expectations on the necessary parameters. You are advised in this case to consult with an expert for the best approach.

Best,
Ben

References

Maxwell, S.E., and Delaney, H.D. (2004). *Designing experiments and analyzing data: A model comparison perspective (2nd ed.)*. Lawrence Erlbaum Associates Publishers.

--
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