Nguồn: https://blogs.worldbank.org/impactevaluations/a-pre-analysis-plan-checklist

A pre-analysis plan checklist

DAVID MCKENZIE

|OCTOBER 28, 2012

A **pre-analysis plan** is a step-by-step plan setting out how a researcher will analyze data which is written in advance of them seeing this data (and ideally before collecting it in cases where the researcher is collecting the data). They are recently starting to become popular in the context of randomized experiments, with Casey et al. and Finkelstein et al.'s recent papers in the QJE both using them. There is also some discussion in political science – see this recent paper by Macarten Humphrey's and co-authors.

There are several goals in specifying an analysis plans, but one important reason is to avoid many of the issues associated with data mining and specification searching by setting out in advance exactly the specifications that will be run and with which variables. This is particularly important for interventions which have a whole range of possible different outcomes, like the CDD programs looked at by Casey et al. They look at 334 different outcomes, and illustrate that they could have picked 7 outcomes that made their program look like it strengthened institutions, or alternatively have picked 6 alternate outcomes that make the program look like it weakened institutions. This is less of an issue in evaluating many other policies in which there are one or two most important key outcome (e.g. profits and sales for a firm intervention, attendance and test scores for a school intervention, or incidence of some disease for some health interventions). But even in those cases there are often many different possible choices of how to measure the key outcome, so some ex-ante discipline on how this outcome is defined can be useful.

I'm new to writing these, but have now done them for four different projects. They take quite a bit of work to put together, but I have found two other really useful results from doing them. First, they help in thinking through questionnaire design. By mapping every equation of interest that we want to estimate to the specific questions in the questionnaire that measure the variables in these equations, we can make sure that we don't inadvertently omit variables that we need to know, as well as thinking more carefully about how to measure outcomes in ways which are most amenable to use in analysis. Second, once the data is collected, data analysis is much quicker and easier, since the pre-analysis plan provides a roadmap to follow through, and you pretty much have half the paper already written.

So what should you include in a pre-analysis plan? Here is my checklist for writing one for an evaluation using a randomized experiment – many of the same items would apply for using other types of data:

1. **Description of the sample to be used in the study**: this should include discussion of how the sample was obtained, what the expected sample size is, how randomization

was done (see my paper with Miriam Bruhn for a checklist on what should be reported on this), and what variables will be included in tests of randomization balance and in tests of survey attrition.

2. **Key data sources:** discussion of what the key sources of data will be for the study, including which surveys are planned, and what types of administrative data are planned.

3. **Hypotheses to be tested throughout the causal chain:** this should specify the key outcomes of interest, the steps along the causal chain to be measured, and the subgroup or heterogeneity analysis that is to be done and the hypotheses that accompany each of these tests. These should be as specific as possible, and **link each outcome specifically to how it will be measured.** For example, rather than just saying the outcome will be employment, you should say that the "outcome will be employment, as measured by question D21 on the follow-up questionnaire which asks whether the individual currently works for 20 hours or more per week."

4. **Specify how variables will be constructed:** this includes, for example, where log or levels of particular variables will be used, how missing variables will be handled, what procedures will be used to deal with outliers, etc. For example, "hours worked per week in last month employed will be measured by question D25 on the follow-up survey. This will be coded as zero for individuals who are not currently working; This will be top-coded at 100 hours per week (99th percentile of baseline response) to reduce influence of outliers. No imputation for missing data from item non-response at follow-up will be performed. We will check whether item non-response is correlated with treatment status following the same procedures as for survey attrition, and if it is, construct bounds for our treatment estimates that are robust to this."

5. **Specify the treatment effect equation to be estimated:** for example, is a difference-in-differences, ancova, or post specification to be used? What controls will be included in the regression? How will standard errors be calculated? The exact equation to be estimated should be written out.

6. **What is the plan for how to deal with multiple outcomes and multiple hypothesis testing?** As noted in my post last week, there are a number of methods for dealing with multiple hypothesis testing. These typically involve either aggregating different measures into a single index – in which case one needs to specify precisely which variables will get included in this aggregate; or what variables will be considered as part of the same family when looking at outcomes within a family of domains.

7. **Procedures to be used for addressing survey attrition:** what checks will be done for attrition, and what adjustments will be made if these checks show that there is selective attrition?

8. **How will the study deal with outcomes with limited variation?** An issue which can arise is that the intent is to look at impacts on an outcome which ex post it turns out that everyone in the control group does and where the intended treatment is meant to increase this outcome. There is no power to be gained from looking at this type of outcome, and including it in a family of outcomes can reduce the power to detect an overall impact. So, for example one can write "In order to limit noise caused by variables with minimal variation, questions for which 95 percent of observations have the same value

within the relevant sample will be omitted from the analysis and will not be included in any indicators or hypothesis tests. In the event that omission decisions result in the exclusion of all constituent variables for an indicator, the indicator will be not be calculated." Likewise, one might pre-determine that outcomes which have item non-response rates above a certain threshold will be omitted.

9. **If you are going to be testing a model, include the model:** in many cases papers include a model to explain their findings. But often these models are written ex post as a way of trying to interpret the results that are found, but presented in a way that makes it seem like the point of the paper is to test this model. Setting out a model in advance makes clear the model the authors have in mind before seeing the data – as well as making sure they collect information on all the key parameters in this model.

10. **Remember to archive it:** Since part of the purpose of doing this is to pre-commit to examining particular measures and outcomes, it is important to make sure this can be verified by others. A couple of examples are the JPAL Hypothesis Registry (which is where I have filed mine so far) and the EGAP (Experiments in Governance and Politics) registry. Note that this does not need to mean that the pre-analysis plans are publicly viewable before the study is released – instead they are time-stamped and released upon the request of the researcher. This is important if you are worried about the possibilities of contamination of the experiment if its details are available online. The AEA and 3ie are currently developing registries of RCTs and developing country evaluations respectively, which will allow, but as far as I know so far not mandate, the option of filing such a detailed plan.

If you are interested in writing one, some examples can be found in the 4 currently publicly viewable examples at the JPAL registry, this example for a program in Afghanistan filed in the EGAP registry, and here is one of mine – for a project that tests vocational training in Turkey.

I should note that there are a number of concerns researchers have about tying their hands too much – Casey et al discuss these trade-offs in some detail, and I think most people agree that the use of these plans is not to rule out the possibility of surprise discoveries (see Bill Easterly's satirical post showing Columbus had no impact since he did not find what he was initially looking for). Their use is still rare in economics, so I am sure we will learn a lot more about how to do them and use them over the coming years. Readers: Any other ideas of key things that should be in the checklist that I am omitting?