Experimental Economics (2023) 26:357–382
https://doi.org/10.1007/s10683-022-09773-8

ORIGINAL PAPER

Improving the statistical power of economic experiments using adaptive designs

Sebastian Jobjörnsson[1] · **Henning Schaak**[2] · **Oliver Musshoff**[3] · **Tim Friede**[1]

Received: 27 January 2021 / Revised: 31 August 2022 / Accepted: 1 September 2022 /
Published online: 25 September 2022

Abstract

An important issue for many economic experiments is how the experimenter can ensure sufficient power in order to reject one or more hypotheses. The paper illustrates how methods for testing multiple hypotheses simultaneously in adaptive, two-stage designs can be used to improve the power of economic experiments. We provide a concise overview of the relevant theory and illustrate the method in three different applications. These include a simulation study of a hypothetical experimental design, as well as illustrations using two data sets from previous experiments. The simulation results highlight the potential for sample size reductions, maintaining the power to reject at least one hypothesis while ensuring strong control of the overall Type I error probability.

Keywords Adaptive design · Multiple testing · Simulation study · Family-wise error rate · Experimental design

JEL Classification C12 · C90

1 Introduction

Imagine an economist being involved in a research project that aims to improve an economic outcome, e.g the productivity of small businesses. Within the project, it is planned to carry out a field experiment to test how their productivity could be

✉ Henning Schaak
henning.schaak@boku.ac.at

[1] Department of Medical Statistics, University Medical Center Göttingen, Göttingen, Germany

[2] Institute of Agricultural and Forestry Economics, Department of Economics and Social Sciences, University of Natural Resources and Life Sciences, Vienna, Austria

[3] Department of Agricultural Economics and Rural Development, Georg-August-Universität Göttingen, Göttingen, Germany

improved. During the planning phase, a number of potential treatments, for example educational interventions, are identified through discussions with project partners, colleagues and stakeholders. Unfortunately, the research funds of the project are too limited to test all treatments with sufficient power. Conventionally, the economist would either have to discard some of the treatments (or even consider only one treatment) or run an under-powered experiment. In practice, the researcher may not be confident enough in the prior beliefs about the treatments' effects to make such a decision. It may also be the case that there are strategic reasons to not leave some treatments untested.

This paper discusses a third option for such cases, which might be considered a compromise between the extremes of selecting treatments prior to the trial and selecting treatments once data of full-sized trials are available, namely to use adaptive designs, which aim to use the available resources to maximize the probability of demonstrating at least one of the expected effects with sufficient power. This is achieved by modifying the experimental design after one or several interim analyses. While such adaptive designs have been applied in other fields (such as clinical trials) for decades (cf. Bauer et al., 2016), the concept has only recently become a topic of interest in the economic literature (Kasy & Sautmann, 2021)[1]. When investigating multiple hypotheses (potentially multiple times, in case of an adaptive design), an additional threat to the validity of the final conclusions stems from an inflation of the Type I error probability. This fundamental issue, also referred to as family-wise error rate (FWER) inflation, often arises in experimental research, where multiple treatments are frequently studied (List et al., 2019). Still, it has gained the interest of experimental economists only recently (List et al., 2019; Thompson & Webb, 2019).

The main goal of adaptive designs is to improve the power of the experiment to detect any potential (at least one) statistically significant effect. Ensuring sufficient statistical power has been acknowledged as an important issue in economic research for some time (De Long & Lang, 1992), and remains a challenge in empirical work (Ziliak & McCloskey, 2004; Ioannidis et al., 2017). In the aftermath of the replication crisis in psychology (see e.g. Pashler and Wagenmakers, 2012; Open Science Collaboration, 2015) the issue of insufficient power has gained additional interest, particularly in experimental economics where researchers have been encouraged to account for it (Canavari et al., 2019; List et al., 2011; Czibor et al., 2019). In principle, higher levels of power can be achieved by simply increasing the sample sizes. However, these are often limited in practice. Researchers may face budget or time-related constraints, or have only a small available pool of potential experimental subjects, either due to a small target population or due to restricted access to the wider population. In experimental economics, methods to investigate the required sample sizes in order to detect hypothesised treatment effects predominantly assume an experiment which is of a fixed size in the planning phase (cf. Bellemare et al., 2016).

[1] Interestingly, methods to improve the design of discrete choice experiments are far more widespread and can be considered a standard practice (see e.g. Reed Johnson et al., 2013).

In this contribution, we demonstrate how to use and evaluate adaptive designs in situations where one wishes to combine the selection of interventions with confirmation of their effects by hypothesis tests including data pre and post adaptation, while still controlling the Type I error probability at a specified level. Although the approach is not new from a methodological viewpoint, its application has mostly been in the area of medical statistics, in particular clinical trials. We wish to widen the scope of application by showing how it can be used to design and analyse economic experiments.

As previously mentioned, the economic literature covering adaptive designs is currently limited. For a review of applications in other fields, see e.g. Bauer et al. (2016). Bhat et al. (2020) develop an approach to classical randomisation in experiments with a single treatment for situations where the treatment effect is marred by the effects of covariates and where individual subjects arrive sequentially. Similarly, for experiments where subjects in the sample are observed for multiple periods Xiong et al. (2022) develop an algorithm for staggered rollouts, which optimises the share of treated subjects at different time points. With respect to experiments with multiple treatments and a binary outcome, Kasy and Sautmann (2021) develop an algorithm for experiments with many data collection rounds.

These methods are tailored to specific settings. The framework to be used here can be applied under a broader set of conditions but imposes stronger changes on the experiment. While the aforementioned approaches adjust the treatment assignment probability continuously, treatments are potentially dropped from the experiment (i.e. reducing the assignment probability to zero) in the following. Still, the framework is flexible with respect to the applied adaptation rules and allows for multiple interim analyses, which can be combined in different ways. Although it allows for an arbitrary number of interim analyses, we will restrict the illustrations to two-stage designs, containing only a single interim analysis. This simplifies our presentation, but is also a reasonable choice in many practical settings. We use available software (the R package `asd`) to illustrate the principal concepts and show how they can be applied to achieve a more efficient use of available resources in settings where other approaches are not feasible.

We want to highlight that the present illustration focuses on settings in which the goal of the analysis is to select and test hypotheses regarding the experimental treatments. The improved power for the selected hypotheses comes at the cost of potential biases of the estimated treatment effects. Also, when hypotheses are tested that were not initially considered in the design, these biases can become more relevant (Hadad et al., 2021). Thus, it is important to note that effect estimates obtained from adaptive designs can require additional adjustments (cf. Sect. 4), which should be considered by investigators when applying an adaptive design.

Throughout the paper, the potential of the method is illustrated by three examples. The scenario from the first paragraph will be used to develop a hypothetical example to illustrate the principle considerations researchers have to consider when designing adaptive experiments. The second and third examples use real data from two experimental studies in order to illustrate the potential of adaptive designs in real-data applications. The second application (Musshoff & Hirschauer, 2014) uses a framed field experiment, based on a business simulation game, to study the effectiveness of different

nitrogen extensification schemes in agricultural production. The experiment was carried out with 190 university students and contained treatments varying in two dimensions: whether nitrogen limits were voluntary or mandatory, and whether the corresponding payments were deterministic or stochastic. The students were asked to act as if they were farmers in the experiment. Although the policy treatments do not differ in their impact on the profitability, the authors find that students respond differently to the treatments. Most importantly, the authors report that penalty-based incentives perform better than allowance-based ones.

The third example (Karlan & List, 2007) uses 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 considered treatments which varied in three different dimensions (the matching ratio, the maximum size of the matching grant and the suggested donation amount). The original study is based on a large sample, including 50,083 individuals. The authors find that matching grant offers increase the response rate and revenue of the solicitations, but that different ratios of the matching grant have no effect overall.

The two studies are chosen as they differ in several aspects relevant for economic experiments. First, they represent different experiment types, namely natural field experiment vs. framed field experiment (cf. Harrison and List, 2004). Although the experiment described by Musshoff and Hirschauer (2014) ensured incentive compatibility, the experiment contained hypothetical decision making, whereas the donation decisions in Karlan and List (2007) were non-hypothetical. Moreover, the study of Musshoff and Hirschauer (2014) analyses data from a so-called convenience group. Second, the experiments differ with respect to their complexity (participants having only to decide upon whether to donate and the respective amount vs. participants having to play a complex multi-period business simulation game). Finally, the studies vary with respect to the sample sizes as well as the reported effect sizes. This allows us to illustrate the general usefulness, and potential pitfalls, of adaptive designs for applications in experimental economics. In order to simplify the presentation and keep the focus on the illustration of the method, we only consider subsets of the original data in the applications.

The remaining content of this paper is structured as follows. Section 2 introduces the necessary concepts for the design and analysis of adaptive designs (with a single interim analysis). This includes the necessary theory for the multiple testing and the combined testing of experimental stages, as well as the general concept of the simulations for the power analyses in the initial design phase. In Sect. 3, we apply and illustrate two-stage designs for the three examples and compare their performance to single-stage designs. Section 4 concludes with a discussion of more advanced variations of adaptive designs further explored in other parts of the literature.

2 The theory of adaptive designs

In order to be able to appropriately consider multiple treatments and combine multiple stages in an adaptive design, there are two main concepts which need to be taken into account. The first one consists of procedures to control the Type I error probability, and

the second concept refers to testing principles for combining data from different stages. In the following two subsections, both concepts are introduced. Assuming that multiple hypothesis testing principles are widely known, their description is kept short while a more detailed description is given in the Appendix. These concepts are required to carry out and analyse experiments with an adaptive design. The general procedure, as well as the simulation principles which are used to carry out power analyses, are outlined in the third subsection.

2.1 Multiple hypotheses testing: Type I error probability control

As mentioned in the introduction, one way to increase the power to find a result of interest is simply to test more hypotheses in the same study. However, unless a proper multiple testing procedure is used, increasing the number of hypotheses tested typically leads to a larger Type I error probability. This is problematic since a goal of many designs is to keep this error under strict control at a certain prespecified level. The phenomenon can be illustrated by the following simple example. Suppose that m different null hypotheses are tested based on independent test statistics, and assume that the tests used have been chosen so as to make the individual (i.e., the nominal) Type I error probabilities equal to α. This means that the marginal probability to not reject a specific null hypotheses equals $1-\alpha$, given that it is true and there is no non-zero effect. By the independence assumption, if all null hypotheses are true, then it follows that the probability to reject at least one of them is $1-(1-\alpha)^m$. Hence, as m increases, the probability to falsely reject at least one null hypotheses converges to certainty.

To what extent it is important to keep the Type I error probability under strict control, and, if so, which significance level to choose, varies across scientific disciplines and applications. In order to choose appropriate values of the Type I error probability and power, one must weigh the value of correctly finding a positive effect against the loss of falsely declaring an effect as positive when it is not.

Before proceeding to discuss multiple testing procedures, note that there have been several different suggestions of how to generalise the concept of a Type I error probability for a single hypothesis to the case of multiple hypotheses. These are referred to as different error rates (see, e.g., Chapter 2 of Bretz et al. (2011)). Most relevant for the present paper is the FWER. Let $H_1, \ldots, H_m$ denote a set of m null hypotheses of interest, m_0 of which are true. The FWER is then defined as the probability of making at least one Type I error. This probability depends on the specific combination of hypotheses that are actually true, which of course is unknown when planning the experiment.

Multiple testing procedures for controlling the FWER are based on the concept of intersection hypotheses. Given the m individual hypotheses $H_1, \ldots, H_m$ and a non-empty subset of indices $I \subseteq \{1, \ldots, m\}$, the intersection hypothesis H_I is then defined as

$$H_I = \bigcap_{i \in I} H_i, \quad I \subseteq \{1, \ldots, m\}. \tag{1}$$

Local control of the FWER at level α for a specific intersection hypothesis H_I holds when the conditional probability to reject at least one hypothesis given H_I is at most α, i.e. when

$$\mathbb{P}_{H_I}\left(\text{Reject } H_i \text{ for some } i \in I\right) \leq \alpha. \tag{2}$$

Strong control of the FWER means that the inequality in Eq. (2) must hold for all possible non-empty subsets I. This more conservative requirement bounds the probability of making at least one Type I error regardless of which hypotheses are true. It is often deemed appropriate when the number of hypotheses is relatively small, while the consequences of making an error may be severe. There are a number of different multiple testing procedures available for local control of the FWER. One of the most widely applicable is the well-known Bonferroni procedure, which rejects H_i at level α if the corresponding nominal test would have rejected H_i at level α/m.

If a test has been defined for the local control of a collection of intersection hypotheses, strong control of the FWER can be attained by employing the closed testing principle. This principle states that if we reject an individual hypothesis H_i if and only if all intersection hypotheses H_I such that $i \in I$ are rejected at local level α, then the FWER is strongly controlled at level α (Marcus et al., 1976). Since this principle is completely general with regards to the specific form of the tests for the intersection hypotheses H_I, it follows that the choice of the local test will determine the properties of the overall procedure.

For many economic experiments, where multiple treatment groups are compared with one control group, a suitable test procedure for testing intersection hypotheses is the Dunnett test (Dunnett, 1955). Using the closed testing principle, the Dunnet test can be used to obtain strong control over the FWER. The test is described in more detail in the Appendix.

2.2 Combined testing in adaptive designs

If we can look at some of the data before spending all available resources, then we could allocate the remaining samples to the most promising treatments. The present section contains a short introduction to the general topic of such adaptive designs, sufficient for the latter applications. As mentioned in the introduction, for reasons of clarity, we will limit ourselves to the case of two-stage designs with a single interim analysis. There are two main approaches. One is referred to as the combination test approach and was introduced by Bauer (1989), and Bauer and Köhne (1994). The other, which makes use of conditional error functions, was introduced by Proschan and Hunsberger (1995) for sample size recalculation, while Müller and Schäfer (2004) widened the approach to allow for general design changes during any part of the trial.

Following Wassmer and Brannath (2016), we assume for the moment that the purpose of the study is to test a single null hypothesis H_0. We will see later how to combine interim adaptations with the closed testing principle when several hypotheses are involved. When the interim analysis is performed, there are two possibilities. Either H_0 is rejected solely based on the interim data T_1 from the first stage, or the study continues

to the second stage, yielding data that we assume can be summarised by means of a statistic T_2. In general, the null distribution of T_2 may then depend on the interim data. For instance, the first stage could inform an adjustment of the sample size of the second stage. Nevertheless, it is often the case that T_2 may be transformed so as to make the resulting distribution invariant with respect to the interim data and the adjustment procedure. In other words, the statistics used to evaluate the two stages will be independent given that H_0 is true, regardless of the interim data and the adaptive procedure.

As an example, suppose that T_2 is a normally distributed sample mean under H_0, with mean 0 and a known variance σ^2/n_2, with σ^2 being the variance of a single observation and n_2 a second stage sample size that depends on the interim data in some way. Then a one-sided p-value for the second stage can be obtained by standardising T_2 according to

$$p_2 = 1 - \boldsymbol{\Phi}\left(\frac{T_2}{\sigma/\sqrt{n_2}}\right). \tag{3}$$

Assuming that the data from the two different stages is independent, it follows that the joint distribution of the p-values p_1 and p_2 from the two stages is known and invariant with respect to the interim adjustment. This is the key property which allows for the flexibility when choosing the adaptive design procedure.

Based on the conditional invariance principle, the design of an adaptive trial controlling the Type I error probability at a level α can be specified in terms of a rejection region for p_1 and a corresponding region for p_2. This results in a sequential test of H_0 which is independent of the adaptation rule. The aforementioned main testing principles (combination tests and conditional error functions) only differ in the way the rejection region for the second stage is specified.

For the combination test approach, suppose that a one-sided null hypothesis $H_0 : \theta \leq 0$. θ is some parameter of interest and is to be tested using a two-stage design with p-values p_1 and p_2. The combination test procedure is specified in terms of three real-valued parameters, α_0, α_1 and c, and a *combination function* $C(p_1, p_2)$. It proceeds as follows:

1. After the interim analysis, stop the trial with rejection of H_0 if $p_1 \leq \alpha_1$ and with acceptance of H_0 if $p_1 > \alpha_0$. If $\alpha_0 < p_1 \leq \alpha_1$, apply the adaptation rule and collect the data of the second stage.
2. After the second stage, reject H_0 if $C(p_1, p_2) \leq c$ and accept H_0 if $C(p_1, p_2) > c$.

By virtue of the conditional invariance principle, it is easily shown that the Type I error probability is controlled at level α by any choice of α_0, α_1, c, and $C(p_1, p_2)$ satisfying

$$\mathbb{P}_{H_0}\big(\text{Reject}H_0\big) = \alpha_1 + \int_{\alpha_1}^{\alpha_0}\int_0^1 \mathbf{1}\big\{C(p_1, p_2) \leq c\big\}\,\mathrm{d}p_2\,\mathrm{d}p_1 \leq \alpha, \tag{4}$$

where $\mathbf{1}\{\cdot\}$ denotes the indicator function. Naturally, α_1 needs to be smaller than α_0. While specific values depend on the individual setting, Bauer and Köhne (1994, p.1031) argue that "a suitable value for α_0 could be $\alpha_0 = .5$". This corresponds

to the situation that the effect goes in the right direction. If this value is chosen for α_0 and the overall level of α is $\alpha = 0.05$, the remaining values are $\alpha_1 = 0.0233$ and $c = 0.0087$ (see Table 1 in Bauer and Köhne (1994) for other typical threshold values). As a special case, if the first stage parameters of Eq. (4) are set to $\alpha_1 = 0$ and $\alpha_0 = 1$, then there will be no early stopping (neither for futility nor with early success) and the null hypothesis will be rejected after the second stage if and only if $C(p_1, p_2) \leq \alpha$.

The other main approach, using conditional error functions, is very similar. The parameters α_0 and α_1 have the same role, but c and $C(p_1, p_2)$ are replaced by a conditional error function $A(p_1)$ such that, after the second stage, H_0 is rejected if $p_2 \leq A(p_1)$ and accepted if $p_2 > A(p_1)$. It is important to note that, as long as one of the procedures described above is followed, any type of rule can be used to update the sample size in an interim analysis. This means that not only can the information on efficacy collected so far be used in an arbitrary manner, but so can also any information external to the trial. For a more comprehensive review of the history and current status of adaptive designs as applied to clinical trials, see Bauer et al. (2016).

In the following, we only consider the inverse normal combination function (Lehmacher & Wassmer, 1999) for combining the p-values from the two stages. Although alternative choices exist it is commonly used and is the default method in the used software implementation. The definition of this function is

$$C(p_1, p_2) = 1 - \boldsymbol{\Phi}\big(w_1 \boldsymbol{\Phi}^{-1}(1 - p_1) + w_2 \boldsymbol{\Phi}^{-1}(1 - p_2)\big), \tag{5}$$

where w_1 and w_2 are pre-specified weights that satisfy the requirement $w_1^2 + w_2^2 = 1$.[2] With p_1 and p_2 independent and uniformly distributed, it follows from the definition that $C(p_1, p_2)$ becomes a random variable of the form $C(p_1, p_2) = 1 - \boldsymbol{\Phi}(Z)$, where Z follows a standard normal distribution. Thus, $C(p_1, p_2)$ can be treated as a p-value that combines the information from the two stages.

2.3 Two stage designs and power simulations

Having briefly introduced the necessary components, we are now ready to describe the step-by-step procedure characterising the class of adaptive designs of interest.

1. Define a set of one-sided, individual null hypotheses $H_1, \ldots, H_m$ corresponding to comparisons of treatment means μ_i against a common control μ_0 (i.e., $H_i : \mu_i \leq \mu_0$).
2. Given the first stage data, apply Dunnett's test and compute a p-value p_1^I for each non-empty intersection hypothesis H_I such that $I \subseteq \{1, \ldots, m\}$.
3. Apply a treatment selection procedure to the first stage data, for example by selecting a fixed number of the best treatments. This results in a subset $J \subseteq \{1, \ldots, m\}$ taken forward to the second stage.

[2] Following Jenkins et al. (2011), the weights $w_1 = \sqrt{n_1/(n_1 + n_2)}$ and $w_2 = \sqrt{n_2/(n_1 + n_2)}$, where n_1 and n_2 are the group sizes for stage 1 and 2, are used in all applications in the following. Note that n_1 for a two-stage design corresponds to n for a single-stage design, as used in Eq. (9) in the Appendix.

4. Given the second stage data, apply Dunnett's test and compute a p-value p_2^I for each non-empty intersection hypothesis H_I such that $I \subseteq J$.
5. Compute combined p-values $p_c^I = C(p_1^I, p_2^I)$ using a normal combination function for each $I \subseteq J$. Reject each H_I at local level α for which $p_c^I \leq \alpha$. Finally, ensure strong control of the FWER at level α by applying the closed testing principle and rejecting the individual hypotheses H_i satisfying $p_c^I \leq \alpha$ for all $I \subseteq J$ such that $i \in I$.

Note that this procedure assumes that those hypotheses dropped at the interim cannot be rejected in the final analysis. This leads to a conservative procedure with an actual Type I error probability that may be lower than the pre-specified level α (Posch et al., 2005).

There are a number of selection rules, which can be applied in this procedure. For example a fixed number (s) of treatments can be taken forward from the first to the second stage, e.g. the single best ($s = 1$) or the two best treatments ($s = 2$). Alternatively, the number of treatments taken forward can be flexible, either taking all treatments within a given range from the best treatment forward (ϵ-rule), or all treatments with a effect estimate at the interim analysis larger than a given threshold. Assuming that in most applications, the total sample size will be fixed due to budget reasons or a limited sample pool, specifying the number of treatments taken forward will often be a sensible choice, as it allows to determine the second stage sample size in the design phase.

Before an experiment is carried out, researchers usually want to investigate its statistical power. For adaptive approaches like the ones outlined above, these analyses build on simulations. All simulations presented in the remainder of the paper were carried out using the open source R package `asd`[3]. 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).

After the treatments have been chosen for consideration in the experiment, the power simulations, as all power analyses, require prior beliefs about the expected effect sizes of the treatments. Further, the researcher needs to specify the number of observations in the different stages of the experiment and the selection rule (i.e. the numbers of groups which transfer to the second stage). The core idea for the simulation is to directly simulate the test statistics for the following hypothesis test, rather than individual observations (see Parsons et al., 2012; Friede et al.,2011 for details). Using a large number of drawn test statistics then allows to assess the expected power. The `asd` package, designed for these purposes, handles the issue of strong control of the FWER. A practical example of the simulation procedure is given in the following section. Additionally, a more principled discussion of the simulation procedure can be found in the supplementary material.

[3] There are commercially available alternatives (e.g., ADDPLAN and EAST). Another R package called `rpact` implements the methods described in the monograph by Wassmer and Brannath (2016) and might be of interest to readers looking to apply the methodology, but is not used here. In addition, Richter et al. (2022) describe another software package that might be of interest.

3 Applications of adaptive designs in experimental economics

As discussed in the introduction, adaptive designs can be applied in many different settings. In order to illustrate the potential, we consider three examples in this section. The first one is the hypothetical setting outlined in Sect. 1 and is used to illustrate the simulation procedure for the planning of an adaptive design. The second and third examples are based on prior research. Based on the data and the results of the original studies, results of simulations for adaptive designs are presented. More detailed descriptions of the experimental designs of the original studies can be found in the supplementary material.

3.1 Application 1: Design of a field experiment

For the scenario outlined in the introduction, imagine that after consultations with stakeholders and colleagues, four potential treatments (+ one control) have been selected for consideration. Further, expected costs for data collection will allow for a total sample size of 300. For the present case, the project group may agree that standardized (Cohen's d) effect sizes of 0.2, 0.2, 0.4 and 0.5 could reasonably be expected for the four treatments, implying that the group thinks that it is realistic that two of the treatments have small sized effects, and that the other two have medium sized effects (noted as "basic" scenario in the following).

One straightforward adaptive design would be that an interim analysis is carried out after 50% of the sample is collected and only the two best treatments are taken forward (implying a group size of 30 observations in the first stage, and 50 in the second stage). In a standard, non-adaptive experiment, the results of an interim analysis would not have any effect on the subsequent data collection, meaning that all four treatments are to be included in the second stage of the data collection. Thus, the power of both such experiments can be simulated with the `asd`-package[4]. Using 10,000 simulations shows that, for the standard design, the power of the experiment to reject at least one of the four Null hypotheses would be 62.27 % (Table 1). Such a design would conventionally be considered to be under-powered.

Table 1 Power simulation for the single stage design

Treatment	Assumed effect size	Selection of the treatment at stage 1 (%)	Hypothesis rejection at endpoint (%)
T1	0.2	100	12.87
T2	0.2	100	12.64
T3	0.4	100	37.72
T4	0.5	100	55.33

Power to reject H1 and/or H2 and/or H3 and/or H4 = 62.27%

[4] For the code see the replication material.

Table 2 Power simulation for the two stage design

Treatment	Assumed effect size	Selection of the treatment at stage 1 (%)	Hypothesis rejection at endpoint (%)
T1	0.2	23.02	6.85
T2	0.2	22.02	6.24
T3	0.4	69.65	45.05
T4	0.5	85.31	67.79

Power to reject H1 and/or H2 and/or H3 and/or H4 = 79.21 %

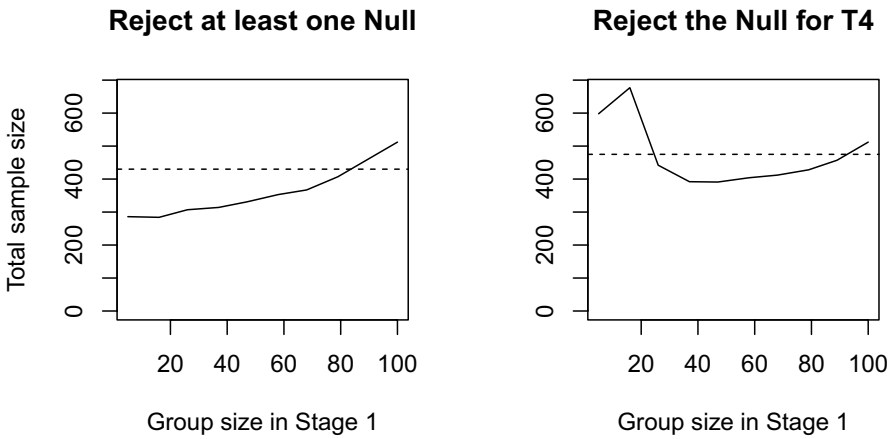

Fig. 1 Total sample sizes to achieve 80% power in the basic scenario, for different first stage group sizes; *Note*: The dashed line indicates the sample size when no treatment selection is carried out

In contrast, when only the two most promising treatments are carried over to the second stage after the interim analysis, the expected power to reject at least one of the Null hypotheses increases up to close to the conventional 80 % threshold (79.21 %, Table 2).

This serves as a first illustration of one of the potential benefits of the adaptive designs: given a fixed sample size, the power to reject at least one of the Nulls is substantially increased. Of course, this represents only one potential design. For example, it may be possible to find a design which also achieves an acceptable level of power, but requires a smaller total sample size, by changing the group size in the first stage. The left side of Fig. 1 shows the relationship between the first stage group size and the total sample size required to achieve 80 % power for rejecting at last one hypothesis. The right side shows the required sample size if the goal would be the rejection of the treatment with the largest assumed effect size (T4)[5].

[5] The depicted sample sizes are the ones which minimize the absolute deviation of the expected power from 80 %.

Table 3 Effect size assumptions used in the power simulations

Scenario	T1	T2	T3	T4
Basic	0.2	0.2	0.4	0.5
Alternative I	0.1	0.1	0.2	0.5
Alternative II	0.1	0.1	0.5	0.8

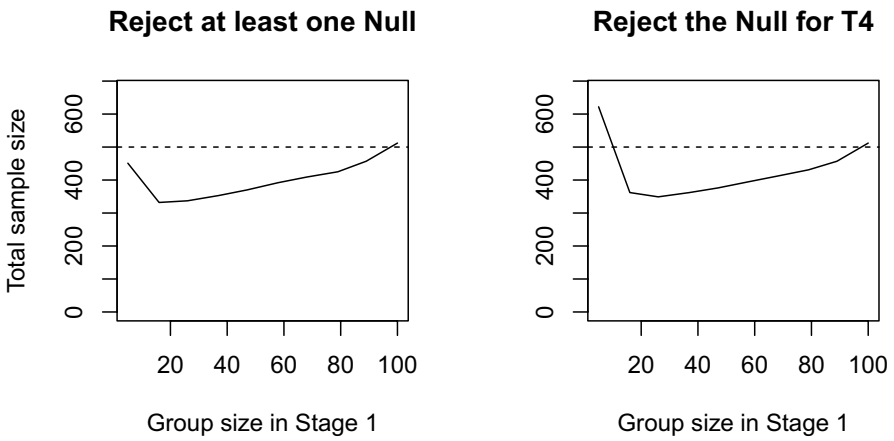

Fig. 2 Total sample sizes to achieve 80% power in scenario "Alternative I", for different first stage group sizes; *Note*: The dashed line indicates the sample size when no treatment selection is carried out

Figure 1 shows that, compared to a single stage design, the required sample size can be decreased in both cases. The required sample size to reject at least one null hypothesis can be reduced to 67 % of the original sample size. When the interest would be solely on rejecting the Null for the largest expected effect, the required minimum sample size is higher, but there is a 16 % reduction in comparison to the non-adaptive sample size. It is noteworthy that when the interest is only to reject the null hypothesis for the largest effect (T4), the group size of the first stage almost doubles, compared to when only aiming to reject one of the null hypotheses (around 40 in the former and around 20 in the latter case; compare the positions of the minima for the total sample sizes in Fig. 1).

As in other power analyses, the outcomes crucially depend on the assumptions about the expected effect sizes. To illustrate this, assume that there are two alternative sets of assumptions about the expected effect sizes held by different groups of researchers ("Alternative I" and "Alternative II"), depicted in Table 3. One group may believe that the "small" effect sizes are actually smaller and the medium effect sizes are more "distinct" than in the basic scenario ("Alternative 1"), while another group may argue that the two treatments will only have small effects, one a medium and the last one a large effect. The required sample sizes of these sets can be simulated for the basic scenario and are depicted in Figs. 2 and 3.

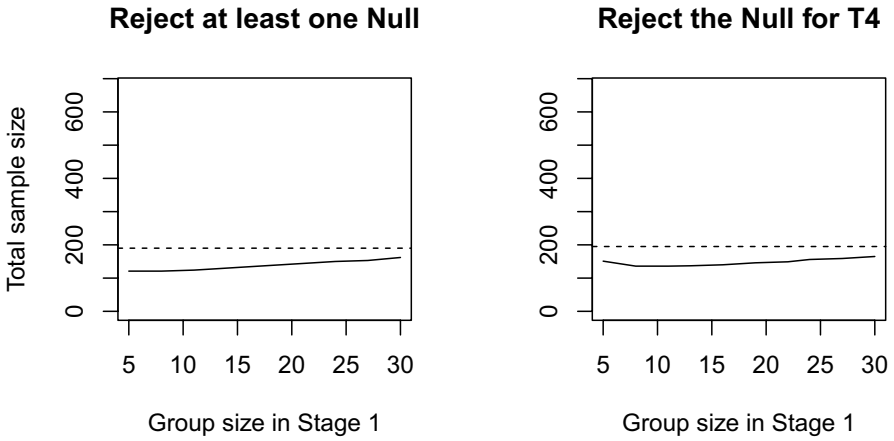

Fig. 3 Total sample sizes to achieve 80% power in scenario "Alternative II", for different first stage group sizes; *Note*: The dashed line indicates the sample size when no treatment selection is carried out

The difference between the expected effect sizes of T3 and T4 in scenario Alternative I is larger than in the basic scenario. At the same time, the expected effect sizes for T1, T2 and T3 are smaller than in the basic scenario (cf. Table 3). This leads to overall larger required sample sizes for Alternative 1, both for single stage designs and two-stage designs (for most of the considered first stage group sizes, compare Figs. 1 and 2). Still, the expected savings in sample size are 33% and 31%, and both optimal group sizes in the first stage around 20–25 observations (cf. the minimum of the solid line in Fig. 2). The assumptions for Alternative II, which include one large expected effect, lead to substantially smaller sample sizes than the basic scenario, for all considered designs (compare Figs. 1 and 3). Still, using an adaptive design allows saving a substantial share of the required total sample size (respectively 35% and 32%).

These simulation results show that the required sample size for a two-stage design could be reduced by about 30% compared to a single-stage design. Still, the sample size savings vary for the different assumptions regarding the effect sizes. It has also be noted that this requires that the group size in the first stage is appropriately chosen; see discussion below.

3.2 Application 2: A business management game for regulatory impact analysis

To further illustrate the application of the adaptive designs, we apply a two-stage design to the experiment of Musshoff and Hirschauer (2014). Here, rather than simulating the experiment, a subset of the actual data from the experiment is used. In their experiment, the impact of four different nitrogen reduction policies on the decision-making of the participants is investigated and compared to a base case. The participants play a multi-period, single-player business simulation game. The

sample consists of data from 190 agricultural university students, taking on the role of farmers deciding on a production plan for a hypothetical crop farm. Each individual plan specifies the land usage shares for the different crops, as well as the amount of nitrogen fertiliser to be used. 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.

In the first 10 production periods, all players receive an unconditional, governmental deterministic agricultural payment. After these initial periods, the players are randomly assigned to one of five different policy scenarios. The first scenario 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. They differ with respect to whether payments are granted (scenario 2 and 3), or whether penalties are charged (scenario 4 and 5) when the fertiliser usage is respectively above or below a given threshold (voluntary vs. prescriptive). Further, they vary in terms of whether the payments or penalties are made without exception (scenario 2 and 4), or whether there is only a certain probability that a player's action is discovered and payments or penalties are applied (scenario 3 and 5; deterministic vs. stochastic). The scenario designs ensure the same expected monetary gain for the players (see Sect. 3 of Musshoff and Hirschauer, 2014 for a detailed description of the experiment).

The main outcome variable of interest is the share of arable land where the players kept the fertiliser levels below a pre-specified threshold referred to as "extensive use", defined to be 75 kg per hectare. In the first ten periods the group mean shares were similar, all between 11.6% and 17.4%. For periods 11 to 20 the shares were, for scenario 1 to 5 in order, 13.6%, 67.5%, 80.7%, 80.9% and 49.4% (see Table 4 of Musshoff and Hirschauer, 2014). Thus, scenarios 2-5 yield much higher share values compared to the reference scenario, and moreover these seem to lead to different outcome means although they all give the same expected profit.

Here, we focus on the subset of data corresponding to scenarios 2, 3 and 4. For each of these groups, there is data for 38 players. As the total sample size increases in the examples below, more and more players are used from each group, in the order they are included in the original data file.[6] We use scenario 2 as the control (rather than scenario 1) and treat 3 and 4 as two different active treatments (treatment 1 and 2). We may interpret this as a smaller study in which a deterministic penalty (scenario 3) and a stochastic penalty (scenario 4) are compared against a deterministic reward (scenario 2). The subset is chosen for illustrative purposes, but is also in close correspondence with one of the original hypotheses of Musshoff and Hirschauer (2014). The null hypothesis for the comparison of scenario 3 vs. 2 is denoted by H_1, while the one corresponding to 4 vs. 2 is denoted by H_2. While scenario 2 (as a policy intervention) would be preferable based on equality arguments, high monitoring costs may make it impractical and warrant alternative interventions. Such are provided by scenarios 3 and 4, but which of them works best when

[6] Table 4 of Musshoff and Hirschauer (2014) can be reconstructed by taking the mean group values of the data used here (for maximum group sizes) and dividing by 400.

Table 4 Hypotheses rejected in Dunnett tests for single- and two-stage designs of increasing sizes, together with mean difference estimates of effect sizes.

(a) Single-stage design

Row	n_1 (N_1)	$H_1 \cap H_2$	Treatment 1	Treatment 2
			H_1 ($\bar{x}_1 - \bar{x}_0$)	H_2 ($\bar{x}_2 - \bar{x}_0$)
1	10 (30)	No	No (8.00)	No (39.30)
2	13 (39)	No	No (35.38)	No (32.54)
3	16 (48)	No	No (32.31)	No (25.63)
4	20 (60)	No	No (36.65)	No (55.75)
5	23 (69)	No	No (39.04)	No (50.22)
6	26 (78)	No	No (41.38)	No (53.19)
7	30 (90)	No	No (55.03)	No (52.60)

(b) Two-stage design

Row	n_1 (N_1)	n_2 (N_2)	$H_1 \cap H_2$	Treatment 1	Treatment 2
				H_1 ($\bar{x}_1 - \bar{x}_0$)	H_2 ($\bar{x}_2 - \bar{x}_0$)
1	5 (15)	7 (14)	No	No (12.83)	No (64.42)
2	7 (21)	9 (18)	No	No (24.28)	No (25.63)
3	9 (27)	11 (22)	No	No (30.77)	No (55.75)
4	10 (30)	15 (30)	No	No (39.00)	No (53.72)
5	12 (36)	17 (34)	No	No (22.24)	No (49.07)
6	14 (42)	19 (38)	Yes	Yes (60.67)	No (29.51)
7	15 (45)	22 (44)	Yes	Yes (58.57)	No (36.00)

Note that the total sample sizes for the corresponding rows of Tables 4a and b have been selected to be as close as possible. Type I error probability $\alpha = 0.025$. n_1 = first stage group size, n_2 = second stage group size. N_1 = total first stage sample size, N_2 = total second stage sample size

taking into account the multiple goals (e.g. income and environment concerns) and bounded rationality of an economic decision maker?

Tables 4a and b provide a comparison showing which hypotheses are rejected for different subsets of the data. As in the previous subsection, groups within the same stage are always of equal size. The subset of the data used consists of one control and two treatment groups (treatment 1 and 2). Thus, the selection rule at the interim analysis study takes only the better active treatment forward. As in the basic scenario in Application 1, approximately half the total sample size is used in the first stage. This design was chosen since it was found to be close to optimal in the simulations performed. It can be observed that the two-stage design is able to reject at least one hypothesis (namely, H_1) once the total sample size N becomes large enough (row 6 and 7 in Table 4b), while the single-stage design never rejects any hypothesis. Hence, the adaptive design increases the power of detecting at least one effect by dropping one scenario at the interim analysis and focusing on the one that seems to have the largest effect.

3.3 Application 3: A field experiment in charitable giving behaviour

Last, we consider the study of Karlan and List (2007), which is concerned with the effect of a matching grant on charitable giving. A matching grant represents an additional donation made by a central donor backing the study, the size of which depends on the amount donated by the regular donors. The sample consists of data from 50,083 individuals, who were identified as previous donors to a non-profit organisation and were contacted by mail.

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)[7]; (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. The authors study both the effects on the response rate as well as the donation amount. However, we will only focus on how the matching ratio affects the response rate. The reason for this simplification is that we want to be able to connect the simulation results in the previous section to the available real-world data without increasing the complexity of the presentation by introducing too many treatment groups.

For illustrative purposes, we will only consider a portion of the data analysed by Karlan and List (2007). Specifically, we will restrict ourselves to the subset consisting of those individuals residing in red states (i.e., states with a republican majority, see Panel C of Table 2A of Karlan and List, 2007). The reason for this is that, as noted by Karlan and List (2007), it is only for this subgroup that the original analysis shows a statistically significant difference between the groups defined by the different matching ratios. Naturally, a real analysis would not ignore data just because it is not statistically significant. However, since our purpose here is to compare methods of data analysis rather than to answer specific research questions, we choose to restrict ourselves to a subset in order to be able to more clearly highlight the difference between the single-stage and two-stage designs. Thus, there are three active treatments, corresponding to the different matching ratios. Again, the sample size is divided equally among the groups in each stage, but we employ a rule that selects the two best treatments at the interim analysis, in contrast to the previous application.

Comparing Table 5a with b, it can be seen that the single-stage design rejects no hypothesis up until a total sample size of $N_1 = 7000$ (row 6 in Table 5a). In contrast, the two-stage design starts to reject a hypothesis, namely H_1 (treatment 1), from $N_1 + N_2 = 6000$ (row 5 in Table 5b), indicating the potential reduction to sample size by 1000 observations.

[7] The ratio $X:$1 indicates that for every dollar the individual donates, the matching donor additionally contributes $X.

Table 5 Hypotheses rejected in Dunnett tests for single- and two-stage designs of increasing sizes, together with mean difference estimates of effect sizes

(a) Single-stage design

Row	n_1 (N_1)	$H_1 \cap H_2 \cap H_3$	Treatment 1 H_1 ($\bar{x}_1 - \bar{x}_0$)	Treatment 2 H_2 ($\bar{x}_2 - \bar{x}_0$)	Treatment 3 H_3 ($\bar{x}_3 - \bar{x}_0$)
1	500 (2000)	No	No (0.004)	No (-0.006)	No (0.002)
2	750 (3000)	No	No (0.008)	No (-0.005)	No (0.000)
3	1,000 (4000)	No	No (0.009)	No (0.000)	No (0.006)
4	1,250 (5000)	No	No (0.010)	No (0.002)	No (0.005)
5	1,500 (6000)	No	No (0.011)	No (0.006)	No (0.007)
6	1,750 (7000)	Yes	Yes (0.014)	No (0.008)	No (0.010)
7	2,000 (8000)	Yes	Yes (0.012)	No (0.008)	Yes (0.010)
8	2,250 (9000)	Yes	Yes (0.010)	No (0.008)	Yes (0.012)
9	2,500 (10,000)	Yes	Yes (0.012)	Yes (0.009)	Yes (0.014)
10	2,750 (11,000)	Yes	Yes (0.013)	Yes (0.009)	Yes (0.014)

(b) Two-stage design

Row	n_1 (N_1)	n_2 (N_2)	$H_1 \cap H_2 \cap H_3$	Treatment 1 H_1 ($\bar{x}_1 - \bar{x}_0$)	Treatment 2 H_2 ($\bar{x}_2 - \bar{x}_0$)	Treatment 3 H_3 ($\bar{x}_3 - \bar{x}_0$)
1	250 (1000)	333 (999)	No	No (0.003)	No (-0.007)	No (0.002)
2	375 (1500)	500 (1500)	No	No (0.010)	No (0.003)	No (0.002)
3	500 (2000)	666 (1988)	No	No (0.009)	No (0.003)	No (0.004)
4	625 (2500)	833 (2499)	No	No (0.009)	No (0.002)	No (0.005)
5	750 (3000)	1000 (3000)	Yes	Yes (0.014)	No (0.005)	No (0.010)
6	875 (3500)	1166 (3498)	Yes	Yes (0.011)	No (0.006)	Yes (0.010)
7	1,000 (4000)	1333 (3999)	Yes	Yes (0.010)	No (0.004)	Yes (0.013)
8	1,125 (4500)	1500 (4500)	Yes	Yes (0.012)	No (0.004)	Yes (0.013)
9	1,250 (5000)	1666 (4998)	Yes	Yes (0.011)	No (0.003)	Yes (0.013)
10	1,375 (5500)	1833 (5499)	Yes	Yes (0.010)	Yes (0.008)	No (0.005)

Note that the total sample sizes for the corresponding rows of Tables 5a and 5b have been selected to be as close as possible. Type I error probability $\alpha = 0.025$. n_1 = first stage group size, n_2 = second stage group size. N_1 = total first stage sample size, N_2 = total second stage sample size

4 Discussion and conclusions

In this paper, we have introduced the basic theory of adaptive designs for multiple hypothesis testing and illustrated it by means of hypothetical use-case scenario and two real-world data sets. Although the basic theory required for the applications is covered by Sect. 2, the general approach of using combination tests and the closed testing principle provides a starting point for more advanced designs. Using the R

package `asd` for our simulation study reflects our aim to introduce and illustrate the methodology, since it is well suited for many applications.

The results presented in Sect. 3.1 illustrate the potential of moving from single-stage to two-stage designs. Clearly, the results on the potential for saving costs by collecting less data, while maintaining a fixed level of power, are specific to the considered scenario. However, they still serve as an indication of what can be gained when moving from a single-stage to a two-stage design. The simulations for the hypothetical examples indicate that the required sample size could be reduced by around 30% by using a two-stage design. While these results show the potential of applying such designs, the decision to apply them require additional practical considerations. Most importantly, the costs associated with the administration and data analysis of the more complex two-stage design compared with a single-stage design have to be taken into account. For example, when field experiments are carried out in the paper-and-pencil-format, digitising and analysing the interim data may significantly increase the study's costs (or even be infeasible due to time constraints). Still, when the data is collected in digital form (e.g. in an economic laboratory), a two-stage design may not impose any additional costs at all.

The relatively simple models considered in the present paper can be extended in a number of different ways. One option would be to adjust the effects for baseline covariates in suitable regression models; the combination test approach illustrated here can still be applied. Another option would be to consider designs with more than two stages. Whether or not this is worthwhile from a practical perspective depends to a high degree on the specific application considered. Again, if experimental data are sequentially collected in multiple sessions and in a form which can be analysed without technical difficulties, multi-stage designs could be a viable option. Here, it is important to emphasise that this will not pose any issues from a theoretical perspective, since the general method of using combination tests and the closed testing principle would still be applicable. Nevertheless, the simulations would require using other (mainly commercial) software packages or additional programming since the `asd` is specifically designed for two stage designs. Also, the main argument for a reduction of the sample size is a reduction in costs (apart from a potentially limited sample pool). This rests on the implicit assumption that all treatments have identical costs per treated observation. In cases where this assumption does not hold, further considerations may be required. In simple cases, treatment effects may be expressed per monetary unit spent, but in more complex cases, specific algorithms may have to be developed. As the goal here was to illustrate a general approach, these considerations are beyond the scope of the present paper.

It is further important to note that alternative selection rules could have been considered in the previous sections. As mentioned earlier, selection rules that take a fixed number of treatments forward to the next stage will often be natural choices, in particular selecting the best treatment only. The specific number of treatments taken further has to be determined based on the number of initial treatments, the expected available total sample size and the relative differences in the expected effect sizes, and it must also be chosen as such that it leads to an acceptable expected power.

The general framework of adaptive designs also allows for essentially any type of treatment selection rule. For example, in settings with a larger number of potential treatments and/or less informative beliefs about relative differences between the effect sizes, research could employ the ϵ-rule and select all the treatments with effect estimates within a certain range of the best one. This mimics decisions by experts who would likely select only a single treatment which is better than others, but might take other treatments further in addition to the best if they come close to the best treatment. Therefore, the ϵ-rule can be used to mimic expert groups in simulations while the decision in the real study might not follow a strict rule but might be made by a group of (independent) experts. Nevertheless, Friede and Stallard (2008) find that in settings where the ϵ-rule is considered, selecting only most promising treatment will often be (close to) optimal. Note that the Type I error rate is still controlled independent of the selection rule. Under the alternative, however, the power will be dependent on the selection rule. Furthermore, researchers may rely on other rules such as a threshold-based selection rule, e.g. only taking to the second stage treatments for which interim results indicate effect sizes of practical relevance.

Related to the choice of the selection rule is the selection of the sample size of the first stage of the experiment. The final decision will depend on the number of planned treatments in the first stage and on the assumptions regarding potential effect sizes as well as differences between treatments. Still, in medical trial settings, research indicates that it will often be optimal to allocate around 50% of the total sample size to the first stage (see e.g. Chataway et al.,2011, Friede et al., 2020, Placzek and Friede 2022). This is within the range of optimal allocations found for the hypothetical example in Sect. 3.1. Here, the simulations suggest to allocate between 25% and 60% of the total sample size to the first stage. This highlights the need for simulations in the planning of the experiments. While these simulations can become time-consuming, recent developments in optimization methods can help to reduce their computational burden (Richter et al., 2022).

The use of Dunnett's test for many-to-one comparisons is suitable for the applications considered here, but it is only one of several possible choices to use for the intersection test when applying the closed testing principle. Hence, alternative applications may require other types of tests. Furthermore, when a researcher wants to carry out large-scale experiments with a very large number of hypotheses tested simultaneously, strong control of the FWER may not be an appropriate criterion. A possible alternative that has been suggested for such cases is the per-comparison error rate (see the Appendix and Bretz et al. (2011)).

In our analysis we focused exclusively on hypothesis testing as the main goal of the experiment. However, formulating and testing null hypotheses is often only partly the goal of similar studies. In addition, the experimenter often seeks to perform some kind of model or parameter estimation. This could consist of reporting estimates and confidence intervals for the different treatments considered. When constructing estimates from the data obtained in an adaptive experiment, it is important to note that these may be biased, as adaptive designs can affect the statistical distribution of the treatment effects (Jennison & Turnbull, 2000).

Although we recognise the importance of estimation problems, the goal here has been to focus on the basic concepts of adaptive designs, particularly the issue of a

strong FWER control in the presence of interim analyses. For a general discussion of bias in adaptive designs and bias-correcting procedures, see Pallmann et al. (2018) and the references therein. More recently, Hadad et al. (2021) discussed approaches specifically in the context of policy evaluation settings. To connect this general issue to one of our applications, consider the analysis by Musshoff and Hirschauer (2014). There, the objective is not only to determine whether a specific set of hypotheses can be rejected or not. The data are also used to fit a regression model, which adjusts for baseline covariates. While the confidence intervals obtained for the estimated parameters are used to test the hypotheses, the value of such a model goes beyond testing since it can also be used for more general predictive statements.

Another issue is potential temporal differences, i.e. changes during the course of the experiment. When the treatment effects change between the different stages of the experiment (e.g. due to seasonal effects in a field experiment), the conclusions drawn from the interim analysis may become biased or unreliable. If systematic heterogeneity between the experimental stages is suspected, researchers may also address these issues by formal testing procedures; see e.g. Friede and Henderson (2009).

There is also a number of other related design approaches, which could be considered for economic experiments. Instead of the many-to-one comparisons that have been our focus here, similar methods can also be applied to the closely related topic of subgroup selection, in which one or more subgroups of treatment units are taken forward to the next stage in a sequential study while the treatment stays the same. These are referred to as adaptive enrichment designs (see, e.g. Chapter 11 of Wassmer and Brannath, 2016). Simulations for a special case of such designs, involving co-primary analyses in a pre-defined subgroup and the full population, are supported by the R-package `asd`.

Another approach, which could be of interest when researchers are interested in (ideally) persistent treatment effects, is adaptive designs which make use of early outcomes. This means that the decision concerning which treatment to take forward to the second stage is based on a measure that need not be identical but only correlated to the one actually used in the final analysis. For example, when a study is interested in the effects of educational interventions, a suitable early outcome could be the result of a test taken (relatively) shortly after the intervention. While such designs are supported by the `asd` package, they require more prior knowledge (respectively assumptions), most importantly about the correlation between the early and final outcomes of the treatments.

An approach that is closely connected to general adaptive designs is that of group-sequential designs. For these, a maximum number of groups is first pre-specified. Subjects are then recruited sequentially and an interim analysis is performed once the results for a full group have been obtained. The value of a summary statistic that includes the information from all groups seen so far is computed and used to decide whether to stop or continue the trial. The main body of the theory is built on the often reasonable assumption of a normally distributed summary statistic, which allows us to use methods to compute the characteristics in a numerically efficient manner (recursive integration formula, Armitage et al. (1969)). A number of different group-sequential designs have been suggested in the literature (e.g. Pocock,

1977, O'Brien and Fleming 1979). These classical designs can be seen as special cases of the more general adaptive approach that we have used in this paper. Whereas the group-sequential approach may lead to a certain degree of flexibility in the total sample size of the trial by allowing for early termination at an interim analysis, it does not allow for things such as sample size reassessment or the dropping of one or several treatment arms. For a comprehensive treatment of the group-sequential approach, see for example Jennison and Turnbull (2000).

In principle, adaptive designs could also be applied for replication studies. Still, some conceptual aspects have to be considered. Like in regular replications, one would first have to consider whether the original study can be considered adequately powered. If the conclusion is positive and there are limited resources, the researcher may maximize the probability of replicating an effect by simply focusing on the treatment with the largest effect (using regular power analysis methods). Similarly, a finding of an adaptive experiment will also typically be confirmed using a regular experiment (e.g. for clinical trials). In relation to this, one may wonder whether the approach outlined in the paper could be used to engage with questionable practices, such as p-hacking, e.g. by omitting the first stage from the presented paper. While outright fraud can never be ruled out, we believe it is not of practical concern. The main reason is that there is not much to gain from hiding the first stage. If data of the first stage would be omitted (which would make it easier to present a consistent description of the data collection), the loss of observations, and thus the loss of power, respectively the increased variability of the estimate, would even weaken the results. Additionally, authors will likely have to preregister their research, which would make this type of behaviour more difficult.

Summarising, the general framework outlined in the present paper is a valuable extension of the methodological toolkit in experimental economic research. It can be applied in many experimental settings using existing, open-source software and has the potential to foster the efficient usage of the resources available to the researcher.

Appendix: Multiple hypotheses testing

In the following, we present an extended, more detailed version of Section 2.1. As mentioned in the introduction, one way to increase the power to find *some* result of interest is simply to test more hypotheses in the same study. However, unless a proper multiple testing procedure is used, increasing the number of hypotheses tested typically leads to a larger Type I error probability. This is problematic since a goal of many designs is to keep this error under strict control at a certain prespecified level. The phenomenon can be illustrated by the following simple example. Suppose that m different null hypotheses are tested based on independent test statistics, and assume that the tests used have been chosen so as to make the individual (i.e., the nominal) Type I error probabilities equal to α. This means that the marginal probability to *not* reject a specific null hypotheses equals $1 - \alpha$, given that it is true and there is no non-zero effect. By the independence assumption, if all null hypotheses are true, then it follows that the probability to reject at least one of them is

$1-(1-\alpha)^m$. Hence, as m increases, the probability to falsely reject at least one null hypotheses converges to certainty.

To what extent it is important to keep the Type I error probability under strict control, and, if so, which significance level to choose, varies across scientific disciplines and applications. In order to choose appropriate values of the Type I error probability and power, one must weigh the value of correctly finding a positive effect against the loss of falsely declaring an effect as positive when it is not. For example, in the regulated area of confirmatory clinical trials for licensing new medical treatments, a Type I error probability of 0.025 or 0.05 is typically required by the authorities (for one-sided and two-sided hypotheses, respectively). In fact, since typically two independent studies would be required to successfully demonstrate an effect, this means that the Type I error probability would actually be much smaller across the studies. If the effect is assumed to be small, this can lead to large and costly studies. Nevertheless, it is often easy to motivate the importance of keeping the Type I error probability this small, since incorrectly allowing non-working medicines on the market could potentially lead to severe negative health effects. Here, our interest does not lie in specifying certain Type I error probabilities as appropriate, or even arguing that they should always be controlled. We assume that this has been deemed appropriate, and aim to demonstrate the methods that, to a large extent, have been developed within the field of medical statistics.

Let $H_1, \dots, H_m$ denote a set of m null hypotheses of interest, m_0 of which are true. Before proceeding to discuss multiple testing procedures, note that there have been several different suggestions of how to generalise the concept of a Type I error probability for a single hypothesis to the case of multiple hypotheses. These are referred to as different *error rates* (see, e.g., Chapter 2 of Bretz et al., 2011). For example, the *per comparison error rate* is the expected proportion of falsely rejected hypotheses, and equals $m_0\alpha/m$ if each hypothesis is tested individually at level α. Here, we will only focus on controlling the *family-wise error rate* (FWER), defined as the probability of making at least one Type I error. This probability depends on the specific combination of hypotheses that are actually true, which of course is unknown when planning the experiment.

In order to give a detailed account of multiple testing procedures for controlling the FWER, we first need to consider the concept of intersection hypotheses. Given m individual hypotheses $H_1, \dots, H_m$ and a non-empty subset of indices $I \subseteq \{1, \dots, m\}$, the intersection hypothesis H_I is defined as

$$H_I = \bigcap_{i \in I} H_i, \quad I \subseteq \{1, \dots, m\}. \tag{6}$$

Local control of the FWER at level α for a specific intersection hypothesis H_I holds when the conditional probability to reject at least one hypothesis given H_I is at most α, i.e. when

$$\mathbb{P}_{H_I}\left(\text{Reject } H_i \text{ for some } i \in I\right) \leq \alpha. \tag{7}$$

Strong control of the FWER means that the inequality in Eq. (7) must hold for *all* possible non-empty subsets I. This more conservative requirement bounds the

probability of making at least one Type I error regardless of which hypotheses are true. It is often deemed appropriate when the number of hypotheses is relatively small, while the consequences of making an error may be severe.

Now, given that tests have been defined for the local control of a collection of intersection hypotheses, how can this be used to attain strong control of the FWER? One option is to employ the widely used *closed testing principle*. This principle states that if we reject an individual hypothesis H_i if and only if all intersection hypotheses H_I such that $i \in I$ are rejected at local level α, then the FWER is strongly controlled at level α (Marcus et al., 1976). Since this principle is completely general with regards to the specific form of the tests for the intersection hypotheses H_I, it follows that the choice of local tests will determine the properties of the overall procedure.

There are a number of different multiple testing procedures available for local control of the FWER. One of the most widely applicable is the well-known Bonferroni procedure, due to the fact that no distributional assumptions are required for the statistical model. In this procedure, it is first assumed that m individual nominal tests have been defined for the hypotheses $H_1, \dots, H_m$, based on ordinary t-tests. The Bonferroni procedure then rejects H_i at level α if the corresponding nominal test would have rejected H_i at level α/m. That this procedure locally controls the FWER for an intersection hypothesis H_I follows immediately from the Bonferroni inequality, since

$$\mathbb{P}_{H_I}\big(\text{Reject } H_i \text{ for some } i \in I\big) \leq \sum_{i \in I} \mathbb{P}_{H_I}\big(\text{Reject } H_i\big) \leq \frac{|I|\alpha}{m} \leq \alpha. \tag{8}$$

Here, $|I|$ denotes the set of true null hypotheses.

In the following, we will use the Dunnett procedure (Dunnett, 1955) for testing intersection hypotheses. This test is tailored to our specific situation, in which several different active treatments are compared to a common control. Specifically, it is assumed that each individual observation belongs to one of $m+1$ different treatment groups. The group with index $i = 0$ is the control group, against which the other groups are compared. The individual observations in each group are used to form group means $\bar{X}_i$. Due to the central limit theorem, we assume these to be normally distributed, with true means μ_i, $i = 0, \dots, m$ and variances σ^2/n, where σ^2 is assumed to be known and n is the group size (all assumed equal). We thus have

$$\bar{X}_i \sim \mathrm{N}\big(\mu_i, \sigma^2/n\big), \quad i = 0, 1, \dots, m. \tag{9}$$

After the experiment, these observations yield the estimates $\bar{X}_i - \bar{X}_0$ of the treatment mean differences relative to control. The actual testing is based on the standardised mean differences,

$$Z_i = \frac{\bar{X}_i - \bar{X}_0}{\sqrt{2\sigma^2/n}} \sim \mathrm{N}\left(\frac{\mu_i - \mu_0}{\sqrt{2\sigma^2/n}}, 1\right), \quad i = 1, \dots, m. \tag{10}$$

These assumptions lead to a certain covariance structure for the joint distribution of the statistics $Z_1, \dots, Z_m$ which is exploited by the Dunnett test. The individual null

hypotheses are $H_i : \mu_i - \mu_0 = 0$. For a given intersection hypothesis H_I, the test statistic is defined as

$$Z_I^{\max} = \max_{i \in I} Z_i. \tag{11}$$

Given that a specific value $Z_I^{\max} = z$ has been observed, a corresponding p-value may be computed as $p_I = 1 - F_I^{\max}(z)$ (Friede & Stallard, 2008), where

$$F_I^{\max}(z) = 1 - \int_{-\infty}^{\infty} \left[\Phi\left(\sqrt{2}z + x \right) \right]^{|I|} \phi(x)\, \mathrm{d}x \tag{12}$$

is the cumulative distribution function of $Z_I^{\max}$, ϕ the standard normal density and Φ the standard normal distribution function. The intersection hypothesis H_I is rejected at level α if $p_I \leq \alpha$. Due to the closed testing principle, the individual hypothesis H_i may be rejected while strongly controlling the FWER if $p_I \leq \alpha$ for all index sets I such that $i \in I$.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10683-022-09773-8.

Acknowledgements The authors thank the editor and two anonymous reviewers for comments that helped to substantially improve the paper.

Funding Open access funding provided by University of Natural Resources and Life Sciences Vienna (BOKU).

Code and data availability Replication material for the study is available at https://osf.io/bk5zy/. The data from Karlan and List (2007) can be found at https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi%3A10.7910/DVN/27853 (Karlan and List, 2014). The data from Musshoff and Hirschauer (2014) was shared by the authors of the study and is included in the replication material.

Declarations

Conflicts of interest The authors declare that they have no conflict of interest.

References

Armitage, P., McPherson, C. K., & Rowe, B. C. (1969). Repeated significance tests on accumulating data. *Journal of the Royal Statistical Society: Series A (General), 132*(2), 235–244.

Bauer, P. (1989). Multistage testing with adaptive designs. *Biometrie und Informatik in Medizin und Biologie, 20,* 130–148.
Bauer, P., Bretz, F., Dragalin, V., König, F., & Wassmer, G. (2016). Twenty-five years of confirmatory adaptive designs: Opportunities and pitfalls. *Statistics in Medicine, 35*(3), 325–347.
Bauer, P., & Köhne, K. (1994). Evaluation of experiments with adaptive interim analyses. *Biometrics, 50*(4), 1029–1041.
Bellemare, C., Bissonnette, L., & Kröger, S. (2016). Simulating power of economic experiments: The powerBBK package. *Journal of the Economic Science Association, 2*(2), 157–168.
Bhat, N., Farias, V. F., Moallemi, C. C., & Sinha, D. (2020). Near-optimal a-b testing. *Management Science, 66*(10), 4359–4919.
Bretz, F., Hothorn, T., & Westfall, P. (2011). *Multiple comparisons using R*. Chapman & Hall/CRC, 6000 Broken Sound Parkway NW, Suite 300, first edition.
Canavari, M., Drichoutis, A. C., Lusk, J. L., & Nayga, R. M. (2019). How to run an experimental auction: A review of recent advances. *European Review of Agricultural Economics, 46*(5), 862–922.
Chataway, J., Nicholas, R., Todd, S., Miller, D. H., Parsons, N., Valdés-Márquez, E., et al. (2011). A novel adaptive design strategy increases the efficiency of clinical trials in secondary progressive multiple sclerosis. *Multiple Sclerosis Journal, 17*(1), 81–88.
Czibor, E., Jimenez-Gomez, D., & List, J. A. (2019). The dozen things experimental economists should do (more of). *Southern Economic Journal, 86*(2), 371–432.
De Long, J. B., & Lang, K. (1992). Are all economic hypotheses false? *Journal of Political Economy, 100*(6), 1257–1272.
Dunnett, C. W. (1955). A multiple comparison procedure for comparing several treatments with a control. *Journal of American Statistical Association, 50*(272), 1096–1121.
Friede, T., & Henderson, R. (2009). Exploring changes in treatment effects across design stages in adaptive trials. *Pharmaceutical Statistics, 8*(1), 62–72.
Friede, T., Parsons, N., Stallard, N., Todd, S., Valdes Marquez, E., Chataway, J., & Nicholas, R. (2011). Designing a seamless phase II/III clinical trial using early outcomes for treatment selection: An application in multiple sclerosis. *Statistics in Medicine, 30*(13), 1528–1540.
Friede, T., & Stallard, N. (2008). A comparison of methods for adaptive treatment selection. *Biometrical Journal, 50*(5), 767–781.
Friede, T., Stallard, N., & 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.
Hadad, V., Hirshberg, D.A., Zhan, R., Wager, S., & Athey, S. (2021). Confidence intervals for policy evaluation in adaptive experiments. In *Proceedings of the national academy of sciences*, 118(15). Publisher: National Academy of Sciences Section: Physical Sciences.
Harrison, G. W., & List, J. A. (2004). Field experiments. *Journal of Economic Literature, 42*(4), 1009–1055.
Ioannidis, J. P. A., Stanley, T. D., & Doucouliagos, H. (2017). The power of bias in economics research. *The Economic Journal, 127*(605), F236–F265.
Jenkins, M., Stone, A., & Jennison, C. (2011). An adaptive seamless phase ii/iii design for oncology trials with subpopulation selection using survival endpoints. *Pharmaceutical Statistics, 10*(4), 347–356.
Jennison, C., & Turnbull, B. W. (2000). *Group sequential methods with application to clinical trials*. Boca Raton: Chapman & Hall/CRC.
Karlan, D., & 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.
Karlan, D. & List, J.A. (2014). Does price matter in charitable giving? Evidence from a large-scale natural field experiment. Type: dataset.
Kasy, M., & Sautmann, A. (2021). Adaptive treatment assignment in experiments for policy choice. *Econometrica, 89*(1), 113–132.
Lehmacher, W., & Wassmer, G. (1999). Adaptive sample size calculations in group sequential trials. *Biometrics, 55*(4), 1286–1290.
List, J. A., Sadoff, S., & Wagner, M. (2011). So you want to run an experiment, now what? Some simple rules of thumb for optimal experimental design. *Experimental Economics, 14*(4), 439–457.
List, J. A., Shaikh, A. M., & Xu, Y. (2019). Multiple hypothesis testing in experimental economics. *Experimental Economics, 22*(4), 773–793.
Marcus, R., Peretz, E., & Gabriel, K. R. (1976). On closed testing procedures with special reference to ordered analysis of variance. *Biometrika, 63*(3), 655–660.

Musshoff, O., & 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.

Müller, H., & Schäfer, H. (2004). A general statistical principle for changing a design any time during the course of a trial. *Statistics in Medicine, 23*(16), 2497–2508.

O'Brien, P. C., & Fleming, T. R. (1979). A multiple testing procedure for clinical trials. *Biometrics, 35*(3), 549–556.

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251):aac4716–aac4716.

Pallmann, P., Bedding, A.W., Choodari-Oskooei, B., Dimairo, M., Flight, L., Hampson, L.V., Holmes, J., Mander, A.P., Odondi, L., Sydes, M.R., Villar, S.S., Wason, J. M.S., Weir, C.J., Wheeler, G.M., Yap, C., & Jaki, T. (2018). Adaptive designs in clinical trials: why use them, and how to run and report them. *BMC Medicine*, *16*(29).

Parsons, N., Friede, T., Todd, S., Marquez, E. V., & 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.

Pashler, H., & Wagenmakers, E. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science, 7*(6), 528–530.

Placzek, M. & Friede, T. (2022). Blinded sample size recalculation in adaptive enrichment designs. *Biometrical Journal*, forthcoming.

Pocock, S. J. (1977). Group sequential methods in the design and analysis of clinical trials. *Biometrika, 64*(2), 191–199.

Posch, M., Koenig, F., Branson, M., Brannath, W., Dunger-Baldauf, C., & Bauer, P. (2005). Testing and estimation in flexible group sequential designs with adaptive treatment selection. *Statistics in Medicine, 24*(24), 3697–3714.

Proschan, M. A., & Hunsberger, S. A. (1995). Designed extension of studies based on conditional power. *Biometrics, 51*(4), 1315–1324.

Reed Johnson, F., Lancsar, E., Marshall, D., Kilambi, V., Mühlbacher, A., Regier, D. A., et al. (2013). Constructing experimental designs for discrete-choice experiments: Report of the ISPOR conjoint analysis experimental design good research practices task force. *Value in Health, 16*(1), 3–13.

Richter, J., Friede, T., & Rahnenführer, J. (2022). Improving adaptive seamless designs through Bayesian optimization. *Biometrical Journal, 64*(5), 948–963.

Thompson, B. S., & Webb, M. D. (2019). A simple, graphical approach to comparing multiple treatments. *The Econometrics Journal, 22*(2), 188–205.

Wassmer, G. & Brannath, W. (2016). *Group sequential and confirmatory adaptive designs in clinical trials*. Springer International, first edition.

Xiong, R., Athey, S., Bayati, M., & Imbens, G. (2022). Optimal experimental design for staggered rollouts. arXiv: 1911.03764 [econ, stat]

Ziliak, S. T., & McCloskey, D. N. (2004). Size matters: The standard error of regressions in the American Economic Review. *The Journal of Socio-Economics, 33*(5), 527–546.