*Journal of the Economic Science Association* (2025), 1–18
doi:10.1017/esa.2025.10019

ORIGINAL PAPER

# The (Statistical) Power of Incentives

Aleksandr Alekseev

Department of Economics, University of Regensburg, Regensburg, Bavaria, BY, Germany
Email: aleksandr.alekseev@ur.de

(Received 14 November 2024; revised 19 March 2025; accepted 25 July 2025)

**Abstract**

I study the optimal design of monetary incentives in experiments where incentives are a treatment variable. I propose a novel framework called the Budget Minimization problem in which a researcher chooses the level of incentives that allows her to detect a predicted treatment effect while minimizing her expected budget. The Budget Minimization problem builds upon the power analysis and structural modeling. It extends the standard optimal design approach by explicitly incorporating the budget as a part of the objective function. I prove theoretically that the problem has an interior solution under fairly mild conditions. To showcase the practical applications of the Budget Minimization problem, I provide examples of its implementation in several well-known experiments. I also offer a practical guide to assist researchers in utilizing the proposed framework. The Budget Minimization problem contributes to the experimental economists' toolkit for an optimal design, however, it also challenges some conventional design recommendations.

**Keywords:** Economic experiments; effect size; experimental design; incentives; power analysis; sample size

**JEL Codes:** C9; D9

## 1. Introduction

Incentives are a cornerstone of experimental economics. The two most common methodological questions about the use of incentives are whether subjects should be paid and how subjects should be paid. Over the years, the field has accumulated a voluminous empirical literature in an attempt to inform the answers to these questions.[1] The theoretical work, on the other hand, has been relatively scarce. Following the early contributions to the question of *whether* to pay subjects (Smith, 1976; 1982), the recent literature has mostly been occupied with the question of *how* to pay subjects, or incentive compatibility of different payoff mechanisms (Cox et al., 2014; Harrison and Swarthout, 2014; Azrieli et al., 2018, 2020; Li, 2021). However, there is another question about incentives that so far has received no theoretical treatment, which is *how much* to pay subjects, or what should be the level of incentives. I attempt to fill in this gap by offering three main contributions. First, I use a simple utility-based framework to formalize the question about the optimal level of incentives. Second, I show theoretically that this question is well known under fairly mild conditions. Third, I illustrate my approach using the data from several well-known experiments and offer a practical guide for implementing it.

[1] For reviews, see Camerer and Hogarth 1999, Hertwig and Ortmann 2001, Gneezy et al. 2011, Cox and Sadiraj (2019), and Voslinsky and Azar 2021. A related strand of literature examines the hypothetical bias in economics (Harrison, 2024; Harrison and Rutström, 2008a; Harrison, 2007; Harrison, 2006; Nape et al., 2003; Cummings et al., 1997; Cummings et al., 1995).

The current approach to how much to pay subjects is typically ad hoc. It usually amounts to setting incentives at some conventional level based on past experiments, a target hourly wage, or lab policies. None of these conventions, however, are standard within the field (Cloos et al., 2023). To put some structure on the problem of choosing an optimal level of incentives, I focus on an important case when incentives are a treatment variable.

Incentives are among the most commonly used treatment variables in economic experiments, both field and lab. Researchers have been studying the effect of incentives on educational outcomes, prosocial behavior, and lifestyle habits (Gneezy et al., 2011), on dishonest behavior (Fischbacher & Föllmi-Heusi, 2013; Gibson et al., 2013; Balasubramanian et al., 2017) and distributional choices (El Harbi et al., 2015), on behavior in social dilemmas (Amir et al., 2012; Rousu et al., 2015; Yamagishi et al., 2016; Mengel, 2017; Leibbrandt and Lynham, 2018) and coordination games (Parravano and Poulsen, 2015), on behavior in dictator (Schier et al., 2016; Larney et al., 2019) and trust (Thielmann et al., 2016) games, on behavior in psychological games (Bellemare et al., 2018) and generic normal-form games (Pulford et al., 2018), on risk preferences (Holt and Laury, 2002), auctions (Smith and Walker, 1993), preference reversals (Grether and Plott, 1979; Cox and Grether, 1996), finance experiments (Kleinlercher and Stöckl, 2018), and performance on various tasks (Araujo et al., 2016; Brañas-Garza et al., 2019; Enke et al., 2023; Alekseev, 2022).

To fix the terms, by (monetary) incentives I understand monetary payments to subjects that are expected to affect their behavior or outcomes. A classic example of that would be experiments where subjects receive a piece rate for completing a real-effort task and the question is whether a higher piece rate induces more effort. My framework also applies to cases when money is a treatment variable, but not an incentive. An example of that would be an experiment that studies the effect of money on happiness and the question is whether a higher monetary transfer leads to greater happiness.

A key factor that enables studying the optimal level of incentives is that researchers are often interested in testing qualitative hypotheses. A typical research question is whether a treatment variable affects subjects' behavior while the specific values of the treatment variable are nuisance parameters. For example, a researcher studying performance pay is more likely to be interested in whether a higher piece rate increases effort rather than whether a specific 2-cent bump in a piece rate increases effort.[2] The qualitative nature of hypotheses creates a degree of freedom that I exploit to pick an "optimal," in a sense precisely defined below, level of incentives.

I introduce a *Budget Minimization problem* in which a researcher chooses the level of monetary incentives that allows her to find a predicted treatment effect for some conventional levels of significance and power while minimizing the total expected budget. The Budget Minimization problem follows from a researcher's utility function and relies on two key ingredients. First, it relies on the power analysis to compute the required sample size for a predicted effect size. Second, it relies on a model (structural or reduced-form) to predict the outcomes in the treatment and control groups for a given level of incentives. The outcome of the Budget Minimization problem is the optimal level of incentives in the treatment group relative to the control, a variable I refer to as the *treatment strength*. The treatment strength pins down the required sample size, expected payoffs per subject, and the total expected budget.

The key tension in the Budget Minimization problem is between a required sample size and expected per-subject payoffs. On the one hand, increasing incentives leads to a higher expected effect size, which in turn drives down the required sample size and hence the expected total budget (the sample-size effect). On the other hand, increasing incentives leads to higher expected per-subject payoffs, which, in turn, lead to a higher expected total budget (the payoff effect). My main theoretical result is that, under fairly mild assumptions, the Budget Minimization problem has a non-trivial solution where the two effects are in the exact balance. I illustrate the properties of a solution using

[2] Setting the appropriate level of a piece rate in performance pay experiments is notoriously difficult (Lazear, 2018; Carpenter and Huet-Vaughn, 2019).

https://doi.org/10.1017/esa.2025.10019 Published online by Cambridge University Press

existing experiments, sketch a practical guide for setting up the problem and solving it for one's own design, and provide a sample R code (see Appendix C).

My main contribution is to offer a disciplined economic approach to the problem of choosing an optimal level of monetary incentives in experiments where incentives are a treatment variable. Experimental budgets are rarely explicitly discussed by researchers. Money, however, is a scarce resource, which makes it natural to ask what is an optimal way to use it. This question is of particular concern to junior scholars and PhD students, whose budgets are usually quite small while the pressure to produce significant results is high, as well as to researchers running expensive large-scale interventions in the field. This question is relevant both for new experiments[3] and replications.[4] Finally, the calculations from the Budget Minimization problem can serve as a convincing justification of the grant money requested from a funding agency.

The Budget Minimization problem is an alternative approach to an optimal experimental design that expands experimental economists' toolkit. The main point of departure from the traditional approach to an optimal design is the explicit inclusion of budget considerations. As an alternative approach, the Budget Minimization problem challenges some received wisdom in experimental design. For example, a common recommendation is to follow the *maximum separation principle*: to set the values of a treatment variable as far apart as possible to ensure a maximum separation between predictions or a maximum variation in the treatment (Friedman and Sunder, 1994; List et al., 2011; Holt, 2019).[5] My approach shows that it may not be optimal to do this if separating the treatment values as much as possible leads to prohibitively high payoffs. Maximizing treatment strength, in other words, is not equivalent to maximizing a researcher's utility.

## 2. Related literature

The present work is most closely connected to the literature that exploits structural modeling to guide experimental design. This literature shows how to use theoretical models to optimize the design of an experiment, typically in terms of statistical power or precision of parameter estimates. Harrison 1989 brought the connection between incentives and power analysis into experimental discourse and showed that subjects' deviations from optimal behavior in auctions lead to small utility losses (a flat-maximum problem).[6] Harrison (1994) extends the flat-maximum critique to experimental tests of the expected utility theory. Moffatt (2007) uses results from the statistical optimal experimental design literature and previous estimates from structural models to optimize (in the sense of maximizing the precision of parameter estimates) experiments that elicit willingness to pay and risk preferences. Rutström and Wilcox 2009 use two different structural models of learning along with previous estimates of their structural parameters to optimize their experiment. Woods (2020) proposes the use of structural (quantal response) model simulations to improve the accuracy of an ex-ante power analysis and to guide optimal design decisions. Monroe 2020 uses simulations to conduct the power analyses of two sets of binary lottery choices designed to classify subjects according to one of two risk preference models. The approach I take is similar to these works in that I also advocate for, and show the benefits of, using theoretical models to guide experimental design. The main difference is that I use a different objective function in the analysis. While the previous work, following the classic optimal experimental design literature (Ford et al., 2018; Atkinson, 2018; Fedorov, 1972; Silvey, 1980), focuses mainly on the statistical properties of a design, I use experimental budget as an objective function. My

[3] Even if a study is not a replication per se, it is common to replicate existing findings to establish a baseline before introducing a new treatment.

[4] An important qualification is that the replication will necessarily be conceptual, rather than direct (Camerer et al., 2019), in this case since the Budget Minimization problem will likely yield treatment values that are different from the ones in an original study.

[5] While this is true in many cases, there are some important exceptions, such as non-linear models (Moffatt, 2015).

[6] Also see the subsequent discussion in Cox et al. 1992, Friedman 1992, Merlo and Schotter 1992, Harrison 1992.

https://doi.org/10.1017/esa.2025.10019 Published online by Cambridge University Press

approach calls for balancing statistical considerations (required sample size) with cost considerations (expected payoff per subject) to find an optimal level of incentives.

This work also contributes to a series of papers that provide tools and guidance for conducting economic experiments. These papers offer general statistical considerations for running an experiment. List et al. 2011 is a concise yet comprehensive guide to experimental design covering the issues of randomization and optimal sample arrangement. Bellemare et al. 2016 develop a statistical package to simulate the power of experiments for parametric and nonparametric statistical tests, different estimation methods, and treatment variables. Vasilaky and Brock 2020 focus on power analysis and provide code examples and tools needed in power calculations. The present work is similar to these papers in that it is also motivated by statistical considerations for running an experiment. The main differences are that I focus on a special, although important, class of experiments in which the treatment variable is monetary incentives and that I supplement statistical considerations with cost considerations in a novel way. Both List et al. 2011 and Bellemare et al. 2016 feature cost and budget considerations: List et al. 2011 provide guidance for sample arrangement in case the sampling costs differ by treatment and Bellemare et al. 2016's package can predict the maximal power an experiment can reach given a specified budget constraint. The present work differs from these papers in that it provides empirical guidance on, and theoretical justification for, how a researcher can optimally choose a level of incentives, in case they are a treatment variable, to minimize the budget.

More broadly, this work connects to the literature that studies the use of incentives in economic experiments. This literature studies the theoretical properties of common payoff mechanisms or proposes new payoff mechanisms that improve upon the existing ones. Cox et al. 2014 discuss the theoretical properties of popular payoff mechanisms, explain which mechanisms are incentive compatible for which theories, and empirically show that different payoff mechanisms significantly affect subjects' revealed risk preferences. Harrison and Swarthout 2014 empirically show that risk preference models that assume violations of the independence axiom cannot be reliably estimated when an experiment assumes the validity of this axiom via the random lottery incentive mechanism. Azrieli et al. 2018, Azrieli et al. 2020 introduce a theoretical framework for analyzing the incentive compatibility of different payoff mechanisms and identify assumptions needed to guarantee the incentive compatibility of the random problem selection mechanism and paying for every period. Li 2021 identifies necessary and sufficient conditions for a payoff mechanism to be incentive-compatible for all risk preference models with complete and transitive preferences and proves that her new payoff mechanism, the Accumulative Best Choice, is the only incentive compatible mechanism in a multiple-task setting. Johnson et al. 2021 introduce the Prince payoff mechanism, which they show to be a transparent and incentive-compatible method for measuring preferences that improves upon popular payoff mechanisms, such as the random incentive mechanism. The main difference of the present work from these papers is that it studies theoretically the optimal level of incentives, or *how much* to pay subjects, in case when incentives are a treatment variable.

## 3. Budget minimization problem

Consider a researcher planning a budget for an experiment. The expected total experimental budget, $b$, depends on the number of subjects in the experiment and expected per-subject payoffs. The researcher plans to use a standard between-subject design with two groups: control ($C$) and treatment ($T$). Let $G = \{C, T\}$ denote the set of experimental groups and $g \in G$ be its generic element. For simplicity, assume that the researcher plans to use an equal number of participants, $n$, in each group.[7]

[7] Using an equal number of participants in the treatment and control groups is optimal when the variances of outcomes are equal in the two groups. When the variances are unequal it is optimal to allocate different numbers of participants to each group. For example, for a $t$-test, the ratio of the participants in each group is equal to the ratio of the standard deviations in these groups (List et al., 2011). There could be other reasons for choosing an unequal allocation between the groups, such as different sampling costs. In all of these cases, one could simply fix $n$ to be the number of participants in the control group and

https://doi.org/10.1017/esa.2025.10019 Published online by Cambridge University Press

The researcher uses monetary incentives as single treatment variable. Depending on the nature of the choice variable in the experiment, the researcher could use, for example, the difference in means or the difference in proportions as the effect of interest.

Let $\tau_g$ denote the value of the treatment variable in group $g$. I denote the difference between the values of the treatment variable in the treatment and control groups as $\tau \equiv \tau_T - \tau_C$, $\tau \in \mathbb{R}_+$ and refer to it as the treatment strength. In some cases it can be of interest to have the treatment strength as a multiplicative factor rather than a difference. The above definition of the treatment strength accommodates these cases by defining the values of the treatment variable on a logarithmic scale. If one defines $\tau \equiv \ln \tilde{\tau}, \tau_g \equiv \ln \tilde{\tau}_g$, then $\tilde{\tau} = \exp(\tau) = \exp(\tau_T)/\exp(\tau_C)$ is the multiplicative treatment strength. I assume that the treatment strength is the only lever the researcher uses to optimize the budget.[8]

The researcher uses the power analysis to determine the required number of subjects in each group. This number will depend on the statistical parameters (significance $\alpha$ and power $1 - \beta$) and on the expected outcomes in each group, $\mu_g$. The researcher sets significance and power at some conventional levels.[9] The expected outcomes can be, for example, the mean choices in each group in case the choice variable is continuous or the proportions of subjects choosing a given alternative in case the outcome is discrete.[10] The expected effect size is then typically a difference in expected outcomes between the treatment and control groups, $\mu_T - \mu_C$, which depends on, although not identical to, the chosen treatment strength $\tau$. To predict the expected effect size, the researcher uses a model parameterized by a vector of parameters $\gamma$. The model can be a structural one, in which case the vector of parameters can include, for example, risk aversion, time preferences parameters, social preferences parameters, the curvature of the cost-of-effort function, etc. Alternatively, the model can be a reduced-form one, in which case the parameters will be regression coefficients. The researcher takes the parameters as given based on prior estimates. The expected outcomes will then depend on the treatment strength, behavioral parameters, as well as any other potential parameters of the experiment lumped in a vector $\delta$: $\mu_g = \mu_g(\tau \mid \gamma, \delta)$. Vector $\delta$ includes things that are not explicitly modeled but that can nevertheless affect behavior, for example, subject pool, number of rounds, framing of the instructions, whether a study is done in the lab or in the field, etc. To make everything a function of $\tau$ only, I use a convention that the level of incentives in the control group, $\tau_C$, is included in vector $\delta$. To summarize, the required number of subjects in each group depends on the parameters as follows: $n = n(\tau \mid \alpha, \beta, \gamma, \delta)$. It is worth emphasizing that the researcher does *not* pick $n$, as is the case in a typical power analysis. Instead, she picks $\tau$ that affects expected outcomes that in turn pin down $n$, conditional on other parameters.

The expected per-subject payoffs in each group, $\pi_g$, will depend on expected outcomes and on the way the outcomes are translated into payoffs. For example, when the outcome is the mean number of problems solved in a real-effort task and the treatment variable is a piece rate, the relationship between outcomes and payoffs takes a separable form: $\pi_g(\tau \mid \gamma, \delta) = \tau_g \mu_g(\tau \mid \gamma, \delta)$. In addition to the payoffs $\pi_g$, subjects in each treatment group receive a participation payment $w$. The total expected per-subject payoff across two groups is then $2w + \pi_C + \pi_T$.

Assume the researcher is risk-neutral and places a prior probability of $\chi \in (0, 1)$ on the existence of the effect she is trying to find. For the sake of illustration, I assume that $m_{neg}$ is her benefit from finding a true negative result, $m_{pos}$ is her benefit from finding a true positive result, and that she

then use the desired factor $k$, often called an *allocation ratio*, (computed based on the variance considerations or others), to compute the number of participants in the treatment group as $kn$.

[8]The researcher can exploit other design parameters to optimize the budget. However, those parameters are likely to be specific to each experiment. Hence, it would be difficult to obtain general results in that case.

[9]While relying on standards of significance thresholds is commonplace, the practice is not without issues (Brodeur et al., 2020; Brodeur et al., 2016).

[10]To be precise, I am calling a choice variable continuous if in the theoretical model it is a continuous function of the treatment variable, and the experiment allows subjects to make their choices among a large set of alternatives.

https://doi.org/10.1017/esa.2025.10019 Published online by Cambridge University Press

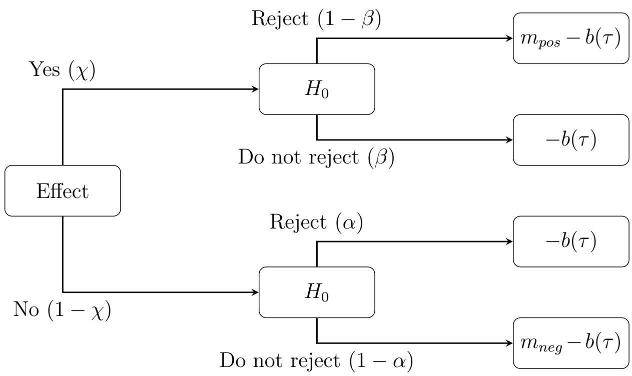


**Fig. 1** Possible outcomes and probabilities

receives zero benefits from making either a Type I or Type II errors. The researcher's budget is $b(\tau)$, which is a function of the treatment strength. In Appendix A, I show that the results hold under arbitrary benefits and arbitrary utility function (as long as it is strictly increasing), including the case of risk aversion. Figure 1 shows all four possible outcomes for the researcher that are contingent on whether the effect exists or not and whether the researcher rejects the null hypothesis or not.

Using Figure 1, it is easy to derive the researcher's expected utility function from conducting the experiment:

$$U(\tau \mid \alpha, \beta, \chi, \gamma, \delta) = \chi(1-\beta)m_{pos} + (1-\chi)(1-\alpha)m_{neg} - b(\tau \mid \alpha, \beta, \gamma, \delta). \tag{1}$$

Since the researcher's utility as a function of $\tau$ equals to the *negative* of the budget, which is also a function of $\tau$, plus a term that does not depend on $\tau$, maximizing the utility function is equivalent to minimizing the budget:

$$\min_{\tau} b(\tau \mid \alpha, \beta, \gamma, \delta) = n(\tau \mid \alpha, \beta, \gamma, \delta)\left(2w + \pi_C(\gamma, \delta) + \pi_T(\tau \mid \gamma, \delta)\right). \tag{2}$$

I refer to this dual problem as the Budget Minimization problem. I formulate this problem without any constraints for simplicity. I discuss constraints in Section 6.

The intuition for why the Budget Minimization problem is well defined is the following. The response of the budget to a change in the treatment strength depends on two effects: the *sample-size effect* and the *payoff effect*. Increasing $\tau$ is expected to increase the difference in outcomes between the treatment and control groups. The predicted effect size will increase, which in turn will drive down the required number of subjects (the sample-size effect). On the other hand, increasing $\tau$ will increase the expected per-subject payoff in the treatment group due to the direct effect of higher incentives and the indirect effect of higher outcomes due to higher incentives (the payoff effect).[11] For example, if the expected per-subject payoff in the treatment group is $\pi_T(\tau \mid \gamma, \delta) = \tau_T \mu_T(\tau \mid \gamma, \delta)$, then the increase in $\tau_T$ is the direct effect of increasing $\tau$, the increase in $\mu_T(\tau \mid \gamma, \delta)$ is the indirect effect of increasing $\tau$, and the total increase in $\pi_T$ is the payoff effect. These two opposing effects—the sample-size effect and the payoff effect—can potentially lead to a point $\tau^*$ where the expected total budget is minimized.

Formally, the following first-order necessary condition must hold at the optimal point $\tau^*$ (to avoid notational clutter, I drop the dependence on the parameters $\alpha, \beta, \gamma, \delta$):

$$-\underbrace{\frac{n'(\tau)}{n(\tau)}}_{\text{sample-size effect}} = \underbrace{\frac{\pi_T'(\tau)}{2w + \pi_C(\tau) + \pi_T(\tau)}}_{\text{payoff effect}}. \tag{3}$$

[11]The indirect, however, will not be present if behavior is insensitive to incentives.

https://doi.org/10.1017/esa.2025.10019 Published online by Cambridge University Press

The condition states, intuitively, that at the optimum the percentage decrease in the required number of subjects due to the higher treatment strength (the sample-size effect) exactly offsets the percentage increase in the per-subject payoffs (the payoff effect). The theoretical question is under what conditions the Budget Minimization problem has a non-trivial solution. Before I turn to the formal analysis of this question, I present two examples of the Budget Minimization problem at work.

## 4. Budget minimization in practice

I illustrate the Budget Minimization problem in two common cases. In the first case, subjects' choice variable is continuous and the effect of interest is the difference in mean choices. In the second case, subjects' choice variable is discrete. In this case, the effect of interest can be either the difference in proportions of subjects choosing a given alternative (binary choice) or the difference in mean choices (more than two alternatives). I focus on the former case when the choice is binary, although a similar logic would apply to the latter case.

### *4.1. Continuous case*

To illustrate the Budget Minimization problem in the continuous case, I use the experiment of DellaVigna and Pope 2018. In the experiment, subjects perform a real-effort task in which they have to repeatedly press two buttons for ten minutes. Subjects receive $w = \$1$ for their participation. A subject's choice variable is the number of button presses, a proxy for a subject's effort. The outcome variable is the average number of button presses.

Suppose that the researcher is interested in testing whether introducing a piece rate in the treatment group increases effort relative to the control group that receives no piece rate, $\tau_C = 0$. The expected per-subject payoff in group $g$ is $\pi_g = \tau_g \mu_g$. Subjects receive a piece rate for each 100 button presses. The goal is to determine the treatment strength $\tau$ that allows one to detect an increase in effort for the conventional levels of significance ($\alpha = 0.05$) and power ($1 - \beta = 0.8$) while minimizing the required budget.

DellaVigna and Pope (2018; P. 1063) propose a model of effort choice that gives the following closed-form solution for the mean effort:[12]

$$\mu_g(\tau \mid \gamma, \delta) = \frac{1}{\eta}[\ln(s + \tau_g) - \ln k]. \tag{4}$$

where $\eta$ and $k$ are the curvature and scale parameters of the cost-of-effort function, respectively, $s$ is an intrinsic reward for performing the task, and $\tau_g$ is a piece rate in group $g$.

One can find the required number of subjects per group conditional on $\tau$ and other parameters using the standard formula for computing the sample size in a two-sided $t$-test for the difference in means:

$$n(\tau \mid \alpha, \beta, \gamma, \delta) = 2\left(z_{1-\alpha/2} + z_{1-\beta}\right)^2 \left(\frac{\sigma}{\mu_T(\tau \mid \gamma, \delta) - \mu_C(\gamma, \delta)}\right)^2, \tag{5}$$

where $z_{1-\alpha/2}$ and $z_{1-\beta}$ are the quantiles of the standard normal distribution, $\mu_C$ and $\mu_T$ are the predicted mean efforts in the control and treatment groups, which can be computed using (4), and $\sigma$ is the standard deviation of effort.[13]

Figure 2 shows how the total number of subjects, the expected payoff per subject, and the total budget change with $\tau$. I compute the total number of subjects across both groups, $2n$, using (5).

[12] Specifically, I use the version of their model with the exponential cost of effort. I make several changes to the authors' original notation to make it consistent with the notation adopted in my paper. In their formula (13), I use $\eta$ instead of $\gamma$ and $\tau_g$ instead of $p$.

[13] The implicit assumption in the formula, which follows the original model in the paper, is that the standard deviation parameter $\sigma$ does not vary with $\tau$. In Section 6 I consider the case when $\sigma$ also varies with $\tau$.

https://doi.org/10.1017/esa.2025.10019 Published online by Cambridge University Press

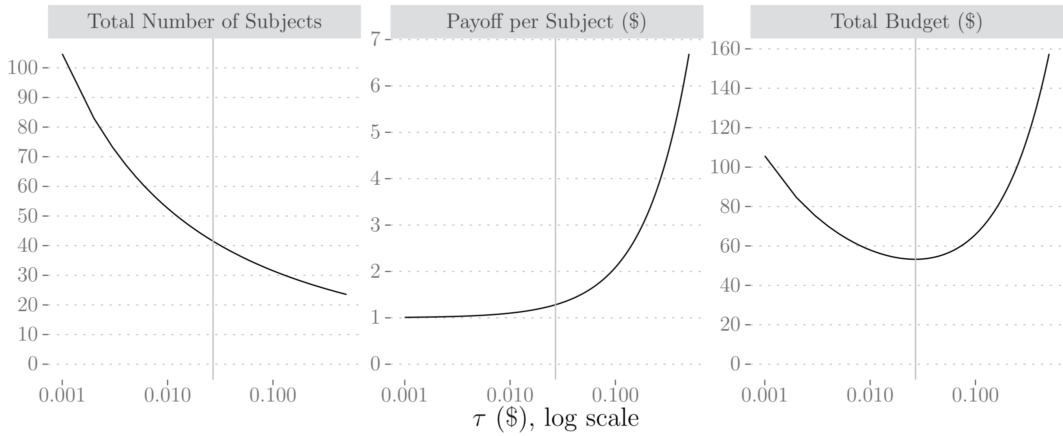


**Fig. 2** Variables of the DellaVigna and Pope 2018 Experiment as a Function of $\tau$
*Note:* The figure shows how the three parameters of the experiment change with the treatment strength $\tau$. The left panel shows the total number of subjects across both treatment groups ($2n$), which is computed using (5) and (4) and the authors' parameter estimates $\eta = 0.015641071$, $k = 1.70926702 \times 10^{-16}$, $s = 3.72225938 \times 10^{-6}$, $\sigma = 653.578104$. The middle panel shows the expected per-subject payoff in \$ across both treatment groups ($w + (\pi_C + \pi_T)/2$), which is computed by plugging in $w = 1$, $\pi_C = \tau_C\,\mu_C = 0$, and $\pi_T = \tau\,\mu_T$, and where $\mu_T$ is computed using (4). The right panel shows the expected total budget in \$ ($b$), which is the product of $2n$ and $w + (\pi_C + \pi_T)/2$. The horizontal axis shows the treatment strength $\tau$(in \$) on a logarithmic scale. The vertical solid line shows the budget-minimizing level of $\tau$.

I plug in the values of the quantiles of the standard normal distribution using the conventional levels of $\alpha = 0.05$ and $\beta = 0.2$. I use the authors' estimate of $\sigma = 653.578104$ (Supplementary Material "NLS_results_Table5_EXPON.csv:"). I use (4) to compute the predicted mean effort in the control group $\mu_C$ by plugging in the authors' estimates of behavioral parameters: $\eta = 0.015641071, k = 1.70926702 \times 10^{-16}, s = 3.72225938 \times 10^{-6}$ (Supplementary Material "NLS_results_Table5_EXPON.csv:") and using $\tau_C = 0$. Finally, I use (4) to compute the predicted mean effort in the treatment group $\mu_T$ using the same estimates of behavioral parameters as for $\mu_C$ but for different values of $\tau_T = \tau_C + \tau = \tau$. As Figure 2 shows, the total number of subjects decreases in $\tau$ since higher incentives increase the expected effect size.

To compute the expected payoff per subject across both groups, $w + (\pi_C + \pi_T)/2$, I first plug in the value of the fixed payment used in the experiment, $w = 1$. I compute the expected payoff per subject in the control group as $\pi_C = \tau_C \mu_C = 0$ (since $\tau_C = 0$). I compute the expected payoff per subject in the treatment group as $\pi_T = \tau_T \mu_T = \tau \mu_T$ for different values of treatment strength $\tau$, where $\mu_T$ is computed as before. The expected payoff per subject increases in $\tau$ since higher incentives increase expected effort, as well as the payoff per unit of effort.

I compute the expected total budget for different values of $\tau$ using (2), which is simply the product of the two previously computed quantities: the total number of subjects, $2n$, and the expected payoff per subject, $w + (\pi_C + \pi_T)/2$. The expected total budget $b$ has a convex shape and reaches a minimum at $\tau^* = 2.7$ cents.

Conducting an experiment with the optimized parameters would be extremely cheap: the experiment would require a total of 42 subjects with an expected per-subject payoff across both groups of \$1.28 and an expected total budget of just \$53. For comparison, the original experiment has 0 and 4 cents treatments, although the total number of subjects in both groups is more than 1000. While the optimal numbers appear small, they are not unreasonable given the large treatment effects found in the data. For instance, the mean effort levels in the 0 and 4 cents treatments are 1521 and 2132, respectively (DellaVigna and Pope, 2018; P. 1045, Table 3). Assuming a common standard deviation of 650, the traditional power analysis would yield 18 subjects per treatment group for the levels of significance (0.05) and power (0.8) assumed in my calculation.

https://doi.org/10.1017/esa.2025.10019 Published online by Cambridge University Press

### *4.2. Discrete choice*

To illustrate the Budget Minimization problem in the discrete-choice setting, I use the classic Holt and Laury 2002 experiment on risk aversion. In this experiment, which popularized the multiple-price-list elicitation method, subjects make a series of binary choices between a safe and a risky lottery. The alternatives are ordered such that a risky lottery gradually becomes more attractive. Experimental treatments involve changing the level of incentives by large factors to see whether this affects the proportion of subjects choosing a safe lottery.

For illustrative purposes, suppose that the researcher is interested in testing whether scaling the payoffs of each lottery up affects the proportion of subjects choosing a safe lottery in just one pair.[14] Suppose the researcher picks pair 5 (Holt and Laury, 2002; P. 1645, Table 1) in which the safe lottery pays \$2 or \$1.6 with equal chances and the risky lottery pays \$3.85 or \$0.1 with equal chances in the control group, and in which the safe lottery pays \$2$\times\tau$ or \$1.6$\times\tau$ with equal chances and the risky lottery pays \$3.85$\times\tau$ or \$0.1$\times\tau$ with equal chances in the treatment group. Here $\tau$ is the multiplicative treatment strength. The expected per-subject payoff in group $g$ is $\pi_g = \tau_g(\mu_g EV_A + (1-\mu_g)EV_B)$, where $EV_A$ and $EV_B$ are the expected values of the safe and risky lotteries, respectively, in the control group ($\tau_C = 1$) and $\mu_g$ is the proportion of subjects choosing the safe lottery in group $g$. The goal is to determine the treatment strength that allows the researcher to detect a change in the proportion of subjects choosing the safe lottery for the conventional levels of significance ($\alpha = 0.05$) and power ($1-\beta = 0.8$) while minimizing the required budget.

Holt and Laury 2002 use the stochastic choice model that specifies the probability of choosing the safe lottery in group $g$ as follows:

$$\mu_g(\tau \mid \gamma, \delta) \equiv \mathbb{P}(A)_g = \frac{U_{A_g}^{1/\lambda}}{U_{A_g}^{1/\lambda} + U_{B_g}^{1/\lambda}}, \tag{6}$$

where $U_{A_g}, U_{B_g}$ are the expected utilities of the safe and risky lotteries, respectively, in group $g$ and $\lambda$ is the noise parameter. The expected utility uses an expo-power utility-of-money function of the form[15]

$$u(x) = \frac{1-\exp(-ax^{1-r})}{a}, \tag{7}$$

where $x$ is a monetary outcome, $a$ is the constant risk aversion parameter, and $r$ is the relative risk aversion parameter.

One can find the required number of subjects per group conditional on $\tau$ and other parameters using the standard formula for computing the sample size in a test for the difference in proportions:[16]

$$n(\tau \mid \alpha, \beta, \gamma, \delta) = (z_{1-\alpha/2} + z_{1-\beta})^2 \frac{\mu_T(1-\mu_T) + \mu_C(1-\mu_C)}{(\mu_T - \mu_C)^2}, \tag{8}$$

where $z_{1-\alpha/2}$ and $z_{1-\beta}$ are the quantiles of the standard normal distribution, and $\mu_C$ and $\mu_T$ are the predicted proportions of subjects choosing the safe lottery in the control and treatment groups, computed as in (6).

Figure 3 shows how the total number of subjects, the expected payoff per subject, and the total budget change with $\tau$. I compute the total number of subjects across both groups, $2n$, using (8). I

[14] Discrete choice does not necessarily imply that the relevant outcome is the proportion of subjects choosing a given alternative. While it is true in the binary choice, in case when there is more than two alternatives a researcher might consider the difference in mean choices. In the context of Holt and Laury 2002, this could be, for example, the mean switching point. The models of stochastic discrete choice, such as the one considered here, can still be used to derive the expected outcomes in the case of more than two alternatives.

[15] I use $a$ instead of $\alpha$ in the authors' original specification ((Holt and Laury, 2002; P. 1653, formula (2))) to avoid confusion with the significance level $\alpha$. I also use $\lambda$ instead of $\mu$ in their formula (1).

[16] To avoid notational clutter, I drop the dependence of $\mu_T$ and $\mu_C$ on $\tau, \gamma, \delta$.

https://doi.org/10.1017/esa.2025.10019 Published online by Cambridge University Press

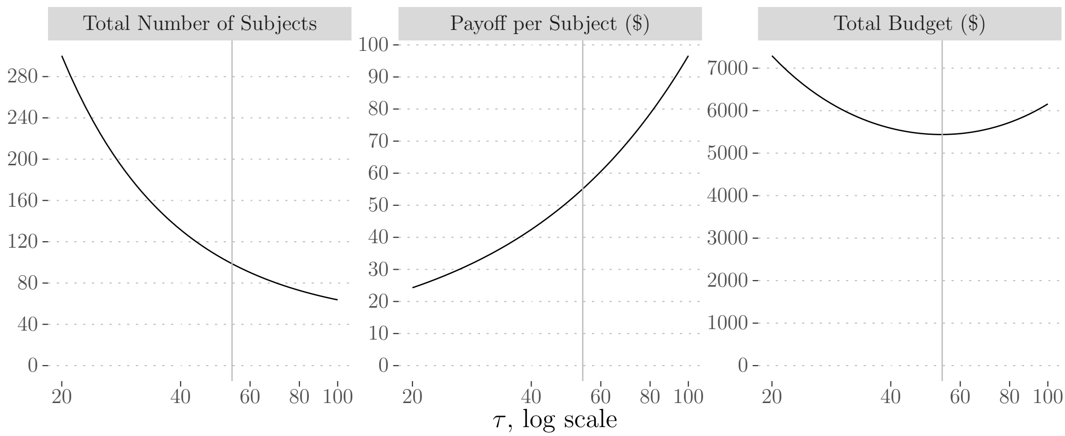


**Fig. 3** Variables of the Holt and Laury 2002 Experiment as a Function of $\tau$

*Note:* The figure shows how the three parameters of the experiment change with the treatment strength $\tau$. The left panel shows the total number of subjects across both treatment groups ($2n$), which is computed using (8), (6), (7) and the authors' parameter estimates $a = 0.029$, $r = 0.269$, $\lambda = 0.134$. The middle panel shows the expected per-subject payoff in \$ across both treatment groups ($w + (\pi_C + \pi_T)/2$), which is computed by plugging in $w = 5$, $\pi_C = \mu_C EV_A + (1 - \mu_C)\, EV_B$, $\pi_T = \tau(\mu_T\, EV_A + (1 - \mu_T)\, EV_B)$, $EV_A = 2 \times 0.5 + 1.6 \times 0.5 = 1.8$, $EV_B = 3.85 \times 0.5 + 0.1 \times 0.5 = 1.975$, and where $\mu_C$ and $\mu_T$ are computed using (6). The right panel shows the expected total budget in \$ ($b$), which is the product of $2n$ and $w + (\pi_C + \pi_T)/2$. The horizontal axis shows the multiplicative treatment strength $\tau$ on a logarithmic scale. The vertical solid line shows the budget-minimizing level of $\tau$.

plug in the values of the quantiles of the standard normal distribution using the conventional levels of $\alpha = 0.05$ and $\beta = 0.2$. In the control group, I compute the expected utility of the safe lottery as $U_{A_C} = u(2)0.5 + u(1.6)0.5$ and the expected utility of the risky lottery as $U_{B_C} = u(3.85)0.5 + u(0.1)0.5$, where the utility function $u$ is given by (7) and the estimates of behavioral parameters are $a = 0.029$, $r = 0.269$ (Holt and Laury, 2002; P. 1653). I then plug in the resulting expected utilities, along with the authors' estimate of $\lambda = 0.134$ (Holt and Laury, 2002; P. 1653), into (6) to compute the predicted proportions of subjects choosing the safe lottery in the control group, $\mu_C$. In the treatment group, I compute the expected utility of the safe lottery as $U_{A_T} = u(2\tau)0.5 + u(1.6\tau)0.5$ and the expected utility of the risky lottery as $U_{B_T} = u(3.85\tau)0.5 + u(0.1\tau)0.5$ for different values of $\tau$. The utility function $u$ is computed as in the control group. I then plug in the resulting expected utilities into (6) to compute the predicted proportions of subjects choosing the safe lottery in the treatment group, $\mu_T$, for different values of $\tau$. Finally, I plug in the resulting values for $\mu_C$ and $\mu_T$ into (8) to get the total number of subjects as a function of $\tau$. As Figure 3 shows, the total number of subjects across both groups decreases in $\tau$.

I compute the expected payoff per subject across both groups as $w + (\pi_C + \pi_T)/2$. While the participation payment is not explicitly mentioned in the text, I assume $w = \$5$, which is a typical amount for laboratory experiments. I compute the expected payoff per subject in the control group as $\pi_C = \tau_C(\mu_C EV_A + (1 - \mu_C)EV_B) = \mu_C EV_A + (1 - \mu_C)EV_B$, since $\tau_C = 1$. The proportion of subjects choosing the safe lottery, $\mu_C$, is computed as above, and the expected values are computed as $EV_A = 2 \times 0.5 + 1.6 \times 0.5 = 1.8$ and $EV_B = 3.85 \times 0.5 + 0.1 \times 0.5 = 1.975$. I compute the expected payoff per subject in the treatment group as $\pi_T = \tau(\mu_T EV_A + (1 - \mu_T)EV_B)$ for different values of treatment strength $\tau$, where $\mu_T$ and the expected values are computed as before. As Figure 3 shows, the expected payoff per subject increases in $\tau$.

I compute the expected total budget for different values of $\tau$ using (2), which is simply the product of the two previously computed quantities: the total number of subjects, $2n$, and the expected payoff per subject, $w + (\pi_C + \pi_T)/2$. The expected total budget has a convex shape, as in the previous example,

https://doi.org/10.1017/esa.2025.10019 Published online by Cambridge University Press

and reaches a minimum at $\tau^* = 54$ (rounded to the nearest digit). This means that the payoffs need to be scaled by more than 50 times.

For the optimized parameters, the experiment would require a total of 99 subjects with an expected per-subject payoff across both groups of \$ 55.1 and an expected total budget of \$ 5439. For comparison, the original experiment does have a 50x treatment, although the number of subjects in this group is only 19.

## 5. Budget minimization in theory

I make two assumptions about the outcome function $\mu_T(\tau)$ to establish a theoretical result.

**Assumption 1 (Continuous Differentiability).** $\mu_T \in C^1$.

**Assumption 2 (Regularity).** $\lim_{\tau \to \tau^{low}} |\mu_T'| < \infty$ *and* $\lim_{\tau \to \infty} d \ln \mu_T / d \ln \tau < 1$.[17]

The first assumption is a technical one. The second assumption takes care of the case when $\mu_T$ is unbounded. In this case, it has to satisfy regularity conditions that require the outcome function (a) not to change too quickly when treatment increases from the lowest value and (b) that the elasticity of the outcome with respect to $\tau$ is small as the treatment strength gets large. Assumption 2 is satisfied automatically if $\mu_T$ is bounded.

**Proposition 5.1.** *If $\mu_T$ satisfies Assumptions 1 and 2 the Budget Minimization problem has an interior solution.*

***Proof.*** *See Appendix B.* □

The idea of the proof relies on the Intermediate Value Theorem and the properties of the two components of the total budget: the sample size and expected payoffs.[18] I consider the limiting behavior of the derivative of the logarithm of the total budget with respect to $\tau$. At the lower limit, when the treatment strength approaches the lower bound, the derivative of the budget goes to negative infinity. The driver behind this result is the required sample size. When the treatment strength is zero (additive case) or one (multiplicative case) the outcomes in the treatment and control groups are identical, which makes the required sample size infinite. Even the smallest increase in the treatment strength is enough to produce an infinitely large decrease in the required sample size. At the lower limit, therefore, the negative sample-size effect dominates the positive payoff effect.[19] When the treatment strength is infinitely large, neither the required sample size nor the expected payoffs change. The derivative of the total budget in the limit is zero. However, one can always find a large enough value of the treatment strength at which the derivative of the total budget is positive. At the upper limit, therefore, the positive payoff effect dominates the negative sample-size effect.[20] The derivative of the total budget is thus negative at the left endpoint and positive at the right endpoint. Since $\mu_T$ is continuously differentiable by Assumption 1, the Intermediate Value Theorem implies that the derivative of the total budget must cross zero. Since the first crossing will occur from below, the First Order Sufficient

[17] Here $\tau^{low}$ denotes the lowest possible value of $\tau$. It is 0 for additive treatment strength and 1 for multiplicative treatment strength.

[18] One might wonder if the Weierstrass theorem would suffice instead. It would not: even if one is willing to impose an upper bound on $\tau$ (which is a priori unclear), the Weierstrass theorem cannot say anything about an interior solution, which is the interesting case.

[19] If the outcome function is unbounded, the first part of Assumption 2 guarantees that.

[20] If the outcome function is unbounded, the second part of Assumption 2 guarantees that.

https://doi.org/10.1017/esa.2025.10019 Published online by Cambridge University Press

Condition for a Minimum implies that the point $\tau^*$ at which this happens must be a minimum point.

The result in Proposition 5.1 is surprisingly general. It applies both in the continuous and discrete cases. The assumptions required for the result are fairly weak. The discrete case effectively only requires Assumption 1, since the outcome is a proportion bounded between zero and one. The continuous case would in addition require Assumption 2 only if the outcome function is unbounded.

Proposition 5.1 explains why the motivating examples work. In the discrete case example, only Assumption 1 needs to be checked. Indeed, since the utility-of-money function (7) is continuously differentiable, so are the expected utility and outcome (6) functions. Proposition 5.1 immediately applies. In the continuous case example, the outcome function (4) is continuously differentiable but unbounded, hence we need to check Assumption 2, as well. First, consider

$$\lim_{\tau\to 0^+} |\mu_T'| = \lim_{\tau\to 0^+} \frac{1}{\eta(s+\tau)} = \frac{1}{\eta s}.$$

The limit is finite, since the estimates of $s$ and $\eta$ are strictly positive. On the other hand,

$$\lim_{\tau\to\infty} \tau(\ln \mu_T)' = \lim_{\tau\to\infty} \frac{\tau}{(s+\tau)\ln\left(\frac{s+\tau}{k}\right)} = \lim_{\tau\to\infty} \frac{1}{\left(\frac{s}{\tau}+1\right)\ln\left(\frac{s+\tau}{k}\right)} = 0 < 1,$$

provided that $k > 0$, which is indeed the case given the model estimates. Hence, Proposition 5.1 also applies.

A few remarks about the theoretical result are in order. The first remark is that Assumptions 1 and 2 are sufficient but not necessary. It might as well be that they are not satisfied but the Budget Minimization problem has an interior solution. The second, and related, remark is that Assumption 1 can have a bite in some cases. It might fail to hold in reference-dependent models, which feature a discontinuity around a reference point. The budget, however, is still likely to have a minimum. The third, and final, remark is that Proposition 5.1 guarantees the existence but not the uniqueness of a solution. It is safe to assume that it should not cause any issues in practice. If there are several minimum points, one can simply compute the budget at each of the candidate solutions and pick the one giving the smallest budget.

## 6. Discussion

In this section, I propose some extensions of the Budget Minimization problem and show that its applicability goes beyond the examples analyzed so far. I also discuss some of the limits of its applicability.

### *6.1. Qualitative hypotheses*

A key assumption that enables studying the optimal level of incentives in the present framework is that a researcher is interested in testing qualitative hypotheses, for example, whether increasing incentives increases a certain behavior or outcomes. The qualitative nature of a hypothesis creates a degree of freedom in the level of incentives that I exploit in the Budget Minimization problem. However, sometimes researchers are interested in specific values of a treatment variable, in which case the present framework is not applicable. For example, researchers might need to use several specific levels of incentives to estimate a structural model or identify a non-linear effect over that range of levels. In these cases, the levels of incentives are determined by identification concerns and cannot be used to optimize the budget. Instead, researchers should use the guidelines for how to optimally arrange their sample across those different levels of the treatment variable (McClelland, 1997; List et al., 2011).

https://doi.org/10.1017/esa.2025.10019 Published online by Cambridge University Press

### 6.2. Continuous treatment variable

Another key assumption is that a researcher can vary the level of incentives in a continuous manner, which enables the use of calculus to optimize the budget. While incentives can be typically varied that way, sometimes researchers might consider a few discrete levels, for example, for procedural reasons. Budget minimization is still possible in this case. A researcher can find the budget-minimizing level of incentives by simply evaluating the expected budget at those few discrete levels and picking the one that minimizes the budget. On the other hand, if a researcher cannot vary the level of incentives at all, the Budget Minimization problem is not applicable. What makes the Budget Minimization problem possible is the trade-off that incentives create between the sample-size effect and the payoff effect. Changing statistical parameters, for example, the power, only affects the sample-size effect but not the payoff effect, hence there will be no optimal level of power.

### 6.3. Strategic settings

Even though the examples I considered are from individual-choice settings, the logic of the Budget Minimization problem carries over to strategic settings. The natural counterpart to the theoretical outcome function $\mu_T$, such as (6), in game theory is the Quantal Response Function (McKelvey and Palfrey, 1995; Goeree et al., 2005). By combining, for instance, the framework developed by Woods (2020) for the quantal response model with the present approach, one can pose and solve the Budget Minimization problem in game-theoretic experiments.

### 6.4. Parameter uncertainty

The solution to the Budget Minimization problem relies on the estimates of the structural parameters of a model. These estimates will have standard errors. The analysis conducted in motivating examples ignores this parameter uncertainty for simplicity. However, the budget-minimizing treatment strength is a function of parameters and hence inherits the uncertainty in their estimates. The optimal treatment strength is unlikely to have a closed-form solution in most cases, hence, using the Delta method would be impossible. A practical solution to deriving the standard errors of the treatment strength would be to use the bootstrap.

### 6.5. Parameter estimates

A related point about parameter estimates is that they have to exist in order to take advantage of the Budget Minimization problem.[21] In the best-case scenario, these estimates could be readily available from the literature. This is likely to be the case for the models of risk and time preferences (Harrison and Rutström, 2008b), lying aversion (Abeler et al., 2019), social preferences (Goeree et al., 2002; Cox et al., 2007; Bellemare et al., 2008), and real-effort tasks (DellaVigna and Pope, 2018). But what should a researcher do when those estimates are not available or cannot be used?

One possibility is that a researcher can use an existing structural model but does not want to use existing parameter estimates. Using existing estimates might not be reliable if, for example, they are derived from a subject pool that is very different from a researcher's subject pool. In other words, a researcher might worry about the portability of the existing estimates. A solution in this case is to run pilot sessions on the subject pool of interest and estimate the parameters of the model using the pilot data. Using pilots to conduct the power analysis is a standard practice in experimental economics, and the only modification to that practice would be the way the data are used. An alternative solution is to exploit an auxiliary variation in the control group that

[21] This is an issue not just for the Budget Minimization problem but for optimal experimental design in general (List et al., 2011; Moffatt, 2015).

https://doi.org/10.1017/esa.2025.10019 Published online by Cambridge University Press

is not related to the treatment variation of interest. For example, experiments on risk and uncertainty preferences involve variation in prospects that allows one to estimate behavioral parameters. A researcher then can use these estimated parameters to optimize the design of the treatment of interest.

### 6.6. Structural model

A more fundamental issue is that an off-the-shelf structural model simply might not exist. In this case, researchers have two possibilities. They can come up with their own model and run pilot experiments, as suggested above, to get initial parameter estimates needed for calculations. Another option would be to use a reduced-form approach instead of a structural approach. The Budget Minimization problem, at its core, relies on knowing how the outcome variable changes with the treatment strength, $\mu_T(\tau)$. Nothing in the logic of the problem requires that this relation comes from a structural model. If there are previous observations on $\tau$ and $\mu_T$, a researcher can use a reduced-form, predictive approach to recover $\mu_T(\tau)$ and then use it in the Budget Minimization problem.

### 6.7. Expected outcomes

The analysis of the Budget Minimization problem has so far focused on the case when the treatment strength affects only expected outcomes. The researcher, however, can also use information on how the treatment strength affects other moments of the distribution of outcomes, or even the whole distribution itself.[22] Using this additional information will make the analysis more efficient. For example, the formula for computing the sample size in a two-sided $t$-test for the difference in means (5) relies on knowing the standard deviation $\sigma$. If the researcher knows how the standard deviation changes with the treatment strength, $\sigma(\tau)$, she can use this information to derive better predictions about how the treatment strength affects the sample size.

A common finding is that higher incentives reduce the standard deviation of outcomes, that is, $\sigma(\tau)$ is likely to be a decreasing function (Camerer and Hogarth, 1999). The sample-size effect will become stronger relative to the case when $\sigma$ does not change with $\tau$. The standardized effect size $\left(\frac{\mu_T(\tau|\gamma,\delta)-\mu_C(\gamma,\delta)}{\sigma(\tau)}\right)$ will increase faster in $\tau$, which will cause the required sample size to decrease faster. In other words, the same increase in $\tau$ will now result in a bigger reduction in the required sample size. The sample-size effect will be present even if the treatment strength affects only the standard deviation and has no effect on the difference expected outcomes, as long as this difference is non-zero: the standardized effect size will still increase in $\tau$.

As an illustration, let us revisit the continuous example from Section 4.1 but now assume that $\sigma$ in the formula (5) is linearly decreasing in $\tau$: $\sigma(\tau) = 653.578104 - 500\tau$. Conducting an experiment with the re-optimized parameters would now require a total of 37 subjects (a decrease from 42 in case of constant $\sigma$) with an expected per-subject payoff across both groups of \$ 1.36 (a slight increase from \$ 1.28) and an expected total budget of \$ 51 (a slight decrease from \$ 53). The budget-minimizing treatment strength would be $\tau^* = 3.4$ cents, which is slightly higher than 2.7 in case of constant $\sigma$.

### 6.8. Constraints

I have presented and analyzed the Budget Minimization problem as an unconstrained problem. In reality, a researcher might face constraints on subjects' payoffs and/or a sample size. Suppose the

[22] Woods (2020) shows, in particular, that the skewness of the distribution of outcomes can have an impact on power analysis.

https://doi.org/10.1017/esa.2025.10019 Published online by Cambridge University Press

sample size at $\tau^*$ is too low to be acceptable (the constraint binds), as we saw in the continuous case example. A researcher can simply tweak the statistical parameters: decreasing $\alpha$ or $\beta$ increases the optimal sample size without changing the optimal treatment strength. Suppose now that the expected per-subject payoffs are too low at $\tau^*$. In this case the optimal treatment strength will have to change. There are several possibilities to satisfy the constraint in that case. One possibility is to change the level of the treatment variable in the control group, re-optimize, and check if the constraint is satisfied. The benefit of this approach is that one can both satisfy the constraint and get an optimal level of $\tau$. Another possibility is to keep increasing $\tau$ until the constraint is satisfied. This approach will distort $\tau$ away from the budget-minimizing level. However, it can be more cost-effective than increasing $\tau_C$. One might also consider changing the participation payment $w$, which will change $\tau^*$. The participation payment, however, is typically set by lab policies and rarely tweaked for the purposes of a particular experiment.[23] On the other end of the spectrum is the case when the expected per-subject payoffs are too high. No simple solution exists in this case, since $\tau^*$ already minimizes the budget and any deviation will only increase it. A researcher would likely have to reconsider other parameters of the design to bring down the budget.

### *6.9. Non-parametric tests*

In practice, researchers often use non-parametric tests, such as the Wilcoxon-Mann-Whitney test, to analyze treatment effects. The reason for relying on parametric tests in my analysis is that they have simple analytical formulas for power calculations and require only minimal predictions about outcomes, such as averages. Power analysis for non-parametric tests, on the other hand, is based either on simulations in which case deriving theoretical results is impossible, or on explicit formulas that require rich predictions about outcomes, such as the entire distribution of outcomes (Rahardja et al., 2009; Happ et al., 2019). One can still pose a practical question about the optimal level of incentives for a non-parametric test, or other more complicated designs, in a given experiment and combine simulations (Bellemare et al., 2016) with the present framework to solve the Budget Minimization problem.

## 7. Conclusion

I study an optimal design of incentives in experiments where incentives are a treatment variable. Using a utility-based framework, I formulate a Budget Minimization problem. In the problem, a researcher chooses a treatment strength that minimizes the expected budget while allowing for the detection of an effect at the given levels of statistical significance and power. The effect of the treatment strength on the budget can be decomposed into two channels: the sample-size effect and the payoff effect. Increasing the treatment strength decreases the required budget via the sample-size effect but increases it via the payoff effect. At a minimum point, the two effects must be in the exact balance. I show theoretically that such a point exists under fairly mild conditions, and thus the Budget Minimization problem is guaranteed to have a non-trivial solution. I illustrate how the Budget Minimization problem applies in practice using existing experiments. The Budget Minimization problem also applies, under certain conditions, to designs where a treatment variable is not monetary incentives.

The main challenge in taking advantage of my approach is having a model of how the outcomes respond to incentives and reliable prior estimates of the model, in other words, good prior data, albeit this is true in general for any optimal design. The main contribution of my analysis is that it takes the guesswork out of the design of the level of incentives and replaces it with a disciplined economic approach. I believe that my approach to the design of incentives will enrich experimental economists' toolkit and help guide future designs. Young researchers on tight budgets and researchers

[23] A notable exception is when the participation payment *is* the treatment variable (Harrison et al., 2009).

https://doi.org/10.1017/esa.2025.10019 Published online by Cambridge University Press

running expensive field interventions will particularly benefit from using the Budget Minimization problem.

**Supplementary material.** The supplementary material for this article can be found at https://10.1017/esa.2025.10019.

**Acknowledgements.** I thank the Editor (Lionel Page) and anonymous reviewers whose detailed suggestions helped significantly improve the quality of the paper. I thank James Bland, Jim Cox, Glenn Harrison, P. J. Healy, Marco Lambrecht, Lily Li, Nate Neligh, and David Rojo-Arjona for valuable suggestions on the early drafts. I thank the seminar participants at the Helsinki Graduate School of Economics, LMU Munich, University of Regensburg and the conference participants at the 2023 ESA World Meeting and 2021 ESA Global Online Around-the-Clock Conference for helpful comments. The replication code is available at https://github.com/aalexee/power_incentives. All remaining errors are my own.

**Competing interests.** I declare that I have no interests, financial or non-financial, that relate, directly or indirectly, to the research described in this paper.

## References

Abeler, J., Nosenzo, D., & Raymond, C. (2019). Preferences for truth-telling. *Econometrica*, *87*(4), 1115–1153.

Alekseev, A. (2022). Give me a challenge or give me a raise. *Experimental Economics*, *25*(1), 170–202.

Amir, O., Rand, D. G., & Gal, Y. K. (2012). Economic games on the internet: The effect of $1 stakes. *PloS one*, *7*(2), e31461.

Araujo, F. A., Carbone, E., Conell-Price, L., Dunietz, M. W., Jaroszewicz, A., Landsman, R., Lamé, D., Vesterlund, L., Wang, S. W., & Wilson, A. J. (2016). The slider task: An example of restricted inference on incentive effects. *Journal of the Economic Science Association*, *2*(1), 1–12.

Atkinson, A. C. (2018). The usefulness of optimum experimental designs. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 59–76.

Azrieli, Y., Chambers, C. P., & Healy, P. J. (2018). Incentives in experiments: A theoretical analysis. *Journal of Political Economy*, *126*(4), 1472–1503.

Azrieli, Y., Chambers, C. P., & Healy, P. J. (2020). Incentives in experiments with objective lotteries. *Experimental Economics*, *23*(1), 1–29.

Balasubramanian, P., Bennett, V. M., & Pierce, L. (2017). The wages of dishonesty: The supply of cheating under high-powered incentives. *Journal of Economic Behavior & Organization*, *137*, 428–444.

Bellemare, C., Bissonnette, L., & Kröger, S. (2016). Simulating power of economic experiments: The powerBBK package. *Journal of the Economic Science Association*, *2*(2), 157–168.

Bellemare, C., Kröger, S., & Van Soest, A. (2008). Measuring inequity aversion in a heterogeneous population using experimental decisions and subjective probabilities. *Econometrica*, *76*(4), 815–839.

Bellemare, C., Sebald, A., & Suetens, S. (2018). Heterogeneous guilt sensitivities and incentive effects. *Experimental Economics*, *21*(2), 316–336.

Brañas-Garza, P., Kujal, P., & Lenkei, B. (2019). Cognitive reflection test: Whom, How, When. *Journal of Behavioral and Experimental Economics*, *82*, 101455.

Brodeur, A., Cook, N., & Heyes, A. (2020). Methods matter: P-hacking and publication bias in causal analysis in economics. *American Economic Review*, *110*(11), 3634–3660.

Brodeur, A., Lé, M., Sangnier, M., & Zylberberg, Y. (2016). Star wars: The empirics strike back. *American Economic Journal: Applied Economics*, *8*(1), 1–32.

Camerer, C. F., Dreber, A., & Johannesson, M. (2019). Replication and other practices for improving scientific quality in experimental economics. In A. Schram, & A. Ule (Eds.), *Handbook of research methods and applications in experimental economics* (pp. 83–102). Cheltenham, UK: Edward Elgar Publishing.

Camerer, C. F., & Hogarth, R. M. (1999). The effects of financial incentives in experiments: A review and capital-labor-production framework. *Journal of Risk and Uncertainty*, *19*(1), 7–42.

Carpenter, J., & Huet-Vaughn, E. (2019). Real-Effort Tasks. In A. Schram, & A. Ule (Eds.) *Handbook of Research Methods and Applications in Experimental Economics* (pp. 368–383). Cheltenham, UK: Edward Elgar Publishing.

Cloos, J., Greiff, M., & Rusch, H. (2023). Editorial favoritism in the field of laboratory experimental economics. *Journal of Behavioral and Experimental Economics*, *107*, 102082.

Cox, J. C., Friedman, D., & Gjerstad, S. (2007). A tractable model of reciprocity and fairness. *Games and Economic Behavior*, *59*(1), 17–45.

Cox, J. C., & Grether, D. M. (1996). The preference reversal phenomenon: Response mode, markets and incentives. *Economic Theory*, *7*(3), 381–405.

Cox, J. C., & Sadiraj, V. (2019). Incentives. In A. Schram A. Ule (Eds.) *Handbook of Research Methods and Applications in Experimental Economics* (pp. 9–27). Cheltenham, UK: Edward Elgar Publishing.

Cox, J. C., Sadiraj, V., & Schmidt, U. (2014). Paradoxes and mechanisms for choice under risk. *Experimental Economics*, *18*(2), 1–36.

Cox, J. C., Smith, V. L., & Walker, J. M. (1992). Theory and misbehavior of first-price auctions: Comment. *The American Economic Review*, *82*(5), 1392–1412.

Cummings, R. G., Elliott, S., Harrison, G. W., & Murphy, J. (1997). Are hypothetical referenda incentive compatible?. *Journal of Political Economy*, *105*(3), 609–621.

Cummings, R. G., Harrison, G. W., & Rutström, E. E. (1995). Homegrown values and hypothetical surveys: Is the dichotomous choice approach incentive-compatible?. *The American Economic Review*, *85*(1), 260–266.

DellaVigna, S., & Pope, D. (2018). What motivates effort? Evidence and expert forecasts. *The Review of Economic Studies*, *85*(2), 1029–1069.

El Harbi, S., Bekir, I, Grolleau, G., & Sutan, A. (2015). Efficiency, equality, positionality: What Do people maximize? Experimental vs. hypothetical evidence from tunisia. *Journal of Economic Psychology*, *47*, 77–84.

Enke, B., Gneezy, U., Hall, B., Martin, D., Nelidov, V., Offerman, T., & van de Ven, J. (2023). Cognitive biases: Mistakes or missing stakes? *The Review of Economics and Statistics*, *105*(4), 818–832.

Fedorov, V. V. (1972). *Theory of Optimal Experiments*. Academic Press.

Fischbacher, U., & Föllmi-Heusi, F. (2013). Lies in Disguise—An experimental study on cheating. *Journal of the European Economic Association*, *11*(3), 525–547.

Ford, I, Torsney, B., & Wu, C. F. J. (2018). The use of a canonical form in the construction of locally optimal designs for non-linear problems. *Journal of the Royal Statistical Society: Series B (Methodological)*, *54*(2), 569–583.

Friedman, D. (1992). Theory and misbehavior of first-price auctions: Comment. *The American Economic Review*, *82*(5), 1374–1378.

Friedman, D., & Sunder, S. (1994). *Experimental Methods: A Primer for Economists*. Cambridge University Press.

Gibson, R., Tanner, C., & Wagner, A. F. (2013). Preferences for truthfulness: Heterogeneity among and within individuals. *American Economic Review*, *103*(1), 532–548.

Gneezy, U., Meier, S., & Rey-Biel, P. (2011). When and Why incentives (Don't) work to modify behavior. *Journal of Economic Perspectives*, *25*(4), 191–210.

Goeree, J. K., Holt, C. A., & Laury, S. K. (2002). Private costs and public benefits: Unraveling the effects of altruism and noisy behavior. *Journal of Public Economics*, *83*(2), 255–276.

Goeree, J. K., Holt, C. A., & Palfrey, T. R. (2005). Regular quantal response equilibrium. *Experimental Economics*, *8*(4), 347–367.

Grether, D. M., & Plott, C. R. (1979). Economic theory of choice and the preference reversal phenomenon. *American Economic Review*, *69*(4), 623–638.

Happ, M., Bathke, A. C., & Brunner, E. (2019). Optimal sample size planning for the Wilcoxon-Mann-Whitney test. *Statistics in Medicine*, *38*(3), 363–375.

Harrison, G. W. (1989). Theory and misbehavior of first-price auctions. *American Economic Review*, *79*(4), 749–762.

Harrison, G. W. (1992). Theory and misbehavior of first-price auctions: Reply. *American Economic Review*, *82*(5), 1426–1443.

Harrison, G. W. (1994). Expected Utility Theory and the Experimentalists. In J. D. Hey (ed.) *Experimental Economics*. pp. 43–73 Heidelberg, Physica-Verlag HD.

Harrison, G. W. (2006). Chapter 3 - Hypothetical Bias Over Uncertain Outcomes. In J. A. List (ed.) *Using Experimental Methods in Environmental and Resource Economics*. Edward Elgar Publishing.

Harrison, G. W. (2007). Chapter 4 - Making Choice Studies Incentive Compatible. In B. J. Kanninen. (ed.) *Valuing Environmental Amenities Using Stated Choice Studies: A Common Sense Approach to Theory and Practice*. pp. 67–110 Dordrecht, Springer Netherlands.

Harrison, G. W. (2024). Real choices and hypothetical choices. In *Handbook of Choice Modelling*. Edward Elgar Publishing, 246–275.

Harrison, G. W., Lau, M. I., & Rutström, E. E. (2009). Risk attitudes, randomization to treatment, and self-selection into experiments. *Journal of Economic Behavior & Organization*, *70*(3), 498–507.

Harrison, G. W., & Rutström, E. E. (2008a). Chapter 81 - Experimental Evidence on the Existence of Hypothetical Bias in Value Elicitation Methods. In C. R. Plott V. L. Smith (eds.) *Handbook of Experimental Economics Results* 1 of *Handbook of Experimental Economics Results* pp. 752–767 Elsevier.

Harrison, G. W., & Rutström, E. E. (2008b). Risk Aversion in the Laboratory. In J. C. Cox G. W. Harrison (eds.) *Risk Aversion in Experiments (Research in Experimental Economics)*. Bingley, Emerald Group Publishing Limited, 12. pp. 41–196.

Harrison, G. W., & Swarthout, J. T. (2014). Experimental payment protocols and the Bipolar Behaviorist. *Theory and Decision*, *77*(3), 423–438.

Hertwig, R., & Ortmann, A. (2001). Experimental practices in economics: A methodological challenge for psychologists? *Behavioral and Brain Sciences*, *24*(3), 383–403.

Holt, C. A. (2019). *Markets, Games, and Strategic Behavior: An Introduction to Experimental Economics*. Princeton University Press.

Holt, C. A., & Laury, S. K. (2002). Risk aversion and incentive effects. *American Economic Review*, *92*(5), 1644–1655.

https://doi.org/10.1017/esa.2025.10019 Published online by Cambridge University Press

Johnson, C., Baillon, A., Bleichrodt, H., Li, Z., van Dolder, D., & Wakker, P. P. (2021). Prince: An improved method for measuring incentivized preferences. *Journal of Risk and Uncertainty*, *62*(1), 1–28.

Kleinlercher, D., & Stöckl, T. (2018). On the provision of incentives in finance experiments. *Experimental Economics*, *21*(1), 154–179.

Larney, A., Rotella, A., & Barclay, P. (2019). Stake size effects in ultimatum game and dictator game offers: A meta-analysis. *Organizational Behavior and Human Decision Processes*, *151*, 61–72.

Lazear, E. P. (2018). Compensation and incentives in the workplace. *Journal of Economic Perspectives*, *32*(3), 195–214.

Leibbrandt, A., & Lynham, J. (2018). Does the paradox of plenty exist? Experimental evidence on the curse of resource abundance. *Experimental Economics*, *21*(2), 337–354.

Li, Y. (2021). The ABC mechanism: An incentive compatible payoff mechanism for elicitation of outcome and probability transformations. *Experimental Economics*, *24*(3), 1019–1046.

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.

McClelland, G. H. (1997). Optimal design in psychological research. *Psychological Methods*, *2*(1), 3–19.

McKelvey, R. D., & Palfrey, T. R. (1995). Quantal response equilibria for normal form games. *Games and Economic Behavior*, *10*(1), 6–38.

Mengel, F. (2017). Risk and temptation: A meta-study on prisoner's dilemma games. *The Economic Journal*, *128*(616), 3182–3209.

Merlo, A., & Schotter, A. (1992). Theory and misbehavior of first-price auctions: Comment. *The American Economic Review*, *82*(5), 1413–1425.

Moffatt, P. G. (2007). Optimal Experimental Design in Models of Decision and Choice. In M. Boumans (ed.) *Measurement in Economics: A Handbook*. pp. 357–377 London, Academic Press.

Moffatt, P. G. (2015). *Experimetrics: Econometrics for Experimental Economics*. Palgrave Macmillan.

Monroe, B. A. (2020). The statistical power of individual-level risk preference estimation. *Journal of the Economic Science Association*, *6*(2), 168–188.

Nape, S., Frykblom, P., Harrison, G. W., & Lesley, J. C. (2003). Hypothetical bias and willingness to accept. *Economics Letters*, *78*(3), 423–430.

Parravano, M., & Poulsen, O. (2015). Stake size and the power of focal points in coordination games: Experimental evidence. *Games and Economic Behavior*, *94*, 191–199.

Pulford, B. D., Colman, A. M., & Loomes, G. (2018). Incentive magnitude effects in experimental games: Bigger is not necessarily better. *Games*, *9*(1), 4.

Rahardja, D., Zhao, Y. D., & Qu, Y. (2009). Sample size determinations for the wilcoxon–mann–whitney test: A comprehensive review. *Statistics in Biopharmaceutical Research*, *1*(3), 317–322.

Rousu, M. C., Corrigan, J. R., Harris, D., Hayter, J. K., Houser, S., Lafrancois, B. A., Onafowora, O., Colson, G., & Hoffer, A. (2015). Do monetary incentives matter in classroom experiments? Effects on course performance. *The Journal of Economic Education*, *46*(4), 341–349.

Rutström, E. E., & Wilcox, N. T. (2009). Stated beliefs versus inferred beliefs: A methodological inquiry and experimental test. *Games and Economic Behavior*, *67*(2), 616–632.

Schier, U. K., Ockenfels, A., & Hofmann, W. (2016). Moral values and increasing stakes in a dictator game. *Journal of Economic Psychology*, *56*, 107–115.

Silvey, S. (1980). *Optimal Design: An Introduction to the Theory for Parameter Estimation*. Chapman & Hall.

Smith, V. L. (1976). Experimental economics: Induced value theory. *American Economic Review*, *66*(2), 274–279.

Smith, V. L. (1982). Microeconomic systems as an experimental science. *American Economic Review*, *72*(5), 923–955.

Smith, V. L., & Walker, J. M. (1993). Rewards, experience and decision costs in first price auctions. *Economic Inquiry*, *31*(2), 237–244.

Thielmann, I, Heck, D. W., & Hilbig, B. E. (2016). Anonymity and incentives: An investigation of techniques to reduce socially desirable responding in the trust game. *Judgment and Decision Making*, *11*(5), 527–536.

Vasilaky, K. N., & Brock, J. M. (2020). Power(ful) guidelines for experimental economists. *Journal of the Economic Science Association*, *6*(2), 189–212.

Voslinsky, A., & Azar, O. H. (2021). Incentives in experimental economics. *Journal of Behavioral and Experimental Economics*, *93*, 101706.

Woods, D. (2020). Improving ex-ante power analysis with quantal response simulations. *Working Paper*, Purdue University. https://woodsd42.github.io/files/JMP.pdf

Yamagishi, T., Li, Y., Matsumoto, Y., & Kiyonari, T. (2016). Moral bargain hunters purchase moral righteousness when it is cheap: Within-individual effect of stake size in economic games. *Scientific Reports*, *6*, 27824.

**Cite this article:** Alekseev, A. (2025). The (Statistical) Power of Incentives. *Journal of the Economic Science Association*, 1–18. https://doi.org/10.1017/esa.2025.10019