Design

Design-Related Questionable Research Practices (Wicherts et al, 2016, Table 1)

1. Creating multiple manipulated independent variables and conditions (i.e., and then *later* selecting only certain conditions for comparison or merging conditions for analysis).
2. Measuring additional variables that can *later* be selected as covariates, independent variables, mediators, or moderators.
3. Measuring the same dependent variable in several alternative ways to increase the likelihood of finding an effect on at least one.
4. Measuring additional constructs that *could* potentially act as primary outcomes.
5. Measuring additional variables that *could* later enable exclusion of participants from the analyses (e.g., awareness or manipulation checks).
6. Failing to conduct a well-founded power analysis.
7. Failing to specify the sampling plan and allowing for running (multiple) small studies.

Guidance:

The key issue here is making decisions that reduce unnecessary complexity in data collection, to limit flexibility during analysis, and evaluation of hypotheses (i.e., confirmatory research). Including multiple measures of the same variable (predictor or dependent variables) in confirmatory research allows for researcher flexibility during the analysis stage. If multiple measures are used as operationalizations of the same construct, be sure to clearly indicate a priori which **one** will be used to evaluate the hypothesis. Switching the measure that is used to evaluate a hypothesis negates the validity of the hypothesis test. Using a measure to evaluate the question underlying a hypothesis that is not specified a priori results in substantially increased Type I error rates. This type of analysis is best considered exploratory – rather than an evaluation of the hypothesis. This same reasoning applies to the use of covariates. It can be challenging to achieve the sample size required to properly power a study. Consequently, you might want to consider programs such as Study Swap as a means of obtaining your requisite sample size. Note that given that most psychology studies typically have statistical power of less than .50, looking at the sample size of a previous study to set your sample size is generally discouraged.

You may find it helpful to read Maxwell and Kelley (2011) prior to planning your sample size:

Maxwell, S. E., & Kelley, K. (2011). Ethics and sample size planning. Handbook of ethics in quantitative methodology, 159-184.

Sample Size / Power Guidance:

A critical aspect of design is determining the sample size that will be used. There are two general approaches:
1) Dynamically setting sample size (i.e., optional stopping)

2) Setting the sample size in advance

Approach 1: Dynamically Setting Sample Size (Optional Stopping)

One approach is to set the sample size dynamically. One periodically examines their data during data collection, and data collection stops when some criterion is achieved (e.g., statistical significance). Historically, this approach has been problematic because it substantially increases Type I errors. Indeed, some authors have noted that, with this optional stopping approach, researchers can always obtain a significant p-value (see Wagenmakers, 2007). Correspondingly, optional stopping (without correction/adjustment) has been classified as a Questionable Research Practice (see Wicherts et al., 2016).

Fortunately, statistical approaches have been devised that allow researchers to use optional stopping (dynamic sample sizes) without engaging in a Questionable Research Practice. One advantage of these approaches is that they do not rely on analyzing power a priori, which can be difficult to estimate accurately. Note, however, that power analyses should still be conducted for other reasons, such as assessing the feasibility of your study given time or financial constraints.

There are two common optional-stopping approaches:

1) Use inferential statistics that directly compare the null and alternative hypotheses, such as the Bayes factor (Rouder, 2014; Schönbrodt & Wagenmakers, 2018; although see de Heide & Grünwald, 2017). The idea here is that you stop data collection as soon as your data provide strong evidence in favour of either the null or alternative, thus avoiding bias for one conclusion over the other.

Rouder, J. N. (2014). Optional stopping: No problem for Bayesians. *Psychonomic Bulletin & Review, 21* (2), 301-308.

Schönbrodt, F. D., & Wagenmakers, E. J. (2018). Bayes factor design analysis: Planning for compelling evidence. *Psychonomic bulletin & review, 25*(1), 128-142.

de Heide, R., & Grünwald, P. D. (2017). Why optional stopping is a problem for Bayesians. arXiv preprint arXiv:1708.08278

2) Optional stopping techniques that involve ‘paying a price’. The simplest version is deciding on the number of times you will peek at your data in advance (e.g., 3) and then applying a Bonferonni correction (alpha / # of peeks) each time you look, instead of alpha equal to .05. Two key articles to read are Lakens (2014) and Sagarin, Ambler, and Lee (2014) that provide less conservative approaches to this problem. Be sure to read these articles and decide determine number of times you will peek at your data before you begin data collection. You might even consider pre-registering the number of times you will peek at your data with this approach.

Lakens, D. (2014). **Performing high-powered studies efficiently with sequential analyses**. *European Journal of Social Psychology, 44(7),*701-710.

Sagarin, B. J., Ambler, J. K., & Lee, E. M. (2014). **An ethical approach to peeking at data**.

Perspectives on Psychological Science, 9(3), 293-304.

Approach 2: Setting the sample size in advance

The key to setting sample sizes in advance is to keep in mind that you **do not set the sample size for the design**(e.g., 2x2 ANOVA). Instead, you **determine the desired sample size for each hypothesis**.

There are two approaches to settings sample sizes in advance.

1. P-*value Approach.*Determining a **desired sample size**for each hypothesis based on **power**(e.g., .80 which is the probability of obtaining a significant result when the alternative hypothesis is true) and **expected effect size**. Examine the desired sample size for each hypothesis and use the largest sample size to ensure all hypothesis meet the desired level of power.

2. *Confidence Interval Approach.*Determine the **desired sample size**for each hypothesis based on the **expected effect size**such that the expected confidence interval will not be larger than the effect size (or some other criterion). For example, if you expect a .30 correlation, the upper bound of the expected confidence interval minus the lower bound should not be larger than .30. In other words, the uncertainty in your effect size estimate should not be larger than the effect itself. Examine the desired confidence interval-based sample size for each hypothesis and use the largest sample size to ensure all hypotheses meet the desired confidence interval width.

In reality, you should probably use both approaches (p-value and confidence interval) to make the most informed sample size plan. Detailed information on both approaches is provided below. We recognize that, following data collection, the **obtained sample size**is often smaller than the **desired sample size**. Therefore, to appropriately interpret *p*-values, you should calculate power after you have finished data collection based on your obtained sample size. Note this is not post hoc power. That is, this power calculation is **not**based on the effect sizes you obtain in your study. Rather, the calculation is based on the effect sizes you specified prior to data collection. This power calculation will allow you and your committee to appropriately interpret your results.

You may also want to calculate the positive predictive value (see description and calculator) which indicates, **given a significant *p*-value**, the probability that the alternative hypothesis is true (details below).

A common problem faced by graduate students is that a thesis must sometimes be submitted/presented prior to the end of data collection. This can be problematic, because it could appear that you are using an optional stopping approach even if that was not your intent. One way to avoid concerns with this course of action is to preregister your planned sample size on the Open Science Foundation website and also preregister that you may need to present a thesis based on a subset of the data prior to end of data collection. Using this approach, you can continue to collect data after you set aside a subset of it to be used for a thesis. Note, you are not stopping data collection - simply setting aside a subset of the data to be used for your thesis. Preregistration and openness make this a viable approach.

1) Expected effect size.

Regardless of whether you are using a confidence interval or *p*-value approach, you will need to have an expected effect size (see calculation details) for each hypothesis. Your expected effect size might be a

specific correlation or standardized mean difference (i.e., d-value). A critical concern is how to pick your expected effect size – see the Hypothesis section of this document in which we outline several strategies.

Example. A past study found *d*= 0.70, *n1*=80, *n2*=80 (relevant to our hypothesis 1) and *r*= .40, *N*= 120 (relevant to our hypothesis 2). We use a safe-guard power approach from this single study and determine expected effect size. Confidence intervals were not reported in the original article. We assume CI's were not reported in the original article and we use the software R to determine the confidence intervals for the effects *d*= .70, 95% CI[0.38, 1.02] and *r*= .40, 95%[.26, .52]. Thus, our conservative *d*-value and correlation expected population effect sizes are 0.38 and .26, respectively.

R code for confidence intervals (assuming **psych**and **MBESS**packages are installed):

```
> library(MBESS)

> ci.smd(smd = 0.70, n.1 = 80, n.2 = 80)

> library(psych)

> r.con(r = .40, n = 160)
```

2. p-value approach to sample size

Setting desired sample size using the power-based approach (i.e., *p*-values will figure prominently in your thesis)

Two tables are illustrated below that should be presented to your committee.

Desired Sample Size Planning:

	a priori expected effect size	Desired power	Overall Sample Size (calculated)
Hypothesis 1	*d*=.38 (CI lower bound)	.80	220 (110 per group)
Hypothesis 2	*r*=.26 (CI lower bound)	.80	113
			Desired N = 220(i.e., pick the higher N)

R code for sample size (assuming **pwr**package is installed):

```
> library(pwr)

> pwr.t.test(d=.38, power=.80)

> pwr.r.test(r=.26, power=.80)
```

Calculating power based on obtained sample size

Actual Power Using Obtained Sample Size:

	a priori expected	Obtained overall sample size	Power based on expected

	effect size		effect size and obtained sample size
Hypothesis 1	d=.38 (CI lower bound)	150 (75 per group)	.64
Hypothesis 2	r=.26 (CI lower bound)	150	.90
etc			

R code for actual power estimate (assuming **pwr**package is installed):

```
> library(pwr)

> pwr.t.test(d = 0.38,  n = 75)

> pwr.r.test(r = .26, n = 150)
```

Calculating positive predictive value based on power

If you report a *p*-value that is significant, a key question is whether the significant *p*-value reflects a "true positive." That is, it would be informative to know the probability that a significant effect reflects a true effect. The number that conveys this information is called positive predictive value (PPV). To understand why most research conclusions in psychology are incorrect and how PPV works, see this video. To calculate PPV for a hypothesis, you need to know alpha (e.g., .05), actual (not desired) power (e.g., .80), and the probability the hypothesis is true. Johnson et al. (2017) found, "the probability that the proportion of experimental hypotheses tested in psychology are false likely exceeds 90%" (p.1). This finding suggests that a .10 value for the "% of true a priori hypothesis" in the link below.
Online PPV Calculator

Johnson, V. E., Payne, R. D., Wang, T., Asher, A., & Mandal, S. (2017). On the reproducibility of psychological science. *Journal of the American Statistical Association, 112(517),*1-10.

3. Confidence Interval Approach to Sample Size

A good approach to setting sample size in advance is to set the required sample size based on the precision you desire in the confidence interval. A good rule of thumb is ensuring the uncertainty in the data is not larger than the effect you are studying. This means the width of a confidence interval (upper bound – lower bound) should not be larger than the effect size (at a bare minimum). Consider the following scenario: You work in a literature with extraordinarily strong effect sizes and your expected effect size is d = 0.38. You would want to set a sample size so that the confidence interval around a d-value of this magnitude is not larger than 0.38. You can do this easily in R with the MBESS package. You simply type the command below (after the package is installed):

R code for sample sized based on confidence interval (assuming **MBESS**package is installed):

```
> library(MBESS)

> ss.aipe.smd(delta=.38, conf.level=.95, width=.38)
> ss.aipe.R2(Population.R2 = .26^2, width = .26^2, p=1)
```

Note 1: We use commands based on regression $R^2$to plan for correlation sample size. So, we need to

use ^2 to indicate the value is squared in the ss.aipe.R2 command above. As well, be aware that p = 1 in the above ss.aipe.R2 commands indicates that the number of predictors is 1.

Note 2: The MBESS package can plan for confidence interval precision for more complex designs – see the documentation. The BUCSS package has many helpful tools for sample size planning especially if you have a **within-participant ANOVA design**. The web apps on the corresponding website Designing Experiments may also be of interest.

Note 3. GPOWERcan also be useful in many scenarios. However, be sure to read the related article in *Behavior Research Methods*for details on how to effectively use GPOWER as well as the follow up articleon correlation and regression designs.

Note 4. Jake Westfall has a number of online power calculators that are helpful: power analysis for crossed random effects, power analysis with two random factors (crossed or nested), and power analysis for general ANOVA designs. This is an excellent source for power analyses for repeated measures designs. Also consider the R package, longpower, for power analyses for repeated measures designs.

Note 5. In terms of Confirmatory Factor Analysis, examine the simsem R package and how it can be used to calculate power under different simulation conditions.

Note 6. If you are using multilevel or nested data the powerlmm R package may for your sample size planning.

Student Check List 2 of 5: Design

We offer a general check list and then an additional checklist for students using dynamic sample size setting.

General:

_____ The student presented a clear rationale and estimate for each expected effect size.

_____ Prior to data collection, the student conducted a thorough power analysis, and has either calculated the needed sample size or committed to a particular “optional stopping” data collection approach.

_____ After data collection, the student is prepared to calculate the observed power (based on expected effect size and obtained sample size) as well as an estimated positive predictive value for each hypothesis.

_____ Correspondingly, the informational value of the study has been discussed with respect to the decision to conduct it.

_____ Estimates of the sample sizes that are needed for confidence intervals that are no larger than the expected effect size were presented for each hypothesis.

_____ The student indicated a commitment to the specific measure that will operationalize each construct with respect to hypothesis testing. (A change of dependent measure for any hypothesis following data collection makes that analysis an example of cherry-picking results and therefore exploratory rather than confirmatory; which implies *p*-values should not be used.)

_____ The student agreed to include all studies conducted as part of the thesis regardless of whether they supported the hypotheses proposed.

_____ Be sure to indicate your intention to share your data in a repository when applying for Research Ethics Board clearance. Wording in the consent form is particularly important in this regard.

Additional: If using the dynamic sample size / optional stopping approach:

_____ The student discussed the advantages and disadvantages of dynamically setting sample size and the approaches for correction.

_____ The student indicated the number of times he/she will peek at the data.

_____ The student indicated the correction approach for peeking that will be used (sequential analysis, *p*-augmented, other).