Improving the Statistical Power of Economic Experiments Using Adaptive Designs: Supplementary material

1 Experimental Instructions

This appendix contains a more detailed description of the experiments serving for the applications 2 and 3.

1.1 Application 2

In the paper used for this application (Musshoff and Hirschauer, 2014), there are five different treatment groups, referred to as scenarios, and 38 students were randomised to each scenario group. The participants were on average 25 years old, and 38% were female. There are three different kinds of crops: wheat, canola and potatoes. Each individual plan specifies the land usage shares for the different crops, as well as the amount of nitrogen fertiliser to be used. This is done under the constraints that a total of 400 ha are available for cultivation, and that each crop must be assigned at least 100 ha. The objective of each player is to maximise his or her bank deposit at the end of the 20 production periods that make up the total game time. The only sources of risk are the market prices, which change over time as stochastic processes.

Each player is provided with a starting capital of 1 million euros. At the beginning of each period, €40,000 are withdrawn for personal living expanses. No interest is paid on positive credit balance at the end of a period. If there are outstanding payments that the player is unable to settle, capital is borrowed at zero interest rate. This is then automatically repaid in later periods if liquid funds then exceed the obligatory withdrawal. Fieldwork is assumed to be carried out by contractors at crop specific cost rates.

The production functions for the three different crops are all of the form

$$y = a + bx - cx^2, \tag{1}$$

where x is the amount of nitrogen used (kg/ha) and y is the yield (dt/ha, decitonnes per hectare). The parameters a, b and c are specific for each crop and participant, determined by drawing from a uniform distribution. The nitrogen price is randomly determined for each participant, and stays fixed during that person's game. At the end of each period, a specific crop price either increases to 120% of what it was in

the previous period or decreases to 80% (each with probability 0.5). By design, no correlation exists between product prices. Start prices, in €/dt, were 20 for wheat, 30 for canola and 10 for potatoes. To incentivise the players, they are informed that 5% of them (randomly selected) will receive a cash prize, the size of which depends on the performance in the game. €500 was given to players who achieved the best possible bank balance, while less successful players were awarded a share of the prize according to their relative performance.

In the first 10 production periods, all players receive an unconditional, deterministic payment of €300/ha. After these initial periods, the players are randomly assigned to one of five different treatment groups, representing different policy scenarios. The first, scenario 1, can be interpreted as "business as usual", in which the deterministic payments are continued. The other four are actual interventions, designed to foster more environmentally friendly production. Scenario 2 is a voluntary scheme in which players receive €300/ha (for each ha individually) if nitrogen levels are kept below 75 kg/ha. Complete visibility of nitrogen use is achieved through a monitoring system. In scenario 3, the transfer payments of €300/ha continues, but players will have to pay a penalty of €300/ha for every hectare where the nitrogen limit of 75 kg/ha is exceeded. Again, there is full visibility due to a monitoring system. Scenario 4 is like scenario 3, except that now compliance to the nitrogen limit is investigated in random sample inspections. The probability to be investigated is 0.1, and upon detection of nitrogen use above the 75 kg/ha limit, a penalty of €3000/ha has to be paid. Finally, in scenario 5, the players receive a transfer of €3000/ha for those areas where the nitrogen use is below the 75 kg/ha limit, but only if they are inspected, which happens with probability 0.1. These different scenarios all ensure the same expected monetary gain for the players, provided that they play optimally. To facilitate comparability, the participants are divided into quintets. Each of the five groups includes one participant who is confronted with the same production functions, the same nitrogen price and the same product price paths as the other members of the quintet in the other scenarios.

The data used in Section 3.2 was provided by the authors and is included in the replication material.

1.2 Application 3

In the paper used for this application, Karlan and List (2007) used a natural field experiment to study the effects of different matching grants on charitable donations. The experiment targeted previous donors of a non-profit organisation and is based on a large sample of 50,083 individuals. Two thirds of the study participants were assigned to some active treatment group, while the rest were assigned to a control group. The active treatments varied along three dimensions: (1) the price ratio of the match ($1:$1, $2:$1 or $3:$1); (2) the maximum size of the matching gift based on all donations ($25,000, $50,000, $100,000 or unstated) and (3) the donation amount suggested in the letter (1.00, 1.25 or 1.50 times the individual's highest previous contribution). Each of these treatment combinations was assigned with equal probability.

Ask amounts were specific for each participant and based on the individual's highest previous contribution. To test for additional effects, such as solicitee income level and political affiliation, the giving data was merged with: (1) demographic data aggre-

gated at zip code level, (2) state and county returns from the 2004 presidential election, and (3) frequency of activities of the organisation asking for donations at the state level. The main response variables of interest for us were the binary decision of whether or not to donate, and the property of belonging to a red state (republican voter).

The original data set used by Karlan and List (2007) was downloaded from its online repository `https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi%3A10.7910/DVN/27853`. The main data file, `AER merged.dta`, was converted to an R file and then used to produce the plots and tables included in our paper. When loaded in R, the table corresponding to `AER merged.dta` contains a column `red0` which contains 1 at a specific row if it corresponds to a respondent from a red state (republican majority). As mentioned above in the main text, we only considered the subset of the data for which `red0 = 1`.

2 Simulation for two stage designs

In order to complement the scenario based simulations for Application 1 in the main text, we present more general considerations about simulations of adaptive designs using the `asd`-package. This simulation package was originally developed for evaluating two-stage adaptive treatment selection designs using Dunnett's test (Parsons et al., 2012), and was later extended to also support adaptive subgroup selection (Friede et al., 2020).

2.1 A general simulation procedure

Based on the assumption of a fixed total budget, all the designs compared are assumed to have the same total sample size. Ignoring any additional complexity costs associated with two-stage vs. single-stage designs, this means that all designs have the same fixed cost. Hence, they may be compared in terms of their power, without any need to relate the power to the cost. Of course, the simulation could also be carried out with goal of reducing the required sample size to obtain a fixed level power, thus reducing the costs for the experiment.

For the two-stage designs, we assume a specific class of selection rules, namely, the ones that always takes a fixed number of treatments forward to the second stage. The number of treatments taken forward is specified by the parameter s. This selection rule works by comparing the Z-values corresponding to the different stage 1 treatments relative to control and selecting the s ones with the largest (absolute) Z-values. Hence, $s = 1$ means that the single best treatment is taken forward, $s = 2$ that the two best are taken forward, and so on. The second parameter characterising the two-stage adaptive designs is the ratio r, defined as the proportion of the total sample size used in the first stage. The following procedure can be used to decide on the values of these parameters:

1. Specify a total sample size for the trial, and a *maximum expected effect size* δ. The parameter δ allows for considering studies with different effect strengths. Typically, higher values of δ means that it will be easier to reach the primary goal and reject at least one hypothesis.

2. Plot the power to reject at least one hypothesis as a function of effect size, for different combinations of the selection rule s and the ratio r. Choose the design which corresponds to the highest power in a neighbourhood of δ. We refrain from specifying exactly how large this neighbourhood is, but some fraction of δ seems reasonable. The idea is that we want to choose a design that is reasonably robust with respect to uncertanties in our knowledge of the true effect sizes of the different treatments.

3. For the specified value of δ, fix the selection rule and ratio found in the previous step and plot the power and selection probabilities in a neighbourhood of the ratio. Adjust the ratio slightly if it leads to better selection probabilities without sacrificing too much power, say at most 1 or 2 percentage points. The idea behind this last step is to improve the performance in terms of a secondary criterion (selection probabilities) without sacrificing too much with respect to our primary criterion (power to reject at least one hypothesis). Of course, in a full formal optimisation problem, the relative importance of these two criteria would have to be specified in detail.

The power in our simulations was always evaluated relative to one of two specific configurations of the treatment effects. In the first, the effects are increasing from 0 to δ in steps of equal length,

$$\mu_i = i\delta/m, \quad i = 0,\ldots,m. \tag{2}$$

In the second configuration, the effect sizes of the active treatments are instead clustered close to δ,

$$\mu_0 = 0, \tag{3}$$

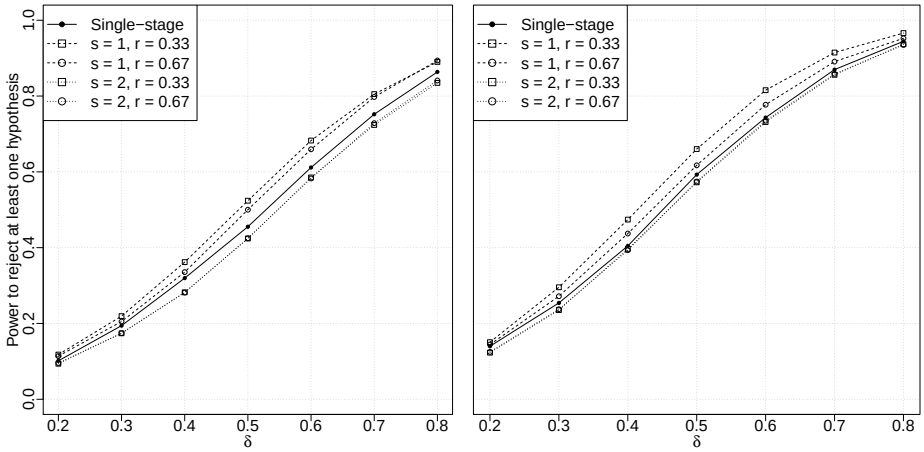
$$\mu_i = \delta - (m-i)\eta, \quad i = 1,\ldots,m, \tag{4}$$

for some small fixed number η. Naturally, we could also consider other types of effect configurations. In a fully specified optimisation problem, we could have a finite number of such configurations, together with an associated prior distribution over them specifying how likely we consider each of them to be a priori.

For illustrative purpose, we consider two different cases. The first is that of a small total sample size and a large effect, while the second instead considers the case of a large total sample size and a small effect. The `asd` package automatically handles the issue of strong control of the FWER. The steps we go through can be seen as a design procedure for selecting a two-stage design based primarily on the power to reject at least one hypothesis, and to a lesser extent on the treatment selection probabilities at the interim analysis. To complement application 1, the procedure represents a more qualitative approach based on plots, rather than an optimization problem.

2.1.1 Two-stage design for a small sample and a large effect

In this section we search for a good two-stage adaptive design with two active treatment groups in the first stage and a small total sample size of 100. The maximum expected



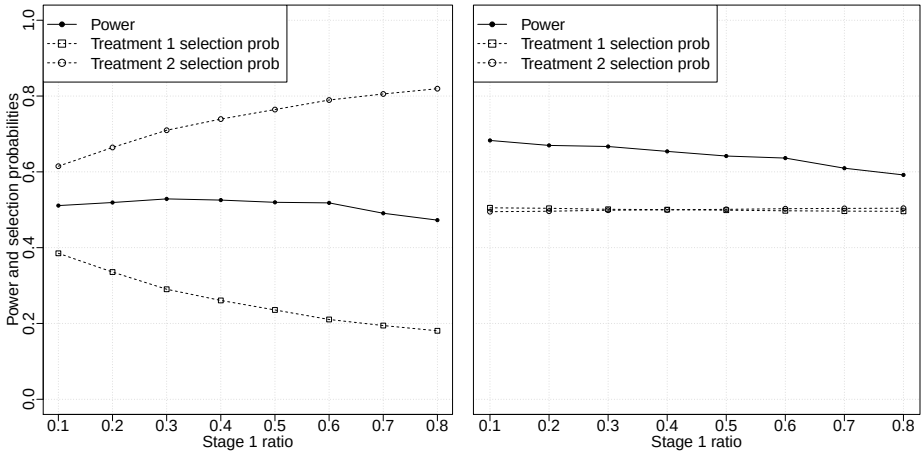
(a) Effect sizes: 0, $\delta/2$, δ **(b)** Effect sizes: 0, $\delta-\eta$, δ, with $\eta = 0.01$

Fig. 1 Power to reject at least one hypothesis as a function of the effect size parameter δ, for two different effect configurations given by Eqs. (2), (3) and (4). Two different selection rules are compared. Either the single best treatment is taken forward to the second stage (single-best-rule), or the two best treatments are taken forward (best-two-rule). There are two options for the proportion of the total sample size used in the first stage ($r = 0.33$ or $r = 0.67$)

effect size is assumed to be $\delta = 0.5$ and our primary aim is to find the design with the highest power in a neighbourhood of this effect. In particular, we would like to see some improvement when comparing against a single-stage design with the same total sample size. Throughout this section and the next, it is assumed that the Type I error probability is controlled at the one-sided level $\alpha = 0.025$.

Figure 1a shows the probability to reject at least one hypothesis as a function of δ for the first type of effect configuration. In addition, it also includes the power of a single-stage design of the same total sample size (solid line). It can be seen that the single-best-rule which selects the better treatment at interim dominates the single-stage design over the entire range $\delta \in [0.2, 0.8]$, for both values of the ratio r. The reason for this is that using the single-best-rule allows for increasing the power by dropping the inferior treatment (with effect $\delta/2$) at the interim, and focus the remaining resources on the superior treatment (with effect δ). Further, the single-stage design in turn dominates the best-two-rule over the entire range $\delta \in [0.2, 0.8]$. This is not surprising, since using the best-two-rule when we actually only have two active treatments will never lead to a treatment being dropped at interim. This means that there can be no power gain from allocating more resources to the most promising treatment. However, when moving from a single-stage to a two-stage design, there will be a loss in power stemming from the adjustments necessary to keep the Type I error probability under control.

For the best-two-rule, it can be seen that the power curves for $r = 0.33$ and $r = 0.67$ are almost equal over the entire range of δ. The small difference observed is only due to the finite samples used in the simulations. The reason is that the inverse normal



(a) Effect sizes: 0, $\delta/2$, δ, with $\delta = 0.5$

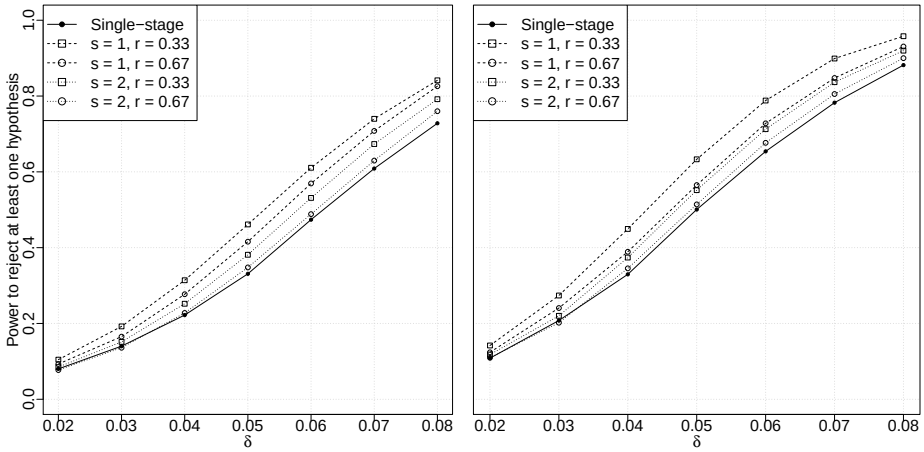
(b) Effect sizes: 0, $\delta - \eta$, δ, with $\delta = 0.5, \eta = 0.01$

Fig. 2 Power to reject at least one hypothesis and selection probabilities as functions of the ratio for a two-stage design which selects the single best active treatment at interim, for the two different effect configurations given by Eqs. (2), (3) and (4)

combination functions, as defined by Eq. (5) in the main text, leads to combined p-values that are invariant in the ratio as long as no hypotheses are dropped at interim. When the first-best-rule is applied, it can be observed that for all values of δ except the very largest (close to 0.8), a ratio value of 0.33 dominates the alternative of 0.67. Hence, based purely on power considerations, we are led to consider a design which selects the best treatment at interim, having a ratio value close to 0.33.

Figure 1b shows the probability to reject at least one hypothesis as a function of δ for the second type of effect configuration. For this configuration the design corresponding to the single-best-rule and $r = 0.33$ clearly dominates the others. As in Figure 1a, the single-stage design does not reach the power for the designs based on the single-best-rule, while it is uniformly better than the designs corresponding to the best-two rule. When comparing to Figure 1a, we can also note that the power is uniformly larger for all curves. The reason is of course that a specific value of δ yields equal effects for the two treatments, while treatments that are second best have effects $\delta - \eta$ and $\delta/2$, respectively. Finally, we note that the power curves corresponding to the best-two-rule are much closer to the single-stage design in Figure 1b. Again, as for the first effect configuration, we are led to consider a design which selects the best treatment at interim, having a ratio value around 0.33.

Next, consider Figure 2a, which shows the power and selection probabilities at the interim analysis as functions of the ratio r for the effect configuration given by Eq. (2). The selection rule is now fixed to only carry over the better treatment and the value of δ is fixed at $\delta = \delta = 0.5$. It can be seen that the power stays approximately constant at a value just above 0.5 until r approaches 0.6. This implies that we can increase the probability of selecting the better treatment at stage 1 without affecting the power too



(a) Effect sizes: $0, \delta/3, 2\delta/3, \delta$

(b) Effect sizes: $0, \delta-2\eta, \delta-\eta, \delta$, with $\eta = 0.001$

Fig. 3 Power to reject at least one hypothesis as a function of the effect size parameter δ, for two different effect configurations given by Eqs. (2), (3) and (4). Two different selection rules are compared. Either the single best treatment is taken forward to the second stage, or the two best treatments are taken forward. There are two options for the proportion of the total sample size used in the first stage ($r = 0.33$ or $r = 0.67$)

much by increasing the value of the ratio from 0.33 to, say, 0.6.

Finally, consider Figure 2b, corresponding to the effect configuration given by Eqs. (3) and (4). As before, the value of δ is fixed at 0.5. Since this effect configuration gives almost equal effects (namely, $\delta - \eta$ and δ), it is no surprise that the treatment selection probabilities are almost equal. Of more interest is the fact that the power is monotonically decreasing. The reason is again the similar effect sizes of the two treatments, which implies that the information obtained in the first stage is essentially worthless when selecting which treatment to take forward. Consequently, the resources are better spent on selecting a treatment as soon as possible and then use all resources to improve the estimate in the second stage. We can therefore conclude that a small value of the ratio should be used for this type of effect configuration.

2.1.2 Two-stage design for a large sample and a small effect

Mimicking the setting in application 3, we next consider the case of a total sample size of 10,000 and an expected effect size of $\delta = 0.05$. Moreover, there are now three active treatments in the first stage instead of just two as in Section 2.1.1.[1] In contrast to Figures 1a and 1b, it can be seen in Figures 3a and 3b that the single-stage design is

[1] Since the means of the Z-statistics have the form $\delta\sqrt{n/2}$, the overall shape of the power curves is the same for the two cases. Therefore, from a purely mathematical viewpoint the two cases are actually very similar. Nevertheless, we choose to illustrate them both using simulation in order to demonstrate the methodology.

dominated by all two-stage designs for all but the smallest effect sizes. This is because, when starting with three active treatments, at least one will always be dropped in the interim analysis. This leads to an improved power to reject at least one hypothesis.

It can be seen that the best design for both effect configurations is clearly the one corresponding to the single-best-rule and $r = 0.33$. As expected, for the case of similar effect sizes shown in Figure 3b, all power curves are shifted upwards (relative to Figure 3a) since the effects are now all clustered close to δ. Since it turns out that behaviour of the selection probabilities is very similar to that observed in Figures 2a and 2b, we do not include the corresponding plots here. Again, we are led to a design which takes a single treatment forward and uses a ratio value that is somewhere between very small and 0.6.

References

Friede, T., Stallard, N., and Parsons, N. (2020). Adaptive seamless clinical trials using early outcomes for treatment or subgroup selection: Methods, simulation model and their implementation in r. *Biometrical Journal*, 62(5):1264–1283.

Karlan, D. and List, J. A. (2007). Does price matter in charitable giving? evidence from a large-scale natural field experiment. *American Economic Review*, 97(5):1774–1793.

Musshoff, O. and Hirschauer, N. (2014). Using business simulation games in regulatory impact analysis - the case of policies aimed at reducing nitrogen leaching. *Applied Economics*, 46(25):3049–3060.

Parsons, N., Friede, T., Todd, S., Marquez, E. V., and Chataway, J. (2012). An r package for implementing simulations for seamless phase ii/iii clinical trials using early outcomes for treatment selection. *Computational Statistics and Data Analysis*, 56(5):1150–1160.