Part 1: Define and Design

In the first 4 steps, the object is clarity. You want to make everything as **clear as possible *to yourself***. The more clear things are at this point, the smoother everything will be.

1. Write out research questions in theoretical and operational terms

A lot of times, when researchers are confused about the right statistical method to use, the real problem is they haven't defined their research questions. They have a general idea of the relationship they want to test, but it's a bit vague. *You need to be very specific*.

For each research question, write it down in both theoretical and operational terms.

2. Design the study or define the design

Depending on whether you are collecting your own data or doing secondary data analysis, you need a clear idea of the design. Design issues are about randomization and sampling:

- Nested and Crossed Factors
- Potential confounders and control variables
- Longitudinal or repeated measurements on a study unit
- Sampling: simple random sample or stratification or clustering

3. Choose the variables for answering research questions and determine their level of measurement

Every model has to take into account both the design and the level of measurement of the variables.

Level of measurement, remember, is whether a variable is nominal, ordinal, or interval. Within interval, you also need to know if variables are discrete counts or continuous.

It's *absolutely vital* that you know the level of measurement of each response and predictor variable, because they determine both the type of information you can get from your model and the family of models that is appropriate.

4. Write an analysis plan

Write your best guess for the statistical method that will answer the research question, taking into account the design and the type of data.

It does not have to be final at this point—it just needs to be a reasonable approximation.

5. Calculate sample size estimations

This is the point at which you should calculate your sample sizes–before you collect data and after you have an analysis plan. You need to know which statistical tests you will use as a basis for the estimates.

And there really is no point in running post-hoc power analyses–it doesn't tell you anything.

Part 2: Prepare and explore

6. Collect, code, enter, and clean data

The parts that are most directly applicable to modeling are entering data and creating new variables.

For data entry, the analysis plan you wrote will determine how to enter variables. For example, if you will be doing a linear mixed model, you will want the data in long format.

7. Create new variables

This step may take longer than you think–it can be quite time consuming. It's pretty rare for every variable you'll need for analysis to be collected in exactly the right form. Create indices, categorize, reverse code, whatever you need to do to get variables in their final form, including running principal components or factor analysis.

8. Run Univariate and Bivariate Statistics

You need to know what you're working with. Check the distributions of the variables you intend to use, as well as bivariate relationships among all variables that might go into the model.

You may find something here that leads you back to step 7 or even step 4. You might have to do some data manipulation or deal with missing data.

More commonly, it will alert you to issues that will become clear in later steps. The earlier you are aware of issues, the better you can deal with them. But even if you don't discover the issue until later, it won't throw you for a loop if you have a good understanding of your variables.

9. Run an initial model

Once you know what you're working with, run the model listed in your analysis plan. In all likelihood, this will not be the final model.

But it should be in the right family of models for the types of variables, the design, and to answer the research questions. You need to have this model to have something to explore and refine.

Part 3: Refine the model

10. Refine predictors and check model fit

If you are doing a truly exploratory analysis, or if the point of the model is pure prediction, you can use some sort of stepwise approach to determine the best predictors.

If the analysis is to test hypotheses or answer theoretical research questions, this part will be more about refinement. You can

- Test, and possibly drop, interactions and quadratic or explore other types of non-linearity
- Drop nonsignificant control variables
- Do hierarchical modeling to see the effects of predictors added alone or in blocks.
- Check for overdispersion
- Test the best specification of random effects

11. Test assumptions

Because you already investigated the right family of models in Part 1, thoroughly investigated your variables in Step 8, and correctly specified your model in Step 10, you should not have big surprises here. Rather, this step will be about confirming, checking, and refining. But what you learn here can send you back to any of those steps for further refinement.

12. Check for and resolve data issues

Steps 11 and 12 are often done together, or perhaps back and forth. This is where you check for data issues that can affect the model, but are not exactly assumptions. These include:

Data issues are about the data, not the model, but occur within the context of the model

- Multicollinearity
- Outliers and influential points
- Missing data
- Truncation and censoring

Once again, data issues don't appear until you have chosen variables and put them in the model.

13. Interpret Results

Now, finally, interpret the results.

You may not notice data issues or misspecified predictors until you interpret the coefficients. Then you find something like a super high standard error or a coefficient with a sign opposite what you expected, sending you back to previous steps.