Guidelines for Presenting Published Paper Research Designs

By now you should have signed up to present, with classmates, three published papers over the course of the semester. This document provides guidance for your team presentations, expanding on the guidance that appears on the syllabus.

For each paper that you are assigned to present, your goal is to present the paper in a manner that focuses on the hypotheses, research design, the interrelation between the two, and statistical power. Imagine that you are presenting the study's research design before it was fielded, and you trying to convince an audience that the study is worth taking to the field.

The presentation will typically require going beyond what is presented in the published paper and examining appendices and other supplementary information. Each presentation should include about 30-40 minutes worth of material, allowing for about 10-20 minutes of interruptions, questions, and discussion.

Theoretical and substantive framing

Provide a theoretical and substantive framing for your study. Present relevant theoretical framework(s) and state your study's hypotheses in terms of theoretical model parameters. It is good to be formal here—that is, to present a formal theoretical framework and crystallize your hypotheses in terms of hypotheses about model parameters.

You also want to discuss the "debate" to which this study is contributing. It is important to frame a study in terms of a debate. Doing so makes it clear that theoretical arguments are inadequate to settle a question and therefore that it is actually necessary to go out and run a field study to address the debate.

The other good thing about framing in terms of a debate is that it sets up the study such that the results are interesting no matter whether they are positive, negative, or null. You never want a study that is interesting *only* if you get a result that goes in one direction or the other. You want to be sure that even a null result is somehow important. At the same time, you want to be sure that any null results are not merely the product of poor power. *That* is bad. If you get a null, you want to be sure it is a "precisely estimated zero" (see the Casey et al. paper for a great example). So, you want a study that estimates effects with a level of precision that is adequate to test for meaningful effects, and you want to motivate the study in terms of a debate such that we learn no matter what the results.

Research design

Explain the research design, including what types of people, communities, or other units you will be studying, how they will be selected (sampling), how experimental treatments are operationalized, and how treatments will be assigned.

Pictures are good here, including randomization trees or tables, maps, etc.

Explain strata, clustering, and other features of the design. Explain the function of these design features. (E.g., are they important for identifying key parameters, or just there for practical reasons?)

Go back to your theoretical framework and explain how your research design targets key parameters in the theoretical framework.

Explain how outcomes are operationalized, how they are measured, and any complications that have to be overcome in obtaining the data.

Discuss any special measurement techniques, indices, etc.

Analysis plan

Present the data analysis specifications (e.g., regression specifications, weighting, etc.). Explain any control variables, justifying their inclusion.

Explain how to interpret the quantities to be estimated from data in terms of parameters in the theoretical framework.

Explain whether your data analysis precisely identifies parameters from the theoretical framework or whether there are some complications or ambiguities in the interpretation.

Explain your hypothesis testing procedures, ways to account for statistical challenges (e.g. multiple testing).

1 Power and minimum detectable effects

What kinds of effects is this study powered to detect? How do those minimum detectable effects relate to what we think would be substantively meaningful for policy or our academic understanding?

For this analysis, you will use the actual outcome data from the study, if available, to do this. So, this will be the true power of the study. Usually you do not have such data to do a power analysis, but rather have to rely on some auxiliary data. This part of the exercise is the one time when you take advantage of the fact that the study has been done and therefore that you can get the outcome data.

There are a few ways you can do this:

Analytical approach

- Get the effective sample size within each treatment/control arm, accounting for design effects:
  - For relatively simple designs, you can do this "by hand":
    - If there is clustering in treatment assignment, first calculate the design effect based on the cluster-variance inflation factors $1 + (m - 1)\rho$ for each potential outcome (using the replication outcome data), where ρ is the ICC for each potential outcome. You can get the rho?s

for each potential outcome by using "loneway" in Stata or "deff" from the Hmisc package in R. Divide the sample size by that to get an effective sample size that accounts for clustering. E.g., of the nominal sample size for the treatment group is 100 households, but they are in clusters of 10 and rho for the treated households is .33, then the effective sample size for the treated is 100/[1+(10-1)*.33] = 25. You could do the same for the controls and for other treatment arms.

  - If there is blocking/stratification, you can compute the design effect for each potential outcome using the equations in pages 12-13 of the lecture notes on design effects.
  - If there are regression controls, then the design effect is given by $1-R^2$ from the regression of the outcome on the controls:
    - E.g., suppose you have just a binary treatment, D, and you have some control variables, X, and then you estimate treatment effects using,

      $$Y_i = \alpha + \beta D_i + X_i'\gamma + \epsilon_i.$$

      Because this is an experiment, we know that (in expectation) D_i is orthogonal to X_i. As a result, the design effect from controlling for X_i is given by $1 - R^2_{Y,X}$, where $R^2_{Y,X}$ is the variance explained in the regression of Y on X (excluding D). That is, if, for example, $R^2_{Y,X} = .25$, then the design effect is $1 - R^2_{Y,X} = .75$, in which case the effective sample size $N/.75 = 1.33N$—that is, regression control that explains this much variation is equivalent to increasing your sample size by 33%. (NB: if you are using weights in the analysis be sure to include them in this regression.)

- Or, you can use a tool like Stata's "svy estat" commands to get design effects. (Presumably R's survey package has similar functionality, but I am not sure.)
  - You compute the design effects for each treatment/control arm.
    - E.g., for the controls, just keep the control data. Then, svyset the data so that you indicate the clusters, strata, and any weights that you need to use in the analysis (e.g., blocking/stratification weights or sampling weights). Then calculate the svy mean, and following that run the estat effects commands.

- With the effective sample sizes, you can calculate MDEs.
  - E.g., suppose there is a control group, treatment group A, and treatment group B. You want to the MDE for the difference in mean outcomes under treatment A (call it $E[Y_A]$) as compared to mean outcomes under control (call it $E[Y_0]$). Given 95% confidence and 80% power, the formula (from the lecture notes) is

    $$MDE = 2.8\sigma_{\hat{\beta}_{A,0}},$$

    where $\beta_{A,0}$ is our estimate of $E[Y_A]-E[Y_0]$. (E.g., with a regression where you have an intercept and then dummies for treatment A and treatment B, this would be the coefficient on the treatment A dummy). Moreover, we know that under a completely randomized experiments,

    $$\sigma_{\hat{\beta}_{A,0}} \leq \sqrt{\frac{S_A^2}{n_A} + \frac{S_0^2}{n_0}},$$

where n_A and n_0 are the sample sizes for treatment A and control. So, the right hand side of the expression above provides a conservative approximation to use for our MDE calculations. Suppose now that we want to express everything in terms of control group standard deviations. That would imply dividing our outcome data by S_0, in which we can write the MDE as

$$\tilde{MDE} = 2.8\tilde{\sigma}_{\hat{\beta}_{A,0}},$$

and the standard error expression becomes,

$$\tilde{\sigma}_{\hat{\beta}_{A,0}} \leq \sqrt{\frac{\psi_A}{n_A} + \frac{1}{n_0}},$$

where $\psi_A = S_A^2/S_0^2$. You can estimate S_A^2 and S_0^2 from the data to calculate ψ_A. (E.g., in Stata, you can get S_A^2 by taking the outcome data for units in treatment A, and then computing the variance, taking into account any necessary weighting. You can do the same with the control group data to get S_0^2.)

Now, this is for a completely randomized experiment with no regression controls, but to the MDE for an experiment that includes clustering, stratification, and regression controls, you just need to replace the n_A and n_0 terms with the *effective sample sizes*. That is, take the nominal n_A and n_0 values for the experiment. Then adjust them using the design effects that you calculated above to account for clustering, blocking/stratification, and regression control. These adjusted values are your effective sample sizes that you can plug into the expression for $\tilde{\sigma}_{\hat{\beta}_{A,0}}$ and, then, $\tilde{MDE}$.

- You can try out tools like Stata's sampsi functions, the Optimal Design software (http://hlmsoft.net/od/), or GPower software (http://www.gpower.hhu.de/en.html) to check your calculations.

Resampling and simulation approaches

An alternative to the analytical approach is to use resampling and simulations. This may be necessary for more complex designs (e.g., if the paper uses something like the Bruhn-McKenzie "big stick") or analytical methods that are more complicated than OLS with treatment dummies. For those situations, it may be too complicated to try to extract the design effects and calculate MDEs.

There are a few ways to go about this:

- Perhaps easiest would be to try the DeclareDesign simulation tools (https://declaredesign.org/).
- Alternatively, you can program simulations yourself. This can be quite informative and rewarding, but will take more programming effort. The process goes, essentially, like this
  - Simulate potential outcomes that represent different effect sizes.
    - You can use baseline data to generate the potential outcome data.
    - You can vary potential outcome variances, perhaps in a manner that corresponds to what you see in the actual data.

  - You can also vary the level of effect heterogeneity.
  - Want you want, in the end of this step, is to have a potential outcome value for each treatment arm, for each subject (e.g., if the treatment arms as control, treatment A, and treatment B, then for each subject, you want a simulated control outcome, Y_0, simulated outcome under treatment A, Y_A, and outcome under treatment B, Y_A).

- Produce replicates of the treatment assignment.
  - Use what the paper and supplemental materials tell you about how treatment was assigned, and then use the design information in the data, to generate new treatment assignments.
  - Collect something like 2500 of these treatment assignments.
- Use the treatment assignments and potential outcomes to generate replicates of the experiment:
  - E.g., suppose we have control, treatment A, and treatment B as treatment arms. Then, suppose in the first treatment assignment replicate, the first unit is assigned to treatment A, the second unit to treatment B, the third unit to treatment A, the fourth unit to control, and so on. Then, the outcome data for this first experiment would correspond to the potential outcomes revealed by this vector or treatment assignments (i.e., $Y_A, Y_B, Y_A, Y_0, ...$).
- Then, for each replicate of the experiment, run the analysis (e.g., run the regressions and tests) and save the p-values for hypothesis tests.
- Use the results to assess MDEs.
  - For potential outcomes representing a given effect size, Δ , the proportion of p-values under .05 is the power of the design (under 95% confidence) to detect an effect of size Δ. The MDE would be the smallest effect size for which 80% of p-values are below .05.

Whether you use the analytical approach or resampling/simulation, the final step is to present the MDEs back on scales that are substantively interpretable. If your MDE is scaled in terms of control group standard deviations, then look at either baseline standard deviations values or control group standard deviations for some of the key outcomes, and then present the MDEs on the scale of those outcomes. E.g., suppose your MDE is .2 control group standard deviations and an outcome of interest is income. Suppose further that the baseline standard deviation for income is USD 500/month. Then, you can express the MDE as .2 * USD 500/month = USD 100/month. Expressing MDEs in the scale of key outcomes allows us to assess whether the study is powered to capture something that is *minimally meaningful*, rather than only being able to capture outlandishly large effects.