J Econ Sci Assoc (2016) 2:157–168
DOI 10.1007/s40881-016-0028-4

EXPERIMENTAL TOOLS

Simulating power of economic experiments: the powerBBK package

Charles Bellemare[1] · Luc Bissonnette[1] ·
Sabine Kröger[1]

Received: 18 March 2016 / Revised: 28 September 2016 / Accepted: 30 September 2016 /
Published online: 13 October 2016

Abstract In this article, we highlight how simulation methods can be used to analyze power of economic experiments. We provide the powerBBK package programmed for experimental economists, that can be used to perform simulations in STATA. Power can be simulated using a single command line for various statistical tests (nonparametric and parametric), estimation methods (linear, binary, and censored regression models), treatment variables (binary, continuous, time-invariant or time varying), sample sizes, experimental periods, and other design features (within or between-subjects design). The package can be used to predict minimum sample sizes required to reach a user-specific level of power, to maximize power of a design given the researcher supplied a budget constraint, or to compute power to detect a user-specified treatment order effect in within-subjects designs. The package can also be used to compute the probability of sign errors—the probability of rejecting the null hypothesis in the wrong direction as well as the share of rejections pointing in the wrong direction. The powerBBK package is provided as

Part of the paper was written at the Institute of Finance at the School of Business and Economics at Humboldt Universität zu Berlin and at the Department of Economics at Zurich University. We thank both institutions for their hospitality. We thank Nicolas Couët for his valuable research assistance. We are grateful to participants at the ASFEE conference in Montpellier (2012), ESA meeting in New York (2012), the IMEBE in Madrid (2013), and seminar participants at the Department of Economics at Zurich University (2013) and at Technische Universität Berlin (2013). Finally, we thank reviewers and co-editors for extremely helpful comments and suggestions.

✉ Charles Bellemare
charles.bellemare@ecn.ulaval.ca

Luc Bissonnette
luc.bissonnette@ecn.ulaval.ca

Sabine Kröger
sabine.kroger@ecn.ulaval.ca

[1] Department of Economics, Laval University, Pavillon J.A.DeSève, Québec, Québec G1V 0A6, Canada

an .ado file along with a help file, both of which can be downloaded here (http://www.bbktools.org).

Keywords Statistical power · Simulation methods · Economic experiments

JEL Codes C8 · C9 · D03

1 Introduction

Underpowered experimental designs can have important consequences for the representativeness of published experimental research (Fanelli and Ioannidis 2013). In particular, it may result in publication bias if papers failing to detect a significant treatment effect face a lower acceptance probability in academic journals (Button et al. 2013; Nosek et al. 2012). This in turn may discourage researchers from even submitting papers reporting insignificant treatment effects. Moreover, underpowered experimental designs can also generate significant treatment effects in the wrong direction (sign error, see Gelman and Carlin 2014). These studies suggest that significant treatment effects in underpowered studies provide little information about the true treatment effects.

Researchers planning an experimental study have to decide among other things about the number of treatment variations, the number of subjects to recruit, the number of experimental periods, and whether to conduct a within or between-subjects design. All these decisions require a careful balancing between the chance of finding an existing effect, the precision with which this effect can be measured, and the available research budget. Statistical power computation using closed-form expressions are typically derived for simple statistical models and tests and tend to be valid under very specific conditions (e.g., large sample sizes and normally distributed errors).[1] Simulation methods, on the other hand, have the advantage to approximate the statistical power of experimental designs under relatively general assumptions about the distribution of model unobservables and experimental design configurations.

In this paper, we illustrate how to simulate power of economic experiments and provide the `powerBBK` package which can be used to perform the simulations in STATA. Power can be simulated for various statistical tests (nonparametric and parametric), estimation methods (linear, binary, and censored regression models), treatment variables (binary or continuous), sample sizes, experimental periods, and other design features (within or between-subjects designs). The `powerBBK` package can be used to achieve different objectives. It can be used to maximize ex ante statistical power of a given design subject to a user-specified budget constraint, taking into account the treatment-specific costs. The package can alternatively be used to simulate the minimal necessary sample size to reach a user-specified level of statistical power. It can also be used to compute the statistical

[1] See List et al. (2011) for results and discussion.

power of a particular design. In doing so, users have the option to predict the probability of detecting a user-specified treatment order effect in the context of within-subjects designs, and the probability of sign error—the probability of rejecting the null hypothesis in the wrong direction as well as the share of rejections pointing in the wrong direction. Finally, the package can be used to conduct ex post power analyses of published results to evaluate their credibility and to get (posterior) estimates of plausible effect sizes. In all cases, `powerBBK` requires that users enter a single command line specifying the desired options and parameters necessary to conduct the simulations.

Table 1 Statistical programs and packages allowing to perform power analysis and/or calculate the optimal sample size

Software/package		Version	Primary audience
Stand-alone software—free			
1	G*Power	3.1.9.2 (2014)	Psychologists and general
2	MorePower	6.0.1 (2013)	Psychologists and general
3	Optimal design	3.01 (2011)	Social scientists
4	PEPI	11.62 (2016)	Epidemiologists
5	PS: power and sample size calculation	3.1.2 (2014)	Biologists and general
Stand-alone software—for purchase			
Prices for academic use: approx. 800$ or annual subscription approx. 400$, 7 or 30 days free trial available			
6	PASS (power analysis and sample size software)	14	General
7	Power and precision	2	Health
8	Systat	13.1	General
Online—free			
9	Power and sample size (very basic)		General
SAS/STAT: add-ons—free or included			
10	UnifyPow (discontinued followed by 11, 12)	2002.08.17a	Biologists, epidemiologists, health
11	PROC POWER (as of SAS 9.1)	SAS/ STAT(R) 13.2	General
12	GLMPOWER (as of SAS 9.1)	SAS/ STAT(R) 13.2	General
STATA: add-ons—included			
13	sampsi (discontinued, followed by 15)	STATA	General
14	stpower (discontinued, followed by 15)	STATA	Survival analysis
15	power (as of STATA 13)	STATA 14	General
The module of this article: STATA add-on—free			
16	powerBBK	STATA 14 (2016)	Economists and general

The electronic version links to web sites of the programs, web sites were retrieved last on March 14, 2016

Other software programs and packages currently available to conduct power analyses are presented in Table 1 (#1–15) along with the current package (#16). They are offered as stand-alone, web-based applications, or STATA / SAS modules and are either free of charge or offered for purchase with free trials. Some of these programs (#1–5,7,10) are adapted to the needs of particular fields, such as psychology, health, epidemiology, biology and education, while others (#1,2,5,6,8,11–15) target a general audience. However, no package currently addresses the special needs of economists. Most of the programs (#1–15) rely on asymptotic approximation and none implements simulation-based methods adapted to the needs of (experimental) economists, e.g., measures power to detect treatment order effects, compares power of within and between-subjects design with multiple periods, proposes an optimal allocation of subjects to treatment and control within a given budget, nor allows for a continuous treatment variable.[2]

The paper is organized as follows. Section 2 discusses the simulation of statistical power and introduces the powerBBK package. Section 3 presents an application to gift exchange experiments. Section 4 concludes.

2 Power computation using powerBBK

The powerBBK package is based on the following treatment effect regression model

$$y^*_{it} = \beta_0 + d_{it}\beta_{1,t} + \mu_i + \epsilon_{it}, \tag{1}$$

where y^*_{it} denotes the latent outcome variable of subject i at period t, $\mathbf{d}_i = [d_{i1}, \ldots, d_{iT},]$ is a vector of time-varying treatment variables, where $d_{it} = 1$ when subject i receives treatment at period t and 0 otherwise. The parameters of interest are $\boldsymbol{\beta}_1 = [\beta_{1,1}, \ldots, \beta_{1,T}]'$. This specification nests as a special case a time-invariant treatment effect model (where all $\beta_{1,t}$ are identical). Treatment variables $\mathbf{d}_i$ are allowed to be either dichotomous or continuous. Time-invariant unobserved heterogeneity is captured by μ_i with corresponding cumulative distribution function F_μ. The remaining errors ϵ_{it} are drawn from a cumulative distribution $F_{\epsilon|\mathbf{d}}(a)$. We allow the errors to be heteroscedastic: the variance of the errors ϵ_{it} can depend on treatment conditions $\mathbf{d}_i$. We denote by $\sigma^2_{\epsilon,\mathbf{d}}$ the variance of ϵ_{it} conditional on treatment. A between-subjects (hereafter *BS*) design implies that $\{d_{it} : t = 1, \ldots, T\}$ does not vary across t. For the case of binary *BS* treatment, a subject is either assigned only to the control condition ($d_{it} = 0$ for all t) or to the treatment condition ($d_{it} = 1$ for all t). The continuous *BS* treatment assigns subjects randomly to a treatment drawn from the researcher specified set of treatment variables. In the presence of homoscedastic errors ϵ_{it}, the noise level $\mu_i + \epsilon_{it}$ is the same for treatment and control conditions. In this case it is reasonable to implement a *BS* design by assigning an equal number of subjects to control and treatment conditions. In the

[2] While out of the scope of this article, others have compared individual features of different statistical power programs for needs of particular sciences, e.g., Thomas and Krebs (1997) for ecology, Dattalo (2009) for health and general, and Peng et al. (2012) for education.

presence of heteroscedastic errors ϵ_{it}, statistical power can possibly be improved by assigning more subjects to the conditions where the noise level is higher. A within-subjects (hereafter *WS*) design implies that $\{d_{it} : t = 1, \ldots, T\}$ varies across t for each subject. In the presence of homoscedastic errors ϵ_{it}, it is reasonable to use a balanced *WS* design with $d_{it} = 0$ for T / 2 periods as long as the expected cost of a subject is approximately the same under both treatment conditions. In the presence of heteroscedastic errors ϵ_{it}, statistical power may be improved by assigning subjects to the noisier conditions for a higher number of periods. Finally, we maintain the assumption that μ_i is independent of all d_{it}. This assumption is typically motivated by the randomization of subjects to treatment conditions.

The powerBBK package considers three leading data-generating processes.

Case 1. Linear model: $y_{it} = y^*_{it}$.
Case 2. Binary choice model: $y_{it} = 1 \mathrm{if} \mathrm{y}^*_{\mathrm{it}} \geq 0$, and 0 otherwise.
Case 3. Model with censoring from below at a: $y_{it} = \max(a, y^*_{it})$,

where the observable outcome variable y_{it} may differ from y^*_{it} according to the case considered. With this parameterization we can generate samples for different sequences $\{d_{it} : t = 1, \ldots, T\}$ given values of $(\beta_0, \boldsymbol{\beta}_1)$ and $(F_\mu, F_{\epsilon|\mathbf{d}})$. Identification of $(\beta_0, \boldsymbol{\beta}_1)$ requires some minimal restrictions on the functions $(F_\mu, F_{\epsilon|\mathbf{d}})$. Mean independence with the treatment indicator is sufficient for the linear model (Case 1). Independence between ϵ_{it} is typically assumed for Cases 2 and 3. Note that Cases 1 and 3 allow the variance of ϵ_{it} to differ between control and treatment conditions. The user can specify any distribution available in STATA for $F_{\epsilon|\mathbf{d}}$ for Case 1. The package implements Case 2 as either a probit or logit model, thus setting F_ϵ to the standard normal or logistic distribution, respectively. The package implements Case 3 by setting $F_{\epsilon|\mathbf{d}}$ to a mean zero normal distribution with variance $\sigma^2_{\epsilon,\mathbf{d}}$, the familiar tobit model. The distribution F_μ is always assumed to be the normal distribution with a user-specified standard deviation, as most panel data models rely on this assumption in the estimation procedure.

The data-generating process described above is relatively flexible in terms of the type of outcome distributions it can capture. This is especially true for Case 1. The package currently does not support other discrete outcomes, notably multinomial choices or ordered responses. The powerBBK is free and open-source, allowing users to extend the package to suit their needs.

The powerBBK package requires the user to specify details concerning the experimental design, such as the number of subjects, number of periods, WS or BS design, balance of WS design and so on. There are options to evaluate the statistical power over a range of values *N* and to assess simultaneously power of both WS and BS designs. The user can specify whether or not to include individual heterogeneity by means of random-effects terms (i.e., the variance of μ_i is greater than 0) or to include treatment-specific heteroscedasticity (i.e., the variance of ϵ_{it} depends on the treatment received). Users can also specify when appropriate (e.g., in linear models) the distribution of errors $(F_\mu, F_{\epsilon|\mathbf{d}})$ they require for their simulations, thus allowing for example heavy-tailed distributions in linear models. The package further permits users to simulate power of nonparametric rank-based tests and can accommodate several common non-linear models (i.e., logit, probit, tobit).[3] Users can use the package to predict the maximal power a design can reach given a user-specified budget constraint with treatment-specific costs. Additional information and examples are available in the help file provided with the package.

Computing power of a given design is straightforward using the following steps.

Step 1 Fix N and T and for a given design (*WS* or *BS*), values of $(\beta_0, \boldsymbol{\beta}_1)$ and choice of $(F_\mu, F_{\epsilon|\mathbf{d}})$ generate a sample $\{\{(y_{it}, d_{it}) : t = 1, \ldots, T\} : i = 1, \ldots, N\}$.

Step 2—parametric Estimate $(\beta_0, \boldsymbol{\beta}_1)$ and the parameters of $(F_\mu, F_{\epsilon|\mathbf{d}})$ and compute $\hat{z}_t = \hat{\beta}_{1,t}/se(\hat{\beta}_{1,t})$ and the corresponding p value of the null hypothesis $H_0 : \beta_{1,t} = 0$ against either a one-sided or two-sided alternative. Here $se(\hat{\beta}_{1,t})$ denotes the standard error of the estimated period t treatment effect.[4]

Step 2—nonparametric Aggregate the individual data over T and use nonparametric rank-based tests (e.g., Wilcoxon rank-sum test for BS data, Wilcoxon signed-rank test for WS data) of the null hypothesis that the distribution of the aggregated values of y are the same under control and treatment conditions and compute the p value of the test.

Step 3 Repeat steps 1 and 2 for a large number of samples. Compute the fraction of p values which are less than the significance level of the test (e.g., 5 %). This represents the power of the test.

Repeating the three steps above for a range of N and T values for each design, enables the researcher to plot power curves for each element of $\boldsymbol{\beta}_1$. Power curves are useful for comparing the designs for a given sample size, for determining the minimal sample size needed to reach a certain statistical power separately for each design, or to look at the effect of the number of periods and how to balance the number of participants in the treatments. The package also offers users the possibility of predicting the maximal power an experimental design can reach given a specified budget constraint. In this case, users are required to additionally specify the expected payoff of a participant in each treatment as well as the total available budget. The package then evaluates the power of a series of user-specified allocations, which easily allows users to determine the allocation that maximizes power. Finally, an issue concerning WS designs is possible treatment order effects. These effects imply that the response depends on whether treatment or control conditions are experienced first. The `powerBBK` package can be used to predict the probability of detecting a user-specified treatment order effect for a given experimental design. This option is currently only implemented for the time-invariant binary treatment effect model where all elements of $\boldsymbol{\beta}_1$ are identical.

[3] In those cases, the distributions $(F_\mu, F_{\epsilon|\mathbf{d}})$ are pre-specified.

[4] Note, that the estimators used in Step 2 will depend on the nature of the outcome variable. The maximum likelihood estimator can be used in all three cases (linear, binary choice or models with censoring). Linear regression with clustered standard errors can also be used in the case of the linear model. This is a popular choice for experimental economists, as the distributions $(F_\mu, F_{\epsilon|\mathbf{d}})$ are often unknown in practice.

3 Illustration: gift exchange in the field

We illustrate the power analysis presented in Sect. 2 with an application in the context of field experiments designed to measure reciprocal preferences of workers. The Appendix provides one of the command lines used to perform this analysis. Our analysis exploits data from two different studies in this area. Gneezy and List (2006) use a *BS* design in the context of a single day spot labor market experiment with a data entry task. They assign 9 workers to their treatment condition (gift) and 10 workers to the control condition (no gift). They estimate a linear random-effects panel data model (Case 1 in Sect. 2) with individual-specific effects μ_i and where t indexes the hour of work within the experimental day. Bellemare and Shearer (2009) use a WS design with 18 workers (tree-planters). They test how workers respond to a gift from their employer. Their WS design is unbalanced: workers planted first for 5 days under control conditions (no gift). Workers then received a gift on the final day of planting on the experimental block. Bellemare and Shearer (2009) estimate a linear fixed-effects panel data model (Case 1) with individual-specific effects μ_i and where t indexes the day of work during the experiment. Both studies use roughly the same total number of subjects and time periods, but the notion of time varies across studies.

We first estimated a random-effects panel data model of Eq. (1) using the Gneezy and List data with the dependent variable being the natural log of productivity. We get $(\widehat{\beta}_0, \widehat{\beta}_1) = (3.674, 0.055)$, $\widehat{\sigma}^2_\mu = 0.088$, $\widehat{\sigma}^2_\epsilon = 0.018$. The corresponding estimates using the Bellemare and Shearer data are $(\widehat{\beta}_0, \widehat{\beta}_1) = (6.955, 0.061)$, $\widehat{\sigma}^2_\mu = 0.046$, $\widehat{\sigma}^2_\epsilon = 0.018$. The estimated treatment effect (β_1) and estimated error variance σ^2_ϵ are very similar for both studies. The estimated value of σ^2_μ (unobserved heterogeneity) on the other hand is twice as high in the Gneezy and List data.[5]

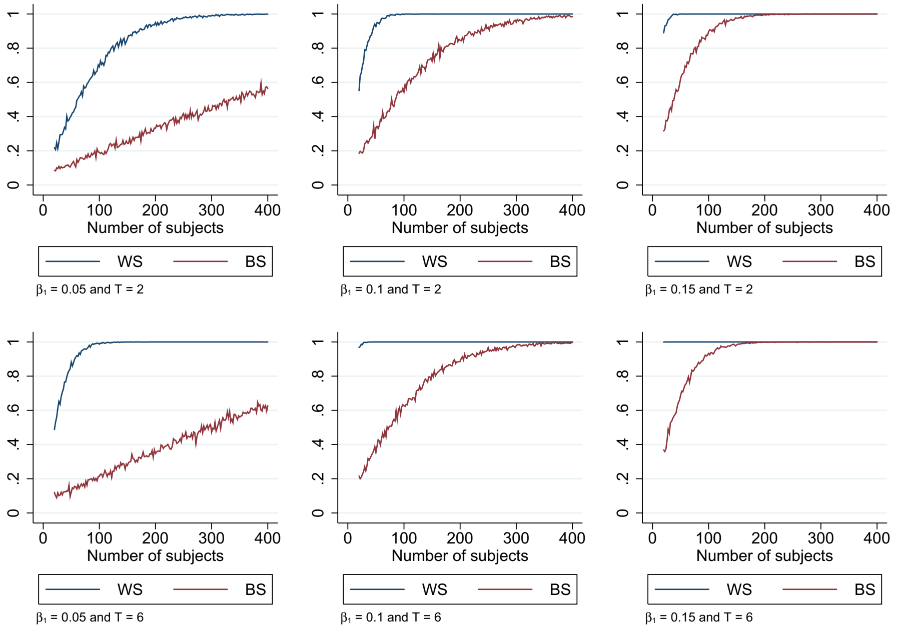
We next used the estimated model parameters from both data sets to simulate power of WS and BS designs for two scenarios, the low-noise and the high-noise scenario. The low-noise scenario sets ($\sigma^2_\mu = 0.045$ and $\sigma^2_\epsilon = 0.02$ while the high-noise scenario sets $\sigma^2_\mu = 0.09$ and $\sigma^2_\epsilon = 0.02$. The variance of μ_i in the high-noise scenario is thus exactly twice the corresponding value for the low-noise scenario. We will consider three values for β_1 (0.05, 0.1 and 0.15) for both scenarios. The value of β_0 plays no role in our analysis and will be set to 7.0 in all our simulations. We will also consider setting T to 2 and 6. Setting $T = 6$ proxies the number of periods used in both studies. The case $T = 2$ is interesting because it proxies experiments which take place with very low number of observations, e.g., for two periods, while still allowing a meaningful comparison of WS and BS designs. It also represents a case where researchers have little information to control for the presence of unobserved individual heterogeneity μ_i. It is straightforward to consider

[5] Both studies estimate regression models using the variable y_{it} in level—they do not use the natural logarithm of productivity as the dependent variable. Using the natural logarithm of productivity simplifies the comparison of the estimated treatment effect of both studies. Bellemare and Shearer (2009) additionally control for weather effects while (Gneezy and List 2006) allow the effect of the gift to vary across time. Estimated model parameters with those additional controls are very similar to the results we report here.

other values of T. We perform a separate power analysis for each scenario for a double-sided test with a 5 % level of significance. We implement the BS design by assigning the same number of subjects to control and treatment conditions. We implement a balanced WS design by assigning subjects to the same number of periods under control and treatment conditions. We also simulated power for an "unbalanced" WS design assigning subjects to the treatment condition for only one out of six periods. Simulated power of the unbalanced WS design was not very different to the power of the balanced WS design. This is to be expected as the variance of the outcome variable is kept constant under control and trial conditions. We thus focus our analysis on the balanced WS design. Finally, we use the OLS estimator with standard errors clustered at the individual level. All our results are very similar when using the (asymptotically more efficient) GLS estimator.

Figure 1 presents the simulated power curves for the low-noise scenario. Several regularities emerge. First, we find that power is systematically higher for the WS design for all 6 combinations of β_1 and T values used. This result is expected given the WS design exploits within subject variation in decisions for a given individual (for a given level of μ_i). This advantage of the WS design over the BS design is well documented [see, e.g., Keren (1993)]. We also find that increasing the number of periods raises power of the WS design, but has relatively minor impact on power of the BS design. The quantitative differences in power between both designs are perhaps more surprising. A natural way to compare both designs is to compare the minimal number of subjects (MNS) required to reach a given level of power. Social scientists often argue that an experiment should aim to correctly detect a treatment effect 80 % of the time (see Cohen 1988) when using a double-sided test along with a 5 % significance level. Table 2 presents the simulated MNS required to reach this power threshold derived from the curves in Fig. 1. We find that the MNS exceeds 400 subjects for the BS design for both values of T when $\beta_1 = 0.05$. In comparison, the MNS of the WS design is 122 subjects when $T = 2$, and 42 subjects when $T = 6$. As expected, the required MNS decrease with β_1. The MNS of the *BS* design when $\beta_1 = 0.1$ are 182 subjects and 162 subjects for 2 and 6 periods, respectively. The corresponding MNS of the WS design are 30 subjects and less than 20 subjects, thus 6–8 times less than the corresponding MNS of the BS design. Finally, MNS of the BS when $\beta_1 = 0.15$ are 84 subjects and 74 subjects for 2 and 6 periods, respectively. Corresponding MNS of the WS design are both below 20 subjects, roughly 4 times less than the BS design.

Figure 2 presents the simulated power curves for the high-noise scenario. Several interesting regularities emerge. First, power curves of the WS design in the high-noise scenario are very similar to those of the WS design in the low-noise scenario. Power of the BS design, on the other hand, is substantially worse under the high-noise scenario than under the low-noise scenario. These regularities are captured by the corresponding MNS of both designs (see Table 2). We find that the MNS of the WS in the high-noise scenario are very similar to the corresponding values in the low-noise scenario. The MNS of the BS design, on the other hand, are considerably higher. In particular, we find that the BS design requires between 286 and 302 subjects to detect a value of $\beta_1 = 0.1$ with power of 80 %. This is roughly 120 subjects more (approx. 65 % more) than required in the low-noise scenario.



β₁ = 0.05 and T = 2

β₁ = 0.1 and T = 2

β₁ = 0.15 and T = 2

β₁ = 0.05 and T = 6

β₁ = 0.1 and T = 6

β₁ = 0.15 and T = 6

Fig. 1 Simulated statistical power of BS and WS designs with $T = 2$ and $T = 6$ for the low-noise scenario. Simulations based on values $\sigma_{\mu}^2 = 0.045$ and $\sigma_{\epsilon}^2 = 0.02$. Results for the BS design are computed by allocating the same number of subjects to control and treatment conditions for all periods. Results for the WS design are computed by assigning all subjects to the same number of control and treatment periods

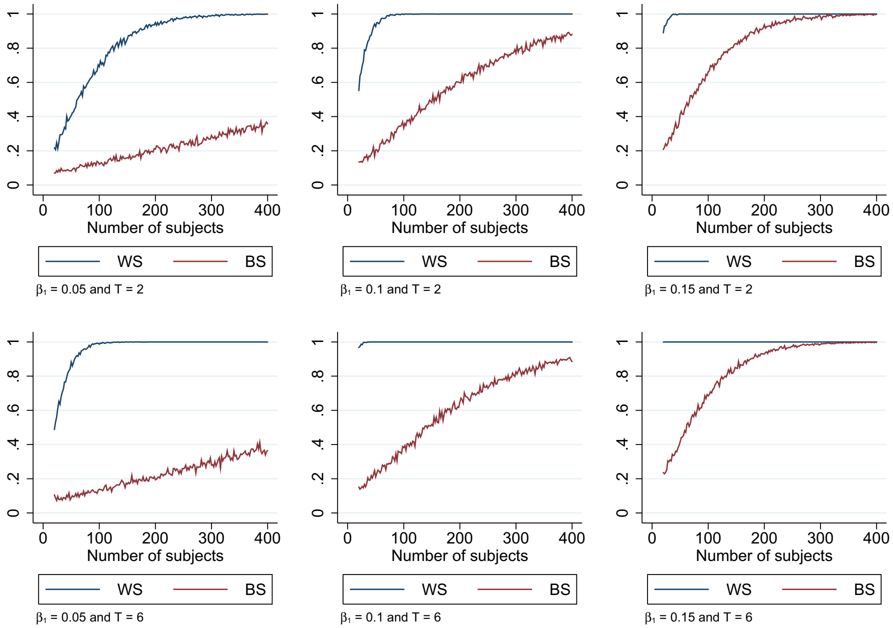
Table 2 Minimal sample sizes (MNS) required to reach a power of 80 % for a double-sided test with 5 % significance level (GLS estimator)

	Low-noise scenario			High-noise scenario		
	$\beta_1 = 0.05$	$\beta_1 = 0.1$	$\beta_1 = 0.15$	$\beta_1 = 0.05$	$\beta_1 = 0.1$	$\beta_1 = 0.15$
Between-subjects design						
$T = 2$	>400	182	84	>400	302	140
$T = 6$	>400	162	74	>400	286	130
Within-subjects design						
$T = 2$	122	30	<20	122	34	<20
$T = 6$	42	<20	<20	44	<20	<20

Simulations for the low-noise scenario based on values $\sigma_{\mu}^2 = 0.045$ and $\sigma_{\epsilon}^2 = 0.02$. Simulations for the high-noise scenario based on values $\sigma_{\mu}^2 = 0.09$ and $\sigma_{\epsilon}^2 = 0.02$. Results for the BS design are computed by allocating the same number of subjects to control and treatment conditions. Results for the WS design are computed by assigning subjects to the same number of control and treatment periods

MNS minimal number of subjects

Similarly, we find that MNS of the *BS* design lies between 130 and 140 subjects when $\beta_1 = 0.15$. This is roughly 60 subjects more (approx. 70 % more) than required in the low-noise scenario. These results suggest that researchers planning to



β₁ = 0.05 and T = 2

β₁ = 0.1 and T = 2

β₁ = 0.15 and T = 2

β₁ = 0.05 and T = 6

β₁ = 0.1 and T = 6

β₁ = 0.15 and T = 6

Fig. 2 Simulated statistical power of BS and WS designs with $T = 2$ and $T = 6$ for the high-noise scenario. Simulations based on values $\sigma_{\mu}^{2} = 0.09$ and $\sigma_{\epsilon}^{2} = 0.02$. Results for the BS design are computed by allocating the same number of subjects to control and treatment conditions for all periods. Results for the WS design are computed by assigning all subjects to the same number of control and treatment periods

conduct *BS* design experiments in this area should carefully consider the level of noise they expect to be present in the data.[6]

In most power analyses, model parameters values are calibrated using data from either existing studies or pilot experiments conducted by the researchers themselves. These values represent estimates of the true but unobserved underlying population parameters and are thus subject to sampling variability. The importance of sampling variability is especially relevant when values are calibrated using small data sets. In these cases, researchers may consider repeating their power analysis for a selected range of values for each model parameter. One straightforward approach would be to draw parameter values from the sampling distribution of the model parameters and evaluate power for each draw, thus approximating the sampling distribution of the predicted power.

[6] We repeated the power analysis using nonparametric rank-based tests (Wilcoxon rank-sum tests and Wilcoxon signed-rank tests). Results are very similar and available upon request.

4 Conclusion

This paper highlighted the usefulness of simulation methods for power analysis of economic experiments and provides the powerBBK package to perform such analyses, taking into account several common design features and the possibility to optimize experimental designs under budget constraints.

Appendix

Consider the following STATA command used to generate the top left panel of Fig. 1.

```
powerBBK pwrexample, budget(40(4)800) t(2) design(both)
          beta(6.3 0.05) muvar(0.045) epsvar(0.02) command(regress) panel
```

The mandatory argument pwrexample is a prefix for the names of variables (to be generated) containing the results of the simulations and can be directly used to reproduce the panel in Fig. 1. We simulate the statistical power for a two-period experiment (set using t(2)) and participants ranging from 20 to 400. This simulation assumes that each period under either treatment or control condition costs 1 unit of currency per participant. We need to consider budgets ranging from 40 to 800 units (the first and last numbers inside the budget command). Samples sizes are incremented by steps of two participants, or 4 budgetary units indicated using (4) in the budget command (the cost of two additional participants over two periods). It is possible to run the simulation for a single budget value, say 800 units, by specifying budget(800). Per period treatment or control condition-specific costs of (say) 2 units can be specified by adding costtreatment(2) and/or costcontrol(2). By default, the package allocates an equal share of participants to control and treatment conditions. Users can alternatively evaluate power for different allocations. Adding the option allocation(0.33 0.5 0.66) will for example produce separate power calculations when 1/3, half, and 2/3 of participants are, respectively, allocated to control conditions.

Simulations are performed using both *WS* and *BS* designs by setting design(both). Setting instead design(within) or design(between) implements only the *WS* or *BS* design. The constant and slope parameters are set using beta(6.3 0.05), which implies $\beta_0 = 6.3$ and $\beta_1 = 0.05$ through the simulations outlined in Sect. 2. The option muvar(0.045) sets $\sigma_\mu^2 = 0.045$ while epsvar(0.02) sets $\sigma_\epsilon^2 = 0.02$. These variance parameter values were previously estimated using the Bellemare and Shearer (2009) data (see Sect. 3). GLS linear panel regression is used by setting command(regress) panel. Alternatively, nonparametric tests can be used by setting command(rank) panel. Tobit and Probit (panel) regressions can be specified using command(tobit) or command(probit), where epsvar(1) ($\sigma_\epsilon^2 = 1$) is the default normalization for

Probit and need not be set by users. Note that `command(tobit)` sets as default censoring from below at 0. These values can be adjusted by users. Detailed information about other features and options of the package can be found in the help file provided with the package.

References

Bellemare, C., & Shearer, B. (2009). Gift giving and worker productivity: evidence from a firm-level experiment. *Games and Economic Behavior*, *67*(1), 233–244.

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., et al. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365–376.

Cohen, J. (1988). *Statistical power analysis for behavioral sciences* (2nd ed.). New Jersey: Routledge Academic.

Dattalo, P. (2009). A review of software for sample size determination. *Evaluation & The Health Professions*, *32*(3), 229–248.

Fanelli, D., & Ioannidis, J. P. A. (2013). US studies may overestimate effect sizes in softer research. *Proceedings of the National Academy of Sciences*, *110*(37), 15031–15036.

Gelman, A., & Carlin, J. (2014). Beyond power calculations: assessing type S (sign) and type M (magnitude) errors. *Perspectives on Psychological Science*, *9*, 641–651.

Gneezy, U., & List, J. A. (2006). Putting behavioral economics to work: testing for gift exchange in labor markets using field experiments. *Econometrica*, *74*(5), 1365–1384.

Keren, G. (1993). Between- or within-subjects design: a methodological dilemma. In G. Keren, C. Lewis (Eds.), *A Handbook for data analysis in the behavioral sciences, Chapter 8*. New Jersey: Lawrence Erlbaum Associates Inc.

List, J., 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.

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific Utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science*, *7*(6), 615–631.

Peng, C. Y. J., Long, H., & Abaci, S. (2012). Power analysis software for educational researchers. *The Journal of Experimental Education*, *80*(2), 113–136.

Thomas, L., & Krebs, C. J. (1997). A review of statistical power analysis software. *Bulletin of the Ecological Society of America*, *78*(2), 126–139.