**Your Study Here: AsPredicted Easy-Use Template & Tips for Pre-Registration Beginners**

**Mallory L. Dobias, Clinical PhD Student**

[**Lab for Scalable Mental Health**](http://www.schleiderlab.org/) **| Stony Brook University**

[**mallory.dobias@stonybrook.edu**](mailto:mallory.dobias@stonybrook.edu) **|** [**@MalloryDobias**](https://twitter.com/MalloryDobias)

**Special Thanks To:** [**Jessica Schleider, PhD**](http://www.schleiderlab.org/labdirector.html)and[**Michael Mullarkey, MA**](https://michaelmullarkey.me/)for providing feedback on the following materials. Thank you!

While there are multiple pre-registration formats, AsPredicted tends to be the simplest format for beginners to pre-registration. The AsPredicted pre-registration format involves answering 8 questions about your study (hypotheses, sample, methods). *Template responses are provided in italicized text* below, with [bracketed words and phrases] representing places for you to add in your specific information. Some key phrases are hyperlinked for more information.

Links to specific pre-registration examples and instructions for submitting a pre-registration are included at the end of this document.

**1. Have any data been collected for this study already?**

This one is relatively straightforward, you just select one of these three options:

* *Yes, at least some data have been collected for this study already*
* *No, no data have been collected for this study yet*
* *It's complicated. We have already collected some data but explain in Question 8 why readers may consider this a valid pre-registration nevertheless.\**

\*note: this third option is appropriate when you are doing a secondary data analysis, but have not yet accessed/seen the data. Later in this pre-registration, you’ll need to explicitly state exactly how, and to what capacity, you’ve been exposed to the data previously. When you are hoping to pre-register, the general rule of thumb is: the less involvement you’ve had with the data, the better.

**2. What's the main question being asked or hypothesis being tested in this study?**

What specific questions are you trying to answer? What do you think you will find? [Your questions should be **specific** and **testable**](https://opentextbc.ca/researchmethods/chapter/generating-good-research-questions/). If you have > 1 hypothesis, you need to write multiple statements (one per hypothesis) using the template(s) below. I find it helpful to write hypotheses in bulleted or numbered format: it forces me to really be clear/concise when referring to them later in question 5 (e.g.*“To answer hypothesis 1, we did this…”*).

Bad Example: “Hypothesis 1: We will evaluate whether experiencing depression impacts treatment seeking.”

Neither "depression" nor "treatment seeking" is defined in this hypothesis - but both can be defined in many different ways (yes/no diagnosis, severity level, or symptom counts for depression; calling a clinic, attending a treatment session, or doing online research for treatment seeking). A good hypothesis allows future researchers to retest your question by measuring your IV and DV in the same ways that you did.

Good Example: “Hypothesis 1: We will evaluate whether higher depression severity is associated with higher levels of mental health treatment access, in a national sample of N = 2,500 adolescents ages 12-17 years, after controlling for income and insurance status.”

For [correlational research](https://opentextbc.ca/researchmethods/chapter/correlational-research/) (are two variables related to one another?):

Template: *We will determine whether [X predictor variable] is related to [Y outcome variable] in a sample of N = [sample size, if available; relevant descriptives, e.g. children/adolescents/adults, age range, sex, race/ethnicity, gender, etc.] after controlling for the effect of [Z variable].*

For [experimental research](https://opentextbc.ca/researchmethods/chapter/experiment-basics/) (does manipulating one variable affect another?):

Template: *We will determine whether participants randomized to [X group of the independent variable] tended to have higher/lower [Y dependent variable] within a sample of N = [sample size; relevant descriptives, e.g. children/adolescents/adults, age range, sex, race/ethnicity, gender, etc.] after controlling for the effect of [Z variable].*

Then, finish with a statement like this:

Template: *This is a [type of analysis; Is this study a randomized experiment? Is it a secondary analysis of an existing dataset?] using data obtained from the [source of data; name of dataset, or name of upcoming study]. The analyses will be performed primarily by [person’s name and/or their role on the project: e.g. first-author],* ***who has not worked with the data or seen it prior to analyses****.\**

\*Important: Please include the bolded/underlined portion only if it is true. To properly pre-register, the person running the analyses needs to not have seen the data before.

**3. Describe the key dependent variable(s) specifying how they will be measured.**

So you think Variable X is somehow related to your dependent Variable Y. What is Variable Y, and how is it measured? Is it with a self-report scale? A behavioral task? You’ll need to describe your Variable Y with full details about how you’re planning on measuring it.

Example: “We will measure our first dependent variable, depression, using the Mood and Feelings Questionnaire (MFQ-Short; Angold et al., 1995). The MFQ asks respondents to rate how they have been feeling and acting in the past two weeks, and is a screening tool for depression. It consists of 13 statements which the participants rates on a 0 to 2 scale of “not true, sometimes, or true.” Total score ranges from 0 to 26, with higher scores reflecting greater depression symptom severity”

Template: *We will measure our first outcome/dependent variable [Y outcome/dependent variable], using [Z measure, e.g. type of self-report scale, behavioral task, etc.]. [details about self-report measure/behavioral task, e.g. range of scores, etc.].*

If the study is a secondary data analysis of an existing dataset, I like to specify the variable labels used in the data itself here (e.g. YMDE\_01). That way, once I download the data, I’ll have a record of which exact variables to analyze, and I can find them very easily.

Depending on how complicated my study is, I sometimes like to define all of my independent variables here as well (i.e. your Variable X’s).

**4. How many and which conditions will participants be assigned to?**

This is for [experimental research](https://opentextbc.ca/researchmethods/chapter/experiment-basics/) only. If you’re working on [correlational research](https://opentextbc.ca/researchmethods/chapter/correlational-research/) without an experimental manipulation, you can skip this section, and put “N/A”. If you are running an experiment (where participants are assigned to undergo one out of two or more conditions), describe each of your conditions here in detail, and specify how participants will be assigned to each condition (Will they be randomized? Will they self-select into conditions?).

Template: *We will randomly assign participants to X of Y possible conditions: [condition 1 name], [condition 2 name], and/or, [condition 3 name].*

*Participants in [condition 1 name] will… [condition description].*

*Participants in [condition 2 name] will… [condition description].*

*Participants in [condition 3 name] will… [condition description].*

**5. Specify exactly which analyses you will conduct to examine the main question/hypothesis.**

This will likely be the longest section of your pre-registration, as you’ll want to specify things like: which models you will run, any tests of assumptions for those models, corrections for multiple tests, and how you will handle missing data. Each of these is described briefly below.

Choosing an analysis:

While you can choose any one of many analysis techniques, here are a few common techniques:

1. **Regression:** examining relationships between your dependent/outcome variable, Y, and your independent/factor variable X

[**Linear regression**](https://en.wikipedia.org/wiki/Linear_regression) is used when your dependent variable is continuous (e.g. weight: pounds). [Link](https://www.youtube.com/watch?v=zPG4NjIkCjc) to a video explaining the basics of linear regression

[**Logistic regression**](https://en.wikipedia.org/wiki/Logistic_regression) is used when your outcome variable is binary, or the observations are sorted into two groups (e.g. t-shirt size: small, or large). [Link](https://www.youtube.com/watch?v=yIYKR4sgzI8) to a video explaining the basics of logistic regression (with some review of linear regression)

**Simple regression** involves predicting your dependent, or outcome variable, using one independent variable. **Multiple regression** involves predicting your dependent variable using multiple independent variables.

1. [**Analysis of Variance (ANOVA)**](https://en.wikipedia.org/wiki/Analysis_of_variance)**:** examining whether there are statistically significant differences between the means of 3 or more different groups, or “levels” of your independent variable. [Link](https://www.youtube.com/watch?v=ITf4vHhyGpc) to a video explaining the basics of ANOVA

ANOVAs are actually just a special kind of regression analysis where the independent variable is categorical, or separated into groups (e.g. low, medium, and high doses of caffeine). Researchers tend to use ANOVAs more often when conducting experimental studies, and they are looking for differences between conditions of their independent variable.

A **one-way ANOVA** determines the relationship between 1 independent variable and the dependent variable (e.g. effect of caffeine dosage - low, medium, or high - on exam score). **Two-way ANOVAs** evaluate the relationship between 2 independent variables and your dependent variable (e.g. the effects of caffeine dosage and time of day - morning, afternoon, or evening - on exam score).

1. [**t-test:**](https://opentextbc.ca/researchmethods/chapter/some-basic-null-hypothesis-tests/) A t-test is very similar to an ANOVA, but it examines whether there are statistically significant differences between the means of 2 or more groups (rather than 3 or more groups). There are **independent samples t-tests**, where the two groups you’re comparing come from independent, separate samples (e.g. male sex vs. female sex; treatment vs. control group). Alternatively, there are **paired samples t-tests**, where you’re comparing means of the same individuals over time (e.g. depression in the full sample at baseline vs. depression in the full sample at follow-up).

Once you have your model selected, you can write it up. Here’s just one example.

Template: *To address Hypothesis 1, we will run a [specific model type: e.g. linear/logistic regression, one-way/two-way analysis of variance, independent/paired samples t-test ,etc.] with [Y variable] included as our outcome/dependent variable, and [X variable] included as our predictor/independent variable. We will enter [A, B, and C variables; e.g. sex, age, gender, race, income, etc.] as covariates in the model.* [*A p-value of less than .05*](https://opentextbc.ca/researchmethods/chapter/understanding-null-hypothesis-testing/) *will indicate a significant effect of the predictor variable(s) on our outcome/dependent variable, [Y variable].\**

\*Link to a brief explanation of p-values

Correcting for multiple comparisons:

If you evaluate multiple comparisons, you increase your chances of finding a false positive result. If you want to conduct more than one statistical test, *or* if your statistical test will yield multiple p-values, we strongly recommend [correcting for multiple tests](https://en.wikipedia.org/wiki/Multiple_comparisons_problem) in order to avoid a super elevated chance of finding false positive results. There are many ways to do this (some methods being more or less conservative), including: [false-discovery rate (FDR)](https://en.wikipedia.org/wiki/False_discovery_rate), [Bonferroni correction](https://en.wikipedia.org/wiki/Bonferroni_correction), [Šidák correction](https://en.wikipedia.org/wiki/%C5%A0id%C3%A1k_correction), and the [Holm-Bonferroni method](https://en.wikipedia.org/wiki/Holm%E2%80%93Bonferroni_method). Generally, people choose based on how conservative they want to be (i.e. how strict they want to be about preventing an elevated chance of false positives).

Example: “To correct for multiple tests, the false discovery rate (FDR) will be applied to identify potential false-positive results. *Q*-values will be computed for *p-*values from GEE models using an online calculator applying Benjamini and Hochberg’s approach (<https://www.sdmproject.com/utilities/?show=FDR>). Results will be considered significant if corrected *q*<0.05 (incurring maximum 5% false-positive rate amongst all tests meeting significance at *p*<0.05).”

Template: *To correct for multiple tests, the [name of method; e.g. false-discovery rate] will be applied to [description of method]. [Description of how the calculation will be performed]. Results will be considered significant if [describe corrected criteria for significance].*

Handling missing data:

Additionally, you will want to specify [how you’ll treat missing data](https://en.wikipedia.org/wiki/Missing_data#Techniques_of_dealing_with_missing_data). Will you use multiple imputation? Omit data/participants with incomplete data? Handling missing data is especially important in longitudinal research, aka: research that collects information across multiple time points. A few common strategies are: [Listwise deletion](https://en.wikipedia.org/wiki/Listwise_deletion), [multiple imputation, and maximum likelihood](https://www.theanalysisfactor.com/missing-data-two-recommended-solutions/) estimation

Example: “Models will include an autoregressive error structure and use full information maximum likelihood (FIML) estimation to address subject-level missing data.”

Template: *To address missing data, we will perform [selected method].*

Checking your assumptions:

To run many of these models, you must first check whether a few key assumptions about your data are met. Typically, this section is included at the beginning of question 5, but I’ve included it here so each analysis method could be described first.

**This is a very important step to complete before running your analyses**, and different tests come with different types of assumptions (e.g. normal distribution of residuals, homogeneity of variance, etc.). Rather than going into deep detail here, I would recommend checking in with a collaborator/mentor/colleague/statistician if you’re not sure which assumptions apply to your chosen analysis method and how best to test them.

Once you’ve determined ways to test your assumptions, you’ll specify these in your pre-registration:

Template: *We will first test if the assumptions necessary to interpret a [analysis type: linear regression, multiple linear regression, analysis of variance, etc.] are met. If not, we will apply the following corrective practices so that the model is interpretable…*

**6. Any secondary analyses?**

Is there anything else that you want to explore that’s not central to your main hypotheses? It’s best if you can specify things up front.

If you run other analyses later that you did not include up-front in your pre-registration, you’ll need to specify these tests are “exploratory tests” (i.e., tests that were not pre-registered), rather than “confirmatory tests” (i.e., tests that were pre-registered) when you write/talk about them. Confirmatory tests can hold more weight in the scientific community, especially within experimental studies—so clarifying your plans here can mean your results hold more weight when it comes down to writing up your results.

**7. How many observations will be collected or what will determine the sample size? No need to justify decision, but be precise about exactly how the number will be determined.**

I usually find it easiest to answer question 8 first, before coming back to answer question 7. Question 8 requires you to outline specific exclusionary criteria (i.e. which participants you will not include in your final analyses), which can impact the number of participants you have in your final N--especially if you are running a secondary data analysis of existing data but aren’t sure how many participants will be eligible for your analysis. In this case, you may want to specify your sample size will be *N* = the initial number of participant responses minus those you exclude based on the criteria you outline below. You may also want to specify a minimum *N* value (e.g., *We will test our primary hypotheses upon reaching an N of [insert number here], inclusive of individuals who provide usable data for the present study [see exclusion criteria below]*).

If you are using an existing dataset and you aren’t sure how many responses meet your eligibility criteria, *N* = the number of participants collected (minus any participants excluded based on your exclusion criteria from question 8).

Template: *Individuals who [exclusion criteria] will be excluded from the present analyses. We will not know our final N until initial analyses are run to identify the number of eligible participant responses across all of our variables of interest. However, our highest possible N can be estimated using the number of participant responses in the existing dataset. While N = [number of responses already collected], some participants will be excluded based on [exclusion criteria variables].*

*Thus, we will start with a maximum sample size of N = [number of responses already collected], and will exclude any individuals who [eligibility criteria]. Final N will be determined by the number of individuals who meet the criteria listed above.*

*We will also ensure a sample size equal to, or greater than, N = [minimum sample size].*

If you are running your own study, specify at what *N* you will stop collecting data. Then, specify if you will be further excluding participants based on your answer to question 8.

Template: *Our final sample size for collected data will be N = [number of participant responses you will collect]. Our final sample size for analyzed data will be N = [number of collected responses] - any participants excluded based on [exclusion criteria].*

**8. Anything else you would like to pre-register? (e.g., data exclusions, variables collected for exploratory purposes, unusual analyses planned?)**

Research studies often collect more participant data than they actually use in their final analyses. It’s important to specify ahead of time the criteria you will use to determine which (if any) data you will not include in your analysis.

You should specify exclusionary criteria decided before the study begins (e.g. participants who need to be a certain age, gender, etc. to participate) as well as criteria you will use to drop people from your analyses after the data is collected (e.g. if a participant’s data is low quality).

Things to consider here:

* Participant exclusions based on demographics (e.g. age, gender, sex, race)
* Participant exclusions based on some eligibility threshold score (e.g. a score of 16 or higher on a particular screening measure)
* If you are using existing data, was there any branching logic in the original study’s design, where participants would have been selected based on specific criteria? (e.g. in order to be prompted with questions about individual depressive symptoms, adolescents had to endorse “Yes” to experiencing either sadness or boredom for the past two weeks)
* Will you exclude participants with low data quality—and how will you define ‘low data quality’ for the purposes of this study? (e.g. Participants who take longer/shorter than the average amount of time to finish a task? Participants who fail an attention check, or reading comprehension test? Participants who failed to answer free-response questions?). If you plan to exclude based on data quality, specify how you will determine who to exclude (e.g. participants who completed the survey > 3 standard deviations from average survey completion time for the whole sample).

Template: *Individuals who [exclusion criteria] will be excluded from the present analyses.*

**Links to Specific Examples on Open Science Framework (in-full):**

Looking for some complete examples? Here are some from the [Lab for Scalable Mental Health](http://www.schleiderlab.org/):

**Secondary Data Analysis:**

[Predicting Adolescent Mental Health Treatment Access](https://osf.io/ag8ez)

**Experiments (Randomized Clinical Trials):**

[Brief Experiences for Teen Resilience (BETR) Project](https://osf.io/dms3f)

[The College Goals Project](https://osf.io/e3f6x/?view_only=006067e2c2404dcf867718cc871321d7)

\**note that RCTs with health-related outcomes must be pre-registered on clinicaltrials.gov or another* [*ICMJE-approved database*](http://www.icmje.org/about-icmje/faqs/clinical-trials-registration/) *prior to the start of data collection (OSF is not yet ICMJE approved). These pre-registrations may include (but are not required to include) the same level of detail as in an OSF-based pre-registration. It is your choice whether to pre-register your clinical trial on clinicaltrials.gov only (including a detailed analytic plan) or clinicaltrials.gov* ***and*** *OSF.*

**Non-randomized, survey-based study:**

[Parent Perceptions of Single Session Interventions Project](https://osf.io/uerv4)

**Non-randomized, participants select into one or more conditions:**

[Project Youth Empowerment & Support (YES)](https://osf.io/e52p3)

**Instructions on Submitting to Open Science Framework:**

1. Visit the [Open Science Framework website](https://osf.io)
2. Create an account (if you haven’t already)
3. Once logged in, go to your Dashboard and select “Create new project”
4. After assigning a title for your project, you can add contributors to the project (“contributors” and “add”); I usually enter contributors in authorship order on the project
5. Select “Registrations” in the grey toolbar above, then “new registration”
6. Select “Preregistration Template from AsPredicted.org”, then “create draft”
7. Copy/paste your answers into the online portal
8. Choose whether to make it public immediately, or wait 24 hours (when it will automatically go public)

**The Big Question: What if You Deviate from Your Pre-Registered Plan?**

Things sometimes don’t go according to plan. In doing science, unpredictable things happen. Participants find new ways to have low data quality that you never would have anticipated. Maybe a particular task didn’t work as well as you’d hoped. Maybe there was an equipment malfunction; your pre-specified sample size proved impossible to recruit; or one of your measures failed to function as you’d expected.

**All of this is okay. The chance that things won’t go as planned should *not* deter you from pre-registering your study!**

The job of the pre-registration is not to eliminate researchers’ problem-solving abilities or creativity. It’s to make the process of arriving at the answers to our questions more credible and transparent. So, what happens when you need to deviate from the plan you pre-specified?

The answer turns out to be relatively simple: in your writing, link to the pre-registration and specify exactly how you deviated from your proposed plan. Provide detailed justification for why you decided to deviate from your original plan. If appropriate, outline any differences in your findings as a result of the deviation. Honest, clear detail is important and appreciated. If reviewers can see the entirety of your decision-making process, they may even be more likely to trust your decisions (e.g. “Yes, I did plan to include these 3 covariates! Check my pre-registration!”). You can also use pre-registration to your advantage in preventing the Run A Thousand Additional Follow-Up Analyses phenomenon per reviewer request (e.g. “unfortunately, we can’t run proposed model #67, as it was not specified in our pre-registration plan.”).

**Other Tips:**

1. Build in structured flexibility to your analytic plan using if/then statements. (e.g. “If data meet assumptions for interpreting multiple regression, then we will… if not, we will…”). If you have built-in backup plans within your pre-registration, there’s a smaller chance you’ll need to deviate from your pre-registration down the road. Obviously, you won’t catch everything this way. But if you find yourself imagining “but what would happen if X?” (e.g. My assumptions aren’t met? There is a difference between condition groups at baseline?), you can address this ahead of time and save yourself the hassle later.
2. Double check your analytic plan before starting data collection using simulated data. This can help you catch other possible if/then plans to specify. While I haven’t done this personally, I definitely plan on doing this next time. There are ways to [simulate data in R](https://cran.r-project.org/web/packages/simstudy/vignettes/simstudy.html) and [SPSS](https://developer.ibm.com/predictiveanalytics/2014/04/15/using-spsss-simplan-to-generate-fake-data/). Additionally, some programs, like Qualtrics, actually offer ways to [simulate data within the platform itself](https://www.qualtrics.com/support/survey-platform/survey-module/survey-tools/generating-test-responses/). This may be particularly useful if you have a really large project coming up and you want to be sure you’ve thought everything through before hitting “submit”.