Statistical Methods in Theses: Guidelines and Explanations

Signed August 2018
Naseem Al-Aidroos, PhD,
Christopher Fiacconi, PhD
Deborah Powell, PhD,
Harvey Marmurek, PhD,
Ian Newby-Clark, PhD,
Jeffrey Spence, PhD,
David Stanley, PhD,
Lana Trick, PhD

Version: 2.00

This document is an organizational aid, and workbook, for students. We encourage students to take this document to meetings with their advisor and committee. This guide should enhance a committee's ability to assess key areas of a student's work.

Context

In recent years a number of well-known and apparently well-established findings have failed to replicate , resulting in what is commonly referred to as the replication crisis. The APA Publication Manual 6thEdition notes that "The essence of the scientific method involves observations that can be repeated and verified by others." (p. 12). However, a systematic investigation of the replicability of psychology findings published in Science revealed that over half of psychology findings do not replicate (see a related commentary in Nature). Even more disturbing, a Bayesian reanalysis of the reproducibility project showed that 64% of studies had sample sizes so small that strong evidence for or against the null or alternative hypotheses did not exist. Indeed, Morey and Lakens (2016) concluded that most of psychology is statistically unfalsifiable due to small sample sizes and correspondingly low power (see article). Our discipline's reputation is suffering. News of the replication crisis has reached the popular press (e.g., The Atlantic, The Economist, Slate,Last Week Tonight).

An increasing number of psychologists have responded by promoting new research standards that involve open science and the elimination of Questionable Research Practices. The open science perspective is made manifest in the Transparency and Openness Promotion (TOP) guidelines for journal publications. These guidelines were adopted some time ago by the Association for Psychological Science . More recently, the guidelines were adopted by American Psychological Association journals (see details) and journals published by Elsevier (see details). It appears likely that, in the very near future, most journals in psychology will be using an open science approach. We strongly advise readers to take a moment to inspect the TOP Guidelines Summary Table.

A key aspect of open science and the TOP guidelines is the sharing of data associated with published research (with respect to medical research, see point #35 in the World Medical Association Declaration of Helsinki). This practice is viewed widely as highly important. Indeed, open science is recommended by all G7 science ministers. All Tri-Agency grants must include a data-management plan that includes plans for sharing: "research data resulting from agency funding should normally be preserved in a publicly accessible, secure and curated repository or other platform for discovery and reuse by others." Moreover, a 2017 editorial published in the *New England Journal of Medicine*announced that

the *International Committee of Medical JournalEditors*believes there is "an ethical obligation to responsibly share data." As of this writing, 60% of highly ranked psychology journals require or encourage data sharing.

The increasing importance of demonstrating that findings are replicable is reflected in calls to make replication a requirement for the promotion of faculty (see details in Nature) and experts in open science are now refereeing applications for tenure and promotion (see details at the Center for Open Science and this article). Most dramatically, in one instance, a paper resulting from a dissertation was retracted due to misleading findings attributable to Questionable Research Practices. Subsequent to the retraction, the Ohio State University's Board of Trustees unanimously revoked the PhD of the graduate student who wrote the dissertation (see details). Thus, the academic environment is changing and it is important to work toward using new best practices in lieu of older practices—many of which are synonymous with Questionable Research Practices. Doing so should help you avoid later career regrets and subsequent public mea culpas. One way to achieve your research objectives in this new academic environment is to incorporate replications into your research. Replications are becoming more common and there are even websites dedicated to helping students conduct replications (e.g., Psychology Science Accelerator) and indexing the success of replications (e.g.,Curate Science). You might even consider conducting a replication for your thesis (subject to committee approval).

As early-career researchers, it is important to be aware of the changing academic environment. Senior principal investigators may be reluctant to engage in open science (see this student perspective in a blog post and podcast) and research on resistance to data sharing indicates that one of the barriers to sharing data is that researchers do not feel that they have knowledge of how to share data online. This document is an educational aid and resource to provide students with introductory knowledge of how to participate in open science and online data sharing to start their education on these subjects.

Guidelines and Explanations

In light of the changes in psychology, faculty members who teach statistics/methods have reviewed the literature and generated this guide for graduate students. The guide is intended to enhance the quality of student theses by facilitating their engagement in open and transparent research practices and by helping them avoid Questionable Research Practices, many of which are now deemed unethical and covered in the ethics section of textbooks.

This document is an informational tool.

How to Start

In order to follow best practices, some first steps need to be followed. Here is a list of things to do:

1. Get an Open Science account. Registration at osf.io is easy!
2. If conducting confirmatory hypothesis testing for your thesis, pre-register your hypotheses (see Section 1-Hypothesizing). The Open Science Foundation website has helpful tutorials and guides to get you going.
3. Also, pre-register your data analysis plan. Pre-registration typically includes how and when you will stop collecting data, how you will deal with violations of statistical assumptions and points of influence ("outliers"), the specific measures you will use, and the analyses you will use to test each hypothesis, possibly including the analysis script. Again, there is a lot of help available for this.

Exploratory and Confirmatory Research Are Both of Value, But Do

Not Confuse the Two

We note that this document largely concerns confirmatory research (i.e., testing hypotheses). We by no means intend to devalue exploratory research. Indeed, it is one of the primary ways that hypotheses are generated for (possible) confirmation. Instead, we emphasize that it is important that you clearly indicate what of your research is exploratory and what is confirmatory. Be clear in your writing and in your preregistration plan. You should explicitly indicate which of your analyses are exploratory and which are confirmatory. Please note also that if you are engaged in exploratory research, then Null Hypothesis Significance Testing (NHST) should probably be avoided (see rationale in Gigerenzer (2004) and Wagenmakers et al., (2012)).

This document is structured around the stages of thesis work: **hypothesizing, design, data collection, analyses, and reporting – consistent with the headings used by Wicherts et al. (2016).**We also list the Questionable Research Practices associated with each stage and provide suggestions for avoiding them. We strongly advise going through all of these sections during thesis/dissertation proposal meetings because a priori decisions need to be made prior to data collection (including analysis decisions).

To help to ensure that the student has informed the committee about key decisions at each stage, there are check boxes at the end of each section.

How to Use This Document in a Proposal Meeting

1. Print off a copy of this document and take it to the proposal meeting.
2. During the meeting, use the document to seek assistance from faculty to address potential problems.
3. Revisit responses to issues raised by this document (especially the Analysis and Reporting Stages) when you are seeking approval to proceed to defense.

Consultation and Help Line

Note that the Center for Open Science now has a help line (for individual researchers and labs) you can call for help with open science issues. They also have training workshops. Please see their website for details.

Hypothesizing

Hypothesizing Questionable Research Practices (Wicherts et al, 2016, Table 1)

1. Conducting exploratory research without any hypothesis (and later characterizing it as confirmatory).
2. Studying a vague hypothesis that fails to specify the direction of the effect (e.g., a two-sided hypothesis when a one-sided hypothesis is appropriate).

Guidance:

A common problem in psychology is specifying hypotheses in a vague way that makes it easier to engage in questionable research practices (also known as *p*-hacking). Indeed, Lakens (2017) noted that, "statistics teaching should focus less on how to answer questions and more on how to ask questions." Consequently, we encourage you to ask your question in a manner that does not facilitate later *p*-hacking as described below.

When hypothesizing you should clearly specify both the direction and the magnitude of an effect. Avoid vague statements like, "there will be an interaction." Instead, specify the exact pattern of the interaction: "For men, there will be a large effect of aggression (approx. d= 0.80), such that men in the high crowding condition will be more aggressive than men in the low crowding condition. For women, the effect of crowding will be smaller and negligible." Note that sample size planning (e.g., power analysis) requires an effect size estimate, so why not incorporate that effect size into your hypothesis?

Effect Size Specification

There are three primary approaches to specifying the effect size you expect for a hypothesis:

1. Published / Pilot Study.

Researchers often use the effect size from a previous study(ies) to guide sample size analysis. Although common, this approach is problematic because the effect size likely will be a substantial overestimate, given the combined effects of sampling error and publication bias. In practice, published effect sizes tend to be double of those from replications (Reproducibility Project). Therefore, one approach is to use an expected effect size that is half the published effect size. Alternatively, you can use a safeguard approach to determining expected effect size. A safeguard approach involves using the lower bound of the effect size confidence interval, which is easy to calculate if not provided. You should also consult Anderson et al. (2017) for additional solutions to this issue, and associated software. Regardless, published effect sizes should not be used "as is" in power estimates given that they are likely overestimates that will result in underpowered studies. Even meta-analytic estimates of the literature are likely biased (e.g., ego-depletion research). Consequently, if you choose this approach, you are well advised to use the lower bound of a meta-analytic effect size confidence interval following the safeguard strategy.

2. Standard/Common effect sizes (small, medium, large).

Another approach is to use small, medium, and large effect sizes. A common practice is to refer to Cohen's standards for small, medium, and large (correlations .10, .30, .50, respectively; d-values 0.20, 0.50, 0.80, respectively; partial-eta squared .01, .06, .14, respectively). However, these classifications are controversial (Ellis, 2010, p. 40). In terms of historical context, these effect sizes are based on an evaluation of effects of a single volume (Volume 61) of the *Journal of Abnormal and Social Psychology*.

The problem with this approach is that what constitutes a small, medium or large effect size varies greatly with research context. More recent evaluations of effect sizes have been done that are more comprehensive in nature – and domain specific. Note that, due to publication bias, the effect size estimates below are likely overestimates. Moreover, it may be difficult to know how an effect size was calculated – so you may want to calculate it on your own from the summary statistics provided in the article.

Industrial Organizational Psychology. A review of approximately 150,000 focal and non-focal

relations over a 30-year period revealed a median correlation of .16 with an interquartile range of .07 (25thpercentile) to .32 (75thpercentile; Bosco, Aguinis, Singh, Field, & Pierce, 2015). Thus, in I-O Psychology, small, medium, and large correlations correspond to .07, .16, and .32.

Social Psychology.A review of 100 years of social psychological research revealed a mean correlation of .21, median correlation of .18, and a standard deviation of correlations across literatures of .15. (Richard, Bond, Stokes-Zoota, 2003). Thus, in Social Psychology, small, medium, and large correlations could be approximated (using the SD) as .03, .18, and .33. Likewise, small, medium, and large d-values correspond roughly to .06, .36, and .69.

Cognitive Neuroscience.A review of approximately 10,000 cognitive neuroscience articles revealed an inter-quartile range of *d*= 0.34 (25thpercentile) to *d*= 1.22 (75thpercentile; Szucs & Ioannidis, 2017). Thus, in cognitive neuroscience a small effect size is *d*= 0.34 and a large effect size is *d*= 1.22.

Clinical Child and Adolescent Psychology. To our knowledge, this field has not undertaken a self-study and so we suggest that student researchers in this area use the effect sizes from the analysis of the I/O or Social Psychology fields, which are comparable.

3. Minimum effect size of interest.

This approach is favoured among some statisticians in psychology due to the fact that the researcher must make it clear the direction and size of the effect in which he/she is interested. For example, let's say that you set the minimum effect size of interest as a *d*-value of 0.80 and use this value in your sample size analysis. In doing so, you are saying that a *d*-value of 0.75 (or anything below 0.80) is not of theoretical importance or interest. Moreover, you are saying that if you obtain an effect of *d*= 0.75 you are "ok" with it being non-significant due to its lack of theoretical importance. This approach gives the researcher the greatest *a priori*flexibility in determining what effect sizes are of interest.

Preregistration of Hypotheses

One solution to the problem of hypothesizing after results are known is the preregistration of hypotheses. This practice prevents researchers from conducting exploratory analyses and later reporting them as confirmatory. As noted above, preregistration of hypotheses before data collection is becoming increasingly important – and a requirement for journals using Level 3 of the TOP guidelines. Note that preregistration of associated data analysis plans is also important. It is discussed in the analysis section of this document.

The committee meeting in which you obtain approval of your thesis is, effectively, a process in which you preregister your thesis hypotheses with the committee. Why not go the extra step and register your hypotheses with the Open Science Foundation (https://osf.io) or at As Predicted (https://aspredicted.org)?

These four links may be of interest: OSF 101, Preregistration: A Plan, Not a Prison, Open Science Knowledge Base, and the Open Science Training Handbook.

Student Check List 1 of 5: Hypothesizing

_____ The student created directional hypotheses.

_____ The student indicated what effect size should be expected for each hypothesis.

_____ The student explicitly indicated which analyses will be exploratory.

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).

Data Collection

Data Collection Questionable Research Practices (Wicherts et al, 2016, Table 1)

1. Failing to randomly assign participants to conditions.
2. Insufficient blinding of participants and/or experimenters.
3. Correcting, coding, or discarding data during data collection in a non-blinded manner.
4. Determining the data collection stopping rule on the basis of desired results or intermediate significance testing. Proviso – unless using a valid dynamic stopping approach (discussed previously).

Guidance:

With respect to random assignment, it is not only important to commit to true random assignment when possible (i.e., by using a random number generator), but it is also important to keep in mind that the goal of random assignment (i.e., equivalence of participants between conditions on all nuisance factors) can be achieved only with much higher sample sizes than previously thought (see article and reflection). The threats to experimental integrity by not blinding participants and/or experimenters are well known (i.e., demand character and inadvertent direction of participants' responding). Performing **any**operation on data in a non-blinded manner could inflate the Type I error rate. As Wagenmakers and others have made clear, it is paramount that there be no possibility of stopping based on observed results. Doing so **will**inflate the Type I error rate – unless appropriate mitigation procedures are used (e.g., *p*-augmented, sequential analysis, or by using a Bayesian statistical approach).

Student Check List 3 of 5: Data Collection

_____ The student thoroughly discussed design considerations and reasoning with the committee.

_____ The student agreed to create power estimates following data collection (see previous section)

_____ The student discussed a data sharing plan with the committee.

_____ The ethics application form appropriately reflects the data sharing strategy on the *Informed Consent*page (if relevant).

Analyses

Analysis Questionable Research Practices (Wicherts et al, 2016, Table 1)

1. Choosing between different options of dealing with incomplete or missing data on ad hoc grounds.
2. Specifying pre-processing of data (e.g., cleaning, normalization, smoothing, motion correction) in an ad hoc manner.
3. Deciding how to deal with violations of statistical assumptions in an ad hoc manner.
4. Deciding on how to deal with outliers in an ad hoc manner.
5. Selecting the dependent variable out of several alternative measures of the same construct.
6. Trying out different ways to score the chosen primary dependent variable.
7. Selecting another construct as the primary outcome.
8. Selecting independent variables out of a set of manipulated independent variables.
9. Operationalizing manipulated independent variables in different ways (e.g., by discarding or combining levels of factors).
10. Choosing to include different measured variables as covariates, independent variables, mediators, or moderators.
11. Operationalizing non-manipulated independent variables in different ways.
12. Using alternative inclusion and exclusion criteria for selecting participants in analyses.
13. Choosing between different statistical models.
14. Choosing the estimation method, software package, and computation of standard errors (SEs).
15. Choosing inference criteria (e.g., Bayes factors, alpha level, sidedness of the test, corrections for multiple testing).

Guidance:

All of the above practices inappropriately (and arguably unethically) allow researchers to make analysis decisions based on the nature of the data obtained. You can avoid these QRPs by constructing a formal **data analysis plan**that concretely addresses all decisions. For example, if you plan to conduct a moderated multiple regression analysis, you should specify (prior to data collection) what alternative procedure you will use if you violate the assumptions of regression (e.g., high reliability of predictors, multivariate normality of errors, etc.). Likewise, if you plan to use a covariate in your regression, you should specify (prior to data collection), not only what the covariate is, but what alternative procedure you will use if the covariate interacts with another predictor. The principle behind doing so is that the researcher will have a clear record of their analysis intentions prior to data collection so that they can demonstrate researcher flexibility was not used during analyses. Some data analysis plans go so far as to have blank templates for the tables and graphs that will be used in the final thesis. Ideally, the data analysis plan is stored before data collection in a repository such as the Open Science Foundation. As per point 3 above, it is crucial to evaluate the assumptions under your analyses (see Osborne, 2017)

When interpreting data, a common practice is to use *p*-values. If reporting *p*-values, report them exactly

and do not round down to meet significance. For example, do not round .054 to .05 (doing so avoids Questionable Research Practice #5). Unfortunately, *p*-values are poorly understood by psychological researchers. Indeed, approximately 80% of psychology professors do not understand the correct interpretation of *p*-values (Haller & Kraus, 2002; Kline, 2009, pp. 120, 125). A correct definition of a *p*-value is available in Kline (2009)—be sure to consult this reference. In addition, there is a long history of criticism of the Null Hypothesis Significance Testing Process (NHSTP) that questions the value of the practice (e.g., Cohen, 1994; Cumming, 2008). Indeed, although most journals accept *p*-values, some have banned them (see Woolston, 2015).

In addition, the American Statistical Association has issued a statement with a few key points about *p*-values (see below). These points were designed to provide "*principles to improve the conduct and interpretation of quantitative science.*" Context for these points is also available.

"*p*-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone"

"By itself, a *p*-value does not provide a good measure of evidence regarding a model or hypothesis."

"A *p*-value, or statistical significance, does not measure the size of an effect or the importance of a result."

"Scientific conclusions and business or policy decisions should not be based only on whether a *p*-value passes a specific threshold."

The consequence of these truths is that a thesis based exclusively or primarily on *p*-values does not represent good science.

How should student researchers proceed in light of these truths? One promising approach is captured in the quotation below from the Executive Director of the American Statistical Association:

"In the post p<0.05 era, scientific argumentation is not based on whether a p-value is small enough or not. Attention is paid to effect sizes and confidence intervals. Evidence is thought of as being continuous rather than some sort of dichotomy."

Ron Wasserstein,
Executive Director,
American Statistical Association, 2016
(Read the complete interview)

Confidence Intervals

The recommendation of the Executive Director of the American Statistical Associations to interpret data using effect sizes and confidence intervals is consistent with APA task force on statistical significance (see PDF links at bottom of the task force webpage). The APA task force report stated "*Always present effect sizes for primary outcomes*" and "*Interval estimates should be given for any effect sizes involving*

principal outcomes" (*p*. 599, Wilkinson, 1999). The 2016 American Statistical Association position goes beyond this by suggesting that confidence intervals and effect sizes should be the primary means of interpretation.

Confidence intervals can be constructed using raw data units (e.g., CI around a mean or mean difference) or around a standardized effect size (e.g., *r* or *d*). A survey of researchers indicated that researchers frequently fail to understand what is conveyed by a confidence interval (Cummings, Williams & Fidler, 2004; also see this document by Howell). Consequently, it may be helpful to review how to interpret confidence intervals "by eye" (Cumming & Finch, 2005; for standard error whiskers see Cumming, Fidler, & Vaux, 2007). In most cases, if dealing with the difference between means, it's easier to interpret a confidence interval for the difference (e.g., *d*-value with CI), rather than the two means.

In short, a confidence interval can be interpreted as a plausible estimate of the range of population-level effects that could have caused the sample effect (see Cumming & Finch, 2005). Population values closer to the middle of the confidence interval are somewhat more likely than those at the extremes. In using confidence intervals, there is a temptation to use them merely as a proxy for significance testing (i.e., in a dichotomous way). This practice is ill-advised. Nevertheless, there is a tendency to do so, as indicated in the article "Editors can lead researchers to confidence intervals but can't make them think: Statistical reform lessons from medicine" by Fidler et al. (2004). That is, many researchers, when switching to confidence intervals, make the error of trying to use them to provide dichotomous evidence (i.e., reject/fail to reject the null hypothesis) rather than continuous evidence. Continuous evidence requires researchers to think about the full range of the confidence interval when interpreting their findings.

Confidence Interval Walk Away Message

We suggest using confidence intervals as the primary basis for your conclusions. When making scientific or applied conclusions ask yourself: "*are my conclusions consistent with the full range of effect sizes in the confidence interval*?" If not, revise your conclusions.

It can be difficult to know what to focus on when reporting confidence intervals. If an effect is significant, it makes sense to discuss the (absolute magnitude) lower bound of the confidence interval to indicate how small the effect could be. Conversely, it makes sense to discuss the (absolute magnitude) upper bound of the confidence interval to indicate how large the (non-significant) effect could be.

In some instances, the confidence interval may be sufficiently wide that few meaningful conclusions appear possible (e.g., the plausible population effect size ranges from near zero to large). In this event, the primary conclusion may be that a larger sample size is needed in that research domain. We provide example text for reporting confidence intervals in the next section.

Practical Significance

In addition to statistical significance, there is an increasing emphasis on the practical significance of findings (e.g., How many fewer days does a major depressive episode last given a certain treatment versus control?). Here is a great example of how to investigate practical significance in the context of testing for an interaction using regression.

Student Check List 4 of 5: Data Analyses

_____ The student has presented the committee with a **data analysis plan**that addresses each of the above points.

_____ This data analysis plan for confirmatory hypothesis testing must be completed prior to data collection.

_____ The data analysis plan clearly indicates the specific analysis that will be used for each hypothesis.

_____ Where relevant, the data analysis plan clearly indicates how the assumptions will be assessed for each hypothesis test and handled if violated.

_____ The data analysis plan clearly indicates how problematic outliers (i.e., points of influence) will be dealt with.

_____ The data analysis plan clearly indicates a strategy for handling missing data in analyses (e.g., pairwise deletion, listwise deletion, etc).

_____ In the event that covariates are to be included in any analysis, the specific covariates for each hypothesis test are mentioned in the pre-data collection analysis plan.

_____ Confidence intervals will be reported for all tests.

_____ Consider avoiding *p*-values when conducting exploratory analyses.

_____ Consistent with the TriAgency position on data management, uploading analysis scripts, descriptions of variables in the data file (i.e., data code book), and the data to an open access platform (e.g., osf.io) was considered.

Reporting

Reportinig Questionable Research Practices (Wicherts et al, 2016, Table 1)

1. Failing to assure reproducibility (verifying the data collection and data analysis).
2. Failing to enable replication (re-running of the study). Poor practices could include inadequate detail in methods sections, failure to make experimental material available, etc.
3. Failing to mention, misrepresenting, or misidentifying the study preregistration.
4. Failing to report “failed studies” that were originally deemed relevant to the research question.
5. Misreporting results and *p*-values.
6. Presenting exploratory analyses as confirmatory (Hypothesizing After Results Known).

Guidance:

Many of the above concerns can be mitigated simply by following the Tri-Agency recommendations for data sharing and management. Beginning in the fall of 2017, all Tri-Agency grants must include a data-management plan: “research data resulting from agency funding should normally be preserved in a publicly accessible, secure and curated repository or other platform for discovery and reuse by others.” Providing a codebook along with your data is a good idea, and see these practical tips for ethical data sharing.

As noted elsewhere, the American Statistical Association states that scientific conclusions should not be based on *p*-values alone. Consequently, effect sizes (raw or standardized) with confidence intervals should be reported for each hypothesis. According to Cumming & Finch (2001) a confidence interval can be interpreted as a plausible range of population effect sizes that could have produced the effect observed in the sample. Be sure to visualize your effect size before writing about it to avoid over interpretation of results (see links for d-values and correlations).

Exploratory analysis should be reported as exploratory. Because *p*-values are only meaningful with an a priori hypothesis, they should not be reported with exploratory analyses (see details in Gigerenzer (2004) and Wagenmakers et al., (2012)). Rather, exploratory analyses should be seen as an empirical process for generating hypotheses to be tested in a subsequent study. Confidence intervals are still appropriate with exploratory analyses.

Many psychology papers have reporting errors that substantially change interpretation of results. A review of 28 years of *p*-values (over 250,000 *p*-values) revealed that the test statistic (e.g., t(28) = 2.60) is inconsistent with the reported *p*-value (e.g., *p* = .0147) more than 50% of the time (Nuijten, Hartgerink, van Assen, Epskamp, & Wicherts, 2015). Interestingly, willingness to share data appears to be associated with fewer statistical reporting errors (Wicherts, Bakker, & Molenaar, 2011). Fortunately, errors can be detected easily with an automated process. A PDF of a thesis can be checked at http://statcheck.io/. Note that some journals (e.g., *Psychological Science*) are now using statcheck on all non-desk rejected articles as part of the review process.

You will likely find the article "Writing Empirical Articles: Transparency, Reproducibility, Clarity, and Memorability" a helpful resource.

New APA Reporting Standards for Quantitative Journal Articles:

In addition, the American Psychological Association has issued new Journal Article Reporting Standards (JARS). We strongly recommend that you read the article and follow its prescriptions.

Example Text:

Below is example text for correlations and *d*-values from the Cumming & Calin-Jageman (2016) textbook. We suggest using a similar style, but also adding *p*-values to the text (and of course ensuring APA-style).

Cumming, G., & Calin-Jageman, R. (2016). Introduction to the new statistics: Estimation, open science, and beyond. Routledge.

Correlation:

“The correlation between well-being and self-esteem was *r*= .35, 95% CI[.16, .53], N = 95. Relative to other correlates of well-being that have been reported, this is a fairly strong relationship. The CI, however, is somewhat long and consistent with anywhere from a weak positive to a very strong positive relationship.” (pp. 324-325)

The correlation between well-being and gratitude was *r*= .35, 95% CI [-.11, .69], N = 20. The CI is quite long. These data are only sufficient to rule out a strong negative relationship between these variables.” (p. 325)

d-value:

"Motor skill for men (M = 33.8%, s = 12.8%, n = 62) was a little higher than for women (M = 29.7% s = 15.8%, n = 89). The difference in performance may seem small in terms of raw scores (M_{mean}- M_{women}= 4.0%, 95% CI [-.7, 8.8]), but the standardized effect size was moderate ($d_{unbiased}$= 0.28, 95% Ci [-.05, 0.61]) relative to the existing literature. However, both CIs are quite long, and are consistent with anywhere from no advantage up to a large advantage for men. More data are required to estimate more precisely the degree to which ender might be related to motor skill."(p. 188)

Ethical Issues in Reporting

Considering and reporting power, positive predictive values, and confidence intervals for all hypothesis tests help to facilitate a method of reporting that draws attention to the statistical inference limitations of a study. In the past, it was common for authors to be in a scenario associated with low power, low positive predictive values, and wide confidence intervals, but not report these statistics and correspondingly draw overly strong conclusions. The process we have outlined in this document is designed to provide readers of your thesis with a realistic view of the inferential limitations of your study and allow them to fairly consider the informational value of the study and other hypotheses. This is consistent with the Canadian Psychological Association's Ethics Guidelines presented below.

III.8 Canadian Psychology Association Code of Ethics

Acknowledge the limitations, and not suppress disconfirming evidence, of their own and their colleagues' methods, findings, interventions, and views, and acknowledge alternative hypotheses and explanations.

III.9 Canadian Psychology Association Code of Ethics

Evaluate how their own experiences, attitudes, culture, beliefs, values, individual differences, specific training, external pressures, personal needs, and historical, economic, and political context might influence their activities and thinking, integrating this awareness into their attempts to be as objective and unbiased as possible in their research, service, teaching, supervision, employment, evaluation, adjudication, editorial, and peer review activities.

Student Check List 5 of 5: Reporting

__ The student used http://statcheck.io/ on the thesis document and provided the committee with the resulting report.

__ The student reported confidence intervals and effect sizes.

__ Interpretation of results was based on the full range of the confidence interval – which conveys the uncertainty of the effect size estimate. Do recognize the center of the confidence interval is more likely than the extremes.

__ Applied recommendations were only made in a manner consistent with the full range of the confidence interval. In particular, keep in mind the lower-bound of the confidence interval.

__ Student has clearly identified exploratory analyses and avoided reporting *p*-values for these analyses.

__ Only a priori confirmatory hypotheses were reported as confirmatory.

__ All analyses conducted were reported.

__ All studies conducted were reported.