Improving Ex-ante Power Analysis With Quantal Response Simulations

Daniel Woods

Purdue University

November 24, 2020

Click here for the latest version.

1 Abstract

Power, the probability of detecting an effect when a true effect exists, is an important but under-considered concept in empirical research. Power analysis, i.e., selecting the number of observations to obtain a given power, is a useful process to avoid issues of low power in experimental research. However, it is often not clear ex-ante what the required parameters for a power analysis, like the effect size and standard deviation, should be. Previous studies, the best guess for these parameters, are likely inaccurate for the novel environments researchers seek to investigate. This paper proposes the use of Quantal Choice/Response (QR) simulations for ex-ante power analysis, as it can map related data-sets into predictions for novel environments. As the QR is a frequently used and successful model of choice behavior, it should improve the accuracy of an ex-ante power analysis. The QR can also guide optimal design decisions, both ex-ante as well as ex-post for conceptual replication studies. This paper demonstrates QR simulations on a wide variety of applications, and finds it is a useful tool for power analysis and experimental design.

2 Introduction

Statistical power, the likelihood of detecting an effect when a true effect exists, is an important consideration in any empirical research, both inside and outside of economics. Low power is problematic by definition, but even when an effect is detected in a low powered study, the finding is either likely to be false (Ioannidis, 2005), or the estimated effect size can be exaggerated or in the wrong direction (Gelman and Carlin, 2014). These

issues contribute to the replication or credibility 'crisis' that affects a wide range of scientific disciplines. It also can lead to research efforts being misplaced, with theories or extensions being proposed on effects that do not actually exist. Power is frequently ignored in economics, although that is becoming less true over time.[1] As a result, low power is pervasive in empirical economics research, which is clearly a situation that needs to be addressed.[2] Unlike other empirical fields, experimental research (such as lab or field experiments) has more control over the power of a study. In particular, the number of observations in a study is a decision variable of the researcher, where increasing observations increases power. Despite this, economics experiments have also historically neglected the consideration of statistical power, and thus also have exhibited low power.[3]

Power analysis is the process of determining an appropriate sample size for an experiment. As a result of previous low power in the field, power analysis is rightfully becoming increasingly important in experimental economics. How a power analysis is often conducted is that estimates of the treatment effect and standard deviations are obtained from previous studies, or inferred from introspection or beliefs. These parameters are then used in a closed-form representation of power (typically from a two-sample t-test) or a statistical software package, which gives the required sample size. The question is, where should these parameters come from? The treatment effect size could be specified by the point predictions from the model in question, however, economic theory is typically more successful in predicting comparative statics than point estimates. In addition, economic theory is typically silent on the expected standard deviation, so it cannot provide guidance on that parameter.[4] A more suitable approach is to use past experimental data as a starting point for these parameters. However, sufficiently close or exhaustive experimental environments may not exist. Even with appropriate data, a novel experiment will differ from past experiments in meaningful ways. For example, a new treatment might be conducted, or different parts of a theory's parameter space might be explored. It is not clear ex-ante what the actual treatment effect size would be, or how much variability subjects would exhibit, in these new treatments.

This paper explores a potential alternative method of power analysis based on a Quantal Choice/Response framework (henceforth QR) of individual behavior over discrete options (McFadden, 1976). The QR, also known as a logistic regression, has been successfully applied to many fields and applications, such as computer science, biometrics, transportation, psychology, and other social sciences, to help explain or model choice behavior. The QR framework is not limited to single decision-maker problems, as the Quantal

[1] Ziliak and McCloskey (2004) report that only 4% of the papers published in the *American Economic Review* in the 1980s mention power, increasing to 8% in the 1990s.

[2] Ioannidis et al. (2017) report a median power of 18% over a wide variety of empirical economics fields.

[3] Zhang and Ortmann (2013) report 1 out of 95 papers in *Experimental Economics* from 2010-2012 mention statistical power, as well as a median power of 25% in experimental studies of the Dictator Game. For reference, a power of 80% is generally considered sufficient in experimental economics (Moffatt (2016, pg. 22) List et al. (2011, pg. 448)).

[4] Notable exceptions exist, such as price dispersion (Burdett and Judd, 1983) where predictions of variability have been experimentally tested (Cason et al., forthcoming).

Response Equilibria (henceforth QRE) extends it to multiple interacting decision-makers (McKelvey and Palfrey, 1995). The QRE has proven successful in explaining subject behavior in strategic settings and is widely used in experimental economics and political science. The QR framework assumes a noise structure in which individuals choose actions in proportion to their relative payoffs, so that better actions are chosen more frequently. Relative payoffs can be presented graphically as a 'payoff hill', which I describe further in Section 3, and these hills intuitively capture subject behavior through the incentives that they face. The method of power analysis that I propose fits a structural QR framework to the most closely related previous experimental data-sets. Given the parameters from that model, the QR framework provides predictions for the probability that each possible action would be taken in the proposed new experiment. In other words, the QR approach takes what is known about subject behavior in previous studies, and maps that onto likely subject behavior in the new experiment that is yet to be conducted. From the predicted probabilities, simulated data-sets can be generated, and it is on these data-sets that power analysis is then conducted. As QR has proven successful in modeling choice behavior, with appropriate parameters inferred from past data, the simulated data-sets should be a reasonable approximation of likely subject behavior. The more accurate the approximation of actual subject behavior is, the more accurate the ex-ante power analysis will be, meaning the study should be reasonably powered ex-post.

This paper describes and exhibits the QR based simulation approach to power analysis. In Section 3 I describe the payoff hills which effectively underlie QR, and arguably contribute to its success. Payoff hills can help explain where empirical treatment effect sizes are likely to differ from theoretical point predictions, as well as provide a measure of likely subject variation, both of which are important considerations for more accurate ex-ante power analysis. In Section 4, I outline the QR simulation approach in more detail, and describe how it could be used to help design experiments. In particular, experiments are typically designed such that theoretical predictions are distantly spaced, i.e. the treatment effect size from point predictions is maximized. However, if theoretical predictions differ substantially from subject behavior, it should be the likely empirical average behavior is distantly spaced instead. The QR framework provides estimates for likely behavior for a variety of possible design decisions. Design decisions with respect to power need to also consider the impact on likely subject variability. The QR framework also provides an estimate of likely subject variability for many possible design decisions. I confirm the importance of considering the treatment effect and standard deviations simultaneously using closed-form expressions of power, as it is not always optimal to maximize the treatment effect size. In Section 5, I demonstrate the application of the QR approach using a motivating example of an experiment in Bayesian Persuasion, and compare it to a more standard approach to power analysis. The Bayesian Persuasion environment is a non-abstract example where maximizing the theoretical treatment effect size is very poor for power. Finally, in Section 6 apply the QR

simulation approach to various high-profile papers. I conduct a thought exercise where I place myself in a similar situation as the original authors, that is, I largely ignore the data in the paper and rely on other sources to conduct the analysis. I find that the QR approach can be applied to a wide range of environments, sometimes with creative modifications to incorporate the operative channel of the treatment effect. However, there are exceptions, namely experiments where no explicit theory is provided that could explain a treatment difference. The ex-ante power analysis from the QR simulations compares favorably to an ex-post power analysis, with the caveat that ex-post is an excessively high standard for ex-ante analysis to meet. I also suggest alternative parameters and treatments to increase power for each paper, which decreases the required number of subjects. Optimally selecting design parameters in such a way could also be conducted ex-post to guide a conceptual replication that is 'efficient', in that it minimizes subject costs.

3 Payoff Hills and Power Analysis

3.1 Payoff Hills

A payoff hill is a graphical representation of the level of payoffs arising from different actions, so-called as the optimum action is the 'peak' of the hill, while the decreases in payoffs from deviating away from the optimum form the 'sides' of the hill. As QR agents exhibit noisy behavior in proportion to their relative payoffs, the QR effectively incorporates the payoff hill, making it a useful visual aid to help explain why QR is successful in explaining choice behavior. Payoff hills are an important design factor that most experimenters are aware of, as they reflect the incentives subjects face to behave optimally. Much attention was brought to this topic by Harrison (1989), who proposes a metric of foregone expected income, which is the inverse of what I describe as a payoff hill. In what is now known as the 'flat-maximum critique', Harrison notes that in an experimental implementation of a first-price sealed bid auction (Cox et al., 1988), large deviations from the optimal action result in only in small decreases in expected payoffs. Therefore, reading too much into such deviations could be problematic. Following this logic, payoff hills that are 'flatter' are likely to exhibit more noisy behavior, because subjects suffer smaller payoff consequences for deviations from the optimal action. The flatness or steepness of a payoff hill can therefore influence the standard deviation, and thus have an impact on power analysis. The Harrison (1989) paper sparked a flurry of discussion, with multiple comments on the article published (e.g. Cox et al. (1992), Kagel and Roth (1992) Merlo and Schotter (1992), Harrison (1992)). Of particular note, Friedman (1992) points out that the flat-maximum critique would predict deviations in both directions if the flatness was symmetrical about the optimal action. In order to predict consistent deviations in one direction (i.e. above or below the optimal action), it would need to be the case that the payoff hill

is asymmetric with one side of the hill being flatter than the other. Subjects will be more likely to make deviations in the direction where they face smaller payoff consequences for doing so. Such directional errors can influence the treatment effect size, especially when the asymmetry of the payoff hill differs by treatment, and thus affect power analysis. The following section investigates the impact of both the steepness/flatness and asymmetry of the payoff hill on the likely treatment effect size and standard deviations.

3.2 Effects of Payoff Hills on Power Analysis

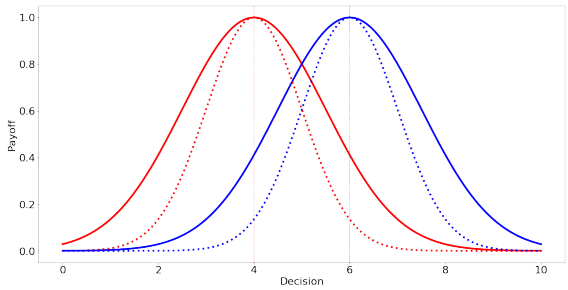
Power analysis is largely driven by three parameters: the treatment effect size τ; the standard deviation of the first treatment σ_1; and the standard deviation of the second treatment σ_2. Due to its impact on likely subject behavior, payoff hills can thus affect all three of these parameters, which is now demonstrated with a constructed example.

For this example, subjects make their decision x over a strategy space of $x \in [0, 10]$. To consider asymmetry, I define the payoff hill over x to be a piece-wise Gaussian function about some maximal payoff $a = 1$ which occurs at point b, while allowing for the possibility for the variable c to be different on either side of the maximum point. By itself, the variable c also allows us to consider steepness of the payoff hill, as it controls the shape of the function when moving away from the optimal point, where lower values of c are steeper.[5] More explicitly, the payoff hill is:

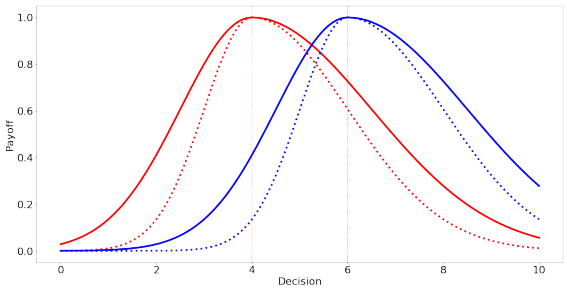
$$\pi(x; a, b, c_{LHS}, c_{RHS}) = \begin{cases} a \exp(-\frac{(x-b)^2}{2c_{LHS}^2}), & \text{if } x < b \\ a \exp(-\frac{(x-b)^2}{2c_{RHS}^2}), & \text{otherwise} \end{cases}$$

Suppose there are two treatments, where the payoff maximizing action is $b_1 = 4$ in Treatment 1, and $b_2 = 6$ in Treatment 2. An initial estimate of the treatment effect size can be derived from the point predictions of standard theory, $\tau = b_2 - b_1 = 6 - 4 = 2$. However, this estimate is likely erroneous, as it is ignorant to the shape of the payoff hill. I consider four potential types of treatment payoff hills, a symmetrical payoff hill where $c_{LHS} = c_{RHS}$ for both treatments, and three different combinations of asymmetric payoff hills. The first type of asymmetric hill has the flatter side of the hill to the same side of the optimum value in both treatments, with $c_{1,LHS} = c_{2,LHS} < c_{1,RHS} = c_{2,RHS}$. The second type of asymmetric hill has the flatter side of the payoff hills facing away each other, with $c_{1,RHS} = c_{2,LHS} < c_{1,LHS} = c_{2,RHS}$. The third type of asymmetric hill has the flatter sides facing towards each other, with $c_{1,LHS} = c_{2,RHS} < c_{1,RHS} = c_{2,LHS}$. These four types of payoff hills can be be considered in conjunction with their relative steepness as encapsulated by c. Diagrams of these four combinations of payoff hills for a 'baseline' hill with $c_{steep} = 1.5$ and $c_{shallow} = 2.5$,

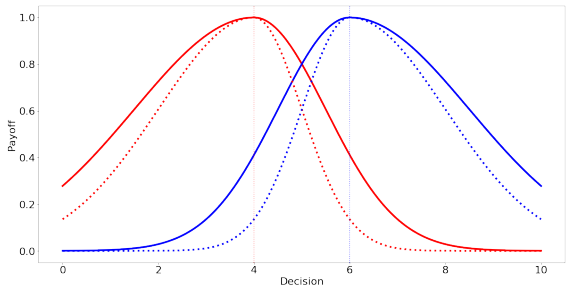
[5] The variable c is analogous to the standard deviation in the normal distribution, where lower standard deviations are 'tighter' about the mean.



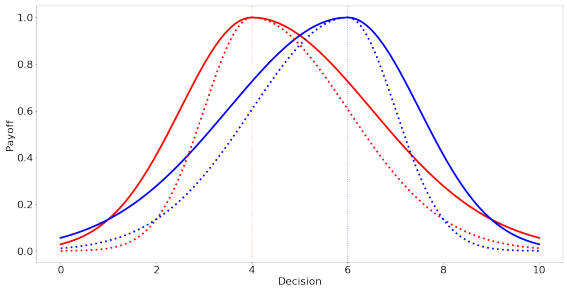
(a) Symmetrical



(b) Asymmetrical: Same



(c) Asymmetrical: Away



(d) Asymmetrical: Towards

Figure 1: Payoff hills with different types of asymmetries.

and a 'steeper' hill with $c_{steep} = 1.0$ and $c_{shallow} = 2.0$ are shown in Figure 1. Similar payoff hills can occur in more natural experiments, but instead of finding experiments that fit, for the purposes of this exercise I assume such payoff hills are constructed exactly in a single decision maker environment.[6]

I now analyze the impact that asymmetry of the payoff hills has on the likely observed treatment effect size. As mentioned previously, an immediate initial estimate of the treatment effect size would be $\tau = 2$, from the point predictions of theory. However, asymmetries in the hills will impact the likely empirical treatment effect size that is observed. This is most evident from the payoff hills displayed in Figures 1c and 1d. Subjects in Figure 1c are more likely to deviate from optimal behavior towards the edges of the action space, increasing the treatment effect size. Whereas those in Figure 1d will be deviating towards the center of the action space, decreasing the treatment effect size. In the limit, this distortion from the point prediction τ would diminish as both $c_{RHS} \to 0$ and $c_{LHS} \to 0$, or as $\lambda \to \infty$. Therefore, it is not only asymmetry that effects the treatment effect size, the steepness of the payoff hill also matters. Ignoring either of these payoff hill features can result in erroneous conclusions about power through incorrect estimates of the treatment effect size. The steepness of the payoff hill more directly impacts the standard deviation rather than the effect size, and the standard deviation is also a determinant of power. Therefore, when considering the impact of the payoff hill on power, these two parameters need to be considered in tandem.

[6]It is typically the case that the experimenter does not have this level of precise control over the payoff hill, however, it is also the case that it is rare that the experimenter has no control at all over the payoff hill.

Parameters	Symmetrical	Asymmetrical: Same	Asymmetrical: Away	Asymmetrical: Towards
$c_{steep} = 1.5$ $c_{shallow} = 2.5$	$\tau = 1.944$ $\sigma_1 = 1.804$ $\sigma_2 = 1.806$ **Power**= 64.6%	$\tau = 1.604$ $\sigma_1 = 2.329$ $\sigma_2 = 2.072$ **Power**= 28.0%	$\tau = 2.147$ $\sigma_1 = 2.069$ $\sigma_2 = 2.072$ **Power**= 60.4%	$\tau = 1.058$ $\sigma_1 = 2.329$ $\sigma_2 = 2.333$ **Power**= 9.1%
$c_{steep} = 1.0$ $c_{shallow} = 2.5$	$\tau = 1.997$ $\sigma_1 = 1.221$ $\sigma_2 = 1.223$ **Power**= 97.1%	$\tau = 1.657$ $\sigma_1 = 2.068$ $\sigma_2 = 1.771$ **Power**= 41.3%	$\tau = 2.735$ $\sigma_1 = 1.769$ $\sigma_2 = 1.771$ **Power**= 95.1%	$\tau = 0.575$ $\sigma_1 = 2.068$ $\sigma_2 = 2.072$ **Power**= 3.5%
$c_{steep} = 1.5$ $c_{shallow} = 2.0$	$\tau = 1.944$ $\sigma_1 = 1.804$ $\sigma_2 = 1.806$ **Power**= 64.6%	$\tau = 1.796$ $\sigma_1 = 2.095$ $\sigma_2 = 1.974$ **Power**= 43.7%	$\tau = 2.150$ $\sigma_1 = 1.971$ $\sigma_2 = 1.974$ **Power**= 65.9%	$\tau = 1.441$ $\sigma_1 = 2.095$ $\sigma_2 = 2.098$ **Power**= 24.4%
$c_{steep} = 1.0$ $c_{shallow} = 2.0$	$\tau = 1.997$ $\sigma_1 = 1.221$ $\sigma_2 = 1.223$ **Power**= 97.1%	$\tau = 1.835$ $\sigma_1 = 1.831$ $\sigma_2 = 1.675$ **Power**= 61.4%	$\tau = 2.707$ $\sigma_1 = 1.673$ $\sigma_2 = 1.675$ **Power**= 96.8%	$\tau = 0.961$ $\sigma_1 = 1.831$ $\sigma_2 = 1.834$ **Power**= 12.7%
$c_{steep} = 2.0$ $c_{shallow} = 3.0$	$\tau = 1.629$ $\sigma_1 = 2.262$ $\sigma_2 = 2.265$ **Power**= 27.3%	$\tau = 1.215$ $\sigma_1 = 2.655$ $\sigma_2 = 2.404$ **Power**= 10.3%	$\tau = 1.385$ $\sigma_1 = 2.401$ $\sigma_2 = 2.404$ **Power**= 15.9%	$\tau = 1.042$ $\sigma_1 = 2.655$ $\sigma_2 = 2.660$ **Power**= 6.7%

Table 1: Effects of Payoff Hill Asymmetry and Steepness on Power

Power is evaluated using a t-test at the $\alpha = 0.01$ level with $N = 15$ observations per treatment.

In order to calculate power for the different combinations of asymmetry and steepness, τ, σ_1, and σ_2 need to be specified, which is where the QR framework comes in handy. I apply the QR framework to map the payoff hills into a probability of choosing a particular action using the logit choice rule $p(a_i) = \frac{e^{\lambda E\pi_{a_i}}}{\sum_{a_j \in A} e^{\lambda E\pi_{a_j}}}$ with an arbitrarily selected $\lambda = 1$.[7] Table 1 reports the output of a simulation based power analysis for various combinations of c_{steep} and $c_{shallow}$. Asymmetric payoff hills do affect the treatment effect size in the manner previously suggested, in that it is suppressed when the shallow sides face towards each other and typically amplified when the shallow sides face away. Increasing the steepness of the payoff hill (decreasing c) unambiguously decreases the standard deviation. There is also an interaction with the boundaries of the action space, as evidenced by the 'Same' column in Table 1 where the treatment effect size is suppressed, which could be lessened if the action space were extended or eliminated if the boundary were removed.[8] This interaction is potentially important as experiments typically have bounded action spaces. Finally, flatter payoff hills overall, as demonstrated in the last row of Table 1, suppress the treatment effect. This is due to subjects getting closer to indifference over all actions, which in the QR framework entails uniform random play, and thus no treatment effect. All of these potential factors arising from the structure of the payoff hills should be considered when evaluating power.

[7] I discretized the action space as 10000 evenly spaced points from 0 to 10 inclusive, and binned adjacent actions depending on their impact on payoffs.

[8] The bounded action space also explains the small deviations from $\tau = 2$ in the Symmetrical case.

4 QR Simulations for Optimal Experimental Design

4.1 Describing the QR Simulation Power Analysis

A simulation approach specifies a data-generating process (DGP), which can then used to generate many simulated data-sets upon which power analysis can be conducted. The QR is an appropriate choice for a DGP of subject behavior, as it has been successful in predicting a wide range of experimental environments, and thus is likely to resemble the true DGP. The specific DGP used in a QR simulation is the logit function: $p(a_i) = \frac{e^{\lambda EU_{a_i}}}{\sum_{a_j \in A} e^{\lambda EU_{a_j}}}$, where EU_{a_i} is the expected utility or payoff of choosing action i, and λ is a decision precision or noise parameter. The expected utility of each action, EU_{a_i}, will depend on the environment in question, and will be a function of the parameters in the experimental design (e.g. monetary payoffs, treatments) as well as any other utility relevant parameters (e.g. risk aversion, other-regarding preferences). In a single decision-maker problem, it is relatively straightforward to calculate EU_{a_i} for each action, whereas in a strategic setting this is more involved as the QRE of the environment in question must be calculated. The main parameter the researcher has to specify is λ, the QR noise parameter that represents how frequently subjects play their best/better actions. Reasonable estimates of λ should be selected by structurally fitting a QR model of the logit function to data from the most closely related previous studies. Additional parameters may need to be specified if the theory requires them (e.g. a hypothesis about reciprocity may require an estimate of other-regarding preferences). These parameters should also be calibrated from previous experiments in the same manner as λ.[9] In the event that no suitable data is available, a small pilot could be conducted and the parameters estimated from that, or other parameters that are reported and considered acceptable by the literature could be used. The parameters inferred from the previous data are then substituted into the QR logit choice rule for the new experimental environment in question. To finalize the DGP, the statistical test that will be used in the experiment needs to be considered. For example, if the final statistical test will cluster its standard errors, then the DGP should have something in it to justify the need for clustering, otherwise the power analysis may be misleading. There are various ways this could be implemented, for example, an individual or session level λ or other parameter could be drawn. This finalizes the specification of the DGP, from which simulated data-sets can be generated. Power is then calculated by performing the desired statistical test on each simulated data-set, and counting the number of times the null hypothesis is rejected. A simulation approach is particularly helpful for more complicated statistical tests that do not have closed-form expressions for their power.

[9] The calibration will likely be imperfect considering that it was probably not the original study's intention to structurally fit such parameters. Instead, the aim is to obtain reasonable estimates and improve upon more arbitrary guessing these parameters ex-ante.

4.2 Optimal Experimental Design

There are many considerations that go into an experimental design, which include the number and levels of treatments, whether to vary treatments between- or within-subjects, whether to use the strategy method, have fixed groups or random re-matching, what payoffs should occur from outcomes, and so forth and so on. In the strictest possible sense, power analysis determines only one design decision, the number of subjects to have in each treatment, when other design decisions are held fixed. However, any design decision could be evaluated and selected on the basis of power. Optimal experimental design refers to making design decisions with respect to maximizing some metric. I focus on the metric of power, and the design decisions of selecting treatment levels and environmental parameters that are fixed by treatment (e.g. payoffs, probabilities, etc.).[10] Traditionally, these parameters would be chosen to increase the difference in theoretical predictions, which usually requires exploring the furthest reaches of the parameter space.[11] This rule of thumb is often successful for two reasons. Firstly, if the effect size is linear in the treatment level, by 'D-optimal design' (Moffatt, 2016, ch. 14) it is always optimal to space the treatment levels as far apart as feasibly possible.[12] Secondly, maximizing the theoretical treatment effect size usually maximizes the actual treatment effect size, which is beneficial for power. However, on the first point, it is not always D-optimal to maximize the difference in treatment levels if the relationship is non-linear, which is effectively the case if the standard deviations depend on the treatment level.[13] On the second point, Section 3 already established that asymmetric payoff hills can result in different theoretical and actual effect sizes, so it is not immediately clear that maximizing one maximizes the other. However, even if it does, the impact that the treatment level or any other parameter has on the standard deviation needs to be taken into account. It could be the case that the standard deviation increases substantially, meaning a given parameter change might not increase power despite increasing the effect size. The rule of thumb proves reasonably successful in a lot of environments, otherwise it would not be accepted as such. However, QR simulations can help identify the environments where these issues could be problematic.

A simulation approach is ideal for optimal experimental design, as it does not require analytical solutions for power. Instead, power is determined by simply calculating the proportion of data-sets where the null was rejected, as a true effect is known to exist in the simulation DGP.[14] Using a QR simulation in particular

[10] Although, just quickly, the strategy method unambiguously increases power over direct response, and doing treatments within-subject is very beneficial for power if there is substantial individual heterogeneity.

[11] For example, from classic textbooks on experimental design: *"A related aspect of calibration is the use of a design in which the predictions of alternative theories are cleanly separated"* (Davis and Holt, 1993, pg. 28), *"Use widely separated levels to sharpen the contrasts."* (Friedman and Sunder, 1994, pg. 31).

[12] D-optimal design maximizes the determinant of the information matrix implied by maximizing the log-likelihood function. This minimizes the confidence intervals on estimated parameters, which is similar to maximizing power.

[13] For an example of treatment level dependent standard deviation on D-optimal design, see Appendix B.

[14] A good example of using (non-QR) simulations to select experiment parameters to improve power is Rutström and Wilcox (2009), who aimed to distinguish between two competing models of subject behavior.

provides reasonable predictions for how treatment levels and other parameters will affect both the treatment effect size and treatment-specific standard deviation. This means that optimal parameters can be selected considering both the likely treatment effect size and standard deviations simultaneously and in a realistic manner.

Note that power is not the only metric that could be used for optimal experimental design. For example, the researcher may want to identify strategies from subject behavior in a probabilistic environment where they may not observe their behavior in all relevant states or for enough periods. A simulation approach would count the likely number of observations that meet the criterion, a metric which can be maximized. Another metric could be the accuracy of estimated parameters compared to their true parameter, which is known in a simulation. This metric can guide experimental design through selecting the most effective questions to identify certain types, as well as confirm the likely identifiability of a structural model given a set of questions and responses.

4.3 Analysis of Closed-form Expressions for Power

The purpose of this section is to confirm that it is not always optimal to maximize the actual treatment effect size when it comes to power, and that standard deviations need to be jointly considered. Restricting the analysis to tests with closed-form expressions for power removes the need for simulations.

4.3.1 Binary Actions

Binary actions are common in economics experiments, particularly as it is often considered good design to have the most simplified environment possible that still tests the conjecture. The random nature of an individual binary decision is a Bernoulli trial with probability p, and multiple observed decisions makes up a Binomial distribution. A common statistical test used for binary data is a test of proportions, either the exact Binomial probability test or the large-sample approximation Z-test of proportions, either of which can be a one-sample (i.e. point prediction) or two-sample (i.e. treatment effect) test. The standard error of the observed proportions is determined by the sample estimate of p, and is $\sigma = \sqrt{\frac{p(1-p)}{N}}$. To decrease σ, one could either increase N, or have p be close to zero or one. The two-sampled, two-sided power has a closed form solution: $\Phi\{\frac{(p_2-p_1)-c-z_{1-\alpha/2}\sigma_P}{\sigma_D}\} + \Phi\{\frac{-(p_2-p_1)-c-z_{1-\alpha/2}\sigma_P}{\sigma_D}\}$, where c is a normal approximation continuity correction ($c = 2/n$ when $n_1 = n_2 = n/2$), $\sigma_P = \sqrt{\bar{p}(1-\bar{p})(1/n_1 + 1/n_2)}$ is the pooled standard deviation and $\sigma_D = \sqrt{p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2}$ is the standard deviation of the difference between proportions (StataCorp, 2013). Power can be increased by increasing $p_2 - p_1$ (i.e. increase the treatment effect), or by moving p_1 or p_2 closer to one or zero (i.e. decrease the standard deviation of either sample), or increasing n_1 and n_2 (again, decreasing the standard deviation). In terms of choosing an optimal treatment

intensity, if possible the optimum would be to set $p_1 = 0$ and $p_2 = 1$, maximizing the treatment effect size and minimizing the standard deviation. However, p_1 and p_2 are provided by the model, and are subject to both theoretical and practical constraints, in that it might not be possible to choose parameters in such a way to achieve both $p_1 = 0$ and $p_2 = 1$. Suppose, for whatever reason, that the maximum possible effect size is $p_2 - p_1 = 0.25$. If this were the case, power would be maximized at $p = 0$, $q = .25$ or $p = 0.75$, $q = 1$, but not at any equivalent interior combination of the same magnitude (e.g. $p = 0.25$, $q = 0.5$). This is an example of where it is not sufficient to only consider the treatment effect size when choosing parameters or treatment intensity, standard deviation needs to be considered as well. It is even the case that smaller treatment effects can be desired over larger ones due to the standard error. One example with a large difference in treatment effect sizes would be the selection of $p_{1A} = 0.0$ and $p_{2A} = 0.3$ over $p_{1B} = 0.3$ and $p_{2B} = 0.7$. Treatment A has a smaller treatment effect (0.3 vs 0.4), but has slightly improved power (84.8% vs 83.2%). Another example with a larger difference in power is $p_{1C} = 0.0$ and $p_{2C} = 0.29$ over $p_{1D} = 0.35$ and $p_{2D} = 0.65$. Treatment C has a slightly smaller treatment effect (.29 vs .3), but has substantially improved power (82.8% vs 54.4%). So even in this relatively simple binary environment (where the standard deviation is provided by the given p), we still must be cognizant of the fact that maximizing treatment effect size alone will not necessarily maximize power.

4.3.2 Non-Binary Actions or Outcomes

Sometimes the parameter of interest is the average outcome that results from binary actions (e.g. election outcomes from a group of binary voters), or that the action set is larger than 2 decisions, and therefore no longer binary. In these cases, we use tests of the differences in means, and the most commonly used test for this is the t-test. The functional form for power of the two-sided version of this test is: $\Phi\{\frac{(\mu_2-\mu_1)}{\sigma_D} - z_{1-\alpha/2}\} + \Phi\{-\frac{(\mu_2-\mu_1)}{\sigma_D} - z_{1-\alpha/2}\}$, where $\sigma_D = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$. This is similar to the functional form for the z-test of proportions considered in the previous section, except now the mean of a treatment has been decoupled from the standard deviation. The ways of increasing power is to either increase the treatment effect size, decrease either standard deviation, or increase either n. However, changing the treatment effect size in an experiment without influencing the standard deviation is unlikely. It is possible that increasing the treatment effect size could actually decrease power, if the standard deviation also increases substantially, as a result of whatever experimental change was required to induce the increase in the treatment effect size. To construct an example, I hold treatment 2 constant, and denote $\tau = \mu_1 - \mu_2$, where $\tau \in [0, 1]$ and let $\sigma_1 = f(\tau)$, and $\sigma_2 = 1$. An example of where maximizing τ does not maximize power is when $f(\tau) = 2\tau^2 + \tau + 1$, where setting $\tau = 1$ yields a power of 27.8%, while power is maximized at $\tau = 0.714$, with a power of 29.9%. More extreme examples could be arbitrarily constructed, but this again illustrates that maximizing the treatment

effect size alone is not necessarily sufficient to maximize power.

4.4 Optimal Design applied to Replications

Replication in economics is an important but under-produced form of research, both inside and outside experimental economics (Maniadis et al., 2014; Duvendack et al., 2017). There have been recent laudable efforts and proposals to increase the provision of this crucial research.[15] Some proposals include increasing the benefit of independent replications by offering coauthorship (Butera et al., 2020), or including the number of replications as a metric for an author's research quality (Maniadis et al., 2015). Camerer et al. (2019) describe two types of replications, direct and conceptual. A direct replication is where the same exact experiment is conducted with as few procedural differences as feasibly possible. A conceptual replication is where the same hypothesis as the original study is tested but using different methods or parameters. The bulk of the encouragement in this area is with regards to direct replication, but conceptual replication is just as important, particularly in experiments that test an explicit economic model. A conceptual rather than direct replication provides more insight on whether the underlying model and its proposed operative channels is an accurate description of behavior, or what extensions or caveats to the model may be required. The suggestions to incentivize more replications tend to focus on increasing the benefits of replication, rather than lowering the costs. I propose the use of QR based simulations to design conceptual replications that economize on the required number of observations. The model parameters and treatment intensities can be altered from the original study to maximize power, which minimizes the required number of subjects and thus the expenditure in terms of both time and money. I call a conceptual replication that explicitly minimizes costs an 'efficient replication'. A QR simulation is particularly well suited for efficient replication, as an appropriate data-set with which to calibrate the required parameters already exists.

5 Power Analysis using Bayesian Persuasion as an Illustrative Example

5.1 Bayesian Persuasion Overview

Bayesian Persuasion (Kamenica and Gentzkow, 2011) is a model of information provision with commitment that has formed the basis for a substantial amount of research.[16] Bayesian Persuasion has been applied to a wide variety of areas, such as finance, optimal feedback, medical testing and research, insurance, and

[15] For example, the establishment of the *Journal of the Experimental Science Association* aims in part to publish under-represented experimental research like replications, and the Experimental Economics Replication Project (Camerer et al., 2016) conducts high-powered replication studies on multiple high-profile papers.

[16] Google Scholar reports more than 1400 citations.

many others too numerous to list here (Kamenica, 2019). Despite this, there have been relatively few experimental studies of Bayesian Persuasion.[17] I use a hypothetical experimental test of Bayesian Persuasion as a motivating example to illustrate the QR approach to power analysis, as well as its potential benefits relative to a more standard approach.

In Bayesian Persuasion, there is a Sender and a Receiver who are paired together. The Receiver has a true unobserved state, either Red or Blue, with a common initial prior p that their state is Red. The Sender designs an information structure that can condition on the Receiver's true state, sending one message that either states r or b. The Sender chooses x and y, where $Pr(b|Blue) = x$ and $Pr(r|Red) = y$. The Receiver observes x and y as well the generated message, updates his belief according to Bayes Rule, and then decides whether to choose Red or Blue. The Receiver gets a payoff of π_m if he chooses the same color as his true state, and π_0 otherwise, with $\pi_m > \pi_0$. The Sender gets a payoff of π_R if the Receiver chooses Red, and π_B if the Receiver chooses Blue, with $\pi_R > \pi_B$. The Sender's optimal strategy is to set the information structure in such a way as to minimally update the Receiver's beliefs such that the Receiver wants to choose Red. If messages are straightforward (i.e. a message is seen as a suggested action), then this involves setting $y = 1$ and $x = \frac{1-\frac{p}{p^*}}{1-p}$, where p^* is the Buyer's reservation threshold (in the current symmetrical setup $p^* = .5$).

Suppose the hypothesis in question is whether Sender's behavior (i.e. their choice of x) changes in response to p. I first consider a power analysis approach of perturbing behavior about the theoretical predictions.

5.2 Normal Distribution DGP

As mentioned previously, there is not a wealth of previous experimental data to help guide a power analysis. For this current example, consider a between-subjects treatment where $p = \frac{1}{5}$ or $p = \frac{1}{3}$. The optimal x's given those p's are $x^* = \frac{3}{4}$ and $x^* = \frac{1}{2}$ respectively. I assume only a one-shot interaction between the Sender and the Receiver for this exercise. A power analysis using a standard rule of thumb would require a treatment effect size, τ, and the level of noise in an individual's behavior in each treatment, σ_0 and σ_1, to be specified. For now, suppose there is no strong a-priori reason to believe that the noise should differ between treatment, so $\sigma_0 = \sigma_1 = \sigma$, and I also assume no heterogeneity.[18] The data-generating process (DGP) is relatively simple: $Y_{iT} = \alpha + \tau T + \epsilon_i$, where $\epsilon_i \sim N(0, \sigma^2)$. If $p = \frac{1}{3}$ is considered $T = 0$, then $\alpha = .5$ and $\tau = \frac{1}{4}$ follows from the theory. All that remains is to specify a σ. Typically this should be chosen based on the most relevant previous experiments. However, given that data is not currently available, and this is just an

[17] I am aware of the following papers: Nguyen (2017), Au and Li (2018), Fréchette et al. (2018), and Wu and Ye (2019).

[18] The impact of differential variance in this environment is presented in Appendix C. In short, more subjects should be allocated to the noisier treatment, with the ratio of subject allocation equal to the ratio of standard deviations (List et al., 2011, pg. 447). However, ex-ante it is not necessarily clear which treatment will have greater variance, something a QR simulation could address.

illustrative example, for this example I set it to $\sigma = 0.5$.[19] Given the assumed normal DGP, the Stata code to yield the required sample size for 80% power for a t-test is: `sampsi .5 .75, sd(.5) p(.8)`, which yields a sample size of $n = 63$, so 63 Sender observations for each treatment.

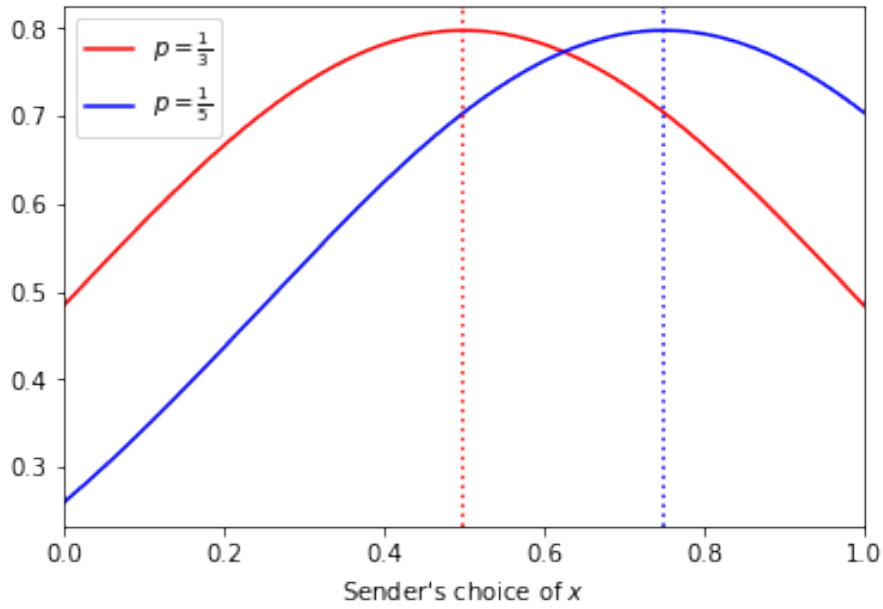


Figure 2: Example Normal DGP for Bayesian Persuasion

This is also a good opportunity to show how this process could be done by generating multiple simulated data-sets of size n from the DGP, conducting the statistical test on each data-set and counting how many times the null is rejected, thus having an empirical measure of power. The code is provided in Appendix A. The general idea is that each data-set is assigned an additional value/column called 'power_block', with which Stata can conduct the statistical test considering only the data within one power block at a time.

The simulation approach using the code in Appendix A returns a power of 79%, which is very close to the 80% power specified in the `sampsi` command. The small deviation is due to the random nature of the simulations, but this is of no practical concern.

5.2.1 Optimal Experimental Design

The initial treatment priors of $p = \frac{1}{3}$ and $p = \frac{1}{5}$ were chosen arbitrarily, but they could instead be chosen to maximize power. A simple way to do so would be to consider the parameters that go into the rule of thumb for the required number of subjects for a t-test, $n = \frac{12.35\sigma^2}{\tau^2}$. From the rule of thumb, it can be seen that increasing the treatment effect size will increase power, while increasing σ will decrease power. Either the treatment effect size should be increased to lower n, or the variance should be decreased. Absent an explicit

[19] So chosen such that it is sufficiently noisy so that the exercise is not trivial (i.e. excessive power for low n).

model for variance, it is not clear how much of a change in the initial prior would change the variance. However, theory does provide an estimate on how the treatment effect would change, so in the absence of anything else, that is the parameter that should be changed.

Following the above, optimal experimental design calls for increasing the separation of the priors in the treatment as much as is practically feasible. For example, using the alternative priors of $p = \frac{1}{10}$ and $p = \frac{2}{5}$ has predicted Sender behavior of $x^* = 0.89(2dp)$ and $x^* = 0.33(2dp)$. This means that $\tau = 0.56$, and that the required sample size would be 13 observations for each treatment. This is substantially less than the 63 observations for each treatment when $\tau = 0.25$.

5.3 QR framework DGP

5.3.1 Motivation

The previous subsection reports a more standard approach for power analysis in a simple one shot experiment, and explores optimal design with choice of treatment levels. However, as discussed previously, ex-ante it is not clear what the standard deviation should be, both in general as well as by treatment. As for the treatment effect itself, the theory should provide an estimate for an appropriate magnitude. Of course, subject behavior may not align with the theory depending on the structure of the payoff hills, as discussed in Section 3. But in what direction should the theoretical treatment effect be adjusted, and by how much?

To illustrate the motivation behind the QR framework, consider the expected profits of a Sender where $\pi_R = 2$ and $\pi_B = 1$, given that a Receiver will choose Red if $q \geq p^*$, which is presented in Figure 3. In other words, Figure 3 represents the Sender's payoff hill over his available actions.

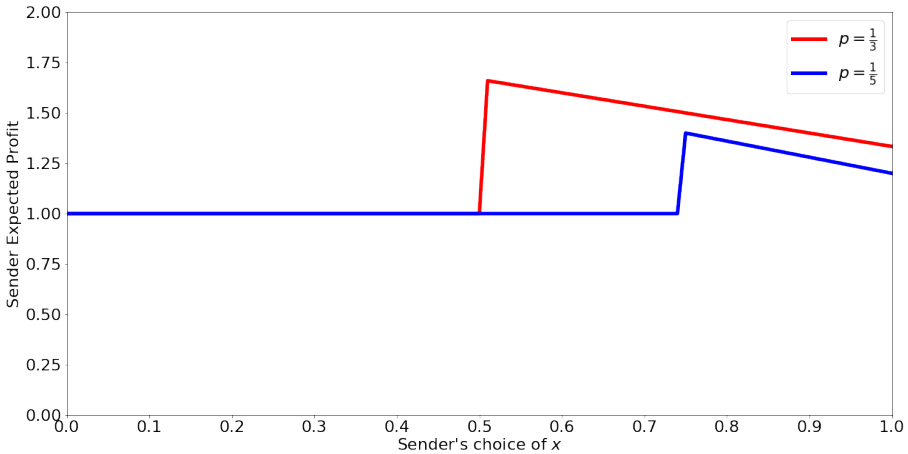


Figure 3: Sender's Expected Profit for $\pi_R = 2$, $\pi_B = 1$, and $q^* = 0.5$

Note two main features of the payoff hill in Figure 3. The first is that it is always more profitable to

choose an x that is higher than the level required to update the Receiver's beliefs above their reservation threshold, than it is to choose an x lower than this level. This means that Senders should be more likely to choose a higher x than the theory predicts than a lower x. In other words, the payoff hills are asymmetric, slightly above the equilibrium value the slope is quite shallow, whereas slightly below the equilibrium value it is very steep. Deviating to a slightly lower x is very costly in terms of expected payoff, while deviating to a slightly higher x is not very costly. Such directional errors provide an adjustment to the likely treatment effect. In this case the treatment effect would likely differ but not excessively so, as both treatments exhibit likely deviations in the same direction, much like the example payoff hill in Figure 1b. Secondly, note that the relative payoff peak is much higher in the red (high-prior) treatment than in the blue (low-prior) treatment. Therefore, in the blue treatment the likelihood of deviating from the optimal action would be greater, as there is less relative incentive to choose that action. In other words, the payoff hill in the blue treatment is overall flatter than the hill in the red treatment. Therefore, subject behavior in the blue treatment should exhibit relatively more noise and thus a higher standard deviation than the red treatment. It is well-established that payoff hills should be considered when designing an experiment, as discussed in Section 3. Payoff hills that are flatter should increase the observed noise in subject behavior. The question is, what effect will the payoff hill have on the magnitudes on the treatment effect and treatment specific subject variance? Despite the fact that payoff hills are considered to be a facet of good experimental design, they have currently have no bearing on power analysis except through the researcher's subjective guess.

5.3.2 The QR approach

Providing some explicit but tractable structure to bring payoff hills into power analysis is the goal of a QR simulation approach. Because the Bayesian Persuasion environment has two players, a QRE needs to be calculated. In this environment, this is not too complicated, as the game is sequential so it can be solved by backwards induction. For each possible belief the state is Red, q, the Receiver could hold after receiving a signal r from x, the expected payoff of choosing Red and Blue is calculated. Then, it is simply a case of using a logit choice rule given λ to calculate the probabilities of choosing Red or Blue for each possible q. Given those Receiver decision probabilities, the Sender's expected profit is calculated for each possible x that could be chosen, taking into account p and the probability of sending r or b messages. From the Sender's expected profits, the logit choice rule is applied to calculate the probabilities of choosing each possible x.

The remaining step is to specify a λ. For the purposes of this example, I select $\lambda_S = \lambda_R = 11.26$ so that a similar number of subjects is required to attain 80% power as the normal DGP ($n = 63$). The QRE simulation code is presented in Appendix A.4. In practice λ should be estimated from the closest experimental data available. Given this λ, the QRE implies a DGP of Sender behavior as is graphically presented in Figure 4.

Something that should be noted is that the information provision x has increased under a QRE relative to standard theory. This is because in the QRE, when the Receiver is indifferent he chooses either action with 50% probability, whereas the standard theory assumes the Receiver would choose Red. Therefore, in a QRE, a Sender optimally provides more information such that the Receiver chooses the Sender's preferred action with a sufficiently high probability given the trade-off implied by the signal structure. This is probably a more realistic assumption about actual human behavior.

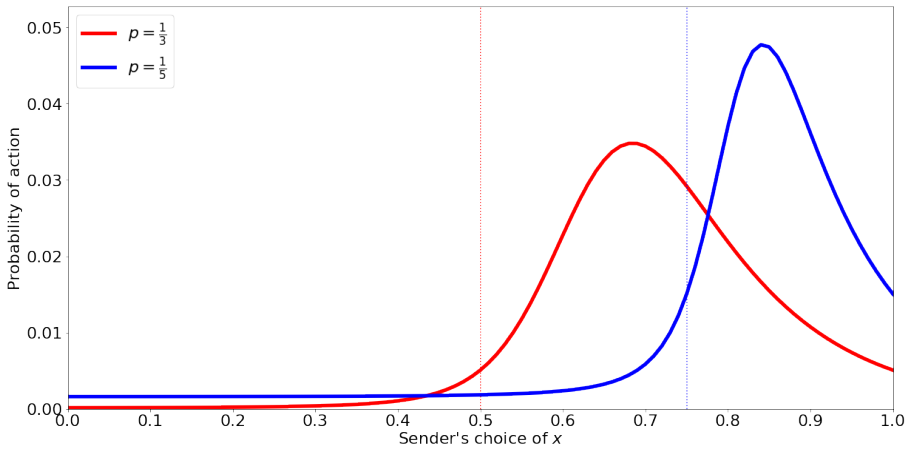


Figure 4: Sender's QRE DGP when $\lambda = 11.26$

5.3.3 Optimal Experimental Design Revisited

In the previous section, more optimal treatment levels were selected from the perspective of power by increasing the theoretical treatment effect size. These were treatments with priors of $p = \frac{1}{10}$ and $p = \frac{2}{5}$ with predicted Sender behavior of $x^* = 0.89(2dp)$ and $x^* = 0.33(2dp)$. Using QRE simulations with $\lambda = 11.26$ and $n = 63$, the statistical power is only 7%, i.e. power is substantially reduced at these 'optimal' treatment levels. To obtain a power of 80%, the design would require $n > 4000$.[20] What went wrong with the previous approach to optimal experimental design? Especially given that it coincides with the common practice of increasing the theoretical treatment effect size. As emphasized in Section 4, it is because the likely actual treatment effect size, as well as treatment-specific standard deviations, were not considered. Maximizing the separation of likely behavior needs to be traded-off against the likely standard deviation. As the QRE provides an explicit prediction for both of these parameters for any treatment level, it can help revolve this trade-off. In particular, the proposed $p = \frac{1}{10}$ treatment substantially increases noise, which is why the statistical power has decreased so much. In a Bayesian Persuasion environment, Senders with a low prior face a low likelihood that the Receiver's true state coincides with the Sender's preferred action. Therefore, regardless of

[20]The code starts to hit RAM limitations above $n = 4000$, which yields a power of 75%.

the level of information x sent, it is unlikely that the Sender will successfully persuade the Receiver to choose Red. As a result, the Sender's expected payoff hill is quite flat, and thus their behavior is likely to be very noisy. Noisy behavior in a QRE tends to uniform play in the limit, so noisier play would move x closer to 0.5, which also can influence the treatment effect size. Both of these factors contribute to the substantially decreased power of these 'optimal' treatment levels.

I now use QRE simulations to select the optimal treatment levels, i.e. the optimal priors (p and q) for each treatment. I assume $p > q$, restrict $p, q \in [.01, 0.4]$, and only let p and q be integer percentages (as would be likely in an actual experiment). I use a grid search and calculate the power for each possible combination of p, q, and select the pair with the highest power. A 3-D surface of power over different combinations of p and q is shown in Figure 5. The power surface is multi-peaked, with the primary peak having a maximum power of 97.1% at $p = 0.25$ and $q = 0.01$, and the secondary peak having a maximum power of 91.9% at $p = 0.4$ and $q = 0.24$. The primary peak seems counter-intuitive, given the discussion about low-priors being very noisy. However, the noise due to an effectively flat payoff hill in a QRE environment entails near-uniform random play, so the mean would be close to 0.5. As the t-test is a test of means, and the mean of the QRE distribution for $p = 0.25$ is $x = 0.787$, having a mean of 0.5 is quite powerful for detecting differences (more-so than the mean at $p = 0.4$, 0.619). However, this conclusion would be erroneous unless the researcher was explicitly testing for a subject's indifference when the payoff hill is very flat. This is an extreme insight from the QRE model, but the design goal is to test the Bayesian Persuasion environment. A more sensible conclusion would arise from the restriction that the mean of p is less than the mean of q (as the theory predicts if $p > q$), which would only yield the secondary peak, suggesting optimal treatment priors of $p = 0.4$ and $q = 0.24$. Such a design requires only 13 Sender observations for each treatment to reach a power of 80%, substantially less than the 63 in the original design.

An advantage of the QRE approach is that it provides structure to map not only any treatment intensity into an estimate of the treatment effect and standard deviations, but that it can also do this for any treatment invariant parameters that affect expected utility. For example, the payoff that a Sender gets for persuading the Receiver to choose Red, π_R, or the payoff that a Receiver gets for his guess matching his true type, π_m, could also be optimally selected in a similar manner as the treatment level. Increasing π_R would steepen the Sender's payoff hill, and similarly increasing π_m would steepen the Receiver's payoff hill. The former has a direct effect on Sender behavior, as they are now more incentivized to choose their optimal action. The latter has an indirect effect on Sender behavior, as Receivers would now more frequently guess the color with the highest belief, which the Sender would respond to. Increasing these payoffs while holding everything else fixed is equivalent to increasing that particular type's λ. A power curve is presented in Figure 6 for increases in π_R and π_m. The resulting curve is as expected, in that increasing π increases the statistical power, and

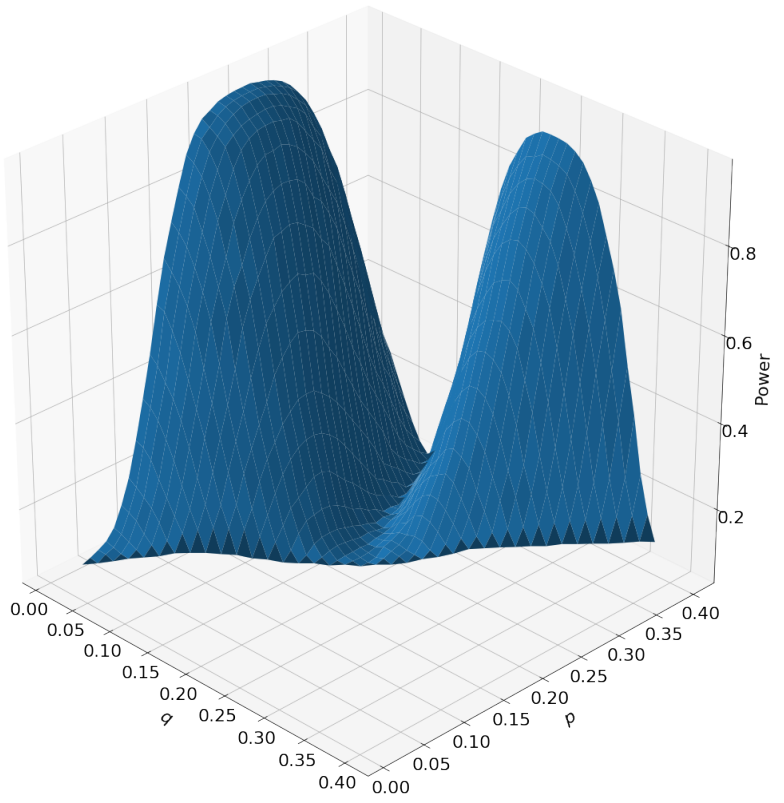


Figure 5: Power Surface for QRE Treatment Priors

this effect is more pronounced for the direct effect on the Sender.

The power curves suggests that π_m and π_R should be increased as much as possible relative to π_B and π_{mm}. There are practical constraints to this, one caveat might be that subjects need to earn a particular amount even in the worst case scenario (so that they do not get too upset), which would cap how large π could practically be relative to the worst-case payoff.

To summarize, a standard approach to power analysis does not explicitly consider the relative payoff hills that subjects face. This can lead to erroneous assumptions about the direction and magnitude of subject deviations from theory, as well as subject variability by treatment. By incorporating a QR framework into a simulation based power analyses, we can more accurately identify likely treatment effects and standard deviations, which can lead to more accurate ex-ante power analysis. Furthermore, we can use the QR framework to guide design decisions, such as selection of optimal treatment levels or other fixed parameters, in order to maximize power.

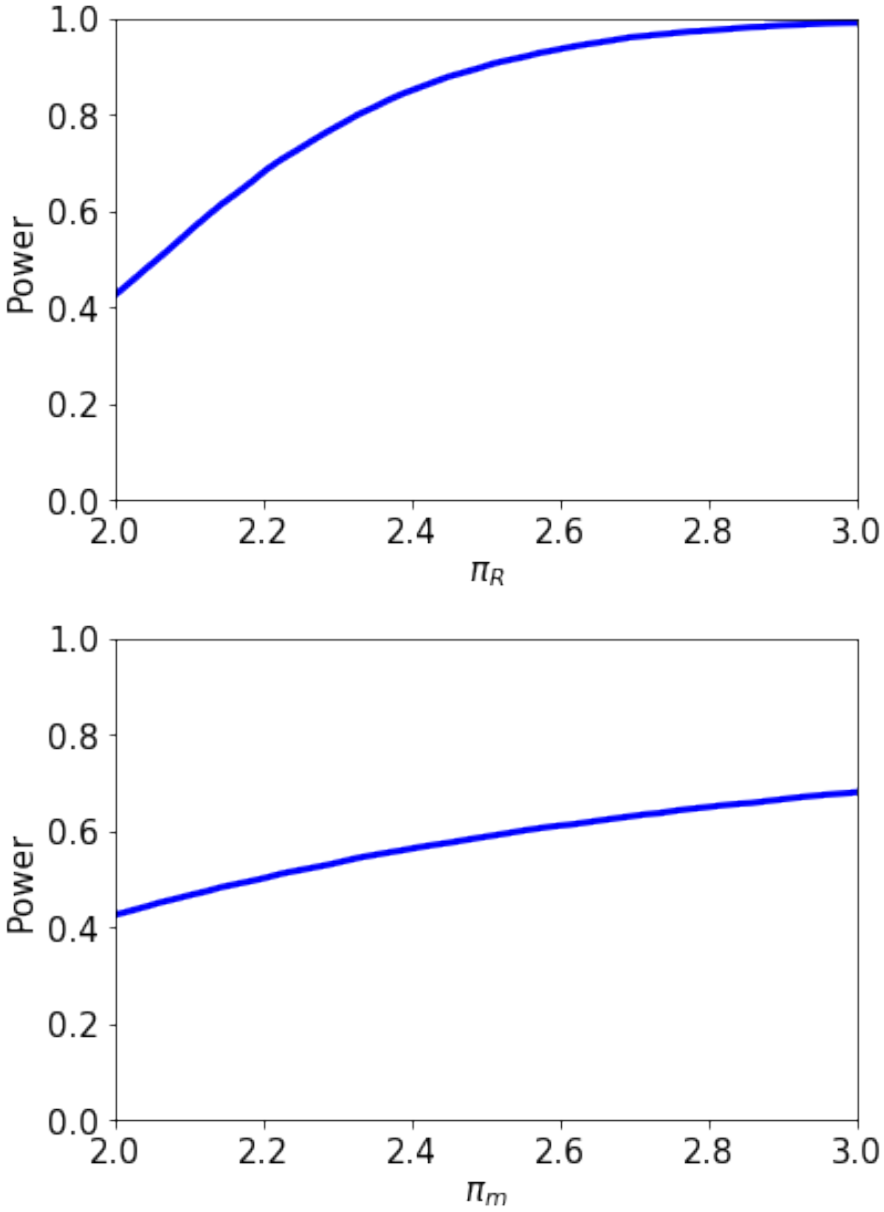


Figure 6: Power Curves for π

6 Ex-ante Power Analyses of Prominent Papers using the QR framework

I demonstrate the application of the QR framework using a thought experiment where I place myself in a similar position as authors of previous papers, and conduct an ex-ante power analysis. In this thought experiment, I take one of the following three approaches. In the first approach, the data-set from the original experiment is ignored and only data from previous related experiments are used. In the second approach a small artificial 'pilot' data-set is sampled from the original experiment. In the third approach, QR parameters are selected where subject behavior is consistent with the spirit of the model in question, which is particularly useful in situations where a structural model proves unidentifiable. All of these situations are similar to the ones a researcher might find themselves in. I only access data that is roughly contemporaneous or previous to the published article and is publicly available.[21] The previous experiments I will use for this exercise come from the ones selected by the Experimental Economics Replication Project (EERP) (Camerer et al., 2016). These are good candidates as they represent a broad range of economics experiments from high-impact journals, and narrows the exercise to the specific 'main' hypotheses selected by the EERP. It should be noted that the QR framework cannot be applied to all of these papers. In particular, there needs to be explicit theory specified that incorporates the channel through which the treatment effect is thought to operate. However, oftentimes small modifications or simplifying assumptions are sufficient to introduce an operative treatment channel when none is explicitly specified. These modifications should not be considered a suggestion of the correct way to model the channel in question, which is far beyond the scope of this exercise. Instead, they are presented as a way to create reasonable predictions of subject behavior ex-ante in environments that would otherwise prove intractable to a simulation approach. The point of the QR framework is to try and improve upon what is currently standard (i.e. subjective extrapolation from previous data, or no ex-ante power analysis at all), rather than provide exact predictions, which is an impossible task ex-ante. I conduct ex-ante power analyses using the QR approach on 7 of the papers in the EERP. I present 5 of the more interesting cases in the following sections, and the remaining 2 in Appendices D.1 and D.2. In addition, where applicable I provide suggestions on what parameters could be changed to increase power. This demonstrates using QR simulations for optimal experimental design, and is also in the spirit of an efficient conceptual replication, as more optimal parameters could be selected to improve power and reduce subject costs.

[21] I define 'publicly available' data as data made available with the journal article or on another website. It is highly likely that authors would be willing to share other data upon request, but technically this requires data transfer agreements to be in compliance with some Human Ethics Committees.

6.1 Abeler et al. (2011)

Abeler et al. (2011) use a real effort task to test a model of expectations-based reference-dependence (e.g. Kőszegi and Rabin (2006)). In the first stage, they have subjects complete a set number of tasks where they counted the number of zeros in a table. In the second stage, it is announced they can complete as many of these tasks as they want at a piece rate w per task, but that this would only be paid to them 50% of the time. Otherwise, they would be paid a fixed sum f that was independent of the number of tasks they had completed. The treatment was to vary f, to either be 3 euros or 7 euros. According to expectations-based reference-dependence, this should make subjects more likely to stop at the point where they have performed enough tasks to earn f in the eventuality they are paid for their tasks. This is because earning more than f from the tasks introduces psychologically costly losses if f eventuates, and vice-versa for earning less than f from the tasks. A standard model of effort provision would predict no differences in f, rather subjects would perform the task up until the point that their marginal cost of performing that task exceeds half of the piece rate.

6.1.1 Ex-ante Power Analysis

An expectations-based reference-dependent agent in this environment faces the following piece-wise utility function:

$$U(e;w,f) = \begin{cases} \frac{we+f}{2} - c(e) + \frac{1}{2}\eta[\frac{1}{2}\lambda(we-f)] + \frac{1}{2}\eta[\frac{1}{2}(f-we)], & \text{if } we < f \\ \frac{we+f}{2} - c(e) + \frac{1}{2}\eta[\frac{1}{2}(we-f)] + \frac{1}{2}\eta[\frac{1}{2}\lambda(f-we)], & \text{otherwise} \end{cases},$$

where e is the number of tasks completed, $\eta \geq 0$ is the sensitivity to reference-dependence, $\lambda > 1$ is loss aversion (not to be confused with λ_{QR}), and $c(e)$ is the cost of performing e tasks.

There are two implementation issues for this environment. Firstly, the variables η and λ are not separately identifiable, as multiple combinations of the two could explain the same decision. Secondly, this particular environment was relatively novel at the time, and there does not appear to be a wealth of previous experimental evidence with which the authors could draw upon. If there were, for example, a similar real-effort task experiment, it might be possible to obtain a reasonable estimate of $c(e)$. Given these limitations, I specify reasonable values for the parameters given the goal of testing a model of reference-dependence. For $c(e)$ I assume $c(e;a) = ae^2$, which encapsulates increasing costs in a simple one-parameter model. I consider what the payoff hills will look like for some various combinations of η, λ, and a, presented in Table 2.

Table 2 highlights that it is important that the costs do not outweigh the reference-dependent and loss

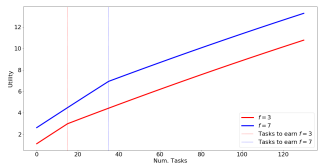
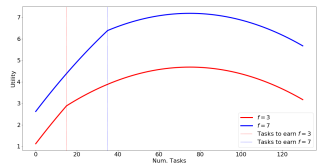
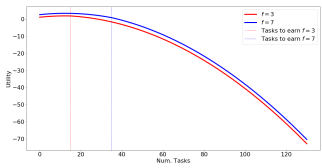
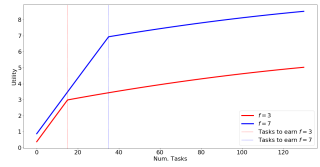
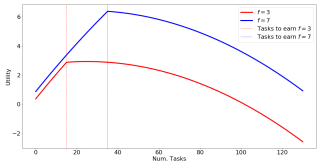
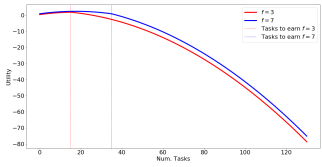
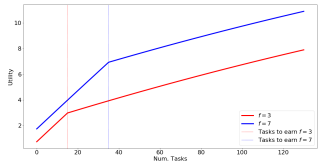
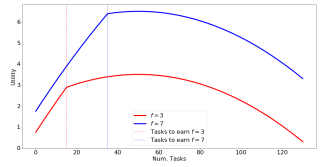
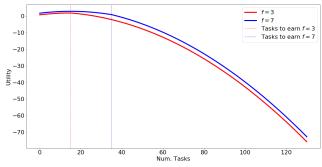
	$a = 0.0001$	$a = 0.001$	$a = 0.01$
$\eta = 0.5$ $\lambda = 2$			
$\eta = 1.5$ $\lambda = 2$			
$\eta = 0.5$ $\lambda = 3$			

Table 2: Payoff hills for various parameters in Abeler et al. (2011)

averse parts of the utility function.[22] If costs are too low, then subjects will want to complete as many tasks as they can, and reference-dependence will only have a small effect. If costs are too high, then subjects will not want to complete many tasks at all, and again, reference-dependence will have little effect. An intermediate cost parameter can still overpower the effect of reference-dependence depending on the parameters, as the center column of Table 2 indicates. Two of the three specifications for $a = 0.001$ result in the same effort point prediction for both treatments, whereas the specification in the middle has a point prediction consistent with reference-dependence.

All of the given combinations would predict some treatment effect in f, but given the focus on reference-dependence, I select $\eta = 1.5$, $\lambda = 2$ and $a = 0.001$ (i.e. the center graph of Table 2). The parameter of λ_{QR} remains to be specified, which is estimated from a small artificial pilot of 10 subjects per treatment.[23] This estimation yields an estimate of $\lambda_{QRE} = 0.741$, which then implies $\tau = 10.297$, $\sigma_{F=3} = 25.039$, and $\sigma_{F=7} = 22.216$. These values are based on the number of tasks completed, which are paid at a piece rate of 0.20 euros. The comparison for τ from the point prediction is therefore $\frac{7-3}{0.20} = 20$ tasks, meaning the QR predicts a treatment effect size of approximately half of that implied by the point prediction. According to a t-test (which is equivalent to an OLS regression with treatment as a dummy variable utilized in the actual

[22] It is not possible to exactly control the cost of the real effort task, however, it is possible to indirectly influence it, by making the task more or less difficult or unpleasant.

[23] Note, it would not possible to fit the other parameters using this pilot data-set, as η, λ, and $c(e)$ are not separately identifiable, especially on such a small data-set.

	$f_{high} = 6.60$	$f_{high} = 7.00$	$f_{high} = 7.40$
$f_{low} = 2.60$	$\tau = 9.930$ $\sigma_{low} = 25.296$ $\sigma_{high} = 22.502$ $Power = 75.9\%$	$\tau = 11.088$ $\sigma_{low} = 25.296$ $\sigma_{high} = 22.216$ $Power = 86.5\%$	$\tau = 12.268$ $\sigma_{low} = 25.296$ $\sigma_{high} = 21.934$ $Power = 91.4\%$
$f_{low} = 3.00$	$\tau = 9.139$ $\sigma_{low} = 25.039$ $\sigma_{high} = 22.502$ $Power = 70.4\%$	$\tau = 10.297$ $\sigma_{low} = 25.039$ $\sigma_{high} = 22.216$ $Power = 80.2\%$	$\tau = 11.478$ $\sigma_{low} = 25.039$ $\sigma_{high} = 21.934$ $Power = 88.8\%$
$f_{low} = 3.40$	$\tau = 8.293$ $\sigma_{low} = 24.769$ $\sigma_{high} = 22.502$ $Power = 62.2\%$	$\tau = 9.451$ $\sigma_{low} = 24.769$ $\sigma_{high} = 22.216$ $Power = 72.8\%$	$\tau = 10.632$ $\sigma_{low} = 24.769$ $\sigma_{high} = 21.934$ $Power = 84.2\%$

Table 3: QR simulations for different combinations of f in Abeler et al. (2011)

paper), with these parameters 83 subjects per treatment to reach a power of 80% at the 5% level.

6.1.2 Optimal Experimental Design

The most obvious candidate to increase power would be to increase the distance between the f's in each treatment, assuming standard deviations do not substantially rise when doing so. However, there are practical constraints on f, for example, if f was quite high, it might induce expectations-based loss-averse subjects to stay in the lab doing the task for excessive periods of time. In addition, the values of f should be 'clean' to present to subjects (e.g. $f = 2.79$ is not great, while $f = 2.80$ would be fine), and in multiples of w to align with the reference-dependent predictions (a preference to match $f = we$). There may also be a design goal of keeping each reference-dependent prediction sufficiently different from standard predictions (here, given w and a, $e^* = 50$). I therefore implement a restricted grid search around f's that were (presumably) carefully selected by the original authors. In particular, the grid search is over combinations of $f_{low} \in \{2.60, 3.00, 3.40\}$ and $f_{high} \in \{6.60, 7.00, 7.40\}$, which is presented in Table 3.

Table 3 confirms that increasing the distance between the f's increases power. This is despite the fact that lowering f_{low} increases the standard deviation of that treatment, as the increase in treatment effect size is clearly offsetting this. The overall optimal configuration is $f_{low} = 2.60$ and $f_{high} = 7.40$, which would require 59 observations per treatment, less than the 83 observations required for the original treatment parameters.

6.2 Charness and Dufwenberg (2011)

Charness and Dufwenberg (2011) experimentally investigate the impact of communication through free-form messages prior to decision making could have on behavior in a hidden information principal/agent game, as displayed in Figure 7. Player A is the principal, and player B is the agent. The agent has a type, either

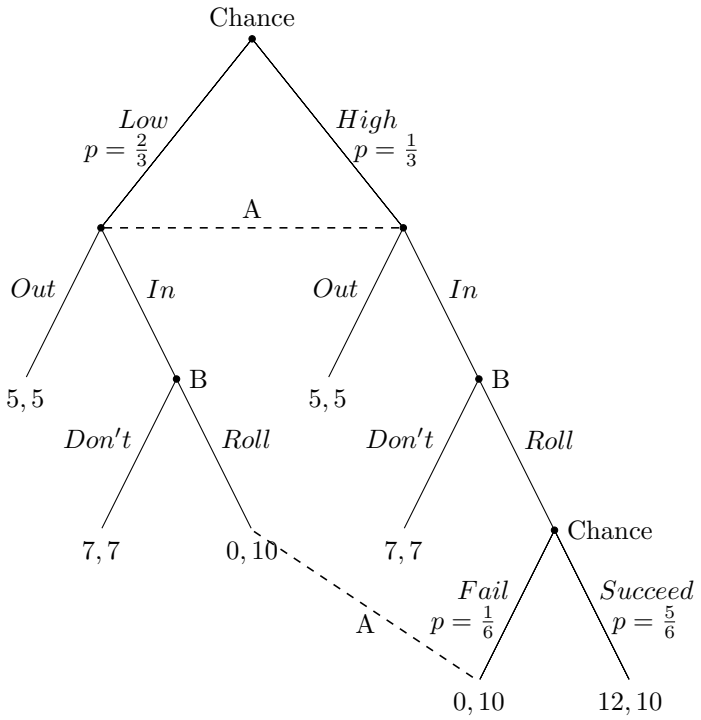


Figure 7: Charness and Dufwenberg (2011)'s '(5,7) Game'

low or high, which determines their suitability for a difficult but high paying task. If a low agent attempts the difficult task, they always fail, but if a high agent attempts the difficult task, they only fail with $p = \frac{1}{6}$. The agent's type is known to them when making their decision, and is never fully revealed to the principal if the type is low. Given this arrangement, the principal must decide whether they want to hire the agent and let them decide what difficulty task to undertake, or collect some outside option instead. The particular treatment effect is whether allowing B's to send a free-form message to their paired A discourages low B's from choosing Roll (or in other words, undertaking the unsuitable high difficulty task).

6.2.1 Ex-ante Power Analysis

The effect of communication is often not explicitly incorporated into theory, but Charness and Dufwenberg (2011) suggest two possible channels, a cost of lying and guilt from blame. I incorporate a simple cost of lying model supplemented by other-regarding preferences, as this is independent of the complications of first- and second-order beliefs present in guilt models. Both types of players have the monetary payoffs of their counterpart enter into their utility function, weighted relative to their own payoff by the parameter α. In other words, A's utility function is $U_A(x_A, x_B) = (1-\alpha)x_A + \alpha x_B$, where x_A and x_B are the monetary payoffs of players A and B respectively. I assume that when given the opportunity, B sends a message indicating that they will choose Don't when they are the low type, and incurs some disutility k if they do not follow through

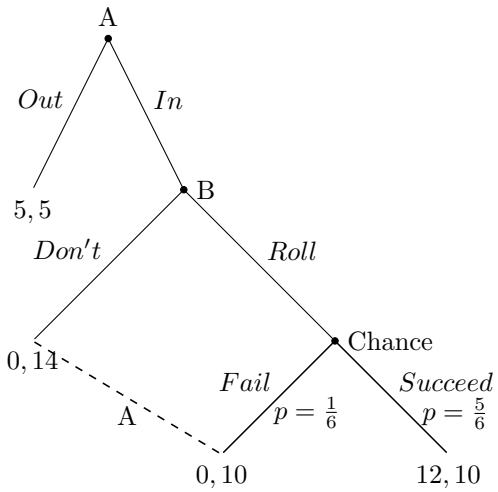


Figure 8: Game in Charness and Dufwenberg (2006)

on this, to reflect a cost of lying. I denote the two treatments as M (Messages) and NM (No Messages), with $k_M > k_{NM} = 0$. So B's utility function if they have told a lie is: $U_B(x_A, x_B) = (1-\alpha)x_B + \alpha x_A - k_M$, and if they have not told a lie: $U_B(x_A, x_B) = (1-\alpha)x_B + \alpha x_A$. For any given α, k_M, and λ, the QRE provides predictions for the probability of a low B choosing Roll, a high B choosing Roll, and A choosing In.

The next step is that the values of α, k_M, and λ need to be specified. The best approach would be to use a closely related study. Fortunately, the game considered in Charness and Dufwenberg (2006) (presented in Figure 8) is a good candidate for this purpose. It has similar forms of communication, subject pools, payoff magnitudes, and a hidden action. I assume other-regarding preferences in the same manner as before, and that B faces the same disutility from lying, except that in this game he now promises to always Roll, and thus faces disutility when choosing Don't. I specify a log likelihood function that depends on the data observed in Charness and Dufwenberg (2006) and the parameters α, k_M, and λ. Using this function, I use maximum likelihood estimation techniques to fit values of the parameters and obtain $\alpha = 0.18$, $k_M = 2.15$, and $\lambda = 4.41$.

I apply the estimated parameters to the QRE of the current game, which provides a prediction for the probabilities for each action in each treatment, summarized in Table 4. With these probabilities, I simulate a data-set of size n, where n is the candidate total subjects needed.[24] It should be noted that the current experimental design calls for a direct response method, so power needs to be considered not only in terms of the treatment effect that is likely to eventuate, but the number of subjects that will actually undertake the relevant decision (i.e. low type B). In order for that to occur, the B's type must be assigned as low by nature, but it must also be the case that the type A player chooses In. Therefore, it is important that A's

[24] I assume that each treatment has equal sample sizes, but this assumption could be relaxed.

	Messages	No Messages
Prob. A In	0.74	0.60
Prob. Low B Don't	0.64	0.34
Prob. High B Don't	0.21	0.21

Table 4: Probability of Actions by Treatment

differential behavior between treatments is modeled given B's cost of lying, despite the fact that I am not directly testing any hypotheses on A's behavior. Due to the direct response method, I simulate each A/B pair sequentially as the game tree indicates, i.e. I draw B's type, A makes their decision, and if that decision is In, B then makes their decision given their type, which is then added to the data-set. I generate 10,000 'power blocks' of data-sets with n total subjects, so $n/2$ subjects per treatment, or $n/4$ pairs of A and B per treatment. On each simulated data-set, I conduct the statistical test on low B decisions, which in this case is a z-test for two-proportions, and count how many times the null hypothesis is rejected. I then adjust n in multiples of 4, until the lowest n that has a power of greater than 80% is obtained. This process suggests a total number of subjects of 388, or 97 pairs of A and B per treatment.

6.2.2 Optimal Experimental Design

I now consider if there are any parameters that could be changed that would increase power. A related question of the optimal treatment intensity is not relevant in this environment, as it is not immediately obvious how the intensity of communication could be varied, and the impact that would have.[25] I instead turn to the other parameters that are fixed in both treatments. I consider that practical constraints exist in the form that numbers should be easily communicable to subjects (e.g. $p = \frac{1}{6}$ is fine, but $p = .17258013$ is not), and that payoffs should not move too far (e.g. increasing one payoff to \$100 may increase power substantially, but would not be affordable). The theory and motivation of the paper also indicates that some parameters should also change together. For example, a low B choosing Roll should have the same payoffs as a high B choosing Roll but failing, or otherwise Roll is not a hidden action.

With that in mind, I consider either raising or lowering each parameter in isolation to the nearest 'clean number' (i.e. integer for payoffs, nearest tenth for probabilities). Table 5 presents the power for raising or lowering one parameter in isolation, holding the other parameters fixed at the levels that were used in the experiment. Table 5 also presents the results of a grid search over every possible combination of the given parameters. The largest increases in power come when lowering the payoff of the outside option or lowering B's payoff in the event that a Roll fails. Both of these factors encourage A to more frequently choose In, which increases power as it increases the number of low B's that actually get to make a decision.[26] Changing

[25] Perhaps if the message is only sent with a certain probability?

[26] This illustrates the advantage of using the strategy method, as a B decision for the case they are low would be observed for

Parameter	Lower	Baseline	Higher
Prob. low $\in \{0.6, \frac{2}{3}, 0.7\}$	78.8%	80.3%	80.5%
Out payoff $\in \{4, 5, 6\}$	87.6%	"	66.6%
Don't payoff $\in \{6, 7, 8\}$	60.4%	"	79.9%
A fail payoff $\in \{0, 1\}$	N/A	"	80.3%
B fail payoff $\in \{9, 10, 11\}$	86.1%	"	63.0%
A succeed payoff $\in \{11, 12, 13\}$	78.4%	"	81.8%
Prob. Succeed $\in \{0.8, \frac{5}{6}, 0.9\}$	79.4%	"	81.8%
Grid Search	0.7,4,6,0,9,13,0.9		93.7%

Table 5: Power for various parameters in Charness and Dufwenberg (2011)

only the most effective parameter in isolation, the outside option, would reduce the total number of required subjects to 320 to reach a power of 80%. The grid search set of parameters would require even fewer subjects again, for a total of 256.

6.3 Chen and Chen (2011)

Chen and Chen (2011) experimentally test the role of group identity on coordination in a two-player minimum effort game. In particular, they induce group identity in an 'enhanced' treatment by not just assigning a group, but reinforcing that identity by having the group help answer a paid question about identifying paintings. They posit that group identity can increase the concern that others have about the payoffs of those in their 'in-group'. The increase in other-regarding preferences for in-group people can change the cost threshold in which a QRE would predict low or high levels of effort in a minimum effort game.

Chen and Chen (2011) provide the following explicit utility function for a two-player minimum effort game:

$$U_i(x_i, x_j; \alpha_j) = a \min x_i, x_j - c[(1 - \alpha_j)x_i + \alpha_j x_j],$$

where a is the benefit of an additional unit of minimum effort, c is the (non-refundable) cost of effort, and α_j is the relative weighting that i placed on j's payoff, which can differ if the other is in the out-group O, or the in-group I (α_O, α_I).

6.3.1 Ex-ante Power Analysis

Chen and Chen (2011) use the $a = 1$, $c = 0.75$ environment from Goeree and Holt (2005), where subjects can choose efforts of $x \in [110, 170]$ in steps of 0.01 (i.e. they can choose from 110, 110.01, 110.02, ..., 169.98, 169.99, 170). They then calculate the QRE for different levels of α given the $\lambda = 0.125$ reported in Goeree

every B, regardless of A's decision and the random draw. However, it may be the case that using the strategy method could change behavior relative to the direct response method (Brandts and Charness, 2000; Casari and Cason, 2009).

Sessions Per Treatment	Power
2	57%
3	65%
4	74%
5	78%
6	81%

Table 6: Ex-ante Power Analysis for Chen and Chen (2011)

and Holt (2005). This is what the QRE approach would have done had it not already been conducted, as Goeree and Holt (2005) is clearly the most closely related environment from which an estimate of λ would be obtained from. The only thing that remains is to specify α_O and α_I, as the hypothesis is whether effort levels differ substantially if the counterpart is in the same group, or in a different group. This would usually be difficult, as there are many different ways of inducing group identities, and this would likely be sensitive to the method used as well as the subject pool.[27] Fortunately, Chen and Li (2009) conduct a near-identical task to induce group identity on a similar subject pool. They obtain estimates of α_I and α_O from tasks relating to social preferences and simple reciprocity games in the style of Charness and Rabin (2002). I take the parameters $\alpha_I = 0.474$ and $\alpha_O = 0.323$ from the Chen and Li (2009) paper.

The statistical test used in this experiment is a panel data regression with clustering at the session level and individual random effects. If the statistical test controls for such things, then the generated data-set should justify the use of these controls, in order to obtain a reasonable estimate of power. To justify session level clustering, I draw a $\lambda \sim U(0.1, 0.15)$ for each session.[28] As for the individual random effect, which is modeled as a normal draw from $N(0, \sigma_e)$, I run a panel regression on the $c = 0.75$ data-set from Goeree and Holt (2005), and obtain an estimate of $\sigma_e = 15$. In the simulation, in each period the subject's action is drawn from the session level QRE distribution, and then the individual's random effect is applied. After this, their effort decision is rounded to the nearest possible valid action.

This particular experimental design calls for exact session sizes of 12 subjects, and so the minimum 'step' for power consideration is 12, of which I will refer to as a session. Table 6 reports that 6 sessions per treatment is required for a power of 80%. This means that 120 subjects would be required in total.

6.3.2 Optimal Improving Power through Experimental Design

The first thing that should be noted is that increasing α_I or decreasing α_O would increase power, however, it is not clear how specific changes in the group identity induction would change the α parameters. So, if there is a better way to induce stronger group identity then it should be implemented, but that this is not something

[27] For example, see the replication report in Camerer et al. (2016) for this paper. The same group identity task seems to have not induced much group identity in a different subject pool, which may be because subjects did not interact as much with the chat task.

[28] Although, drawing a session-level α would also be a valid approach.

	Baseline	$c = 0.7$	$c = 0.8$	$x \in [115, 165]$	$x \in [105, 175]$	x step= 1	x step= 10
τ	6.61	5.12	7.29	4.03	9.94	6.04	4.06
σ_O	18.32	18.16	18.23	16.18	20.28	18.48	19.62
σ_I	17.74	17.56	17.97	15.90	19.22	18.04	19.66
Power at 2	57%	52%	61%	38%	76%	51%	38%
Power at 3	65%	60%	75%	37%	86%	62%	39%
Power at 4	74%	63%	75%	48%	92%	66%	49%
Power at 5	78%	70%	88%	52%	97%	72%	54%

Table 7: Ex-ante Experimental Design for Chen and Chen (2011).

τ is the empirical average treatment effect size by power block for the 5 session data-set, σ_O is the empirical average standard deviation of the outgroup's effort levels by power block for the 5 session data-set, similarly for σ_I but for the ingroup, and finally 'Power at X' is the simulated power for the X sessions per treatment data-set.

the QR framework can address without some estimate of the induced α. The remaining parameters that could vary are a and c. I focus on the cost c, as $a = 1$ is convenient to explain to subjects. I consider only small changes in c, to $c = 0.7$ and $c = 0.8$, as $c = 0.75$ was originally chosen as it tended to result in low effort in experiments without group identity, and the authors posited that that the introduction of group identity might overcome this. Another design consideration is the set of permissible effort levels that can be chosen. I consider the range of actions, in that I expand and reduce the action space by 10 effort points. Additionally I consider making the action space coarser, by considering integer steps and steps of 10, instead of the original step size of 0.01. In this exercise, I hold all other parameters at the same original level. The output is summarized in Table 7.

What Table 7 reveals is that increasing c increases power, because it decreases effort in the outgroup treatment by more than the ingroup treatment.[29] It also reveals that increasing the range of permissible actions also increases power, which is because the higher efforts will be more likely to be played by in-group players (and vice versa), which in the QRE can have a reinforcing effect. The increase in the treatment effect dominates the increase in standard deviation brought about by the additional noise due to having more available actions. Finally, increasing the coarseness of the strategy space decreases power, so this should be avoided. Figure 9 illustrates the effect of the action space coarseness, where increasing the size of the step causes substantially more low effort decisions for both treatments, compressing the treatment effect. Of the possible design changes, increasing the range of actions would be the most beneficial for power. Utilizing the increased strategy space design would require only 3 sessions per treatment to attain 80% power, less than the 6 sessions per treatment required under the original parameters.

[29] This effect would not be monotonic though - increasing $c \geq 1$ for instance would result in a decreased treatment effect size, as both treatments would begin to choose the lowest possible effort.

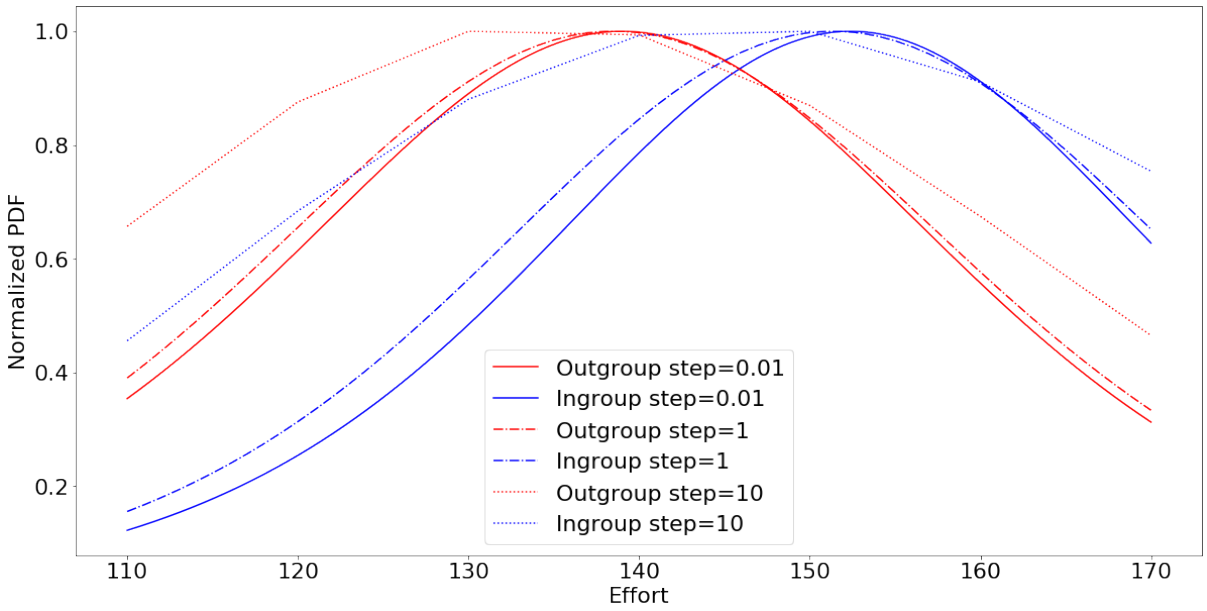


Figure 9: Effect of increasing step size in the Minimum Effort Game

6.4 Fehr et al. (2013)

Fehr et al. (2013) consider the effects of delegation in a principal agent game where there are a number of projects, one of which must be selected to be implemented. There is a known 'outside option' project, while the states of the all of other projects are unknown, but can be revealed through costly effort. The principal is initially endowed with the decision right on what project to implement. If the principal retains this decision right, the agent can provide a recommendation for which project to do, while the principal makes his decision based on his own knowledge and/or the recommendation. If the principal delegates this decision right, the roles are reversed, the principal can provide a recommendation but it is the agent who makes the final decision based on the agent's knowledge and/or the recommendation. The game starts with the principal deciding whether to delegate or not, which is followed by simultaneous effort decisions by both players. The effort level chosen by an individual is the probability that the true state of all projects will be revealed to that individual. The final stage is the controlling party deciding which project to implement, given their own information (if the true state was revealed to them), and the recommendation from the other party. In this particular setup, there are 36 projects, 1 of which is always known and pays out 10 to each subject. Of the remaining 35, 33 of the projects pay out 0 to each subject. And, of the remaining 2, one of these projects is preferred by the principal, and the other is preferred by the agent (although both of these projects are preferred to the known outside option). The cost of effort function is $c(e_i) = 25e_i^2$, where $e_i \in [0, 1]$, and the true state is revealed to i with probability $p(e_i) = e_i$. Fehr et al. (2013) consider a range of payoff values for projects, such that the principal is predicted to either want to delegate or not. However, they find that

Num. Groups per Treatment	Power
$N = 3$	64.4%
$N = 4$	77.7%
$N = 5$	84.7%
$N = 6$	88.0%

Table 8: Fehr et al. (2013) Power Analysis

delegation is substantially lower than predicted.

6.4.1 Ex-ante Power Analysis

Given the low delegation rates from the initial experiments, Fehr et al. (2013) conduct additional treatments to test whether the low delegation rates were driven by a non-pecuniary cost of being overruled. They propose a treatment where whichever party is not in control cannot exert any effort, and therefore cannot provide a recommendation that could be overruled. This is the treatment effect that the EERP considers, that the delegation rates in this treatment are higher than an environment which has the possibility of being overruled but has similar Nash equilibrium returns to delegation. Because this experiment was designed after observing the initial results, I can place myself in the exact position of the authors, by restricting myself to only consider their baseline set of data.

I fit a QRE to the baseline data with one modification to the model. In particular, I add an additional parameter to be estimated, that is a disutility D of being overruled. I assume that a subject is overruled when they do not have control over the final decision, but that they have full information (i.e. the true state is revealed to them) and thus provide a recommendation which is not then followed. This changes effort levels in equilibrium, as well as the delegation decision. Fitting this model yields $D = 41.22$ and $\lambda = 0.216$. The disutility parameter is substantial, and implies that being overruled is disliked so much as to more than wipe out even the highest pecuniary incentive. I then apply these parameters to a QRE of the two additional treatments, and conduct a power analysis on the generated data-sets. Because the conducted statistical test incorporates a learning trend, I increment λ linearly by period, such that the estimated $\lambda = 0.216$ is obtained halfway into the experiment. As a higher λ is associated with less noisy play, this is an appropriate way to model a learning effect. In addition, the statistical test clusters at the individual level. I justify clustering at this level by having a subject's disutility of being overruled to be drawn from $N(41.22, 10)$. The results are presented in Table 8, and suggest that 5 groups per treatment (each group had 10 subjects in it, so a total number of 100 subjects), would be sufficient to reach 80% power.

6.4.2 Optimal Experimental Design

The two additional treatments were carefully and explicitly designed so that the Nash equilibrium returns to delegation were very similar. To facilitate this, and because the two environments are procedurally different, the payoffs for different outcomes are necessarily different between the two treatments. Changing these payoffs would mechanically induce a difference in delegation rates quite trivially (due to the equilibrium returns to delegation), but it would not be due to the mechanism that is proposed. This particular hypothesis is not a good application of the QRE approach to optimal experimental design. Therefore, I do not suggest alternative parameters to increase power for this paper.

6.5 Ifcher and Zarghamee (2011)

Ifcher and Zarghamee (2011) conduct an experiment where time preferences are elicited directly after viewing a video clip designed to induce a positive mood. As compared to a neutral video clip, they find that subjects exhibit more patience (i.e. discount the future less) when in the induced positive mood.

6.5.1 Ex-ante Power Analysis

The challenge in this environment is to specify an explicit functional form with which positive affect can operate on time preferences, as well as coming up with a reasonable estimate as to the magnitude of the effect that positive affect would have. A brief foray into the experimental economics literature that induces positive affect in a similar way yields some papers with other-regarding preferences and strategic environments (Capra, 2004; Kirchsteiger et al., 2006; Drouvelis and Grosskopf, 2016). I relate other-regarding preferences with time preferences by assuming people see the present-self and the future-self as being different entities, as posited by Pronin and Ross (2006). I therefore assume that actions towards the future-self are similar to actions towards other people in the present. In particular, I assume that subjects exhibit quasi-hyperbolic discounting, $D(y,t) = y\beta_i e^{-rt}$ if $t > 0$, where the present bias β_i can differ depending on what mood the subject is in.

Ifcher and Zarghamee (2011) elicit time preferences by asking for what value of $\$p$ would make them indifferent between receiving $\$p$ today and receiving $\$m$ in t days time. They consider 30 combinations of m and t, and incentivize the reporting of the indifference level of p using a discrete BDM mechanism. In this mechanism, balls labeled $b \in \{1, 2, ..., m+1\}$ are placed into an urn and then a ball is drawn at random.[30] If the ball drawn b is less than or equal to p, then the subject gets paid m in t days time, whereas if $b > p$, the subject gets paid $\$b$ today. The discretization of the BDM mechanism means that ranges of p result in

[30] In the case where $m+1$ is not an integer, the final ball is the next highest integer above m.

Subjects per Treatment	Power
$N = 30$	74.1%
$N = 31$	74.5%
$N = 32$	77.0%
$N = 33$	78.2%
$N = 34$	80.1%
$N = 35$	81.8%

Table 9: Ifcher and Zarghamee (2011) Power Analysis

the same payoff outcome, so I discretize the subject's p decisions as the mid-points of these ranges.[31]

A good place to obtain a reasonable estimates for β_N (neutral mood), r and λ would be from Benhabib et al. (2010), who elicit time preferences in a very similar manner, just without any mood induction. However, this data-set is not currently publicly available, and would not give an estimate for β_G (induced good mood) anyway. Instead, I assume a small initial pilot is conducted, and fit the model to that. I simulate such a pilot by taking 10 subjects per treatment from the full data-set. For the artificial pilot, the estimated parameters are $\beta_G = 0.814$, $\beta_N = 0.701$, $r = 0.002$, and $\lambda = 5.633$. Finally, the statistical test that is conducted is a regression clustered at the individual level. In order to justify individual level clustering, I assume that the induced mood has differential effects on each subject, such that $\beta_i \sim U[\beta - 0.025, \beta + 0.025]$. Table 9 presents the results of this QRE simulation power analysis. The required number of subjects per treatment to reach 80% power is 34 subjects per treatment, so 68 subjects in total.

6.5.2 Optimal Experimental Design

The obvious initial change here would be to induce a bad mood rather than a neutral one, which has been shown to change other-regarding preferences in strategic settings (Drouvelis and Grosskopf, 2016). I assume a similar decrease from β_N to β_B as the increase from β_N to β_G, and set $\beta_B = 0.60$. This increases power substantially, for example at $N = 34$ power increases from 80.1% to 100%. Such a design change would then require only $N = 8$ subjects per treatment to reach a power of 80%. Another design change could be to reduce the discretization of the BDM mechanism to provide a richer action space with which subjects could exhibit their true preferences. I specifically consider adding additional balls such that $b \in \{0.1, 0.2, 0.3, ...m + 0.9, m + 1\}$, substantially decreasing the coarseness of the effective strategy space in p. However, the effect on power is small, driven by small decreases in the standard deviation due to the finer space. The only minor improvement of power justifies the original experimental design, as it may be easier to explain the BDM mechanism with discrete integer ball drawing.

[31] For the unbounded above upper range, I add the difference between the midpoint and lower-bound of the other ranges to the lower-bound of this range.

6.6 Summary

Table 10 presents a summary of the total number of required subjects required by the ex-ante QR simulation approach using the same parameters that were selected in the experiment, as well as if 'optimal' parameters in terms of power were selected. It also presents the total number of subjects that were actually used in the original experiments, and an ex-post approach that uses the same power analysis method as the EERP (except for 80% power). Table 10 shows that the QR simulation approach is typically more conservative than the actual experiments. This is not necessarily surprising given that these papers are from a time when power analysis was less common. It is not exactly clear from the listed papers how these subject numbers were determined. The QR framework also compares somewhat favorably to the required number of subjects from an ex-post approach, where all data are known. This is a very high bar to clear, so having the QR framework be anywhere near those numbers is promising, and suggests the QR framework can be a reasonable way to improve ex-ante power analysis. Finally, in some environments there exists room to increase power through changes in the experimental design, as indicated by the lower subject numbers required when optimal parameters are selected.

Paper	Ex-ante QR (Original Params.)	Ex-ante QR (Optimal Params.)	Ex-ante Original	Ex-post* (Camerer et al., 2016)
Abeler et al. (2011)	166	118	120	237
Charness and Dufwenberg (2011)	320	256	162	192
Chen and Chen (2011)	120	72	72	125
De Clippel et al. (2014)	200	40	158	115
Fehr et al. (2013)	100	N/A	60	73
Huck et al. (2011)	60	N/A	120	114
Ifcher and Zarghamee (2011)	68	16	58	98

Table 10: Required total number of subjects

*Using the z-approximation formula for 80% power, rather than the 90% used in the EERP.

7 Conclusion and Future Research

Power analysis is an important, but until recently under-considered, factor in experimental economics research. Power analysis should be the sole determinant of the number of subjects to be used in an experimental study. However, the correct way to conduct a power analysis is only clear when enough data already exists in the same treatments that are to be conducted, a situation a researcher only finds themselves in when conducting a replication. Instead, the researcher looks to the most closely related previous experiments, and must extrapolate from there what behavior might look like in the novel environments that they are

considering. The question is how this extrapolation is done, as there is substantial room for subjectivity in this process, which can reduce the accuracy of an ex-ante power analysis. To address this issue, I propose the use of a QR simulation based approach to power analysis. The QR framework assumes that subjects make proportional errors in line with the relative payoff of each action, or in other words, their payoff hill. I demonstrate that the steepness of the payoff hill as well as potential asymmetry can influence the parameters used in a power analysis, and that assuming normally distributed noise about theoretical point predictions can result in under-powered studies. The QR simulation approach to power analysis uses structural techniques on previous but related data-sets to calculate appropriate estimates of the QR noise parameter λ, as well as any additional parameters that are required for the operative channel of the treatment effect to work. With those parameters, the QR framework provides predictions for behavior in novel environments. Those predictions can then be used to generate simulated data-sets, with which power analyses can be conducted even in situations where a closed-form solution for power does not exist.

I demonstrate the use of the QR simulation approach using a motivating example of Bayesian Persuasion. That particular experimental environment shows the potential failings of a more standard rule of thumb due to asymmetric payoff hills that vary in steepness by treatment. It is not always the case that treatment differences should be maximized when considering the objective of maximizing power. I then use the QR framework to conduct ex-ante power analyses on various high-profile papers. From an ex-ante perspective, QR simulations usually perform well, getting reasonably close to the objectively correct ex-post approach. For selected papers, I provide an ex-ante suggestion for the optimal selection of parameters or treatments that would improve power. This results in a reasonable decrease in the number of required subjects, potentially saving both time and money. Such a process can also be conducted ex-post, when considering replication. Instead of the more standard direct replication, the researcher could conduct a conceptual replication with the goal of minimizing the cost of the replication, an 'efficient' replication. Lowering the costs of replication could result in more provision of this currently under-provided research.

Future research should take the following messages from this paper. Firstly, an ex-ante power analysis should be included to justify the number of subjects in any experimental paper, and ideally this should be pre-registered. The profession as a whole should discuss phasing this in so that in the future a clear ex-ante power analysis plan would be a requirement for publication. Secondly, the QR framework is applicable to a wide variety of experimental environments, and can not only be used for power analysis in novel environments, but it can also be used as a tool to guide experimental design both in terms of power as well as any other conceivable metric. As for future research on this particular methodological topic, further confirmation that this approach provides reasonable predictions from an ex-ante perspective would be helpful. In addition, conceptual replications in the spirit of being efficient from a power perspective would be an important

contribution to supplement more traditional direct replications.

A Example Stata and Python Code for a Simulation based Power Analysis, with a Normal DGP

A.1 Stata Code

```
clear all

// num_sims = number of generated data-sets (the bigger the better)
scalar num_sims = 1000
// n - number of observations per treatment_vector
scalar num_obs = 63

// DGP params:
scalar intercept = .5
scalar tau = .25
// subscript 0, T=0, similarly 1, T=1
scalar std0 = .5
scalar std1 = .5

Python:

# importing packages
import numpy as np
import pandas as pd
# importing stata interface package,
# so that both stata and Python have access to variable 'num_sims'
from sfi import Scalar

# n - short for num_obs - reading in variable from stata
n = int( Scalar.getValue('num_obs') )
# reading in variables from stata
num_sims = int( Scalar.getValue('num_sims') )
intercept = float( Scalar.getValue('intercept') )
```

```
tau = float( Scalar.getValue('tau') )
std0 = float( Scalar.getValue('std0') )
std1 = float( Scalar.getValue('std1') )

treatment_vector = np.repeat([0,1],n*num_sims)

# set seed for consistent results, comment out or change seed otherwise
np.random.seed(1)

# drawing noise terms
noise0 = np.random.normal(0,std0,n*num_sims)
noise1 = np.random.normal(0,std1,n*num_sims)
# putting the noise together in a manner consistent with treatment vector
noise_vector = np.append(noise0,noise1)

# generating the outcome according to the specified DGP
outcome_vector = intercept + tau*treatment_vector + noise_vector

# the data-set is put into 'power blocks',
# so one data-set is where power_block==0,
# another where power_block==1, etc.
power_block_vector = np.tile(np.repeat(np.arange(num_sims),n),2)

# putting it into a format that stata will like
matrix = np.stack([treatment_vector,outcome_vector,power_block_vector],axis=-1)
df = pd.DataFrame(data=matrix, columns=["treatment", "outcome", "power_block"])
df.to_stata('pa_sims.dta')

end

use "pa_sims.dta"

// so we can count the number of tests rejected
```

```
gen rejected = 0

// a is desired alpha level
local a = .05

// needed for loop
local ns = num_sims-1

forval i = 0 / `ns' {
// perform the t-test
quietly ttest outcome if power_block == `i', by(treatment)
// save the p_value
scalar p_val = r(p)
// record if the null was rejected
quietly replace rejected = 1 if p_val <= `a' & power_block == `i'
}



quietly summ rejected
// the mean of the rejected variable is the power
display as text "Power: " as result r(mean)

// // results from non-simulation ttest for comparison
scalar mean1 = intercept + tau
sampsi `=scalar(intercept)' `=scalar(mean1)', ///
 sd1(`=scalar(std0)') sd2(`=scalar(std1)') p(.8)
```

A.2 Discussion

Notice the lack of loops in the Python code that generates the data-sets, vectorizing these improves speed performance substantially. In practice, one should avoid loops as often as practically possible, unless there are memory issues or parallellization techniques (think multi-core processors) can be easily implemented. Unfortunately Stata offers little opportunities for vectorization, and thus must loop over each power block

to conduct the statistical test. This loop process is very parallelizable as the results of one loop do not affect the others so they could be conducted simultaneously. Stata does offer easy access to parallelization techniques through its (non-standard, costly) MP version, but does not currently support parallelization over loops. There are also community created packages for parallel computing in Stata, however, these are (understandably) much more difficult to implement for a regular user. One last thing to mention is that the Python section of the code has exported the generated data-set to a Stata dta file which Stata then reads back in, which is not strictly necessary if the package "`from sfi import Data`" is utilized. However, for more complicated examples Stata does not seem to be very stable when running Python code, and thus it is recommended to run the Python code separately, export the data in a Stata readable file, and then run the do file that uses that data.

It is possible to conduct this particular example in Python, and as it is vectorized it is several orders of magnitude faster than the given Stata code. This code is provided in the section below. However, for more complicated statistical tests, Python does not have readily available and user friendly packages for many tests Stata can conduct (or if they do, they are not vectorizable), thus requiring the use of looping in Stata.

A.3 Python Code

```
import numpy as np


mean1 = .5
std1 = .5
mean2 = .75
std2 = std1


n=63
num_sims = 1000
num_reject = 0
alpha = .05


np.random.seed(1)


d1 = np.random.normal(mean1,std1,(n,num_sims))
d2 = np.random.normal(mean2,std2,(n,num_sims))
```

```
p_vals = sps.ttest_ind(d1,d2,axis=0)[1]
rejected = np.logical_not(p_vals>=alpha)

print('power:',np.mean(rejected))
```

A.4 Python Code for QRE approach for Bayesian Persuasion

```
import scipy.stats as sps
import numpy as np

# functions for bayesian persuasion:

# returns the probability that a red message would be generated
def prob_succeed(prior,x):
    return prior*1.0 + (1.0-prior)*(1-x)

# can be used to calculate the optimal x to get receivers to a certain posterior
def optimal_x(prior, posterior_improvement):
    # note this function takes in the prior as a percentage (e.g. 35% instead of .35)
    # for backwards compatibility reasons
    # if you'd like to change that, remove the '/100' from below
    p = prior/100
    q = (prior + posterior_improvement)/100
    out = (p/q-p)/(1-p) - 1
    return out*-1

# prob succeed but with x instead of change in prior
# prior and x are between 0 and 1 (i.e. probabilities, not percentages)
def prob_succeedX(prior,x):
    return prior*1.0 + (1.0-prior)*(1-x)

# for a given x, calculates the new prior be conditional on a red message
```

```
def update_prior(prior,x):
    return prior/(prior+(1-x)*(1-prior))

# calculates the expected profit for a receiver,
# for each possible posterior, what is the expected profit of choosing red or blue, given the payoffs
# returns a 2*(len(posteriors)) matrix, first row is ep_red, second_row is ep_blue
def receiver_ep(posteriors,pay_match,pay_mismatch):
    red_vec = posteriors*pay_match+(1-posteriors)*pay_mismatch
    blue_vec = posteriors*pay_mismatch + (1-posteriors)*pay_match
    return np.vstack((red_vec,blue_vec))

# calculates the sender's ep for each possible x decision he is allowed to make,
# given the qre probability matrix of the reciever for each induced posterior
# (should map 1to1 to x decisions) (receiver_probs)
def sender_ep(possible_x_decisions,receiver_probs,pay_red,pay_blue,receiver_prior):
    prob_red = receiver_probs[0,:]
    prob_succs = prob_succeedX(receiver_prior,possible_x_decisions)
    return prob_succs*(prob_red*pay_red+(1-prob_red)*pay_blue) + (1-prob_succs)*pay_blue

# given a matrix of expected profits, where each row represents a decision,
# and each column represents a state,
# will return a probability matrix with the probability of each decision given the state
# according by the logit QRE function with the parameter lamb
# (lamb=lambda, but that word is reserved in python)
# (also seems to work just fine for a vector)
def epmat2qre(ep_mat,lamb):
    ep_lamb = ep_mat*lamb
    mx = np.max(ep_lamb,axis=0)
    exp = np.exp(ep_lamb-mx)
    return exp/np.sum(exp,axis=0)

# assume senders can set x=0,.01,.02, ... , .99, 1. i.e., integer percentages
all_actions = np.array([.01*i for i in range(101)])
```

```
# ... meaning there is a finite number of posteriors they can induce
# this function calculates the induced posterior for each x they could choose
# priors and x's should be between 0 and 1
# possible x's should be a list
def possible_posteriors(prior,possible_xs):
    return update_prior(prior,np.array(possible_xs))

# calculates the qre equilibrium via backwards induction
# takes in the receiver's expected payoff matrix - calculate using receiver_ep function above
def qre_equilibrium(pay_red,pay_blue,sender_lamb,receiver_lamb,receiver_prior,receiver_ep):
    receiver_probs = epmat2qre(receiver_ep,receiver_lamb)
    sep = sender_ep(np.array(all_actions),receiver_probs,pay_red,pay_blue,receiver_prior)
    sqre = epmat2qre(sep,sender_lamb)
#     weighted_x = np.sum(np.array(possible_xs)*sqre)
    return sqre


# setting the bayesian persuasion parameters
pay_high = 2
pay_low = 1
prior_1 = 1/3
prior_2 = 1/5
pay_match = 2
pay_mismatch = 1

# setting the qre parameters
lambs = 11.26
lambr = lambs

# this particular function provides similar functionality to np.random.choice
# but instead takes in an entire matrix (to be applied row-wise)
# IN: choices N*M matrix, probs N*M matrix, OUT: choices N*1 vector
def choice_vec(choices,probs):
```

```
    cumprobs = probs.cumsum(axis=1)
    draws = np.random.random(probs.shape[0])
    oneminus = cumprobs-draws.reshape(draws.shape[0],1)
    cols = np.argmax(oneminus>0,axis=1)
    rows = np.arange(cumprobs.shape[0])
    return choices[rows,cols]

# first calculate receiver's expected profit
full_rep1 = receiver_ep(np.array(possible_posteriors(prior_1,all_actions)),pay_match,pay_mismatch)
full_rep2 = receiver_ep(np.array(possible_posteriors(prior_2,all_actions)),pay_match,pay_mismatch)

# now calculate sender's noisy BR to the receiver actions
dist1 = qre_equilibrium(pay_high,pay_low,lambs,lambr,prior_1,full_rep1)
dist2 = qre_equilibrium(pay_high,pay_low,lambs,lambr,prior_2,full_rep2)

num_sims = 1000
n = 63

# set seed for consistent results
np.random.seed(1)
actions1 = choice_vec(np.tile(all_actions,(num_sims*n,1)), np.tile(dist1,(num_sims*n,1)))
np.random.seed(1000000)
actions2 = choice_vec(np.tile(all_actions,(num_sims*n,1)), np.tile(dist2,(num_sims*n,1)))

# actions are currently a vector, which is handy for stata
# however, for vectorization, putting it as a matrix
actions1 = actions1.reshape(n,num_sims)
actions2 = actions2.reshape(n,num_sims)

alpha = .05

# conducting the t-tests and counting the number of rejections
p_vals = sps.ttest_ind(actions1,actions2,axis=0)[1]
```

```
rejected = np.logical_not(p_vals>=alpha)

print('power:',np.mean(rejected))
```

B D-optimal Example with Treatment-specific Standard Deviation

Consider the following (extreme) illustrative example. Assume an artificial DGP of the form $y = b + cx_i + (x_i + 1)^{2.5}\epsilon_i$, $\epsilon_i \sim N(0, \sigma^2)$, and where the possible treatment levels are $x_i \in [0, 1]$. From the DGP it can be seen that a higher x_i results in noisier data. To ease notation, consider a case where 2 subjects will be assigned to each treatment (the general pattern holds for any distribution of n over treatments), and denote the two treatment intensities to be chosen as w and x, with $x > w$. The Log Likelihood would be as follows:

$LL(x, w) = 4k - \frac{2}{(w+1)^{5.0}}(-b - cw + y)^2 - \frac{2}{(x+1)^{5.0}}(-b - cx + y)^2$

Differentiating with respect to b and c (the parameters we seek to accurately estimate) and forming the information matrix yields:

$det(I) = -\left(-\frac{4w}{(w+1)^{5.0}} - \frac{4x}{(x+1)^{5.0}}\right)^2 + \left(-\frac{4w^2}{(w+1)^{5.0}} - \frac{4x^2}{(x+1)^{5.0}}\right)\left(-\frac{4}{(w+1)^{5.0}} - \frac{4}{(x+1)^{5.0}}\right)$

Utilizing a grid search where $w, x \in [1e-10, 1]$, the determinant is maximized where $w = 1e-10$ and $x = 0.668$. The smallest possible value is selected for w, as this is where noise is minimized. As for x there exists a trade-off, increasing it increases separation between the treatments, but also increases the noise of the observations in the x treatment, meaning it is not optimal to increase x all the way to 1.

C Differential Treatment Variance in Bayesian Persuasion

This section explores the impact of each treatment having different subject variability in the Bayesian Persuasion environment specified in Section 5. If the standard deviation differs substantially between treatments, then for optimal power more subjects should be allocated to the more noisy treatment, in proportion with the relative variances (List et al., 2011). Table 11 presents the required sample size to hit a power of 80% for a selection of larger and smaller variances across each treatment. As evident in Table 11, a higher variance requires more subjects, and assigning subjects efficiently by allowing the sample size to differ across treatments results in a lower total number of required observations. There are two questions that would need to be answered ex-ante in order to allocate different numbers of subjects to each treatment. Firstly, what treatment would have more variation in subject behavior? Secondly, what would be the ratio of the treatment standard deviations? Ex-post this should be less of an issue, as there should be reasonably reliable

estimates of the variation in subject behavior by treatment, but this insight would only really be of use to replication studies. Ex-ante, the QR framework can provide estimates of the anticipated standard deviation of each treatment. If the estimated standard deviations differ substantially, then consideration of non-equal subject allocations may be prudent.

	$\sigma_1 = .25$	$\sigma_1 = .5$	$\sigma_1 = .75$
$\sigma_0 = .25$	16,16	40,40 24,48	79,79 32,96
$\sigma_0 = .5$	40,40 48,24	63,63	103,103 79,119
$\sigma_0 = .75$	79,79 96,32	103,103 119,79	142,142

Table 11: Effects of Unequal Variance

D Ex-ante Power Analysis of Omitted Papers

D.1 Ex-ante Power Analysis of De Clippel et al. (2014)

De Clippel et al. (2014) experimentally analyze two rules for the selection of different outcomes (motivated in the paper as arbitrators), the short-list (SL) and veto-rank (VR) methods. In the SL method, one player is randomly selected to create a list of $n < N$ (here $n = 3$, $N = 5$) options, of which the other player chooses the final outcome from. In the VR method, both players remove $m < N$ options from contention (here $m = 2$), and then rank the remaining options. The selected option is the option with the lowest sum of ranks that has not been removed/vetoed.

D.1.1 Ex-ante Power Analysis

The particular treatment effect is that the SL method induces higher average payoffs than the VR method for the preference orderings [a,b,c,d,e] and [b,a,c,d,e] for players 1 and 2 respectively (denoted $Pf2$ in the paper). Unfortunately, this experiment was relatively novel, and there does not appear to be any closely related experiments from which to obtain estimates of λ. For the purposes of this exercise, I will consider the data from the other three preference orderings that were also reported in De Clippel et al. (2014), but that are not involved with the given hypothesis. This generates a 'past experiment' of sorts, and somewhat resembles the problem an experimenter might face.

I fit the QRE λ separately on the VR and SL environments, as despite their similarities as both being methods to determine an outcome for the group, the strategic environments differ substantially. The SL environment is sequential, with the second mover only needing to choose which of the 3 options to take,

and the first mover only needing to choose out of 10 possible short-lists. Whereas the VR environment is simultaneous, with each player needing to choose from 60 possible veto rank combinations. The VR environment has multiple Nash equilibria, so I consider only the logit path in this game.[32] Fitting the QRE model to all of the preference orderings except for the one of interest yields $\lambda_{VR} = 10.00$ and $\lambda_{SL} = 2.49$. For the preference ordering of interest, 500 total observations per treatment are required for a power of 80%. This is obtained by having 50 pairs of subjects observed over the 10 rounds in the experiment (so 200 subjects in total).[33]

D.1.2 Optimal Experimental Design

There are a few different factors that could be varied here, such as the number of options, the size of the short-list, the number of vetos, the payoff for each action, and the preference orderings. I focus on the latter only, as fitting a QRE to the VR condition takes a substantial amount of computing time due to the logit path restriction, making considering all possible dimensions intractable. I consider 10 randomly sampled preference orderings that were not considered in the original experiment. Like the original environment, I hold player 1's preference ordering to be $[a, b, c, d, e]$, while player 2's preference ordering changes. For comparison, I also include the four original preference orderings, denoted $Pf1-4$, and denote the newly considered preference orders as $PfX1-10$. The resulting analysis is presented in Table 12, which reveals a clear separation between low to moderately powered preference orderings, and very high powered preference orderings with a simulation power at or near 100%. I focus on the high powered preference sets, and seek to differentiate between them by setting $N = 50$, $\alpha = 0.01$, and by considering a closed form solution of power that is less subject to random noise. This is presented in Table 13, and yields a 'winner' of $PfX10 - [d, a, e, c, b]$. It should be noted that here is another example of not just maximizing the treatment effect size. $PfX10$ gains its edge over $PfX7$ despite having a lower τ, due to the decrease in standard deviation in the SL treatment. Using $PfX10$ instead of $Pf2$ would result in a required number of observations of $N = 22$ per treatment, which I round up to the minimum of one session for each treatment with a session size of 20, so $N = 100$ observations per treatment, or 40 subjects in total.

[32] The logit path is the unique path along λ that connects the uniform random play when $\lambda = 0$ to one particular Nash equilibrium when $\lambda \to \infty$. The Nash equilibrium that is converged to is called the logit solution, and is a refinement of the Nash equilibrium.

[33] It should be noted, as there is random re-matching within the session, and the statistical test used is the t-test, then this test really should be conducted on the average at the session level. However, I report this approach as this is how the statistical test was conducted in the original paper.

Player 2 Preference Ordering	μ_{SL},μ_{VR} τ σ_{SL}, σ_{VR}	Power with $\alpha = .05$
$Pf1$ - [e, d, c, b, a]	1.0,1.0 0.0 0.0,0.0	0.0%
$Pf2$ - [b, a, c, d, e]	1.46,1.54 0.08 0.51,0.41	0.582%
$Pf3$ - [c, b, a, d, e]	1.33,1.37 0.04 0.43,0.34	29.6%
$Pf4$ - [e, c, a, b, d]	1.16,1.29 0.13 0.32,0.31	99.8%
$PfX1$ - [e, b, a, c, d]	1.23,1.34 0.11 0.38,0.32	96.4%
$PfX2$ - [e, d, c, a, b]	1.02,1.03 0.01 0.11,0.11	20.1%
$PfX3$ - [b, c, d, e, a]	1.34,1.59 0.25 0.48,0.31	100%
$PfX4$ - [d, c, e, b, a]	1.09,1.17 0.08 0.27,0.17	99.9%
$PfX5$ - [c, e, b, a, d]	1.2,1.22 0.02 0.36,0.28	11.6%
$PfX6$ - [a, c, b, d, e]	1.53,1.86 0.33 0.57,0.31	100%
$PfX7$ - [a, d, e, b, c]	1.49,1.87 0.38 0.61,0.31	100%
$PfX8$ - [c, b, a, e, d]	1.32,1.41 0.09 0.44,0.29	81.3%
$PfX9$ - [b, e, c, a, d]	1.32,1.58 0.25 0.48,0.32	100%
$PfX10$ - [d, a, e, c, b]	1.26,1.6 0.34 0.47,0.3	100%

Table 12: De Clippel et al. (2014) - Power of Preference Orderings at $N = 300$

Player 2 Preference Ordering	μ_{SL},μ_{VR} τ σ_{SL}, σ_{VR}	Power with $\alpha = .01$
$Pf4$ - [e, c, a, b, d]	1.13,1.29 0.15 0.34,0.31	42.4%
$PfX1$ - [e, b, a, c ,d]	1.23,1.34 0.11 0.38,0.32	15.9%
$PfX3$ - [b, c, d, e, a]	1.33,1.59 0.25 0.48,0.31	71.0%
$PfX4$ - [d, c, e, b, a]	1.09,1.17 0.09 0.27,0.17	25.8%
$PfX6$ - [a, c, b, d, e]	1.53,1.86 0.33 0.57,0.31	86.2%
$PfX7$ - [a, d, e, b, c]	1.49,1.87 0.38 0.61,0.3	91.2%
$PfX9$ - [b, e, c, a, d]	1.32,1.58 0.25 0.47,0.32	70.2%
$PfX10$ - [d, a, e, c, b]	1.27,1.6 0.34 0.47,0.29	95.5%

Table 13: De Clippel et al. (2014) - Power of Selected Preference Orderings at $N = 50$

D.2 Ex-ante Power Analysis of Huck et al. (2011)

Huck et al. (2011) consider an experimental test of the Lazear model of deferred compensation (Lazear, 1979). In their experiment, the firm initially sets the wage profile W_1, W_2, and W_3, which is the worker's wage in periods $t = 1$, $t = 2$, and $t = 3$ respectively. Whether the firm can commit to this wage profile is the treatment of interest. The worker makes effort decisions at $t = 1.5$ and $t = 2.5$, and can choose $E_t \in \{L, M, H\}$. If the worker chooses L, he is fired with probability p, and does not receive the wages that would follow (i.e. either W_2 and W_3 if fired at $t = 1.5$, and W_3 if fired at $t = 2.5$). Both the firm's payoff and the worker's cost is increasing in the chosen effort level. If the worker is fired at $t = 1.5$, then the firm receives the benefit of L effort at $t = 2.5$, at no cost to either worker or firm. The particular parameters chosen for the experiment were $p = 0.5$, costs of effort $C_L = 0$, $C_M = 20$, $C_H = 40$, firm revenues $Z_L = 50$, $Z_M = 100$, $Z_H = 140$, and wage offers could be $W_t \in \{0, 1, 2, ..., 119, 120\}$. The equilibrium is solved via backwards induction, and under commitment the firm's equilibrium behavior is to set a deferred compensation contract of $W_1 = 0$, $W_2 \in [0, 20]$ and $W_3 = 60 - W_2$, which induces the worker to choose M at both effort decisions. Without commitment, the equilibrium is that the firm offers no wages and the worker always chooses L effort. The hypothesis in question is whether the deferred compensation portion of the wage profile $W_2 + W_3$ differs between the commitment and no commitment cases.

D.2.1 Ex-ante Power Analysis

It is relatively straightforward to fit a QRE to either of these sequential environments, so all that remains is to specify a reasonable λ. However, unfortunately there does not appear to be any tangentially related experimental environments with which to obtain this from. Instead, for the purposes of this exercise, I conduct a thought experiment where I only observe data from one of the treatments, and then use that to fit the λ. This is another way to try and make an analogous situation to what an experimenter might face, using a previous experimental setup as a baseline, and new environment. I run this exercise both ways, i.e. I assume I only have the commitment data to fit a λ to then generate predictions for the no commitment experiment, and vice versa. From either approach I obtain similar estimates of λ, $\lambda_{FCT} = 0.0414$ and $\lambda_{NCT} = 0.0552$. This yields predictions as presented in Table 14. The QRE simulation has performed quite well in terms of the treatment effect size, but less so in terms of point predictions. It has predicted the differential standard deviation well, but not so much from the FCT treatment. The statistical test that was conducted was a Mann-Whitney test on the average deferred wages by session, which consisted of groups of 10, 5 firms and 5 workers, who interacted over 20 rounds. The averaging by session reduces the noise substantially, and I find that in conjunction with to the large treatment effect size, the test is very strongly powered even for only a

Treatment	λ	μ_{NCT}	σ_{NCT}	μ_{FCT}	σ_{FCT}	τ	Sessions per Treatment
FCT	0.0414	45.75	2.93	73.29	3.66	27.54	3
NCT	0.0552	34.95	2.34	68.63	3.20	33.68	3
Actual	0.0480*	21.31	3.65	54.74	12.61	33.43	6**

Table 14: Huck et al. (2011)

*QRE fit on both data-sets. **Number of sessions used in Huck et al. (2011).

small number of independent observations.

D.2.2 Optimal Experimental Design

As the test is already strongly powered even for very low number of observations, via simulation there are issues with 'perfectly powered' tests with a power of 100%. It is not possible to improve from a perfectly powered test, therefore I cannot investigate changes in power due to changes in the experimental design. Furthermore, as the statistical test is so strongly powered, it is of little practical importance to improve power. Therefore, for this particular paper, I do not conduct any thought exercise with regards to changes in the experimental design.

References

Johannes Abeler, Armin Falk, Lorenz Goette, and David Huffman. Reference points and effort provision. *American Economic Review*, 101(2):470–492, 4 2011. ISSN 00028282. doi: 10.1257/aer.101.2.470.

Pak Hung Au and King King Li. Bayesian Persuasion and Reciprocity: Theory and Experiment. *SSRN Electronic Journal*, 6 2018. ISSN 1556-5068. doi: 10.2139/ssrn.3191203. URL https://www.ssrn.com/abstract=3191203.

Jess Benhabib, Alberto Bisin, and Andrew Schotter. Present-bias, quasi-hyperbolic discounting, and fixed costs. *Games and Economic Behavior*, 69(2):205–223, 7 2010. ISSN 08998256. doi: 10.1016/j.geb.2009.11.003.

Jordi Brandts and Gary Charness. Hot vs. Cold: Sequential Responses and Preference Stability in Experimental Games. *Experimental Economics*, 2:227–238, 2000.

Kenneth Burdett and Kenneth Judd. Equilibrium Price Dispersion. *Econometrica*, 51(4):955–969, 7 1983.

Luigi Butera, Philip J. Grossman, Daniel Houser, John A. List, and Marie Claire Villeval. A New Mechanism to Alleviate the Crises of Confidence in Science With An Application to the Public Goods Game. 2020.

Colin F. Camerer, Anna Dreber, Eskil Forsell, Teck Hua Ho, Jürgen Huber, Magnus Johannesson, Michael Kirchler, Johan Almenberg, Adam Altmejd, Taizan Chan, Emma Heikensten, Felix Holzmeister, Taisuke Imai, Siri Isaksson, Gideon Nave, Thomas Pfeiffer, Michael Razen, and Hang Wu. Evaluating replicability of laboratory experiments in economics. *Science*, 351(6280):1433–1436, 3 2016. ISSN 10959203. doi: 10.1126/science.aaf0918.

Colin F. Camerer, Anna Dreber, and Magnus Johannesson. Replication and other practices for improving scientific quality in experimental economics. In Arthur Schram and Aljaž Ule, editors, *Handbook of Research Methods and Applications in Experimental Economics*, chapter 5, pages 83–102. Edward Elgar Publishing, 2019.

C Mónica Capra. Mood-Driven Behavior in Strategic Interactions. *The American Economic Review: Papers and Proceedings*, 94(2):367–372, 2004.

Marco Casari and Timothy N. Cason. The strategy method lowers measured trustworthy behavior. *Economics Letters*, 103(3):157–159, 6 2009. ISSN 01651765. doi: 10.1016/j.econlet.2009.03.012.

Timothy N. Cason, Daniel Friedman, and Ed Hopkins. An Experimental Investigation of Price Dispersion and Cycles. *Journal of Political Economy*.

Gary Charness and Martin Dufwenberg. Promises and Partnership. *Econometrica*, 74(6):1579–1601, 2006.

Gary Charness and Martin Dufwenberg. Participation. *American Economic Review*, 101(4):1211–1237, 6 2011. doi: 10.1257/aer.101.4.1211.

Gary Charness and Matthew Rabin. Understanding Social Preferences with Simple Tests. *The Quarterly Journal of Economics*, 117(3):817–869, 2002. URL https://www.jstor.org/stable/4132490.

Roy Chen and Yan Chen. The potential of social identity for equilibrium selection. *American Economic Review*, 101(6):2562–2589, 10 2011. ISSN 00028282. doi: 10.1257/aer.101.6.2562.

Yan Chen and Sherry Xin Li. Group identity and social preferences. *American Economic Review*, 99(1): 431–457, 3 2009. ISSN 00028282. doi: 10.1257/aer.99.1.431.

James C. Cox, Vernon L. Smith, and James M. Walker. Theory and individual behavior of first-price auctions. *Journal of Risk and Uncertainty*, 1(1):61–99, 3 1988. ISSN 08955646. doi: 10.1007/BF00055565.

James C Cox, Vernon L Smith, and James M Walker. Theory and Misbehavior of First-Price Auctions: Comment. *The American Economic Review*, 82(5):1392–1412, 1992.

Douglas D. Davis and Charles A. Holt. *Experimental Economics*. Princeton University Press, 1993.

Geoffroy De Clippel, Kfir Eliaz, and Brian Knight. On the selection of arbitrators. *American Economic Review*, 104(11):3434–3458, 11 2014. ISSN 00028282. doi: 10.1257/aer.104.11.3434.

Michalis Drouvelis and Brit Grosskopf. The effects of induced emotions on pro-social behaviour. *Journal of Public Economics*, 134:1–8, 2 2016. ISSN 00472727. doi: 10.1016/j.jpubeco.2015.12.012.

Maren Duvendack, Richard Palmer-Jones, and W. Robert Reed. What is meant by "Replication" and why does it encounter resistance in economics? In *American Economic Review: Papers & Proceedings*, volume 107, pages 46–51. American Economic Association, 5 2017. doi: 10.1257/aer.p20171031.

Ernst Fehr, Holger Herz, and Tom Wilkening. The lure of authority: Motivation and incentive effects of power. *American Economic Review*, 103(4):1325–1359, 6 2013. ISSN 00028282. doi: 10.1257/aer.103.4.1325.

Guillaume Fréchette, Alessandro Lizzeri, and Jacopo Perego. Rules and Commitment in Communication. 2018.

Daniel Friedman. Theory and Misbehavior of First-Price Auctions: Comment. *American Economic Review*, 82(5):1374–1378, 1992. doi: 10.2307/2117485.

Daniel Friedman and Shyam Sunder. *Experimental Methods: A Primer for Economists*. Cambridge University Press, 1994.

Andrew Gelman and John Carlin. Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors. *Perspectives on Psychological Science*, 9(6):641–651, 11 2014. ISSN 17456924. doi: 10.1177/1745691614551642.

Jacob K. Goeree and Charles A. Holt. An experimental study of costly coordination. *Games and Economic Behavior*, 51(2 SPEC. ISS.):349–364, 2005. ISSN 08998256. doi: 10.1016/j.geb.2004.08.006.

Glenn W Harrison. Theory and Misbehavior of First-Price Auctions. *The American Economic Review*, 79 (4):749–762, 1989. URL https://www.jstor.org/stable/1827930.

Glenn W Harrison. Theory and Misbehavior of First-Price Auctions: Reply. *The American Economic Review*, 82(5):1426–1443, 1992.

Steffen Huck, Andrew J. Seltzer, and Brian Wallace. Deferred compensation in multiperiod labor contracts: An experimental test of Lazear's model. *American Economic Review*, 101(2):819–843, 4 2011. ISSN 00028282. doi: 10.1257/aer.101.2.819.

John Ifcher and Homa Zarghamee. Happiness and time preference: The effect of positive affect in a random-assignment experiment. *American Economic Review*, 101(7):3109–3129, 12 2011. ISSN 00028282. doi: 10.1257/aer.101.7.3109.

John P.A. Ioannidis. Why most published research findings are false. *PLoS Medicine*, 2(8):2–8, 7 2005. ISSN 15491277. doi: 10.1371/journal.pmed.0020124.

John P.A. Ioannidis, T. D. Stanley, and Hristos Doucouliagos. The Power of Bias in Economics Research. *Economic Journal*, 127(605):F236–F265, 10 2017. ISSN 14680297. doi: 10.1111/ecoj.12461.

John H. Kagel and Alvin E. Roth. Theory and Misbehavior of First-Price Auctions: Comment. *The American Economic Review*, 82(5):1379–1391, 1992.

Emir Kamenica. Bayesian Persuasion and Information Design. *Annual Review of Economics*, 11(1):080218–025739, 8 2019. ISSN 1941-1383. doi: 10.1146/annurev-economics-080218-025739. URL https://www.annualreviews.org/doi/10.1146/annurev-economics-080218-025739.

Emir Kamenica and Matthew Gentzkow. Bayesian Persuasion. *American Economic Review*, 101(6):2590–2615, 10 2011. ISSN 0002-8282. doi: 10.1257/aer.101.6.2590. URL http://pubs.aeaweb.org/doi/10.1257/aer.101.6.2590.

Georg Kirchsteiger, Luca Rigotti, and Aldo Rustichini. Your morals might be your moods. *Journal of Economic Behavior and Organization*, 59(2):155–172, 2 2006. ISSN 01672681. doi: 10.1016/j.jebo.2004.07.004.

Botond Kőszegi and Matthew Rabin. A Model of Reference-Dependent Preferences. *The Quarterly Journal of Economics*, 121(4):1133–1165, 11 2006. doi: 10.1093/qje/121.4.1133. URL https://academic.oup.com/qje/article-lookup/doi/10.1093/qje/121.4.1133.

Edward P Lazear. Why Is There Mandatory Retirement? *Journal of Political Economy*, 87(6):1261–1284, 1979. URL https://www.jstor.org/stable/1833332.

John A. List, Sally Sadoff, and Mathis Wagner. So you want to run an experiment, now what? Some simple rules of thumb for optimal experimental design. *Experimental Economics*, 14(4):439–457, 11 2011. ISSN 13864157. doi: 10.1007/s10683-011-9275-7.

Zacharias Maniadis, Fabio Tufano, and John A. List. One swallow doesn't make a summer: New evidence on anchoring effects. *American Economic Review*, 104(1):277–290, 1 2014. ISSN 00028282. doi: 10.1257/aer.104.1.277.

Zacharias Maniadis, Fabio Tufano, and John A. List. How to make experimental economics research more reproducible: Lessons from other disciplines and a new proposal. *Research in Experimental Economics*, 18:215–230, 2015. ISSN 01932306. doi: 10.1108/S0193-230620150000018008.

Daniel L McFadden. Quantal Choice Analysis: A Survey. *Annals of Economic and Social Measurement*, 5 (4), 1976.

Richard D. McKelvey and Thomas R. Palfrey. Quantal Response Equilibria for Normal Form Games. *Games and Economic Behavior*, 10(1):6–38, 7 1995. ISSN 0899-8256. doi: 10.1006/GAME.1995.1023. URL https://www.sciencedirect.com/science/article/pii/S0899825685710238.

Antonio Merlo and Andrew Schotter. Theory and Misbehavior of First-Price Auctions: Comment. *The American Economic Review*, 82(5):1413–1425, 1992.

Peter G. Moffatt. *Experimetrics : Econometrics for Experimental Economics*. Palgrave, 2016. ISBN 0230250238.

Quyen Nguyen. Bayesian persuasion: Evidence from the laboratory. 2017.

Emily Pronin and Lee Ross. Temporal differences in trait self-ascription: When the self is seen as an other. *Journal of Personality and Social Psychology*, 90(2), 2006. ISSN 1939-1315. doi: 10.1037/0022-3514.90.2.197.

E. Elisabet Rutström and Nathaniel T. Wilcox. Stated beliefs versus inferred beliefs: A methodological inquiry and experimental test. *Games and Economic Behavior*, 67(2):616–632, 11 2009. ISSN 08998256. doi: 10.1016/j.geb.2009.04.001.

StataCorp. *Stata 13 Power and Sample Size Reference Manual*. Stata Press, College Station, TX, 2013.

Wenhao Wu and Bohan Ye. Competition in Persuasion: An Experiment. Technical report, 2019.

Le Zhang and Andreas Ortmann. Exploring the meaning of significance in experimental economics. 2013.

Stephen T. Ziliak and Deirdre N. McCloskey. Size matters: The standard error of regressions in the American Economic Review. *Journal of Socio-Economics*, 33(5):527–546, 11 2004. ISSN 10535357. doi: 10.1016/j.socec.2004.09.024.