Sample size calculation in economic experiments

Sven Gruener

Martin Luther University Halle-Wittenberg

09.09.2019

Abstract

Clinical studies and economic experiments are often conducted with randomized controlled trials. In clinical studies, power calculations are carried out as a standard. But what's about economic experiments? After describing the basic idea of the calculation procedure, I tackle the practice of sample size calculations in the realm of experimental economics by considering the publications of 5 economic journals in the period 2000–2018. These are two top-ranked economic journals (*Quarterly Journals of Economics* and *American Economic Review*), the leading field journals in the area of experimental economics (*Experimental Economics*) and behavioral sciences (*Journal of Economic Behavior and Organization*), and a leading field journal in environmental economics (*Environmental and Resource Economics*). In contrast to clinical drug trials, sample size calculations have rarely been carried out by experimental economists. But the number of power calculations has slightly increased in recent years, especially in the leading journals of economics. However, this can be partly explained by the fact that field experiments (in which scholars pay more attention to power analyses than in lab experiments these days) play an important role in these journals.

Keywords

Economic experiments; sample size; power, multiple testing

JEL

C90; C10

Article type

Under debate

1 Introduction

Randomization and double-blinding are the key design techniques in clinical trials. Blinding is used to mask the intervention patients receive. In practice, double-blinded studies are commonly used in which information that could reveal to which group the patients belong to are not only withheld from the patient but also from the investigator. Randomization is the random assignment of subjects to either the control or one of the treatment groups. It aims to ensure that the characteristics of subjects between the groups are as similar as possible, i.e. statistical independence between both known and unknown confounders (e.g. age, gender or other personal characteristics) and groups (e.g. Dunning 2012). Randomization achieves balance of covariates in expectation. However, in certain and particularly small samples, randomization may be prone to not achieving a satisfactory covariate balance and thus jeopardizing valid causal inference. A standard procedure in clinical drug trials is to calculate the optimal sample size before collecting the data (e.g. Whitley and Ball 2002; Noordzij et al. 2010). If too few subjects are recruited (i.e. the study is underpowered) it may not be possible to detect a meaningful treatment difference between groups if such a difference exists (Nguyen and Landais 2017). This runs the risk that the results do not give a reliable answer to the research question of the clinical trial. Whereas recruiting too many subjects (i.e. the study is overpowered) leads to waste of resources, including monetary costs for both the experimenter and the sponsor of the study, and a possible delay in data collection. Ethical considerations may also be problematic. In contrast, the question of the optimal number of subjects has received much less attention in the social sciences.¹ What methodical differences can there be between clinical studies and experimental economics? Randomized controlled clinical trials are rather expensive and directly linked to the health of the patients (experimental subjects). In addition, clinical trials are aimed to make inference from the sample to the population. Thus, the negative consequences of erroneous results due to low external validity are far-reaching. Roughly speaking, there are two categories of economic experiments: laboratory and field experiments. Many laboratory experiments are carried out with university students because they are easy to recruit and have low opportunity costs (Guala 2005; Levitt and List 2007). These experiments go hand in hand with low monetary costs and are easy to replicate. Some of the laboratory experiments

¹ Two important exceptions are List et al. (2011) and Maniadis et al. (2014).

look for causal relations within the sample but do not claim to be generalizable (e.g. experiments that test assumptions/predictions of standard economics). In contrast, field experiments are carried out in a real-world setting. The high external validity (generalizability) is often considered as an advantage of field experiments (e.g. Roe and Just 2009). However, they are often cost intensive, not easy to replicate, and ethical considerations are an issue. All in all, field experiments seem to have some similarity to clinical trials (e.g. complexity, costs). It is therefore conceivable that power analyses are at least not completely meaningless in field experiments.

The rest of the paper is structured as follows. After presenting a small tutorial on how to calculate the sample size in economic experiments (Section 2), I shall describe the practice of sample size calculation in 5 selected economic journals (Section 3). Section 4 concludes.

2 Sample size calculation—Parameters and experimental applications

2.1 Fundamental parameters of power analyses

To conduct sample size calculations, researchers must consider the following five parameters: type I error, type II error, power, MRD (i.e. **Minimal Relevant Difference**), and the variability of the outcome variable (e.g. Flurkey et al. 2007; Noordzij et al 2010; cf., Table 1):

- Potential for **Type I error** exists if a statistical test indicates differences between two groups. Type I error is the probability of falsely rejecting H_0 (false positive result). It is usually denoted by the Greek letter α and often set at 0.05.
- There is potential for **Type II error** if a statistical test does not indicate real differences between two groups. Type II error is the probability of falsely accepting H_0 (false negative result), i.e. missing the real effect. It is usually denoted by the Greek letter β and often set at 0.20. If negative findings (“no difference”) are especially important for the conclusion of a study, lower β ’s are to be used (e.g. 0.1).
- **Power** is the (pre-study) probability of correctly rejecting H_0 (true positive result), i.e. the chance of detecting the effect. It is inverse to Type II error (i.e. $1 - \beta$). Power is often set at 0.8 (i.e. $\beta = 0.2$) for a pivotal clinical trial. Power calculations should be carried out before collecting data. For example, Senn (2002) argues that ex-post calculations should not be performed because power is not about analyzing data.

- Another variable that is required to calculate the optimal sample size deals with the question of what difference the investigator wants to detect. The investigator needs information about the minimum difference between treatment and control group (i.e. effect size of the treatment) deemed important. The **minimal relevant difference (MRD)** must be “clinically relevant and biologically plausible“ (Noordzij et al. 2010:1389) in the context of medical studies; and economically relevant and plausible with spite of human behavior in the context of economic experiments. For this purpose, previous studies, former publications or expert opinions can be used. MRD is a numerical difference if the outcome variable is continuous, whereas event rates are used in case of binary outcomes.
- Moreover, assumptions about the **variability** of the sample data are required. These are initially unknown and must be taken from pilot studies or similar previous studies.² Similarity relates among other things to methods, subjects, and the design of the experiment. The variability of the sample data is measured in terms of the population standard deviation if the outcome measure is continuous, or as a formula of event rates (depending on group assignment and outcome) if the outcome measure is binary.

Table 1 Parameters to calculate the sample size

α (Probability Type I error)	Probability, falsely rejecting H_0 (=false positive result)
β (Probability Type II error)	Probability, falsely accepting H_0 (=false negative result)
Power = $1 - \beta$	(Pre-study) probability, correctly rejecting H_0 (=true positive result)
Minimal [Clinically/Economically] Relevant Difference (=MRD)	Difference between treatment and control group the investigator wants to detect (=effect size of the treatment)
Variability	Variability of the outcome variable

Determining parameters such as α and β directly influences the result of the sample size calculation. Some researchers have recently proposed to use 0.005 instead of 0.05 as the default p-value threshold (cf., Benjamin et al. 2017). However, fixed thresholds are arbitrary to a certain extent in any case. Moreover, assumptions about the minimal relevant difference and the variability seem to be challenging. As it will be shown by simple examples in the following section, and can

² Zelmer (2003) notes that meta-analysis results are helpful to conduct a priori power calculations.

be illustrated with the help of sensitivity analyses in which assumptions are systematically varied, even small changes can alter the required sample size substantially. However, thinking about the optimal sample size gives the investigator at least an idea under which circumstances more or less subjects are necessary to answer the research question at hand.

2.2 Calculating sample sizes in economics—Two simple (fictitious) experimental examples

There formulas to calculate the sample size depend on the type of response variable (e.g. continuous, ordinal or binary), study design (e.g. cross-over trial, parallel group trial), and aim of the trial (e.g. superiority trials, non-superiority trials, non-inferiority trials) (Noordzij et al 2010; Christensen 2007). However, it is relatively straightforward to perform power analyses with the help of software tools (Hickey et al. 2018). It can be assisted by stand-alone power and sample size software, such as G*Power, PS, nQuery or PASS (the first two of these examples are for free). However, general-purpose statistical software with built-in power analysis capabilities (e.g. Stata, Statistica) is suitable for many applications nowadays. For example, the comprehensive Stata manual on sample size calculations does not only provide details on how to run such analyses but also gives the reader an impression about the many types of calculations that are implemented as standard. Beyond standard power calculations, simulations can be used to deal with more complex models (e.g. structural equation models, multilevel models). Huber (2019) describes, for example, how to calculate power/sample size using Monte Carlo simulations. It is not within the scope of this paper to give a comprehensive overview of these formulas, as there are many good overviews elsewhere. The interested reader may refer to Julious (2004), Julious and Campbell (2012) or fundamental manuals of biomedical research. Instead, I would like to present the required components, briefly provide the intuition behind the calculations, and discuss whether this is a practical way in experimental economics. With this in mind, I illustrate the procedure using a randomized controlled trial design³ in which the experimenter compares only two groups of equal sample size.

2.2.1 Continuous outcome variable

If the outcome measure is continuous, the required sample size for *each* group reads as follows:

³ Complexity is somewhat greater with other designs. The formulas used in this section to calculate the optimal sample sizes are widely distributed among fundamental papers that deal with clinical trials or biomedical analyses. They can be found, for example, in Noordzij et al (2010: 1390) or Flurkey et al. (2007).

$$N_{1,2} = \frac{2(a + b)^2 \sigma^2}{(\mu_1 - \mu_2)^2}$$

To calculate the sample size one needs the standardized Z-scores (multipliers) for α (denoted as a in the formula above) and β (denoted as b in the formula above). The multiplier a can be both one-tailed and two-tailed. In two-tailed tests, the direction of the difference between the mean of the treatment group and the control group is not specified a priori. In contrast, with one-tailed tests the direction is specified a priori. For example, the multiplier of $\alpha = 0.05$ equals 1.96 if the test is two-tailed and 1.645 if the test is one-tailed; the multiplier for $\beta = 0.2$ equals 0.842 (cf., Table 2). For a given power, a larger sample size is required for two-tailed tests (Flurkey et al. 2007). If there is multiple testing, the probability of Type I error should be adjusted (Flurkey 2007).⁴ For example, the Bonferroni-adjustment proposes to divide 0.05 through the number of tests (cf., e.g. Holm 1979). After adjusting for multiple testing the standardized Z-score is obtained from the normal distribution. The difference of $(\mu_1 - \mu_2)$ is the treatment effect the investigator wants to detect. Since this term is squared, even small differences influence the sample size considerably. The same holds for the variance of the population σ^2 . For continuous dependent variables the population standard deviation is used.

Table 2 Some multipliers for one- and two tailed tests

α (a) β (b)	Two-tailed tests			One-tailed tests		
	0.10 (1.645)	0.05 (1.96)	0.01 (2.576)	0.10 (1.283)	0.05 (1.645)	0.01 (2.326)
0.2 (0.842)	6.2	7.9	11.7	4.5	6.2	10.0
0.1 (1.283)	8.6	10.5	14.9	6.6	8.6	13.0

Note: The table depicts the multipliers (Z-scores) for conventional values of α (one-sided or two-sided) and β . For example, $\alpha = 0.05$ (two-tailed test) equals 1.96. The multipliers of α and β are denoted as a and b , respectively. The calculations of $(\alpha + \beta)^2$ are shown in the table because they are necessary to calculate sample size. For more details see, for example, Snedecor and Cochran (1980:104) or Flurkey et al. (2007:659).

⁴ Experimental economists often test multiple null hypotheses simultaneously (e.g. multiple treatments, subgroups, and outcomes). If researchers do not take account of multiple testing, the evidence against the null is inflated. According to List et al. (2016), the problem is largely ignored in experimental economics studies.

Example: Dictator Experiment

Let us assume that the experimenter would give player A €100. Player A is then asked about the number of units he is willing to transfer to player B. There are two treatments: (1) player A and player B are anonymous and cannot see each other, and (2) player A and player B can see each other. Due to prior knowledge, the investigator believes that there is an inverse relationship between pro-social behavior and anonymity. Thus, he or she considers the amount of sharing to be higher in treatment 2. The investigator wants to detect a difference of €10, which he or she supposes to be meaningful. Previous studies with similar subjects have shown that the variability is approximately normally distributed with a standard deviation of 25. Moreover, α and β are assumed to be 0.05 and 0.2, respectively.

It is assumed that the sum of money transferred to player B is higher in treatment 2. Thus, we use a one-tailed test. As can be seen in Table 2, using the multipliers of $\alpha = 0.05$ and $\beta = 0.2$ gives $(\alpha + \beta)^2 = (1.645 + 0.842)^2 = 6.2$. The variance σ^2 is 625 and the minimal relevant difference $(\mu_1 - \mu_2)$ equals 10. Putting these values into the formula above gives $N_{1,2} = \frac{2*6.2*25^2}{(10)^2} = 77.5$. To answer the research question, the required sample size for each group is 78. Doubling the minimal relevant difference from 10 to 20 reduces c.p. the required sample size to $N_{1,2} = 19.375 = 20$. The investigator needs only 1/4 of the subjects (differences are due to rounding). Halving the standard deviation from 25 to 12.5 leads to the same conclusion.

2.2.2 Binary outcome variable

If the outcome measure is binary, the required sample size for *each* group reads as follows:

$$N_{1,2} = \frac{(a + b)^2(f_1 * \bar{f}_1 + f_2 * \bar{f}_2)}{\mu^2}$$

Similar the continuous outcome scenario described earlier, the standardized Z-scores for α ($= a$ in the formula above) and β ($= b$ in the formula above) have to be considered to calculate the sample size. Two groups and outcomes have to be distinguished: f_1 ($\bar{f}_1$) denotes the fraction of people with (without) outcome 1 in treatment group 1, whereas f_2 ($\bar{f}_2$) denotes the fraction of people with (without) outcome 1 in treatment group 2. The difference the investigator supposes to be meaningful and wants to detect is captured by μ . This term is squared and has an important influence on the sample size calculation. For binary dependent variables event rates are used.

Example: One-shot prisoner's dilemma

The investigator conducts a one-shot prisoner's dilemma game. He or she wants to examine whether information about the opponent influences the decision to cooperate or defect. In treatment 1 there is no information about the opponent, whereas in treatment 2 it is known that the opponent raised in rural areas. Suppose that the investigator knows from earlier studies that 30% of the subjects cooperate without having any information about their opponent, whereas 40% of the subjects tend to cooperate if the opponent raised in rural areas. The investigator deems a 20% difference in cooperation to be meaningful. Moreover, α and β are assumed to be 0.05 and 0.2, respectively. The investigator wants to run a two-sided test.

The fraction of subjects who cooperate is captured by f_1 and f_2 , and non-cooperation is labeled as $\bar{f}_1$ and $\bar{f}_2$, with the lower superscript numbers indicating the number of the treatment group: $f_1 * \bar{f}_1 + f_2 * \bar{f}_2 = 0.4(1 - 0.4) + 0.3(1 - 0.3) = 0.45$. The minimal relevant difference μ^2 is $0.2^2 = 0.04$. Assuming a two-sided test gives $(a + b)^2 = (1.96 + 0.842)^2 = 7.9$. The formula to calculate the required sample size for each group gives $N_{1,2} = \frac{7.9 * 0.45}{0.04} = 88.875$. To answer the research question 89 subjects should be recruited for each of the groups. If we would halve the difference deemed to be meaningful to 0.1, the required sample size would be quadrupled. A total of 356 subjects per group (the unrounded number is 355.5) would be required to answer the research question.

3 Practice of sample size analysis in economic experiments: a literature review

3.1 Data basis and search strategy

To obtain an approximate picture of the current practice on how often power analysis/sample size calculations are performed in economic experiments, I review publications from 5 journals covering the period 2000–2018. These are two top-ranked economic journals (*Quarterly Journals of Economics* and *American Economic Review*⁵), the leading field journals in experimental economics (*Experimental Economics*) and the behavioral sciences (*Journal of Economic Behavior and Organization*), and a leading field journal in environmental economics (*Environmental and Resource Economics*).

⁵ The Papers and Proceedings sections (which are based on the AEA Annual Meetings) are excluded from consideration.

The first exercise is to identify contributions in which an experiment has been carried out by the researchers. To put it differently, it is about studies in which an experimenter actively generates data and investigates revealed behavior. After clicking on the respective journal contributions, I took a brief look at the title and abstract to gain an initial impression of the paper. Then I searched the papers for keywords. I started with “experiment.” If the authors do not explicitly refer to the term “experiment,” I extended the search and looked for “game.” Furthermore, it was examined if the authors speak of “subjects” / “participants” / “recruit,” and “design” and, in case they do, if the context can be interpreted as an experimental study. If all these search activities were not successful, I concluded that the study at hand is not experimental.

In addition to the keywords search, it was necessary to take a closer look at various sections of the papers. To establish that an experiment had actually been carried out⁶, I took an in-depth look at the experimental design (“design”), experimental subjects (“participants”, “people”, “subjects”) and data section (“experimental findings”).⁷ However, using the noun “experiment” cannot automatically be equated with the presence of an experiment. In particular, the following nouns used in the papers are different from experiments I was looking for: *thought experiment*, *social experiment*, *Monte Carlo experiment*, *historical experiment*, *policy experiment*, *numerical experiment*, *hypothetical experiment*, *computational experiment*, *quasi experiment*, *natural experiment*⁸, *simulation experiment*, and *discrete choice experiments*⁹. In addition, several authors used the term “experiment” as a verb, which, however, was often not in line with the research goal of this article and thus dropped from the analysis: *experiment with numbers*, *experiment with certain kind of regression type*, *experiment with alternative econometric specifications*, *experimented with a number of other predictors*, *experimented with bootstrapping the standard errors*, *experimented with quadratic weighting functions*, *experimented with alternative age cutoffs*, *experimented with other historical definitions*, *experimented with lagged wages*, and *experimented with several other controls*. Moreover, I excluded papers that summarize existing studies in systematic reviews and meta-analyses.

⁶ I excluded papers that adopted the data of other authors because there was no influence on the number of subjects. Moreover, the scholars did not generate their own data with own experiments.

⁷ For example, Benati and Surico (2009) have a section in their paper “VAR Analysis and the Great Moderation” labeled as “The Experiment Design,” which is however far away from what experimental economists would describe as an experiment.

⁸ Natural experiments are observational studies because there is no active experimenter (e.g. Dunning 2013).

⁹ Discrete choice experiments are statements in surveys.

After identifying experimental papers, I had to explore in which contributions power analyses and sample size calculations were performed. For this purpose, I searched for the keywords “power” and “sample size” in the papers. However, the authors could have carried out such analyses in earlier versions of their papers without mentioning it in the journal publication. Thus, I screened for *register*, *registration*, *appendix*, and *supplementary*. In case it was available, I looked at it. If the authors of a study calculated the power (with respect to the number of subjects) or calculated the optimal sample size of the experimental study, then the paper was counted as “*experimental paper with power/sample size calculations*” (cf., Table 3). I did not count papers in which the authors equated more subjects with more power.

3.2 Evidence of sample size and power calculations

The experimental papers usually mention the number of the recruited subjects. But do the authors also justify the number of experimental subjects? In Table 3, I summarize the fraction of experimental papers that use power analysis / sample size calculations in the period 2000–2018 in the 5 economic journals (Quarterly Journals of Economics, American Economic Review, Experimental Economics, Journal of Economic Behavior and Organization, and Environmental and Resource Economics¹⁰) under consideration. Besides the overall picture (2000-2018), I present data for the periods 2000–2009 and 2010–2018, and provide numbers for the very recent years of the period under consideration (i.e. 2015–2016 and 2017–2018). The purpose of these demarcations is to give a rough idea of whether the relevance of the number of subjects has changed over time. Only a small fraction of the experimental papers performs power analysis / sample size calculations. During the period 2000–2018 the share is particularly low for EE, JEBO, and ERE. The topic seems to be of greater relevance for the leading economic journals, but the level is also quite low here. However, the table indicates that (with the exception of ERE) the fraction of experimental papers performing power calculations is higher for the period 2010–2018 than in 2000-2009.¹¹ The changes are relatively high with AER and especially with QJE. The very last years of consideration (2015–2016 & 2017–2018) indicate that power analyses / sample size calculations in general are of little relevance at present, which also applies to the leading field

¹⁰ Confer the Appendix for some typical statements made in the journal articles about power and sample size calculations.

¹¹ Power analyses seem to play a subordinate role in this journal. As we can see from the table, there is no increase in power analysis over time. The number of power analyses seems to be somewhat erratically distributed over time (2007, 2009, and 2013).

journal for experimental economics. However, the relevance seems to be higher for leading economic journals than for field journals at first glance.

Table 3 **Fraction of experimental papers with power / sample size calculations in the five selected journals (numbers in percent)^(a)**

Period		Journal	QJE	AER	EE	JEBO	ERE
Overall		2000–2018	8.64	5.38	1.68	2.64	2.67
Decade		2000–2009	0.00	2.10	0.00	1.44	4.08
		2010–2018	13.46	7.81	2.77	3.10	1.58
Recent years		2015–2016	11.11	11.11	1.38	6.15	0.00
		2017–2018	16.66	11.11	5.19	4.61	0.00

^(a) QJE (Quarterly Journals of Economics), AER (American Economic Review), EE (Experimental Economics), JEBO (Journal of Economic Behavior and Organization), ERE (Environmental and Resource Economics)

What did the authors do with the insights obtained from their calculations? Several papers only mention the results of their calculations. Typical statements are “Conventional power calculation [...] for the data reported in Fig. 4b yields the power of 0.357” (Jacquemet et al. 2018) or “A sample size analysis indicates that 1411, 965 and 38407 observations would be needed to obtain a significant result for [...]” (Dreber et al. 2011). In some papers, the findings are used to recruit a particular number of subjects. For example, Deming et al. (2016) argue that their original data collection provided insufficient power to carry out the planned analyses. As a consequence, further data were collected and analyzed once the collection was completed. Higuchi et al. (2015) gave monetary restrictions as a reason not to follow their calculations. They instead conducted follow-up studies. Filiz-Ozbay et al. (2018) argue in a similar way. They were requested by a referee to conduct an ex post power analysis and mention that it helps to determine whether additional sessions have to be conducted or not. Power analyses may influence the researcher in his or her data analysis. For example, Giné et al. (2012) justify their number of groups with power calculations. In addition, power analyses play a role in interpreting and discussing findings. For example, Bettinger et al. (2012) use the term “underpowered and imprecise” to describe subgroup estimates in their study. Persson (2018) argues “One limitation of the experiment is the relatively low power of the tests.” Other researchers, such as Schechter (2007), note that some of their

research questions, which cannot be analyzed due to insufficient power, may be of interest for future investigations. Irrespective of individual studies, power analyses have a teaching effect: the readership of papers gets a rough idea of the number of subjects, which may serve as reference for further studies.

3.3 Discussion

3.3.1 Don't experimental economists care at all about power analyses and sample calculations?

As we have seen above, power analyses are rarely carried out in experiments. However, it cannot be concluded that the topic is perceived to be entirely irrelevant for the authors. Besides performing calculations, several authors discuss issues that are related sample size and power calculations. I want to briefly describe typical categories and give a few examples for each category.

(1) Equating power and sample size

A number of experiments equate a small (high) sample size with a low (high) power. For example, Barham et al. (2014) argue: “Our sample size is rather small, so we do not have a lot of power. But we find some interesting suggestive evidence.” Another example is Baird et al. (2011): “As the sample size for this group is quite small [...], dividing the treatment group into a CCT and a UCT group would yield an experiment with low statistical power.” Quite similar arguments can be found with Resnick et al. (2006): “First, it would reduce the sample size quite a bit, reducing the statistical power.” and Callen et al. (2014): “Given our small sample size, we have limited power to test for effects by subgroups within the violence-affected sample.”

(2) Equating power and statistical significance

Some authors argue that the absence of statistical significance is due to a lack of statistical power. For instance, Gjedrem and Rege (2017) argue: “We can see in Column 1 that, due to power issues there are no significant differences [...]” Another example is Aguiar-Conraria et al. (2016): “Part of the explanation for this lack of statistical significance may be attributed to the low power of the statistical effects [...]”. According to Dickhaut et al. (2013): “[...] it is important to show that differences if present could in principle be discovered with our sample size and that the inability to reject the null hypothesis is not due to insufficient statistical power.” In a similar way, Roux and Thöni (2015) claim: “Insignificant results must, of course, be interpreted with great caution. If the empirical setup does not offer sufficient power to detect the effect the lack of significant results

cannot be interpreted in a meaningful way.” Overall it seems that power analyses may trigger to think in terms of statistically significant vs. statistically non-significant.¹²

(3) Experimental design and statistical power

A limited fraction of the authors motivate their experimental design through statistical power. Most notably, they argue in favor of a within-subject design. For example, Fehr and Gächter (2000) remark: “This has the advantage that, in addition to across-subject comparisons, we can make within-subject comparisons of cooperation levels, which have much more statistical power.” Similarly, Cummings et al. (2009) write: “We adopted a within subject design for several reasons. It increases statistical power since the characteristics of the participants are held constant while the decision treatment is altered.” Riyanto and Jonathan (2018) made another point: “Since the sample size was smaller in this experiment, we opted for more periods in order to increase statistical power.”

(4) Manipulation of variables and statistical power

To increase the statistical power, some authors manipulate variables. For example, Casari et al. (2007) “[...] pool groups (a) and (b) above to create a high bankruptcy group. The two groups have very similar (high) bankruptcy rates and pooling yields essentially the same results as the control group alone, but with substantially more statistical power (given that combining the two groups essentially doubles the sample size).” Another example is Cadsby et al. (2010): “[...] we pooled the data from all three treatments together to allow for greater statistical power in analyzing the other possible manifestations of a gender effect.” Similarly, Morgan et al. (2018) argue: “To gain sufficient power to identify screen location effects at longer distances, we pool screen locations of distance 3 or greater.”

(5) Further points

A few researchers described the researchers’ budget as a limiting factor for the number of subjects included in the study. Ziegelmeyer et al. (2012), for example, argue that “[...] recruiting more subjects in Experiment 5 was beyond our financial budget.” Other authors refer to comparable

¹² In contrast, p-values should be defined as “continuous measure of the strength of evidence against the null hypothesis” (e.g. Amrhein et al. 2017). The frequently made dichotomous interpretation of the p-value (based on arbitrary significance thresholds) as well as the use of the term statistically significant is somewhat controversial (Hirschauer et al. 2019; Wasserstein et al. 2019).

studies as a benchmark. For example, Candelo et al. (2018) note: “This sample size is comparable to previous work in the field using social networks (e.g. Karlan et al. 2009).”

3.3.2 Publications in top-ranked journals: more awareness for power analyses?

The literature review indicates that top-ranked journals publish more experimental papers with power analyses / sample size calculations than field journals do. Can we therefore infer that top-ranked journals attach greater importance to power analyses (e.g. due to stricter selection criteria for accepting contributions)? Not necessarily. Another explanation is product differentiation. Product differentiation arises, for example, when the ratio of laboratory tests to field tests is systematically different among journals. The Quarterly Journal of Economics puts great emphasis on field experiments. The American Economic Review publishes both laboratory and field experiments. The latter are of high importance. The publications of the field journals *Experimental Economics*, *Journal of Economic Behavior and Organization*, and *Environmental and Resource Economics* deal primarily with laboratory experiments. In field experiments, researchers seem to be more inclined to deal with sample size calculation. The reason may be that field experiments are often relatively expensive compared to laboratory experiments. Filiz-Ozbay et al. (2018) provide another argument in favor of power calculations in field experiments: “Power calculations [...] are less frequent in advance in laboratory experiments than in field experiments because it is much easier to run additional lab sessions.”

4 Conclusion

In economic experiments, power analyses and sample size calculations have relatively rarely been carried out. This is astonishing because economics itself is about dealing with scarce resources and optimization to coordinate them is so widespread. The literature review has shown that power analyses are more likely to be published in the top-ranked journals, which was partly associated with the higher priority attached to field experiments. However, there seems to be more attention given to power analysis in the last few years, although the level is still quite low. Power analyses can assist the researcher not to over-interpret findings of their own studies as well as studies from the literature. For example, Kahneman admitted that he “[...] placed too much faith in underpowered studies.”¹³ He is probably no exception, since people tend to think in terms of stories

¹³ <https://retractionwatch.com/2017/02/20/placed-much-faith-underpowered-studies-nobel-prize-winner-admits-mistakes/>

rather than probabilities. But it runs the risk to misdirect new studies because of a biases literature background and thus hypotheses.

Filiz-Ozbay et al. (2018) argue that researchers put much more effort in sample size calculations in field experiments because lab experiments could be replicated easily. Do we need sample size calculations in lab experiments or is it sufficient to aggregate evidence through replication? In some cases extensive replications can be observed. For example, the Ultimatum bargaining game, introduced by Güth et al. (1982), has been replicated many times. Another example is the experimental work of Kahneman and Tversky. There are many robustness checks of their initial studies, which makes it easy to conduct systematic reviews and meta analyses—a starting point for further research. However, what kind of experiments are likely to be replicated? Probably the ones that are easy to replicate and do not cause much costs. Therefore, replication cannot fully serve as a substitute for sample size calculation. Things are even worse because the overall incentives to replicate experiments are relatively low (Galiani et al. 2017). For example, Camerer et al. (2016) provide evidence for the insufficient reproducibility of studies in the empirical social sciences. They attempted to replicate 18 economic experiments, published in the American Economic Review and The Quarterly Journal of Economics in the 2011–2014 journals. The authors find “a significant effect in the same direction as in the original study for 11 replications (61%); on average, the replicated effect size is 66% of the original.”

If there is no solid data basis (e.g. in case of an entirely new experiment), sample size calculations require the researcher to refer to expert estimates, carry out pilot studies or at least to argue about standardized effect sizes. Of course, studies without clear pre-specified hypotheses should be identified as such. For example, Berry (2016) proposes to explicitly declare them as exploratory and to point out that there is no claim of generalizability. Cooper et al. (1999) distinguish between hypotheses that were specified before the data was collected (*ex ante hypotheses*) and hypotheses that were discovered after seeing the data (*ex post hypotheses*). The former can be classified as confirmatory and the latter as exploratory analysis.

There is at least one indirect, but crucial advantage attached to sample size calculations. Thinking about the required sample size necessitates to make precise assumptions about both the null and the alternative hypothesis. While the former should “capture the status quo line of thinking” (Brandon and List 2015), it is often unrealistically assumed to be zero in economics regression

analysis. But this disregards prior knowledge and former research findings.¹⁴ Besides methodical similarities (e.g. using RCTs) sample size calculations are standard practice in clinical, but not economic trials. As described above there are some good arguments to conduct them in experimental economics research, too. One of the main arguments is the objectivity of research. If other scientists read a paper, it is important for them to be informed about issues like power and the optimal sample size. Published research is at least input for new studies and we need to know how to evaluate their findings.

Acknowledgment

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – 388911356.

References

- Aguiar-Conraria, L., Magalhães, P.C., Vanberg, C.A. (2016): Experimental evidence that quorum rules discourage turnout and promote election boycotts. *Experimental Economics* 19(4): 886–909.
- Amrhein, V., Korner-Nievergelt, F., Roth, T. (2017): The earth is flat ($p > 0.05$): significance thresholds and the crisis of unreplicable research. *PeerJ* 5: e3544. doi:10.7717/peerj.3544.
- Baird, S., McIntosh, C., Özler, B. (2011): Cash or Condition? Evidence from a Cash Transfer Experiment. *The Quarterly Journal of Economics* 126(4): 1709–1753.
- Barham, B.L., Chavas, J.-P., Fitz, D., Salas, V.R., Schechter, L. (2014): The roles of risk and ambiguity in technology adoption. *Journal of Economic Behavior & Organization* 97: 204–218.
- Benati, L., Surico, P. (2009): VAR Analysis and the Great Moderation. *The American Economic Review* 99(4): 1636–1652.
- Benjamin, D.J. et al. (2017): Redefine Statistical Significance. *Human Nature Behavior*. <https://www.nature.com/articles/s41562-017-0189-z>
- Berry, D.A. (2016): P-Values Are Not What They’re Cracked Up to Be. Online Discussion: ASA Statement on Statistical Significance and P-values. *The American Statistician* 70(2): 1–2.
- Bettinger, E.P., Long, B.T., Oreopoulos, P., Sanbonmatsu, L. (2012): The Role of Application Assistance and Information in College Decisions: Results from the H&R Block Fafsa Experiment. *The Quarterly Journal of Economics* 127(3): 1205–1242.
- Brandon, A., List, J.A. (2015): Markets for replication. *Proceedings of the National Academy of Sciences of the United States of America* 112(50): 15267–15268.

¹⁴ To be fair, clinical investigators have a comparative advantage to economists. Economists do have databases to search for literature (e.g. Web of Science, RePEc), but they are not in that luxurious situation to have access to high level established standards. Clinical investigators can use PubMed to have easy and efficient access to biomedical literature.

- Cadsby, C.B.; Servátka, M., Song, F. (2010): Gender and generosity: does degree of anonymity or group gender composition matter? *Experimental Economics* 13(3): 299–308.
- Callen, M., Isaqzadeh, M., Long, J.D., Sprenger, C. (2014): Violence and Risk Preference: Experimental Evidence from Afghanistan. *The American Economic Review* 104(1): 123–148.
- Camerer, C. et al. (2016): Evaluating replicability of laboratory experiments in economics. *Science* 351: 1433–1436.
- Candelo, N., Croson, R.T.A., Eckel, C. (2018): Transmission of information within transnational social networks: a field experiment. *Experimental Economics* 21(4): 905–923.
- Casari, M., Ham, J.C., Kagel, J.H. (2007): Selection Bias, Demographic Effects, and Ability Effects in Common Value Auction Experiments. *American Economic Review* 97(4): 1278–1304.
- Christensen, E. (2007): Methodology of superiority vs. equivalence trials and non-inferiority trials. *Journal of Hepatology* 46(5): 947–954.
- Cooper, D.J., Kagel, J.H., Lo, W., Gu, L.Q. (1999): Gaming Against Managers in Incentive Systems: Experiments with Chinese Students and Chinese Managers. *American Economic Review* 89(4): 781–804.
- Cummings, R.G., Martinez-Vazquez, J., McKee, M., Torgler, B. (2009): Tax morale affects tax compliance: Evidence from surveys and an artefactual field experiment. *Journal of Economic Behavior & Organization* 70(3): 447–457.
- Deming, D.J., Yuchtman, N., Abulafi, A., Goldin, C., Katz, L.F. (2016): The Value of Postsecondary Credentials in the Labor Market: An Experimental Study. *American Economic Review* 106(3): 778–806.
- Dickhaut, J., Houser, D., Aimone, J.A., Tila, D., Johnson, C. (2013): High stakes behavior with low payoffs: Inducing preferences with Holt–Laury gambles. *Journal of Economic Behavior & Organization* 94: 183–189.
- Dreber, A., von Essen, E., Ranehill, E. (2011): Outrunning the gender gap—boys and girls compete equally. *Experimental Economics* 14(4): 567–582.
- Dunning, T. (2012): *Natural Experiments in the Social Sciences: A Design-based Approach*. Cambridge: Cambridge University Press.
- Fehr, E., Gächter, S. (2000): Cooperation and Punishment in Public Goods Experiments. *American Economic Review* 90(4): 980–994.
- Filiz-Ozbay, E., Ham, J.C., Kagel, J.H., Ozbay, E.Y. (2018): The role of cognitive ability and personality traits for men and women in gift exchange outcomes. *Experimental Economics* 21(3): 650–672.
- Flurkey, K., Curren, J.M., Harrison, D.E. (2007): Mouse Models in Aging Research. In: Fox et al. *The Mouse in Biomedical Research* 3: 637–672.
- Galiani, S., Gertler, P., Romero, M. (2017): Incentives for replication in economics. Tech. rept. National Bureau of Economic Research. <https://www.nber.org/papers/w23576.pdf>
- Giné, X., Goldberg, J., Yang, D. (2012): Credit Market Consequences of Improved Personal Identification: Field Experimental Evidence from Malawi. *The American Economic Review* 102(6), 2923–2954.
- Gjedrem, W.G., Rege, M. (2017): The effect of less autonomy on performance in retail: Evidence from a quasi-natural field experiment. *Journal of Economic Behavior & Organization* 136: 76–90.
- Guala, F. (2005): *The Methodology of Experimental Economics*. Cambridge University Press, Cambridge.
- Güth, W., Schmittberger, R., Schwarze, B. (1982): An experimental-analysis of ultimatum bargaining. *Journal of Economic Behavior & Organization* 3(4): 367–388.

- Hickey, G.L., Grant, S.W., Dunning, J., Siepe, M. (2018): Statistical primer: sample size and power calculations—why, when and how? *European Journal of Cardio-Thoracic Surgery* 54(1): 4–9.
- Higuchi, Y., Nam, V.H., Sonobe, T. (2015): Sustained impacts of Kaizen training. *Journal of Economic Behavior & Organization* 120: 189–206.
- Hirschauer, N., Grüner, S., Mußhoff, O., Becker, C. (2019): Twenty steps towards an adequate inferential interpretation of p-values in econometrics. *Journal of Economics and Statistics* 239(4): 703–721.
- Holm, S. (1979): A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* 6(2): 65–70.
- Huber, C. (2019): https://blog.stata.com/2019/01/10/calculating-power-using-monte-carlo-simulations-part-1-the-basics/#disqus_thread
- Jacquemet, N., Luchini, S., Shogren, J.F., Zylbersztejn, A. (2018): Coordination with communication under oath. *Experimental Economics* 21(3): 627–649.
- Julious, S.A. (2004): Tutorial in Biostatistics. Sample sizes for clinical trials with Normal data. *Statistics in Medicine* 23(12): 1921–1986.
- Julious, S.A., Campbell M.J. (2012): Tutorial in biostatistics: sample sizes for parallel group clinical trials with binary data. *Statistics in Medicine* 31(24): 2904–2936.
- Levitt, S.D., List, J.A. (2007): What Do Laboratory Experiments Measuring Social Preferences Reveal about the Real World? *The Journal of Economic Perspectives* 21(2): 153–174.
- List, J.A., Sadoff, S. & Wagner, M. (2011): So you want to run an experiment, now what? Some simple rules of thumb for optimal experimental design. *Experimental Economics* 14: 439–457.
- List, J.A., Shaikh, A.M., Xu, Y. (2016): Multiple Hypothesis Testing in Experimental Economics. NBER Working Paper No. 21875.
- Maniadis, Z., Tufano, F., List, J. (2014): One Swallow Doesn't Make a Summer: New Evidence on Anchoring Effects. *American Economic Review* 104(1): 277–290.
- Morgan, J., Ong, D., Zhong, Z.Z. (2018): Location still matters: Evidence from an online shopping field experiment. *Journal of Economic Behavior & Organization* 146: 43–54.
- Nguyen, T.-L., Landais, P. (2017): Randomized controlled trials: significant results—fragile, though. *Kidney International* 92(6): 1319–1320.
- Noordzij, M., Tripepi, G., Dekker, F.W., Zoccali, C., Tanck, M.W., Jager, K.J. (2010): Sample size calculations: Basic principles and common pitfalls. *Nephrol Dial Transplant* 25: 1388–1393.
- Persson, E. (2018): Testing the impact of frustration and anger when responsibility is low. *Journal of Economic Behavior & Organization* 145: 435–448.
- Resnick, P., Zeckhauser, R., Swanson, J., Lockwood, K. (2006): The value of reputation on eBay: A controlled experiment. *Experimental Economics* 9(2): 79–101.
- Riyanto, Y.E., Jonathan, Y.X.W. (2018): Directed trust and trustworthiness in a social network: An experimental investigation. *Journal of Economic Behavior & Organization* 151: 234–253.
- Roe, B.E., Just, D.R. (2009): Internal and External Validity in Economics Research: Tradeoffs between Experiments, Field Experiments, Natural Experiments, and Field Data. *American Journal of Agricultural Economics* 91(5): 1266–1271.
- Roux, C., Thöni, C. (2015): Do control questions influence behavior in experiments? *Experimental Economics* 18(2): 185–194.
- Schechter, L. (2007): Traditional trust measurement and the risk confound: An experiment in rural Paraguay. *Journal of Economic Behavior & Organization* 62(2): 272–292.

- Senn, S.J. (2002): Power is indeed irrelevant in interpreting completed studies. *British Medical Journal* 325(7375): 1304.
- Snedecor, G.W., Cochran, G.W. (1980): *Statistical methods*. The Iowa State University Press: Ames.
- Wasserstein, R.L., Schirm, A.L., Lazar, N.A. (2019): Moving to a World Beyond “ $p < 0.05$ ”. *The American Statistician* 73: 1–19.
- Whitley, E., Ball, J. (2002): Statistics review 4: sample size calculations. *Critical Care* 6(334): 335–341.
- Zelmer, J. (2003): Linear Public Goods Experiments: A Meta-Analysis. *Experimental Economics* 6(3): 299–310.
- Zieglmeier, A., Schmelz, K., Ploner, M. (2012): Hidden costs of control: four repetitions and an extension. *Experimental Economics* 15(2): 323–340.

Appendix: Power / sample size calculations

Appendix 1 Power / sample size calculations in the QJE

Authors	Type of experiment	Evidence
Bertrand et al. (2010)	field	“Standard power calculations show that identifying a content feature effect [...] would require over 300,000 observations.”
Fryer (2011)	field	“An important potential limitation in our set of field experiments is that they were constructed to detect effects of 0.15 standard deviations or more with 80% power.”
Bettinger et al. (2012)	field	“With a control mean of 0.2, the sample size gives us about 80% statistical power to detect a 3 percentage point difference in FAFSA submission rates at the 5% significance level.”
Kroft et al. (2013)	field	“Our power calculations called for 12,000 résumé submissions.”
Lewis & Rao (2015)	field	“The “experiment multiplier” tells us how much larger an experiment would have to be in terms of new, independent individuals to achieve adequate power, which we define as an expected t-statistic of 3, as this produces power of 91% with a one-sided test size of 5%.”
Atkin et al. (2017)	field	“We chose the number of firms to be treated based on calculations for statistical power to pick up spillover effects.”
de Ree et al. (2018)	field	“These zero effects are precisely estimated; the small standard errors of 0.025σ provide us adequate power to detect effects as low as 0.05σ at the 5% level.”

References

- Bertrand, M., Karlan, D., Mullainathan, S., Shafir, E., Zinman, J. (2010): What’s Advertising Content Worth? Evidence from a Consumer Credit Marketing Field Experiment. *The Quarterly Journal of Economics* 125(1): 263–306.
- Fryer, R.G. (2011): Financial Incentives and Student Achievement: Evidence from Randomized Trials. *The Quarterly Journal of Economics* 126(4): 1755–1798.
- Bettinger, E.P., Long, B.T., Oreopoulos, P., Sanbonmatsu, L. (2012): The Role of Application Assistance and Information in College Decisions: Results from the H&R Block Fafsa Experiment. *The Quarterly Journal of Economics* 127(3): 1205–1242.
- Kroft, K., Lange, F., Notowidigdo, M.J. (2013): Duration Dependence and Labor Market Conditions: Evidence from a Field Experiment. *The Quarterly Journal of Economics* 128(3): 1123–1167.
- Lewis, R.A., Rao, J.M. (2015): The Unfavorable Economics of Measuring the Returns to Advertising. *The Quarterly Journal of Economics* 130(4): 1941–1973.
- Atkin, D., Chaudhry, A., Chaudry, S., Khandelwal, A.K., Verhoogen, E. (2017): Organizational Barriers to Technology Adoption: Evidence from Soccer-Ball Producers in Pakistan. *The Quarterly Journal of Economics* 132(3): 1101–1164.
- de Ree, J., Muralidharan, K., Pradhan, M., Rogers, H. (2017): Double for Nothing? Experimental Evidence on an Unconditional Teacher Salary Increase in Indonesia. *The Quarterly Journal of Economics* 133(2): 993–1039.

Appendix 2 Power / sample size calculations in the AER

Authors	Type of experiment	Evidence
Plott & Zeiler (2007)	lab	“The proportions are not statistically significantly different ($p = 0.31$; power = 0.0675).”
Choi et al. (2007)	lab	“The power of the experiment is very sensitive to the number of observations for each subject. To illustrate this point, we simulated the choices of random subjects [...]”
Caplin et al. (2011)	lab	“[...] for 76,000 simulated subjects with the same number of switches as our subjects but who choose at random—a measure of the power of our test [...]”
Voors et al. (2012)	field	“[...] the probability of a type II error is now less than 0.14 (power is 0.86), which is below the 0.20 (above 0.80) threshold routinely assumed in empirical analysis.” & “We used G*Power3 software to conduct the power tests; see Faul et al. (2007).”
Giné et al. (2012)	field	“Given the number of farmers in our study, it was infeasible to implement this design because power calculations suggested we could have at best two groups.”
Maniadis et al. (2014)	lab	“Two other considerations—the statistical power of the test and the fraction of tested hypotheses that are true associations—are key factors to consider when making appropriate inference.”
Eriksson & Rooth (2014)	field	“The fractions were chosen to ensure that we should be able to estimate any economically significant effects of these attributes (i.e., based on power calculations).”
Bhargava & Manoli (2015)	field	“Treatments were randomized with equal sample weights with three exceptions: The control notice was over-sampled ($\times 4$) to heighten the statistical power for pair-wise comparisons; the benefit display notices were over-sampled ($\times 3$) to power tests of differentiation across listed benefit amounts [...]”
Carvalho et al (2016)	lab	“A power analysis of Study 2 indicates that with a sample size of 2,700 participants and a 98 percent compliance rate (i.e., the fraction of the before-payday group who completed the follow-up survey before payday) we can detect a <i>before-versus-after-payday difference</i> of 0.11 of a standard deviation in a two-sided test with power 0.8 and a significance level of 5 percent.”
Deming et al. (2016)	field	“We sent out health resumes from April to July 2014 as well. But the much smaller number of health job postings (Observations = 1,460 through July 2014) did not provide us with adequate statistical power.”
Blattman et al. (2017)	field	“Based on the pilot, we estimated that the Minimum Detectable Effect for the full 1,000 (with a quarter for each treatment) would be a 0.12 standard deviation change in a standardized dependent variable for a two-tail hypothesis test with statistical significance of 0.05, statistical power of 0.80, an intra-cluster correlation of 0.25, and the proportion of individual variance explained by covariates as 0.10.” (→ Online-Appendix)
Friebel et al. (2017)	field	“Power calculations reveal that the sample size is more than sufficient: on the basis of 27 months of observations pretreatment (January 2012 to March 2014) and three months of observations post treatment (April to June 2014), we would need 70 shops in each group to detect a 3 percent treatment effect at a 5 percent significance level with the probability 0.9.”

References

- Plott, C., Zeiler, K. (2007): Exchange Asymmetries Incorrectly Interpreted as Evidence of Endowment Effect Theory and Prospect Theory? *The American Economic Review* 97(4): 1449–1466.
- Choi, S., Fisman, R., Gale, D., Kariv, S. (2007): Consistency and Heterogeneity of Individual Behavior under Uncertainty. *American Economic Review* 97(5): 1921–1938.
- Caplin, A., Dean, M., Martin, D. (2011): Search and Satisficing. *The American Economic Review* 101(7) 2899–2922.
- Voors, M.J., Nillesen, E.E.M., Verwimp, P., Bulte, E.H., Lensink, R., Van Soest, D. (2012). Violent Conflict and Behavior: A Field Experiment in Burundi. *The American Economic Review* 102(2) 941–964.

- Giné, X., Goldberg, J., Yang, D. (2012): Credit Market Consequences of Improved Personal Identification: Field Experimental Evidence from Malawi. *The American Economic Review* 102(6), 2923–2954.
- Maniadis, Z., Tufano, F., & List, J. (2014). One Swallow Doesn't Make a Summer: New Evidence on Anchoring Effects. *The American Economic Review* 104(1): 277–290.
- Eriksson, S., Rooth, D.-O. (2014): Do Employers Use Unemployment as a Sorting Criterion When Hiring? Evidence from a Field Experiment. *The American Economic Review* 104(3): 1014–1039.
- Bhargava, S., Manoli, D. (2015): Psychological Frictions and the Incomplete Take-Up of Social Benefits: Evidence from an IRS Field Experiment. *American Economic Review* 105(11): 3489–3529.
- Carvalho, L.S., Meier, S., Wang, S.W. (2016): Poverty and Economic Decision-Making: Evidence from Changes in Financial Resources at Payday. *American Economic Review* 106(2): 260–284.
- Deming, D.J., Yuchtman, N., Abulafi, A., Goldin, C., Katz, L.F. (2016): The Value of Postsecondary Credentials in the Labor Market: An Experimental Study. *American Economic Review* 106(3): 778–806.
- Blattman, C., Jamison, J.C., Sheridan, M. (2017): Reducing Crime and Violence: Experimental Evidence from Cognitive Behavioral Therapy in Liberia. *American Economic Review* 107(4): 1165–1206. [Online Appendix]
- Friebe, G., Heinz, M., Krueger, M., Zubanov, N. (2017): Team Incentives and Performance: Evidence from a Retail Chain. *American Economic Review* 107(8): 2168–2203.

Appendix 3 Power / sample size calculations in EE

Authors	Type of experiment	Evidence
Dreber et al. (2011)	field	“A sample size analysis indicates that 1411, 965 and 38407 observations would be needed to obtain a significant result for the performance change in running, jumping and dancing respectively. The basis for the power calculation is a significance level of 5% and a power of 80%.”
List et al. (2011)	[Theoretical]	“In calculating optimal sample sizes an experimenter must consider three key elements: (1) the significance level, (2) the power of the subsequent hypothesis test, and (3) the minimum detectable effect size.” & “A simple rule of thumb to maximize power given a fixed experimental budget naturally follows: the ratio of the sample sizes is equal to the ratio of the standard deviations of outcomes.”
Dreber et al. (2014)	lab	“A sample size analysis indicates that 2037 observations would be needed to obtain a significant result for the gender gap in competition choice in word search. The basis for the power calculation is a significance level of 5 % and a power of 80 %.”
Stoop, J. (2014)	lab & field	“Because the main result is based on the non-rejection of the null hypothesis, power analyses were carried out. A priori, a sample of 80 allows an effect size of at least 40 percent to be detected with a significance level of 5 percent and a power of 80 percent (Faul et al. 2007). The differences in behavior reported in Fig. 1 are so small that even if they were statistically significant, they would be economically insignificant. The StuLab and CitLab experiments show a difference with CitField of approximately 7.5 percent points. To make such a difference statistically significant (with a significance level of 5 percent and a power of 80 percent), the number of observations in each experiment would have to be 577.”
Jordan et al. (2016)	lab	“We also conduct a power analysis to assess the smallest effects of our endowment and strategy method manipulations that we could have detected with 80% probability in Experiment 1.”
Andersson et al. (2017)	lab	“To estimate the statistical power and sample size needed for the main study, we used the observed standard deviation of 4.85 from the final pilot experiment with $n = 44$ (i.e., the standard deviation of the average donation across the different charities). We wanted to have a sufficient sample size to be well powered to detect a medium-sized effect (i.e., Cohen's $d = 0.5$; Cohen 1992) when testing Hypothesis 2 (the test in the sample with a universalism value above the median). ¹⁴ We decided to include approximately 300 subjects in total, which implies a sample size of 150 for testing Hypothesis 2. This provides us with 86 % power to detect a medium-sized effect for Hypothesis 2; for Hypothesis 1, where we include the total sample, the power is 99 % of detecting a medium-sized effect (but if the universalism prime only affects donations in

		individuals with high universalism, this would decrease the expected effect size and consequently the power of the test of Hypothesis 1).”
Koppel et al. (2017)	lab	“A sample-size calculation based on means and standard deviations from a previous study from our lab (Kirchler et al. in press) and with 70% power showed that 50 participants were needed in each condition.”
Filiz-Ozbay et al. (2018)	lab	“A referee raised the question of getting some idea of the increases in sample size that would be needed to determine if some of the Big 5 characteristics that had large, but statistically insignificant values, were in fact significant.” & “If we triple the sample to 138 women, this increases to 0.25 chance of finding that it is significantly different from zero, and if we increase the sample size by a factor of five, it increases to 0.26. These results imply that the insignificant coefficient on openness in column (2) of Table 2 may reflect a lack of power, but that very large increases in the sample size are needed to increase power substantially. One can do power calculations for other variables with large but insignificant coefficients in an analogous manner.”
Jacquemet et al. (2018)	lab	“Conventional power calculation (based on a standard one-sided proportion test supporting the significance of the effect of the oath on the truth-telling ratio, with $\alpha = 5\%$) for the data reported in Fig. 4b yields the power of 0.357. The same test performed on the data reported in Fig. 4a has a power of 0.838. We thank a referee for pointing this out.”

References

- Dreber, A., von Essen, E., Ranehill, E. (2011): Outrunning the gender gap—boys and girls compete equally. *Experimental Economics* 14(4): 567–582.
- List, J.A., Sadoff, S., Wagner, M. (2011): So you want to run an experiment, now what? Some simple rules of thumb for optimal experimental design. *Experimental Economics* 14(4): 439–457.
- Dreber, A., von Essen, E., Ranehill, E. (2014): Gender and competition in adolescence: task matters. *Experimental Economics* 17(1): 154–172.
- Stoop, J. (2014): From the lab to the field: envelopes, dictators and manners. *Experimental Economics* 17(2): 304–313.
- Jordan, J., McAuliffe, K., Rand, D. (2016): The effects of endowment size and strategy method on third party punishment. *Experimental Economics* 19(4): 741–763.
- Andersson, O., Miettinen, T., Hytönen, K., Johannesson, M., Stephan, U. (2017): Subliminal influence on generosity. *Experimental Economics* 20(3): 531–555.
- Koppel, L., Andersson, D., Morrison, I., Posadzy, K., Västfjäll, D., Tinghög, G. (2017): The effect of acute pain on risky and intertemporal choice. *Experimental Economics* 20(4): 878–893.
- Filiz-Ozbay, E., Ham, J.C., Kagel, J.H., Ozbay, E.Y. (2018): The role of cognitive ability and personality traits for men and women in gift exchange outcomes. *Experimental Economics* 21(3): 650–672.
- Jacquemet, N., Luchini, S., Shogren, J.F., Zylbersztejn, A. (2018): Coordination with communication under oath. *Experimental Economics* 21(3): 627–649.

Appendix 4 Power / sample size calculations in JEBO

Authors	Type of experiment	Evidence
Mattei (2000)	lab	“[...] the power of the test is just 0.432.”
Norman et al. (2003)	lab	“For all results except Group 9 the power of the test is >0.8. With our worst case scenario for the coefficient of a modified bubble sort, the power of the test, shown in above Table, is still greater than 0.7 for these cases except Groups 6 and 7. For Group 6 combined with Group 7, the worse case scenario power of the test is >0.9.”
Haruvy & Stahl (2007)	lab	“This likelihood ratio test has asymptotic power of 1, and Monte Carlo simulations show that for our sample size, it has a power of 0.9 or higher [...]”
Schechter (2007)	Lab	“In order to approach a power level of 0.70, we would have needed 1535 observations in the gender regression, 5430 in the education regression, and 570 in the language regression. While the sample size here does not permit such subsample analysis, it leaves an interesting area for further examination.”
Martin & Randal (2008)	field	“Thus, failure to reject a null hypothesis that busyness has no influence on behaviour could be a result of one of the following possibilities: busyness is in fact not relevant, busyness is important but the effects cancel out, or simply a lack of statistical power.”
Ismayilov & Potters (2013)	lab	“If we hypothesize that the deception rate under no disclosure is about 0.44 (based on the two closest treatments in Gneezy, 2005) and that it increases by 50% to 0.66 with disclosure, then the power of our test for the effect of disclosure is almost 80% (two-sided test, no continuity correction).”
Bronchetti et al. (2015)	field	“Our power calculations suggest that the sample size (and the size of each experimental arm) is sufficient to detect a 3.2-percentage point increase in flu vaccination, relative to the baseline vaccination rate for the control group of 9 percent, after correcting for multiple comparisons.”
de Haan & van Veldhuizen (2015)	lab	“Finally, as we show in Appendix A, our null result does not seem to be due to our study having a low power. [...] For example, for the attraction effect task assuming a true effect size similar to the effect observed by Pocheptsova et al. (2009), our power is estimated to be in the 0.8–1 range, which is considered to be high to very high.”
Hong et al. (2015)	field	“According to the rules of thumb provided by List et al. (2011), in our experimental design, our sample size is sufficient to detect about a 1/3 difference of winning probability and about 0.7-standard deviation difference of productivity, with power of 0.8 and significance level of 0.05.”
Beltramo et al. (2015)	lab	“Based on power calculations and the minimum detectable effect for experimentally testing marketing messages, a total of 36 parishes were selected.”
Karlan et al. (2015)	field	“We focus our discussion of statistical power on the profit results. The estimate for the effect of the consulting only treatment on stated income is an increase of 0.9 (se = 21.4) cedis over a control group mean of 146 (Column 1 of Table 6). Thus, the upper bound of the 95% confidence interval is 41 cedis, a 28% increase over the control group mean.”
Katuščák et al. (2015)	Lab	“Note that, given the number of subjects we use in 4H and 4HR, if the effect of loser vs. minimal feedback identified by FO were real and if their data provides a good description of the idiosyncratic variance in bidding behavior, a two-tailed test would have a power of over 0.95 against the null hypothesis of no effect. If we instead use the variance implied by our data, the power of the test ranges from 0.73 to 0.95, depending on the protocol and the outcome measure.”
Higuchi et al. (2015)	field	“Given the large variance in business performance at baseline, the number of samples needed to detect 10% change in business performance with 90% statistical power is over one thousand, which is far beyond our training and survey budget.”
Barton et al. (2016)	field	“We performed initial power calculations for the districts pooled, which indicated that we would detect a 5 percent difference in turnout rates at $p = 0.05$ between the negative and positive groups with a power of 0.85. When looking at each district individually, however, the effect size needs to be roughly ten percentage points to have an 80% chance of finding it at $p = 0.05$.”

Ericson & Kessler (2016)	lab	“In the first and second wave, some participants were also assigned to a subsidy, or status quo condition. These conditions were discontinued as a result of a power calculation and because the status quo eventually became a mandate.”
Czermak et al. (2016)	lab	“Second, we calculated the power of our test and found that a power of 0.80 for a significance level of $\alpha = 0.05$ could be achieved with our current sample size of $N = 191$ already if the range of the relative frequency of choosing Nash is less than 9 percentage points (assuming a mean of 45%).”
Kriss et al. (2016)	lab	“In the former case, this gives us 94% power to detect a proportion that differs from 50% by $\pm 25\%$ at $p < 0.05$ in a two-sided test; in the latter case we have 97% power (Chow et al., 2008).”
Lindeboom et al. (2016)	field	“The original proposal for the field experiment including power analysis is available at http://personal.vu.nl/b.vander.klaauw/OpzetCIZOnderzoek.pdf [in Dutch].”
Bhanot (2017)	field	“The sample size for this study was determined with the loan repayment rate outcome in mind, based on an estimated ex-ante repayment rate of 80%, power of 0.80, and an estimated effect size of five percentage points (informed primarily by the findings in Shu et al. (2012) and Karlan et al. (2010)).”
Galeotti et al. (2017)	lab	“To estimate the sample size necessary to uncover these hypothesized effects, we conduct an a priori power analysis. This type of analysis would be of limited value without a reasonable estimate of the treatment differences we would expect in our experiment. Such estimates are typically based on the results of previous similar studies. The studies which are most similar to ours are Gino and Pierce (2009, 2010).”
Persson (2018)	lab	“One limitation of the experiment is the relatively low power of the tests. For example, pooling the three treatments against the <i>Control</i> treatment and hypothesizing that the punishment probability is 0.3 in the former group (where $N = 82$) and 0.1 in the latter group (where $N = 24$), using a significance level of 0.05 the power of a one-sided Fisher’s exact test is 0.58; for comparisons across the treatments it is lower.”
Andersen et al. (2018)	lab	“A power calculation using a two-sided t-test for simplicity shows that we would need a sample size of 470 individuals per treatment cell to detect a significant effect if power is set to 80%, significance level is set to 5%, and given the responses of our subjects.”
Prasada & Bose (2018)	lab	“We make the standard assumptions of confidence level of 0.95 for type 1 error, 80% for power of the test. We specify a conservative desired difference of 0.1, which is about 25% of the standard deviation of conflict effort (3.7). Then according to Altman (1981), the minimum required number of independent observations is 214.”
Heggedal et al. (2018)	lab	“Moreover, given a 5% significance level and the observed variances of the D and N treatments, this test has a power of 99%.” & “The power computation was carried out using the simulation routine of Bellemare et al. (2016).”
Ding et al. (2018)	lab	“We report the power-analysis results in the last two columns of the table below. [...] Our results indicate that our study has more than 80% power to detect differences in Amplitude, APD, and RAD, regardless of whether we set the difference at 30% or 50%.”
Brookins et al. (2018)	lab	“In view of the null result for part (a), we also conducted an <i>ex post</i> power calculation. The effect size in our data is 0.17, which is conventionally considered small. With this effect size, at 80% power it would take 208 independent clusters (matches) per treatment – a sample size well beyond feasible – to obtain a significant difference in total output between [...]”
Rahwan et al. (2018)	lab	“We simulated 1,000 random samples for differently-sized groups, and determined the proportion of samples for each group that found a significant positive correlation between the number of coin flips reported as matched and negative moral affect using a one-sided Spearman test at the 5% significance level. Using this procedure, we determined that a sample size of at least 460 participants per condition was required to achieve 80% power in the main experiment. We decided to round up to 500 participants per condition to further increase power.”

References

- Mattei, A. (2000): Full-scale real tests of consumer behavior using experimental data. *Journal of Economic Behavior & Organization* 43(4): 487–497.
- Norman, A., Ahmed, M., Chou, J., Fortson, K., Kurz, C., Lee, H., Linden, L., Meythaler, K., Rando, R., Sheppard, K., Tantzen, N., White, I., Ziegler, M. (2003): An ordering experiment. *Journal of Economic Behavior & Organization* 50(2): 249–262.
- Haruvy, E., Stahl, D.O. (2007): Equilibrium selection and bounded rationality in symmetric normal-form games. *Journal of Economic Behavior & Organization* 62(1): 98–119.
- Schechter, L. (2007): Traditional trust measurement and the risk confound: An experiment in rural Paraguay. *Journal of Economic Behavior & Organization* 62(2): 272–292.
- Martin, R., Randal, J. (2008): How is donation behaviour affected by the donations of others? *Journal of Economic Behavior & Organization* 67(1): 228–238.
- Ismayilov, H., Potters, J. (2013): Disclosing advisor's interests neither hurts nor helps. *Journal of Economic Behavior & Organization* 93: 314–320.
- Bronchetti, E.T., Huffman, D.B., Magenheimer, E. (2015): Attention, intentions, and follow-through in preventive health behavior: Field experimental evidence on flu vaccination. *Journal of Economic Behavior & Organization* 116: 270–291.
- de Haan, T., van Veldhuizen, R. (2015): Willpower depletion and framing effects. *Journal of Economic Behavior & Organization* 117: 47–61.
- Hong, F., Hossain, T., List, J.A. (2015): Framing manipulations in contests: A natural field experiment. *Journal of Economic Behavior & Organization* 118: 372–382.
- Beltramo, T., Blalock, G., Levine, D.I., Simons, A.M. (2015): The effect of marketing messages and payment over time on willingness to pay for fuel-efficient cookstoves. *Journal of Economic Behavior & Organization* 118: 333–345.
- Karlan, D., Knight, R., Udry, C. (2015): Consulting and capital experiments with microenterprise tailors in Ghana. *Journal of Economic Behavior & Organization* 118: 281–302.
- Katuščák, P., Michelucci, F., Zajíček, M. (2015): Does feedback really matter in one-shot first-price auctions? *Journal of Economic Behavior & Organization* 119: 139–152.
- Higuchi, Y., Nam, V.H., Sonobe, T. (2015): Sustained impacts of Kaizen training. *Journal of Economic Behavior & Organization* 120: 189–206.
- Barton, J., Castillo, M., Petrie, R. (2016): Negative campaigning, fundraising, and voter turnout: A field experiment. *Journal of Economic Behavior & Organization* 121: 99–113.
- Ericson, K.M. & Kessler, J.B. (2016): The articulation of government policy: Health insurance mandates versus taxes. *Journal of Economic Behavior & Organization* 124: 43–54.
- Czermak, S., Feri, F., Glätzle-Rützler, D., Sutter, M. (2016): How strategic are children and adolescents? Experimental evidence from normal-form games. *Journal of Economic Behavior & Organization* 128: 265–285.
- Kriss, P.H., Weber, R.A., Xiao, E. (2016): Turning a blind eye, but not the other cheek: On the robustness of costly punishment. *Journal of Economic Behavior & Organization* 128: 159–177.
- Lindeboom, M., van der Klaauw, B., Vriend, S. (2016): Audit rates and compliance: A field experiment in care provision. *Journal of Economic Behavior & Organization* 131, Part B: 160–173.
- Bhanot, S.P. (2017): Cheap promises: Evidence from loan repayment pledges in an online experiment. *Journal of Economic Behavior & Organization* 140: 246–266.
- Galeotti, F., Kline, R., Orsini, R. (2017): When foul play seems fair: Exploring the link between just deserts and honesty. *Journal of Economic Behavior & Organization* 142: 451–467.
- Persson, E. (2018): Testing the impact of frustration and anger when responsibility is low. *Journal of Economic Behavior & Organization* 145: 435–448.
- Andersen, S., Gneezy, U., Kajackaite, A., Marx, J. (2018): Allowing for reflection time does not change behavior in dictator and cheating games. *Journal of Economic Behavior & Organization* 145: 24–33.

- Prasada, D.V.P., Bose, G. (2018): Rational conflict and pre-commitment to peace. *Journal of Economic Behavior & Organization* 149: 215–238.
- Heggedal, T.-R., Helland, L., Joslin, K.-E.N. (2018): Should I Stay or should I Go? Bandwagons in the lab. *Journal of Economic Behavior & Organization* 150: 86–97.
- Ding, S., Lugovsky, V., Puzzello, D., Tucker, S., Williams, A. (2018): Cash versus extra-credit incentives in experimental asset markets. *Journal of Economic Behavior & Organization* 150: 19–27
- Brookins, P., Lightle, J.P., Ryvkin, D. (2018): Sorting and communication in weak-link group contests. *Journal of Economic Behavior & Organization* 152: 64–80.
- Rahwan, Z., Hauser, O.P., Kochanowska, E., Fasolo, B. (2018): High stakes: A little more cheating, a lot less charity. *Journal of Economic Behavior & Organization* 152: 276–295.

Appendix 5 Power / sample size calculations in ERE

Authors	Type of experiment	Evidence
Burton et al. (2007)	lab	“Pre-experiment power analyses indicated a power of 0.8 using very conservative assumptions about differences between treatments. As differences between treatments become greater, the power approaches one rapidly.”
Mitani & Flores (2009)	lab	“The calculated power for a gender effect in Real Payment model is improved from 0.378 in Mitani and Flores (2007) to 0.9 in the current analysis [...]”
Araña & León (2013)	field	“For the largest price of 60 € the proportion of an acceptance response is 0.36 for the opt-in treatment and 0.38 for the opt-out treatment, which leads us to interpret that we can not reject the null hypothesis of equal proportions among samples with a significance level of 0.69 and a power of the contrast of 0.84.”

References

- Burton, A.C., Carson, K.S., Chilton, S.M., Hutchinson, W.G. (2007): Resolving questions about bias in real and hypothetical referenda. *Environmental and Resource Economics* 38(4): 513–525.
- Mitani, Y., Flores, N.E. (2009): Demand Revelation, Hypothetical Bias, and Threshold Public Goods Provision. *Environmental and Resource Economics* 44(2): 231–243.
- Araña, J.E., León, C.J. (2013): Can Defaults Save the Climate? Evidence from a Field Experiment on Carbon Offsetting Programs. *Environmental and Resource Economics* 54(4): 613–626.