Calculating statistical power

1. **Option 1**: The researchers may state the power, or minimal detectable effect size (MDE), in the paper or pre-registration (because they conducted this prior to the experiment, to work out the sample size).
    a. Check that the effect size they used to calculate power seems reasonable to you; researchers have a motive to ensure the calculation produces a sample size that they have the resources to collect data from.
    b. Check that the MDE size encompasses your approximation of the likely effect size.

2. **Option 2:** use PowerUp (this is also GiveWell's solution).
    a. Use either the excel, shiny app, R or Python version. I could not get the excel version to work (I think because it is an xslm file, and this doesn't seem to work in google sheets?), so I am using the shiny app below. Note that the excel version additionally allows for power analyses for regression discontinuity and interrupted time series designs.
    b. Select the effect that you are calculating power for. Typically you will be looking at the main effect; moderators identify on whom and under what circumstances treatments have different effects, while mediators identify why and how treatments have effects.[1]
    c. Select the appropriate design.
        i. IRT; individually randomised trial. Individuals are allocated to either treatment at random.
        ii. CRT; cluster randomised trial. Pre-existing groups ("clusters") are randomly allocated to either treatment at random.
        iii. Level; this refers to the nesting. Most of the designs we might come across are two-level; e.g. study subjects are nested into clusters. Some designs have additional levels, because there is an additional layer of nesting. E.g. school children are nested within schools (sub clusters), which are nested within different teaching districts (clusters); randomisation is carried out at the teaching district level, while outputs are collected from the school children.
        iv. Random, fixed or constant; people use the terms 'fixed effects' and 'random effects' in different ways. We are concerned with how the treatment effects are assumed to vary across clusters/ subclusters. Select constant if the models (testing the RCT results) assume constant treatment effects across blocks, fixed if the models assume that each block has a specific treatment effect, and random if treatment effects are assumed to randomly vary across blocks. Usually the paper will state which kind of design it used; my best understanding is that most RCTs we

[1] See Kraemer et al. 2002 for more detail.

examine are likely to assume fixed effects, but some will use random effects.

d. Fill in the effect size, type 1 error rate (usually 0.05), and two tailed test
    i. Post-hoc power tests have been rightly criticised for providing an inflated assessment of statistical power, when they rely on the experiment's effect size (Hoenig & Heisey, 2001). To avoid this issue, use an effect size external to the data at hand- i.e. sourced from an estimate independent to the data at hand.
e. Fill in the number of covariates at a given level.
    i. The 'level' refers to which level of nesting (e.g. school children are nested within schools, nested within teaching districts. A level 1 covariate would be a covariate at the level of the child; i.e. their previous test results. A level 3 covariate would be a covariate at the level of the teaching district; i.e. the amount of funding that goes to each teaching district).
f. Fill in the proportion of variance, at each level.
    i. This should be listed in the paper, often called the intracluster correlation coefficient/ intraclass correlations, or ρ (rho). This is the fraction of the total variance in outcome that lies within clusters. It is a value between 0 and 1; lower scores are associated with higher statical power.
    ii. E.g. the ICC from level 2 covariates (for the school child example above), is the fraction of total variance coming from between schools. The ICC from level 3 covariates is the fraction of total variance coming from between teaching districts.
    iii. If the ICC is not reported, I recommend estimating it. Do you think that the outcome result is likely to differ by cluster/ subcluster, irrespective of the experimental condition? ICCs will be between 0 and 1, and you can look for baseline ICC estimates online (e.g. between 0.1 to 0.3 for various health-based metrics across different clinician practices in the USA here).
g. Fill in the sample size metrics according to the paper, according to the paper.
h. Press 'update'. Voila!

3. **Option 3:** use simulations. For more complex designs, you may need to run simulations. J-PAL have a helpful resource here, which includes links to their code. Note that simulations may actually be easiest to get correct, for people with coding skills.

4. [For all options] **Sanity check**; in psychology the median power is estimated at around 36% (Stanley et al., 2018; note that this value is significantly lower within social psychology, and higher in cognitive science), in political science articles ~10%, in economics ~18%, (Ioannidis et al., 2017). We can expect that some experiments (e.g. developmental economics experiments with huge sample sizes, and a large expected effect size) will have significantly higher power. Does your calculated effect size seem reasonable?