# Traversing the landscape of experimental power *

J Michelle Brock † Kathryn N Vasilaky ‡

February 7, 2016

**Abstract**

We present an overview of the use of power calculations in experimental economics as well as other disciplines. We review the methodology proposed by the field of economics as well the pitfalls in failing to incorporate power calculations in lab and field experiments. We write this note to further draw attention to the issue, and to make a case that details of power calculations should be reported in experimental economics papers. This note should serve as a reference and overview to researchers in experimental economics on power calculations.

JEL: C9

Keywords: Power, Experiments, Design, Significance

*Thanks to contributions from the ESA discussion forum.

†European Bank for Reconstruction and Development(e-mail: BrockM@ebrd.com)

‡Earth Institute, Columbia University(e-mail: katyav@iri.columbia.edu)

# 1 Introduction

> In spite of years of teaching and using statistics, we had not developed an intuitive sense of the reliability of statistical results observed in small samples. Our subjective judgments were biased: we were far too willing to believe research findings based on inadequate evidence and prone to collect too few observations in our own research. (Kahneman, 2011)

While economists learn about statistical power in introductory econometrics courses, the concept is more or less ignored when in reporting of results. The significance of a test is central to the discussion of results in studies published in peer reviewed journals, but its cousin, power, rarely makes it into published articles or seminar presentations. This holds both in analyses of secondary data as well as in analyses of controlled lab and field experiments. Calculating the optimal sample size or minimum detectable effect has become secondary (or even tertiary) to finding a significant correlation.

For many of us, our econometrics training focused on tools for ex-post analysis of secondary data rather than statistical analysis of experimental data. When using data that has already been collected, power calculations seem unnecessary and researchers are not expected to perform them. In lieu of power calculations, experimental economists tend to apply rules of thumb (e.g. N>30) for determining sufficient sample sizes. Rules of thumb are not without statistical underpinning ((Berenson et al., 1988), pg 227), but power calculations are important in designing surveys and laboratory experiments, and experimental economists are either not doing enough of them, or simply not reporting them.

Harrison and List (2004)'s paper on experimental design describes, in detail, the importance of power. Power matters in experimental and survey design to:

1. Replicate: Contribute to the replicability of studies, comparisons of studies, and ultimately, to the external validity of results[1]

[1]Prediction markets in which past studies are replicated and the results are bid upon is one such route (Gelman and Carlin, 2014; Nosek, 2015).

2. Detect an effect: Ensure that researchers give themselves the best possible chance to detect an effect

3. Minimize costs

4. Increase efficiency: Researchers optimize on the number participants to assign to each treatment group depending on the variance of the outcome variable across groups, as well as the correlation in outcomes between participants.

With the growing trend of field experimentalists (where power is king due to demands of granting institutions[2]) and lab experimentalists working together, we doubt that power will continue to be neglected by experimental economists. We write this note to further draw attention to the issue, and to make a case that details of power calculations should be reported in experimental economics papers. Power calculations are important to the discipline. The gains to standard reporting of power calculations outweigh the (small) effort of using them over using rules of thumb.

# 2 What is power?

Power is the probability that the significance of a coefficient is correctly detected, and is inversely related to Type II error. In intervention research, this is referred to as sensitivity - the ability to detect a difference between the treatment and control conditions on some outcome of interest (Lipsey, 1990). It is the other side of the coin to significance, but is rarely cited in most economics papers, if it is at all consulted ex-ante. We suspect that this is in part due to the fact that empirical economists often rely on pre-existing data, with no ability to change sample size and consequently, no control over power.

Let us derive power in the context of an intervention, where $\bar{\mu}$ is the mean of the outcome variable of interest, and $\mu$ is the average of the outcome variable in the population of interest.

[2]See a review by granting institution 3ie of the pitfalls of underpowered studies (White, 2014)

For the remainder of the discussion, we invoke the central limit theorem on our sample averages, which are therefore normal.

Before running any intervention, one must first specify a hypothesis. A hypothesis has two parts, a null hypothesis and an alternative hypothesis.

**Null Hypothesis**:

$H_0 : \mu = \mu_0$

**Alternative Hypothesis**:

$H_1 : \mu = \mu_1$

For every hypothesis, there are two possible realities - either $H_0$ is true or $H_1$ is true. Statistical tests are always in terms of $H_0$: one either rejects $H_0$ or fails to reject $H_0$. When deciding whether to reject $H_0$ it is possible to make two different kinds of errors. These are referred to as Type I and Type II errors.

**Type I Error**:

A Type I error, or a false negative, occurs when one rejects the null hypothesis that $\mu = \mu_0$ when it's in fact true. The probability of a Type I error is

Prob(reject $H_0|H_0$)=Probability$[\ \frac{\bar{\mu}-\mu_0}{\frac{\sigma}{\sqrt{n}}}] \leq B_\alpha = \alpha$

where $B_\alpha$ denotes the cutoff value associated with the $\alpha$ portion of the standard normal distribution.

**Type II Error**:

A Type II error, or a false positive, occurs when one does not reject the null hypothesis

$\mu = \mu_0$ when it is in fact false. The probability of a Type II error is usually denoted as $\beta$.

Prob(reject $H_0|H_1$)=Probability[ $\frac{\bar{\mu}-\mu_1}{\frac{\sigma}{\sqrt{n}}}$] $\leq B_\beta = \beta$

where $B_\beta$ denotes the standard normal critical value associated with a $\beta$ critical level.

**Power**:

Power is 1- $\beta$, or 1 minus the probability of a Type II error. It's the probability that a researcher makes the right decision when the null is not correct (i.e. we correctly reject it). And, of course, the lower the Type II error, the higher the power.

Prob(reject $H_0|H_1$)=Probability[ $\frac{\bar{\mu}-\mu_1}{\frac{\sigma}{\sqrt{n}}}$] $\geq 1 - B_\beta = \beta$

When we compute power we are looking for the minimum sample size, n, that balances a desired probability of a type I error and power, such that:

$\bar{\mu} \geq B_\alpha \frac{\sigma}{\sqrt{n}} + \mu_0$

and

$\bar{\mu} \geq 1 - B_\beta \frac{\sigma}{\sqrt{n}} + \mu_1$

Setting the expressions for $\bar{\mu}$ equal to each other and solving for n:

$$n = (B_\alpha + B_{1-\beta})^2 - \frac{\sigma^2}{(\mu_1 - \mu_0)^2} \tag{1}$$

Null hypotheses are typically stated in terms of an effect of a treatment or intervention being equal to zero (i.e. ex ante assuming that there is no statistically significant effect). If a test is "well-powered" then there should be enough data to reject the null if there truly is an

effect from the intervention (i.e. be able to detect a statistically significant effect, where the change in the outcome is not equal to zero, or some known starting point). It is generally accepted to aim for a power of 0.8.

# 3 Why test for Power?

Researchers care about power in order to optimize their chances of detecting significant changes in outcomes. A well-powered test is one in which n is chosen for a specified level of power and statistical significance, given an anticipated change in the outcome variable.[3] Imagine one would like to test the relationship between X and Y. Suppose there is in fact a strong positive relationship. Not having sufficient power means that the probability of finding that relationship is low. A researcher runs the chance of not finding a significant relationship between X and Y when it actually exists. This is particularly important in research using clinical trials. Imagine X is a medication and Y is a health outcome. Researchers would like to find a positive effect if it exists, but they also need to know if the drug has a negative effect, because a negative health effect implies that the medication not only does not work, but it may in fact harm patients. They cannot simply fail to reject a null hypothesis that the drugs effect is equal to a certain value; they need a two tailed hypothesis with enough power to detect any relationship that may exist. In a low powered study, where a drug's side effects are statistically different from zero, one may fail to detect this. Ozier (2010) provides a useful table.

Accounting for power lowers the probability of falling into the upper right quadrant of this table.

Laboratory experiments in economics are not harmful, the way drug side effects may be.

[3]Most empirical economists use at least 30 observations as a rule of thumb. However, "thirty" most likely became a popular reference number because it is the threshold number for inference and for invoking the Central Limit Theorem on our averages (and betas). Thirty may be a sufficient number given the central limit theorem, but this is under perfect circumstances: uncorrelated error terms between subjects, and low "enough" standard errors. Further, a well-powered test is unrelated to the proof that the statistical average approaches a normally distributed random variable.

| | Test Result | |
|---|---|---|
| | **Reject Null-Find Effect** | **Accept Null-No Effect** |
| **Truth: There is an effect** | A) Great | B) Type II Error (low power) |
| **Truth: There is no effect** | C) Type I Error (test size) | D) Great |

Generally, the lack of detecting a statistically significant effect of an economic intervention in the laboratory does not compare with the lack of detecting a potentially lethal side effect of a drug. So then why should experimental economists care about power calculations? Taking power into account when designing a study is important for economists because doing so will improve replicability and external validity, increase the focus on economically meaningful effect sizes and help to minimize the cost of data collection.

## 3.1 Improving Replicability and External Validity

Power matters for experimentalists and surveyors because using simple rules of thumb for determining samples sizes reduces potential replicability of a study. Replication is easier when there is a standardized approach to determining and reporting sample sizes. Standardization also allows researchers to more easily compare the validity of results across studies, and adds to the scientific legitimacy of economists' research. Without replicability and comparability across studies, the external validity of each study cannot be effectively tested, and economists reduce their opportunity to generalize their empirical findings with theory. (Daniel Kahneman called for a similar requisite from the field of psychology in his open letter (Yong, 2012).) This is particularly important for development microeconomics, in which small-scale studies, conducted in specific village settings, are questioned for their generalizability. Suppose a study conducted in India finds significant effects of an intervention, and the study falls within quadrant A of our table. Researchers and policy makers conclude that the intervention is effective. A similar intervention conducted in Uganda fails to find a significant effect with the same sample size, not because the intervention is ineffective but because the effect size is smaller. This second study thus turns out to be underpowered, so

the results fall within quadrant B. If neither study reports power calculations or discusses power in their results, we cannot draw accurate conclusions about whether this intervention works across contexts. On the other hand, if we know the power of both, we can more easily compare results across studies and make statements about the external validity of the India study.

Finally, reporting power and, in particular, the elements that enter into calculating power, forces authors to be clear about their experimental design, as one needs information on experimental design to run power calculations. Further, reporting details of power calculations can contribute to a pool of validated statistics available for future experiments. We refer again to Equation 1. How do we know what values to use for $\mu$ and $\alpha$? In other disciplines, researchers generally take them from other studies. Thus, when researchers report their power calculations, they not only validate their own work, but also provide sample statistics for other researchers to use.

## 3.2 Finding Economically Meaningful Effect Size

Using power as a factor in study design can improve focus on effect sizes that are economically meaningful. Effect size, $\mu_1 - \mu_0$, is in the denominator of the power calculation. The output of the calculation is the sample size needed to maximize chances of actually finding an effect *of that size*. Putting a study in terms of economically significant effects facilitates communication of results, especially with policy makers. Moreover, choosing a sample size that allows one to detect *any* significant change may be misguided if effect size matters for interpreting results. By way of example, consider, increasingly popular use of "big data" for analyzing internet-based experiments, also known as A/B testing. Some A/B testing is conducted through automated programs, and increasing sample size comes at a very low cost, as the data (e.g. click through rates) are often collected at a high frequency. Hence, one can easily detect significant effects that are nonetheless quite small. Practitioners should question whether these small detectable differences are economically significant (e.g. is a change in

click through rates of 0.00001% of economic significance to overall sales?). The advantage of power calculations in this setting is that one can define the economically significant effect size and tailor the sample size accordingly. Note, however, that sufficient power does not guarantee the detection of a pre-specified change, nor does it guarantee that the change will be significant (especially, if the pre-specified standard errors were inaccurate). Nonetheless, power calculations based on an economically significant effect can inform an optimal sample size choice and reduce criticism over significance that may result from use of unnecessarily large samples ((Nuzzo, 2014), pg 151) for the purpose of asterisk hunting (Gelman and Carlin, 2014; Lenth, 2001).

On the opposite end of the spectrum is detecting effects in small samples. One may think that if a significant effect is detected in an underpowered study then power calculations are simply perfunctory. On the contrary, with small sample sizes the sampling distribution of the outcome tends to exhibit greater variation or wider tails. As a result, studies with smaller sample sizes and lower power actually have a greater probability of exhibiting a larger effect size (Gelman and Weakliem, 2009). Therefore, power calculations are relevant to detecting meaningful effects in large samples and believable effects in small samples.

## 3.3 Minimizing Cost

Related to detecting a desired effect size is the corresponding cost of the sample. To ensure that there is a statistically significant effect from a change in drugs, social programs, or even the placement of ads on a website, scientists need a sufficiently large sample of participants. Since it is costly to run interventions with very large samples, power calculations are used to estimate the minimum sample needed to detect a specified change. Power is thus an essential element of any grant proposal for a field experiment - these are expensive studies, and calculating the minimum number of subjects necessary to detect an economically significant effect is what determines field studies' budgets. Grantors also look at power as a signal of whether the estimated cost is appropriate and of the risk that a study will be a good place

to put their money.

While the cost per participant is small in laboratory experiments relative to field experiments, calculating the optimal sample size is still a valuable exercise. Laboratory experiments require fewer subjects to find an effect because it is a more controlled environment. The early literature in experimental economics set a precedent of "take what you can get". Most of the seminal papers in experimental economics have sample sizes of 30-40 participants per treatment, but the power of the overall experiment size is rarely discussed. Earlier work such as Andreoni (1995a,b) are examples of experiments that use 40 subjects per treatment with a total of 120 and 80 total subjects respectively.

More recently experimentalists are able to aim for much larger samples, in the hundreds rather than the tens.[4] With this ability to pump up experimental samples, power calculations become more useful and more necessary for containing cost.

## 3.4 Technical Considerations: Study design and efficiency

Many experimental designs call for more than one treatment group, and most analyses of experiments conduct more than one comparison or hypothesis test. Both the number and size of treatment arms, and the number of hypotheses to be tested must be factored into power calculations. If the researcher plans to involve multiple treatment arms, she should account for this in her power calculations. If the variance of the outcome variable is expected to differ across treatment arms, this should also be accounted for in the power calculations. Finally, if the researcher plans to conduct multiple hypothesis tests, such as between each treatment and the control, she should account for this in power calculations.

Multiple treatments

[4]The exception to this may be those who do classroom-based experiments in the field due to small subject pools (e.g. teachers, farmers in small villages, etc.). These kinds of experiments also require smaller sessions because using paper and pencil for collecting data and the relative inexperience of participants are unwieldy with too many subjects per session.

List et al. (2011) provide a derivation of the power calculations for a treatment and control group with unequal variances in their equations 6 and 7. This can be extended to M treatment arms. However, most studies with several treatment groups uniformly distribute 30 subjects into each cell, even when the variance of the outcome variable across treatment groups varies. With such suboptimal designs (equal sizes across all treatment groups) a larger sample size is required to achieve the same power. Ideally, treatment arms that exhibit a lower variance in the outcome would require a smaller sample size.

Another correction that may be necessary in power calculations is a within group correction for session-level clustering. If the experiment allows for interaction between members of a particular session and treatment arm, then power calculations should be adjusted for clustering standard errors. This accounts for the fact that individual error terms are not independently distributed. List et al. (2011) provide a derivation of the power calculations for within group clustering in their equation 9.

Multiple Testing

Power calculations should also account for the number of hypothesis tests that will be conducted on an outcome variable. As the number of hypothesis tests increases, say 20 or more, the probability that one of them will be significant rapidly increases. After M independent tests, the probability of a type I error is $1-(1-\alpha)^M$. So after 50 tests, and 5% significance, the probability of falsely rejecting the null is already 92%. There are two paths to account for multiple testing: before the experiment by adjusting the power calculations with a corrected Type I error rate, or after the experiment by adjusting inference using a Bonferroni correction.

In terms of documenting efforts to calculate power prior to data collection, pre-registration is one practice that is beginning to take hold. Pre-registration essentially forces a researcher to publicize her intended hypothesis tests and the needed sample size for those tests. This prevents the researcher from data mining, or running dozens of hypothesis tests, while only

reporting the one or two results that she found to be significant. Anderson (2008) proposed this methodology and provides the Stata code used to account for multiple inference (also known as FWER, family-wise error rate) in his pre-registered study, and economists Katherine Casey, Rachel Glennerster and Edward Miguel implement the methodology in Casey et al. (2011).

A Bonferroni correction accounts for multiple testing after the experiment is conducted, at the hypothesis testing stage. If M independent comparisons are conducted on the data, the significance level of each test is 1/n of the significance level if only one test were conducted.[5] For the reasons we stated earlier regarding detection of effect size in both large and small samples, we think it is preferable that researchers account for multiple testing within their power calculations rather than at the hypothesis testing stage.

# 4 Who currently checks for power?

There are inconsistencies across fields as to whether practitioners report power calculations, or calculate their sample size needed to detect x effect, before they run an experiment or a survey. We looked at a handful of fields' use of experiments, where their respective terminology for controlled experiments is in parentheses, followed by our understanding of whether they run power analyses at the study design stage:[6]

1. MDs (clinical trials, randomized controlled trials), Definitely
2. Biostatisticians, randomize controlled trilas, Definitely
3. Epidemiologists (RCTs, field trials and community trials), Definitely
4. Experimental economists (laboratory experiments), Sometimes

[5]Bonferroni correction for multiple testing is applicable when tests are independent as well as when they are not independent. When tests are not independent, the correction is more complicated.

[6]Researchers may choose not to run power calculations when they have no control over N. But the exercise can still be meaningful, especially when looked at in conjunction with potentially insignificant results.

5. Development economists (randomized control field trials), Definitely

6. Psychologists (laboratory experiments, clinical trials), Generally

7. Social psychologists (laboratory experiments), Generally

8. Evolutionary dynamics (laboratory, field experiments), Generally

9. Web analytics, data science (A/B testing), Sometimes

In reality, not all of these fields are consistent about running and reporting power calculations. Micro-development economics, perhaps because of the cost imposed on collecting large sample datasets, are more inclined to report their power calculations (Duflo et al., 2008). However, in psychology, a field with similar, if not higher, costs to increasing sample sizes, field studies often omit discussions of power. A 2008 survey of prominently published RCT papers in psychology finds that as much as 50% of published papers neglect to mention power (Faulkner et al., 2008). That being said, there is a difference between doing a power analysis to aid experimental design and reporting those details in a manuscript. So who reports details of power calculations in published articles?

1. MDs (clinical trials), Definitely

2. Epidemiologists, Definitely

3. Experimental economists (lab experiments), Usually do not [7]

4. Development economists (randomized control trials), Usually do not

5. Social psychologists (lab experiments), Not always

6. Evolutionary dynamics (lab and field experiments), Not always

[7]For example, Zhang and Ortmann (2013) tracked the frequency of reported power calculations among all dictator game studies published in Experimental Economics from 2010 to 2012 and find that only one reported on power.

7. Web analytics, data science (A/B testing), Sometimes[8]

Part of the lack of reported power calculations among economists may be because researchers may not know how to compute them, or that built in tools in statistical packages like R or Stata are quite limited in the number of parameters that can be adjusted. For instance, Stata's sampsi command doesn't allow for heterogeneity in the standard deviation of the outcome variable across multiple treatment groups, something that List et al. (2011) also speak to (page 447). One way around this is by using simulations to estimate study power (Arnold et al., 2011). Simulations can accommodate complex designs. When the study calls for multiple tests, several treatment arms with unequal variances, and dependent observations, simulated power calculations can be quite useful. Using simulations for this purpose is common for statisticians and scientists (Cristofolini and Testoni, 2000; van der Sluis et al., 2008), but less common for social scientists. Essentially, the researcher simulates data generated from a distribution that accounts for the anticipated effect size, N times. She then returns power as the proportion of p-values that are less than alpha.[9] User added programs do exist within Stata and R (Luedicke, 2013; Smart, 2013), as well as stand alone code that has not been packaged (Yoeli, Yoeli; York University, York University). Power calculations by bayesian methods also exist using the Stan package (Gelman, Gelman). For analytic power calculations in Stata and R see the packages PSS and PWR respectfully. For a simple overview of power calculations more generally see the Jameel Poverty Action Lab's note on power calculations in the course "Evaluating Social Programs" (JPAL, JPAL).

[8]Large organizations that are running A/B tests do run simulated power calculations. Smaller startups may use built in tools to run automated A/B tests. Automated tools tend to increase the sample size as the experiment is being conducted, a faulty practice in itself, as it increases the Type I error rate of the experiment.

[9]If the p-value is greater than alpha, that means we fail to reject the null. However, the data is generated with a pre-determined effect size between (or within) classes. So the number of times we fail to reject the null but should is the Type II error, or 1-power.

# 5 Do ex post power calculations tell us anything?

In using existing data sets, where the sample size is pre-determined, there may seem little to be gained from power calculations. Indeed, ex post low power may just mean that the researcher had wrong assumptions about the standard errors or reasonable effect size ex ante. But power can inform the discussion of what it means when there is no significant effect. With the well-known publication bias toward significant results, discussions about the probability that a correlation may in fact exist, for which the power was insufficient to detect, are rare. Similarly, when coefficients are significant, there may not be a demand to report the probability that such a result should be found given the outcome variable's standard deviations and given sample size.

Reporting power could help reduce publication bias if we were to give power the similar deference as that we give to statistical significance. The failure to detect significance in a well-powered study, where others may have found a significant effect in a low powered study, is an important part of getting closer to the truth. For example, Zethraeus et al. (2009) look at the relationship between hormones and economic behavior in the lab. The authors are explicit about the power of their study, which is sufficiently high at over 90%. The hormone intervention is implemented as a (double-blind) medical trial. Participants are randomized into different hormone treatment groups and then play a series of games. Importantly this study involves women. The authors find no significant effect, a contradiction to existing correlative results using men (e.g. Apicella et al. (2008); Burnham (2007)) and an important contribution to this growing area of research. But would such a result ever be published in an economics journal? Would the reviewers have appreciated the power analysis as an endorsement of the non-result?

# 6 Conclusion

The purpose of this note has been to discuss the value of power calculations and reporting of power in published articles in economics, in particular among experimental and development economists. We discussed why we test for power and expounded briefly on the costs of not testing for power. Testing for power is increasingly popular and we hope to see not only the results but the inputs into the calculations more and more in papers and presentations. Power calculations help to discipline research design. Sharing details of power calculations will help the profession to develop accepted standards for how inputs (i.e. standard error estimates) should be decided in the absence of empirically motivated, context specific priors. We reiterate that the gains to standard reporting of power calculations outweigh the (small) effort of using them over using rules of thumb.

# References


Andreoni, J. (1995a). Cooperation in Public-Goods Experiments: Kindess or Confusion? *The American Economic Review 85*(4), 891–904.

Andreoni, J. (1995b). Warm-Glow Versus Cold-Prickle: TheEffects of Positive and Negative Framing on Cooperation in Experiments. *The Quarterly Journal of Economics 110*(1), 1–21.

Apicella, C., A. Dreber, B. Campbell, P. Gray, M. Hoffman, and A. Little (2008, November). Testosterone and financial risk preferences. *Evolution and Human Behavior 29*(6), 384–390.

Arnold, B. F., D. R. Hogan, J. M. Colford, and A. E. Hubbard (2011, January). Simulation methods to estimate design power: an overview for applied research. *BMC medical research methodology 11*(1), 94.

Berenson, M. L., D. M. Levine, and D. Rindskopf (1988). *Applied statistics: a first course*. Englewood Cliffs: Prentice Hall.

Burnham, T. C. (2007). High-testosterone men reject low ultimatum game offers. *Proceedings of the Royal Society 274*(July), 2327–2330.

Casey, K., R. Glennerster, and E. Miguel (2011). Reshaping Institutions: Evidence on AID Impacts Using a Pre-analysis Plan.

Cristofolini, L. and M. Testoni (2000). The importance of sample size and statistical power in experimental research. A comparative study. *Acta of Bioengineering and Biomechanics 2*(1).

Duflo, E., R. Glennerster, and M. Kremer (2008). Using Randomization in Development Econimcs Research: A Toolkit. *Handbook of Development Economics 4*(07), 3895–3957.

Faulkner, C., F. Fidler, and G. Cumming (2008, February). The value of RCT evidence depends on the quality of statistical analysis. *Behaviour research and therapy 46*(2), 270–81.

Gelman. MultilevelPowerCalculationUsingFake-DataSimulation.

Gelman, A. and J. Carlin (2014). Beyond Power Calculations : Assessing Type S ( Sign ) and Type M ( Magnitude ) Errors. *Association for Psychological Science 9*(6), 641–651.

Gelman, A. and D. Weakliem (2009). Of Beauty , Sex and Power. *American Scientist 97*.

Harrison, G. W. and J. A. List (2004). Experiments. *Journal of Economic Literature 42*(4), 1009–1055.

JPAL. How to do Power Calculations in Optimal Design Software.

Kahneman, D. (2011). *Thinking, Fast and Slow*. New York: Penguin Books.

Lenth, R. V. (2001). Some Practical Guidelines for Effective Sample Size Determination. *The American Statistician 55*(3), 187–193.

Lipsey, M. (1990). *Design Sensitivity: Statistical Power for Experimental Research*. Newbury Park, CA: Sage.

List, J. a., S. Sadoff, and M. Wagner (2011, March). So you want to run an experiment, now what? Some simple rules of thumb for optimal experimental design. *Experimental Economics 14*(4), 439–457.

Luedicke, J. (2013). Simulation-based power analysis for linear and generalized linear models. In *Stata Conference*, pp. 1–25.

Nosek (2015). Estimating the reproducibility of psychological science. *Science 349*.

Nuzzo, R. (2014). Statistical errors. *Nature 506*, 150–152.

Ozier, O. (2010). Sample Size and Power Calculations. Technical Report July, JPAL, Limuru, Kenya.

Smart, F. (2013). R-bloggers Welcome ! Jobs for R- users Popular Searches. In *July 24, 2013*, Number 552, pp. 1–8.

van der Sluis, S., C. V. Dolan, M. C. Neale, and D. Posthuma (2008, March). Power calculations using exact data simulation: a useful tool for genetic study designs. *Behavior genetics 38*(2), 202–11.

White, H. (2014). Improving development policy and practice Requiring fuel gauges : A pitch for justifying impact evaluation sample size assumptions.

Yoeli, E. Stata Code for Simulated Power Calculations.

Yong, E. (2012). Nobel laureate challenges psychologists to clean up their act. *Nature News*.

York University, S. C. MATH 6643 Summer 2012 Applications of Mixed Models/Simulation for Power.R.

Zethraeus, N., L. Kocoska-Maras, T. Ellingsen, B. von Schoultz, A. L. Hirschberg, and M. Johannesson (2009, April). A randomized trial of the effect of estrogen and testosterone on economic behavior. *Proceedings of the National Academy of Sciences of the United States of America 106*(16), 6535–8.

Zhang, L. and A. Ortmann (2013). Australian School of Business Working Paper.