
**IFPRI Discussion Paper 01273**

**June 2013**

# The Logic of Adaptive Sequential Experimentation in Policy Design

**Haipeng Xing**

**Xiaobo Zhang**

**Development Strategy and Governance Division**

## INTERNATIONAL FOOD POLICY RESEARCH INSTITUTE

The International Food Policy Research Institute (IFPRI), established in 1975, provides evidence-based policy solutions to sustainably end hunger and malnutrition and reduce poverty. The Institute conducts research, communicates results, optimizes partnerships, and builds capacity to ensure sustainable food production, promote healthy food systems, improve markets and trade, transform agriculture, build resilience, and strengthen institutions and governance. Gender is considered in all of the Institutes work. IFPRI collaborates with partners around the world, including development implementers, public institutions, the private sector, and farmers organizations, to ensure that local, national, regional, and global food policies are based on evidence. IFPRI is a member of the CGIAR Consortium.

## AUTHORS

**Haipeng Xing, State University of New York at Stony Brook**
Associate Professor, Economics

**Xiaobo Zhang, International Food Policy Research Institute**
Senior Research Fellow, Development Strategy and Governance Division
z.zhang@cgiar.org
**National School of Development, Peking University, China**
Professor, Economics

## Contents

## ACKNOWLEDGMENTS

Haipeng Xing thanks the World Bank for financial support on this research. Haipeng Xing also gives special thanks to Professor Yifu Lin for discussing various phenomena in China's economic development and ideas in new structural economics, which provide the economic basis for the statistical model developed in this paper. The authors are also grateful for the helpful comments from Chih Ming Tan, Junfu Zhang, and seminar participants at Clark University, Peking University, and the China Meeting of Econometric Society in 2013.

## ABSTRACT

Inspired by the wide adoption of rigorous randomized controlled trials (RCTs) in medical research, economists and other social scientists have increasingly used RCTs in their research. As researchers pick up projects amenable to the RCT methodology, they likely leave out important questions to which RCTs cannot be directly applied. As a result, RCTs have been criticized for the proclivity of addressing trivial questions. As a matter of fact, in medical research RCTs are an integral part of adaptive sequential experiment design—a few steps must be taken to screen out drugs that have toxins and strong side effects before running any RCTs on humans. In this paper, we argue that economists can learn a great deal from the design principles implemented in medical research. We develop a theoretical model to show the logic of adaptive sequential experiment design in the presence of uncertainty over negative effects and discuss how to choose samples in a population to minimize the experiment cost. We also point out the applications of our proposed framework in the economic domain, such as economic reforms and new product design.



**Keywords: Chinese economy, experimentation, randomized controlled trials, reform**

# 1. INTRODUCTION

It has been increasingly recognized that evidence-based research plays an important role in policymaking (Ravallion 2009). However, there is considerable debate as to what constitutes good evidence. Inspired by the wide adoption of rigorous randomized controlled trials (RCTs) in various science fields, in the past couple of decades the application of RCTs in economic research has blossomed. To a significant extent, RCTs have revitalized the way of conducting empirical research in economics (Duflo, Glennerster, and Kremer 2008). Evidence based on RCTs is even regarded as the "gold standard"(Banerjee 2008). However, as the adoption of RCTs spreads, some of their flaws have been also exposed. For example, a lack of external validity (results found in one context may not be applicable elsewhere) has been voiced as a major concern (Ravallion 2009; Deaton 2010; Rodrik 2009). In reality, issues of general equilibrium and political economics pose challenges for external validity (Acemoglu 2010). Without demonstrating external validity, it would be difficult to scale up successful experiments shown in RCTs. Some other weaknesses have also been mentioned in the literature, including but not limited to failing to capture general equilibrium effects, neglect of heterogeneity, the nonrandomness of selecting projects for RCTs, and ethical constraints (Barrett and Carter 2010; Deaton 2010; Ravallion 2012).

In any scientific field, "no harm" to the experiment subjects is a key principle for the application of RCTs. For example, in medicine field drugs with a toxin are not allowed in human clinical trials. Testing for toxins is an integral part of drug development and must be conducted prior to human clinical trials. Economists have implicitly followed the same principle. They normally select RCT projects, such as bed nets in Africa, that are known to have no major deleterious effect on the survey subjects (Cohen and Dupas 2010). Of course, many economists have implicitly screened their RCT projects through field observations or qualitative investigations, but unlike scientists writing in the medical literature, they rarely explicitly discuss the initial screening steps.[1]Because of the strong bias in favor of publishing papers based on RCTs, researchers tend to select those projects that are suitable for RCTs, that is, posing no obvious harm to the survey subjects. Consequently, many of the chosen projects tend to focus on "small questions" that may not always be of interest to policymakers (Rodrik 2009; Lin 2011; Ravallion 2012).

We fully recognize the merits of using RCTs in economics research. However, as shown in the medical literature that our economists aim to emulate, RCTs are just one integral part of sequential experiment design.[2] For example, drug development encompasses several steps—pathological analysis, toxin testing, animal trials, and clinical trials on humans. Even for human trials, many drugs initially are tested only on terminally ill patients. Only those drugs that have been shown to have no harmful effects on humans are allowed into RCTs on human beings. In other words, RCTs are just one step of the scientific discovery process. If the focus is only on the RCT and ignores other necessary steps in experiment design, our profession would likely run the risk of spending limited resources on relatively trivial questions that can be safely randomized, at the expense of more important and policy-relevant questions (Rodrik 2009).

For many proposed economic policies, policymakers face considerable economic and political uncertainty. When facing choices never seen before, it is extremely risky for agents to make radical decisions before seeing solid evidence. Learning by experimentation is a key strategy to uncover the actual payoffs and costs associated with proposed changes. In reality, experimentation does not necessarily start with RCTs from the very beginning because of the uncertainty over potential failures and resultant negative spillover effects. Furthermore, RCTs measure the average effect on the treatment groups compared with the control group, ignoring heterogeneity and the distributional effect (Deaton 2010), while policymakers, apprehensive of compromising stability, are much more concerned about potential adverse distributional effect for some segment of the population (Kanbur 2001). Instead of solely using a RCT,

[1]As a matter of fact, Banerjee and Duflo, two of the most influential advocates of applying RCTs in economics research, rely heavily on observations and field interviews to identify RCT projects, as narrated in their book, *Poor Economics* (Banerjee and Duflo 2011). However, these rich stories vanish in the published economics papers.

[2]There is also strong debate as to what constitutes the best evidence in medicine (Worrall 2007). No matter whether RCTs are used or not, no harm is a general principle underlying experiment designs in medicine.

therefore, it is more sensible for policymakers to adopt a sequential experiment design approach, following the example of medical literature.

Our paper is also related to the extensive literature on experimentation in the areas of transition economics and fiscal federalism. When there is significant uncertainty about the benefits and costs of a proposed policy reform, small-scale incremental experimentation has been shown to be an effective way to reveal information and potentially convince skeptics to adopt the reform measures, especially in economies with an M-form organizational structure (many similar, self-contained, sublevel governments) (Qian, Roland, and Xu 2006). Experimentation is also a key feature of fiscal federalism. Fiscal decentralization enables interjurisdictional competition, which induces local governments to experiment with new policies on a small scale. The information generated from the experiments brings about a great deal of externality (Oates 1999; Besley and Case 1995). In contrast, the spillover effect is normally not taken into account in RCTs (Ravallion 2012).

In this paper, we first elaborate the idea of adaptive sequential experiment design using three concrete examples—drug development, China's economic reforms, and industrial product design. Next, we develop a conceptual model to demonstrate the logic of adaptive sequential experiments. The paper ends with conclusions.

# 2. APPLICATIONS OF ADAPTIVE SEQUENTIAL EXPERIMENTATION

## Drug Development

In the United States, drug regulations have evolved largely in response to crises. One milestone is the accidental deaths of 107 people in 1939 from taking the medicine "Elixir sulfanilamide" (Routledge 1998). At the time, drug manufacturers were not required to conduct toxicity testing. A well-intentioned chemist mixed diethylene glycol with sulfanilamide to make a liquid formulation, unaware of the adverse effects of diethylene glycol. In responding to the fatalities, Congress enacted the Federal Food, Drug, and Cosmetic Act of 1938, empowering the US Food and Drug Administration (FDA) to regulate and oversee the process of developing drugs. One key objective of the FDA is to ensure drug safety and prevent similar incidents from happening again.

Under the current FDA regulations, drug development encompasses two main steps—the preclinical phase and the clinical phase (Lipsky and Sharp 2001). In the preclinical phase, the first task is to identify a promising new chemical entity (called the *candidate compound*) based on scientific advances in understanding a disease. The next step is to undertake toxicology and safety studies for the identified compound using experimental animals prior to use in humans. For most drug candidates, the FDA requires tests on at least two laboratory animal species, rodent (rat or mouse) and nonrodent (rabbit, dog, or monkey), to identify the observable signs of toxicity and determine safe dose levels. Even the animal tests follow a step-by-step approach: "The group sizes for the early range-finding studies may consist of only a few animals and one sex (one animal per dose level). Once a suitable dose range is identified, group sizes are increased to at least three per sex per dose level to allow statistical comparison" (Steinmetz and Spack 2009, 5).

After the compound passes the toxin test on animals, researchers can gather the preclinical testing information and submit an investigational new drug (IND) application to regulatory authorities (the FDA in the United States). If the application is approved, drug development can move to the clinical phase. On average, only about four percent of the compounds tested on animals qualify for human tests (Stratmann 2010).

The clinical phase is further divided into three or four subphases. Phase I tests human toleration limits and safe dose levels, normally based on use in a small group of healthy volunteers (Friedman, Furberg, and DeMets 2010). Testing follows a cautious procedure, starting with very low doses, which are gradually increased, as described by Friedman, Furberg, and DeMets: "In estimating the maximally tolerate dose, the investigator usually starts with a very low dose and escalates the dose until a prespecified level of toxicity is obtained. Typically, a small number of participants, usually three, are entered sequentially at a particular dose. If no specified level of toxicity is observed, the next predefined higher dose level is used. If unacceptable toxicity is observed in any of the three participants, an additional number of participants, usually three, are treated at the same dose. If an additional unacceptable toxicity is observed, then the dose escalation is terminated and that dose, or perhaps the previous dose, is declared to be the maximally tolerated dose" (2010, 5). Typically, two-thirds of tested compounds are safe enough to enter the next phase (Lipsky and Sharp 2001).

Phase II determines a drug's efficacy and measures its side effects. To avoid potential unknown harmful effects on a healthy population, this phase tends to recruit a small group of patients (such as terminally ill cancer patients) that the drug is intended to treat. Within Phase II, a two-stage design is frequently practiced (Friedman, Furberg, and DeMets 2010). In the first step, investigators aim to eliminate drugs that have a harmful effect or show little or no biologic activity. For example, if the toxin level exceeds a certain prespecified threshold or the drug does not show any activity in more than a certain predefined proportion of participants, the experiment will be stopped. Otherwise, more participants will be added to obtain a better estimate of the response rate in the second stage. It takes about four to five years to finish Phase II testing. In the process, some drugs are weeded out because of ineffectiveness or unacceptable side effects. In the end, only about one-third of INDs survive this phase.

Phase III is the step to demonstrate effectiveness, determine the best dosage, and further check safety. The FDA requires randomized controlled trials (RCTs) on a larger human population in this phase. This step is more time consuming than the first two clinical phases, normally lasting six to eight years. On average, only about 27 percent of INDs eventually pass this stage and receive FDA approval (Lipsky and Sharp 2001).

Drugs frequently pass the first two phases of clinical trials but falter at the third stage (FierceBiotech 2012; Arrowsmith 2011). One example of third-phase failure is the case of Dimebon, an Alzheimer drug that was originally developed as an antihistamine. Initial tests were conducted using laboratory rats and, later, a pilot study of 14 Alzheimer's patients. Positive results from these initial trials were published and received attention from both researchers and investors. However, because most of the third-phase trials showed no significant differences between treatment and control groups, the major investor, Pfizer, pulled out and declared the experiments a failure (Carroll 2012).

Even well-designed RCTs still have the potential for flaws. For example, the human subjects used in the drug tests may be different from the general population in the real world. There is also the possibility that the effectiveness shown in Phase III cannot be externally validated in practice. In addition, some drugs could have long-term side effects, which are not necessarily discoverable in Phase III testing. The FDA uses two methods to remedy these problems. First, sometimes the FDA requests that a sponsor conduct a Phase IV test on a different population so as to verify the validity externally. Second, the FDA has set up a hotline (1-800-FDA-1088) to keep track of serious adverse reactions related to the use of new drugs even after they are approved and released. Drug manufacturers are required to report side effects every quarter for three years after a drug is approved.

It is worth mentioning that if safety is a concern, both Phase I and Phase II can be based on a small number of subjects and need not be randomized and controlled (Karlberg and Speers 2010). The drugs tested in clinical Phase III are largely known to pose no major harms after passing the scrutiny of the first two clinical phases. As a result, RCTs can be run on a larger population in Phase III (and Phase IV if called for).

Looking at the whole drug development process, it is clear that the FDA follows the principle of "safety first". RCTs are just one of the several approaches used in drug development. Friedman, Furberg, and DeMets summarized the basic requirement for conducting RCTs in the following paragraph: "Before conducting a trial, an investigator needs to have the necessary knowledge and tools. He must know something about the safety of the intervention and what outcomes to assess and have the techniques to do so. Well-run clinical trials of adequate magnitude are costly and should be done only when preliminary evidence of the efficacy of an intervention looks promising enough to warrant the effort and expense involved" (2010, 11).

## China's Economic Reforms

Pragmatism, trial and error, evidence-based policymaking, and experimentation with small-scale policy reforms that are later scaled up are all defining features of China's reforms in the past several decades. The course of China's rural reform clearly illustrates this point. At the end of the Cultural Revolution (1966-1976), China faced serious food shortages, largely due to the collective farming system featured in the era of the planned economy. To avert potential food shortages, in 1978 Xiaogang Village in Anhui Province contracted collective land to farmers, considerably boosting crop yields and farmers' income. After hearing about this success, researchers at the Research Center for Rural Development (RCRD) at the state council paid a visit to the village, evaluated the practice, and proposed to scale it up nationwide.

However, when they first proposed the household responsibility system (HRS) reform based on the Anhui experience at a conference with seven major agricultural provinces (including Anhui Province) organized by the National Agricultural Commission, five out of the seven provinces opposed it. At the time, public ownership had been in place for more than two decades, and many policymakers were used to the order of the collective farming system and concerned about the potential chaos stemming from this

reform. More importantly, the HRS seemed to forsake the socialist principles embedded in the minds of most officials.

Facing the impossibility of accomplishing the reform in one fell swoop, Du Runsheng, the head of RCRD, came up with an ingenious idea and submitted it to Deng Xiaoping, China's supreme leader at the time. Du proposed to conduct a trial of HRS in a few impoverished mountainous regions, based on the fact that these regions were already facing a shortfall of food grains and posed a heavy burden on the state; hence, if the trials failed, the impact would be confined to these limited regions. After hearing the proposal, Deng made the following remarks: "Hardship regions are allowed to carry out the HRS. If it turns out to be mistaken and they come back in, it is nothing special. Rich regions that have enough to eat do not need to start right away" (Du 2010, 18).

Following Deng's instructions, different forms of the agricultural production responsibility system were allowed as experiments in different regions. Impoverished areas carried out the full HRS; developed coastal regions could keep the collective production modes but with specialized contracts linking wages to output. Intermediate regions could freely choose. After one year the test results came out, overwhelmingly showing that the HRS was more effective than other responsibility systems. The impoverished areas running the trials had enough food to eat and no longer relied on the central government for food grain subsidies. The compelling evidence easily convinced most decisionmakers in the government, and the HRS was fully rolled out nationwide just a few years later during the 1980s. The success of the rural reform laid a foundation for subsequent rapid economic growth and the most massive poverty reduction in human history (Lin 1992).

Not only has the rural reform followed a step-by-step experimental approach, but other major economic reforms in China have followed suit. The creation of the Shenzhen Special Economic Zone is another telling example. Shenzhen was a very small town by the border between Hong Kong and mainland China with a population of 30,000 people in the late 1970s. In 1979, Yuan Geng, director of the China Investment Promotion Bureau in Hong Kong, proposed to setting up a special industrial zone in the Shekou area of Shenzhen as a pilot for market reforms, taking advantage of the proximity to Hong Kong. The state council quickly approved the proposal, earmarking 2.14 square kilometers for the zone and granting it special rights to test the applicability of the market economy in the zone. The industrial zone turned out to be an instant success. The investment from Hong Kong quickly filled in the limited land, generating hundreds of thousands of jobs.

After observing this success, in 1980 the central government established a larger Shenzhen Special Economic Zone, which encompasses 1,953 square kilometers, to carry out full-fledged market reforms on a larger scale. This is the first citywide special economic zone in China. For fear of any negative spillover to other regions, all the reforms were confined to the special economic zone. Initially even Chinese citizens had to apply for a special travel document to enter Shenzhen. In 1980, Shenzhen's gross domestic product (GDP) was only 0.3 percent of that in Hong Kong. Due to the special policy of opening up and reform, its GDP is now nearing 70 percent of Hong Kong's and is projected to surpass that of Hong Kong in 10 years. At this time, Shenzhen boasts being one of the richest cities in China with a population of more than 13 million. The Shenzhen experience illustrates that capitalism is not as dangerous as was taught during the era of the planned economy, effectively erasing the ideological taboo about capitalism.

Following the success of Shenzhen, 3 additional city-level special economic zones (Zhuhai, Shantou, and Gongbei) were established in the next few years. In 1984 China further opened up 14 more coastal cities (Tianjin, Shanghai, Dalian, Qinhuangdao, Yantai, Qingdao, Lianyungang, Nantong, Ningbo, Wenzhou, Fuzhou, Guangzhou, Zhanjiang, and Beihai). Although these "opened-up" cities did not enjoy the full privileges of special economic zones, they still received considerable discretionary powers in attracting foreign direct investment and exploring market reforms. In 1990, the central government designated Shanghai Pudong as a special economic and technological development zone, allowing it to conduct various economic reforms. In 2005, Pudong was further classified as a pilot area for integrated reforms (beyond just economic reforms). The development process of special zones illustrates the step-by-step experimental approach commonly seen in Chinese reforms.

Experiments yield information to help policymakers understand what works and what does not. Thus, even failures can be helpful because they can lead to the elimination of unfavorable options. Still, a large-scale mistake may be irreversible and therefore may undermine the credibility and stability of the political leadership, thus weakening overall learning capacity. The invention of the dual-track price reform (allowing state-owned enterprises to sell their unused quota of raw materials to town and village enterprises at market prices) provides a good example on this point.[3] After the success of rural reform, price reform became more urgent. There were two schools of thought regarding price reform. One school was in favor of a big-bang type of price reform, instantly liberalizing planned prices to market prices. Another school proposed to improve the determination of prices through better planning. Facing uncertainty, the RCRD sent a few young researchers to conduct a pilot on radical price reform in Hebei Province. Luo Xiaopeng was one of the researchers sent to the field. However, the experiment of course was purposely not made known to the outside at the time. Luo (2010) later reported that the failure of this laissez-faire price reform experiment helped him come up with the idea of dual-track price reform.

Such experimentation has been particularly important in overcoming several major obstacles to effective reform in China, related to its size, its diversity, and the history and hierarchical structure of its political system. For a large and diverse economy like China's, it is very difficult to derive a single one-size-fits-all blueprint for reform simply by applying textbook economic theories. Instead, trial-and-error processes can help researchers discover local best practices. Moreover, the basis for formulating sound market-oriented policies in 1978 was limited. Few bureaucrats had any formal training in orthodox economics, nor even substantial experience of living in a market economy. Chinese reformers therefore felt compelled to use experimentation as a collective learning mechanism.

In fact, the decentralized experiment of "proceeding from point to surface" (*you dian dao mian*) is a pervasive feature in China's economic transformation, dating back even to the Chinese Communist revolutionary era (Heilmann 2008).

### Industrial Product Design

Similar to policy experimentation, strategic experiments in business allow companies to test business decisions, including adding new products or altering existing products, before full-scale implementation. A step-by-step, trial-and-error approach allows for more frequent updating of prior assumptions to align with current results. Imperfect information and fear of potential failures motivate the need for a trial-and-error approach to learning in business (Govindarajan and Trimble 2004). Experiments allow companies to test consumer preference, competitors' reactions, profitability, and feasibility of full-scale implementation (Anthony 2009). The goal of these experiments is to provide a low-cost method of testing these elements, thereby reducing risks associated with full-scale implementation. For practical reasons, strategic experimentation in business does not follow a strict RCT-type experimental approach (Govindarajan and Trimble 2004).

One example of innovative product design is the case of Proctor & Gambles development of Align, an over-the-counter supplement that had the potential to treat irritable bowel syndrome (IBS), a condition that restricts a person's food choices and behavior patterns (Anthony 2010). Although the product showed great promise, the corporate leadership was concerned about several potential risks. First, although IBS was a relatively common condition, no current market existed for the new product. Second, developing a new brand is expensive. Third, the projected returns seemed to be so small as to make those among the leadership skeptical about embracing this new product.

To uncover the potential payoffs and risks of this new product, the company engaged in a series of low-cost, information-rich experiments. After initial analysis to develop assumptions on which a full-scale launch would rely, pilot tests were conducted. In the pilot, Internet marketing strategies were used to gauge customer responses and refine product and marketing design. One positive result of these tests was the manner in which the product was ultimately sold. Instead of being packaged in a generic bottle, the

[3] See Lau, Qian and Roland (2000) for details about the logic of dual-track reform in China.

pills were packaged individually, much like chewing gum, and had markings that corresponded to days of the week, offering customers a helpful way to remember to take the medicine. Based on these experimental steps, Align was launched nationally and shows promise of success.

Thomke (2003) described one example that highlights the role of sequential experimentation in product development, the case of design changes in Bank of America branches. Because of the bank's extremely large volume of transactions every day, senior management was very concerned about the negative impact on customers of potential glitches associated with national rollouts of untested systemwide new designs of branches. To reduce the risk of large-scale failure, Bank of America used an experimental, step-by-step approach, beginning collecting ideas from employees at various levels about how to innovate the service experience. The goal in this step was to collect as many as ideas as possible. Only a small number of ideas were selected for experimentation.

Next, after choosing the best ideas for experimentation, the bank set up a "prototype center" in Charlotte, North Carolina, designed to mimic a real branch. Some staff members were invited to rehearse as customers, interacting with actual bank hosts in the newly designed environment. The purpose of the rehearsal in the prototype center was to filter ideas.

After an idea passed the rehearsal process, the bank launched it as an experiment in some of the 25 innovative market branches (the bank's "living laboratory") in Atlanta. The experiments normally ran for 90 days. However, if customers liked an idea, the experiment might become permanent practice in the branch after the trial period. For example, in one experiment, the research team used two approaches to reducing noise in a branch. First, they frequently ran an experiment in multiple branches to average out individual noises. Second, the team compared the performance between the experimental branches and similar branches running under normal conditions in the same city.

Based on the results of the test, the final step produced recommendations for scaled-up concept implementation. This kind of experimental approach is used not only in the financial sector but also in many other sectors, such as automobile and yacht design, as described by Thomke (2003). Because it is too costly to build a new prototype car for crash tests or a yacht for tank and tunnel tests, companies have adopted a sequential experimental approach to reduce the cost. Computer-aided design (CAD) is widely used to simulate the inner workings and safety of a new product. After the CAD stage, a scaled-down prototype is often built for real tests (crash tests for cars, tank and tunnel tests for yachts). A full-size prototype won't be built until the small-scale prototype passes the necessary tests. Through this process, bad ideas can be identified and ruled out earlier, avoiding larger, more wasteful mistakes in the later stage.

In short, this kind of small-scale, sequential experimental process provides innovative recommendations for implementation on a large scale at a minimum of cost and risk to current business practices.

# 3. ADAPTIVE DESIGN IN SEQUENTIAL EXPERIMENTS

## The Logic of Sequential Decision Theory

The spirit of sequential experimentation is embedded in human nature and dates back to thousands of years ago. "Perhaps the earliest proponent was Noah, who on successive days released a dove from the Ark in order to test for the presence of dry land during the subsidence of the Flood" (Jennison and Turnbull 2000, 4). This spirit was first conceptualized by statisticians during 1920s and 1930s to tackle various sequential decisionmaking problems, such as sequential sampling inspection procedures (Dodge and Romig 1929); quality control charts (Shewhart 1931); two-stage experiment design (Thompson 1933); and multistep, large-scale survey sampling (Maha 1940).

In response to the need for efficient testing of antiaircraft gunnery during the World War II, Abraham Wald, the founder of sequential analysis, and his collaborators developed the *sequential probability ratio test* (SPRT) in 1943, which reduces the number of samples without sacrificing the reliability of the terminal decisions (Wald 1947). Let $X_1, X_2, \ldots$ be independent and identically distributed observations successively sampled from a common distribution $P_\theta$ or density function $f_\theta(x)$, where $\theta$ is an element of the parameter space $\Theta$ and can be considered as the state of nature governing the outcome of a process. To test the null hypothesis, $H_0 : \theta = \theta_0$, versus the alternative, $H_1 : \theta = \theta_1$ $(\theta_1 \neq \theta_0)$, the SPRT stops sampling at stage

$$T = \inf\{t \geq 1 : R_t \geq a \text{ or } R_t \leq b\}, \tag{1}$$

where $R_t = \prod_{i=1}^{t} \left[ f_{\theta_1}(X_i)/f_{\theta_0}(X_i) \right]$ is the likelihood ratio, and $(b,a)$ are stopping boundaries $(a > 1 > b > 0)$. When $R_n$ goes beyond the stopping boundaries, stopping occurs and $H_0$ (or $H_1$) is accepted if $R_t \leq b$ (or $R_t \geq a$). If $R_t \in (a,b)$, the sampling procedure should be continued by observing $X_{t+1}$. The choice of $a$ and $b$ are determined by the error probabilities $\alpha = \text{Prob}(R_T \geq a | H_0 \text{ is true})$ and $\beta = \text{Prob}(R_T \leq b | H_1 \text{ is true})$. Wald and Wolfowitz (1948) showed that the SPRT minimizes the expectations of $T$ under both $H_0$ and $H_1$ among all tests in which the number of samples has a finite expectation under $H_0$ and $H_1$ and whose error probabilities satisfy

$$\text{Prob}(\text{Reject } H_0 | H_0 \text{ is true}) \leq \alpha \tag{2}$$

and

$$\text{Prob}(\text{Reject } H_1 | H_1 \text{ is true}) \leq \beta. \tag{3}$$

The optimality of the SPRT is actually closely related to the discussion on optimal solutions of sequential decisionmaking problems in Arrow, Blackwell, and Girshick (1949). In this discussion, an experimenter observes a sequence of random variables $X_1, X_2, \ldots$ from the distribution $P_\theta$, when $\theta$ is the true parameter or the state of nature, and is required to choose some action $a$ from an action space $\mathscr{A}$ consisting of all available actions to be chosen. He also incurs a loss function $L(\theta, a)$, representing the loss $L(\theta, a)$ when $\theta$ is the parameter value and action $a$ is chosen. However, he doesn't need to choose an action immediately. Instead, he may decide to select a subset of sequence $\{X_i\}$ to obtain partial information about $\theta$ so that a wiser selection of action can be made. Let $\mathbf{X}_t = (X_1, \ldots, X_t)$. A sequential procedure $T \in \mathscr{T}$ is a sequence of disjunct sets $\mathscr{X}_0, \mathscr{X}_1, \mathscr{X}_2, \ldots \mathscr{X}_t, \ldots$, where $\mathscr{X}_t = \{\mathbf{X}_t\}$ represents the sampling procedure that terminates with the observation $\mathbf{X}_t$. By the definition of the sequential procedure, we require that $\sum_{t=0}^{\infty} \text{Prob}(\mathscr{X}_t) = 1$. Let $a_t(\mathbf{X}_t)$ denote an action function of observing $X_1, \ldots, X_t$. A *sequential decision rule* $d_t$ is a function $d_t : \mathscr{X}_t \longrightarrow \mathscr{A}$ takes action $d_n(\mathbf{X}_t)$ when $\mathbf{X}_t$ is observed.[4]

We shall notice that a loss $L\left[\theta, d_t(\mathbf{X}_t)\right]$ is incurred when the sampling procedure $T \in \mathscr{T}$ stops at stage $t$ and $\mathbf{X}_t$ is observed. Then given a prior distribution $\pi(\theta)$ on the parameter space $\Theta$, the average loss of choosing decision $d_t$ for the observed sample $\mathbf{X}_t$ over the parameter space $\Theta$ is

[4]The statistical specification here follows the Bayesian setting described by Arrow, Blackwell, and Girshick (1949).

$$E\{L[\theta, d_t(\mathbf{X}_t)]\,|\mathbf{X}_t\} = \int L[\theta, d_t(\mathbf{X}_t)]\,\pi(\theta)d\theta. \tag{4}$$

Then for the entire sampling procedure $T$ and the sequence of decision rules $d = \{d_t\}$, the total cost is given by

$$R(T,d) = \sum_{t=0}^{\infty} \int_{\mathscr{X}_t} \Big[E\{L[\theta, d_t(\mathbf{X}_t)]\,|\mathbf{X}_t\} + c_t(\mathbf{X}_t)\Big] d\text{Prob}(\mathbf{X}_t), \tag{5}$$

in which $c_t(\mathbf{X}_t)$ is the cost of the sampling procedure stopped at stage $t$. Arrow, Blackwell, and Girshick (1949) showed that for all sampling procedures $T \in \mathscr{T}$, there exists a fixed sequence of decision rules $d^m = \{d_t^m; t = 0,1,2,\dots\}$ such that the minimum of the total cost $R(T,d)$ can be attained, that is,

$$R(T,d^m) \longrightarrow w(T) := \inf_d R(T,d), \quad \text{for all } T \in \mathscr{T}. \tag{6}$$

They further showed that the optimal sampling procedure can be constructed and an associated sequence of decision rules can be found; hence the total cost $w(T)$ can be minimized.

Arrow, Blackwell, and Girshick (1949) showed that SPRT provides an optimal stopping rule from the perspective of making sequential decisions. Specifically, for the two hypotheses $H_0$ and $H_1$, let $w_{ij}$ be the nonnegative loss incurred in accepting the hypothesis $H_j$ when $H_i$ is in fact true, and assume that $\theta$ can take on $\theta_0$ and $\theta_1$ with prior distribution $\pi$ and $1-\pi$, respectively, where $0 < \pi < 1$. Then the stopping rule of SPRT conFceived by Wald and Wolfowitz (1948) is equivalent to the sampling procedure suggested by Arrow, Blackwell, and Girshick (1949). The SPRT marks the birth of sequential analysis (Ghosh 1991) and has motivated many important developments and breakthroughs in sequential analysis in the last several decades; see Siegmund (1985), Sen and Ghosh (1991), and Lai (2001) for general reviews of the subject.

## Adaptive Sequential Methods in Experiment Design

Note that the conclusions drawn by Wald and Wolfowitz (1948) and Arrow, Blackwell, and Girshick (1949) suggest that under certain circumstances, decisions based on sequential observations are better than those based on fixed samples. This provides an important implication in the design of experiments, that is, *depending on the loss function $L(\theta,d)$, the prior distribution $\pi(\theta)$ on the parameter space (or the space of the state of nature) $\Theta$, and the distribution of samples (or population) $P_\theta$, experiments with fixed samples or sequential observations can be optimally selected before the inception of experimentation*. This implication further motivates the development of *multistage* (or *group sequential*) design, which is widely used in today's clinical trials (Jennison and Turnbull 2000, Section 1.2).

However, in the real world, policy designers often do not know the prior distribution on the parameter space for a proposed experiment. Therefore it is difficult to derive a stopping threshold, as for the SPRT, based on the prior distribution. In this paper, we propose to use posterior distribution to proxy the prior distribution in multistage experiment design.

In multistage experiment design, the experiment is carried out in several stages, and at each stage a decision is made to continue or abort based on the results collected at previous stages. To see this process, we consider the following decisionmaking problem for the designer of an experiment, who needs to choose an action $a$ from the action space $\mathscr{A}$ based on observations with a fixed sample size. Suppose the experiment generates an observation $X$, which is sampled from the distribution $P_\theta$, where $\theta \in \Theta$ is the state of nature governing the outcome of the experiment. For given $\theta$ and an action $d \in \mathscr{A}$, the designer of the experiment incurs a loss $L(\theta,d)$. Given a prior distribution $\pi(\theta)$ on the parameter space $\Theta$, the *risk* of taking action $d$ for parameter $\theta$ can be evaluated as

$$R(\theta,d) = E_\theta L(\theta, d(X)) = \int L(\theta, d(x))dP_\theta(X) = \int L(\theta, d(X))f_\theta(X)dX, \tag{7}$$

and the Bayes risk for decision rule $d$ is

$$R(d) = \int R(\theta, d) d\pi(\theta) = \int \Big[ \int L(\theta, d(X)) f_\theta(X) \pi(\theta) d\theta \Big] dX \tag{8}$$

(see Lai and Xing 2008, Section 4.3.2). To evaluate this risk through sampling, one should sample $X$ from its marginal distribution

$$\widetilde{f}(X) = \int f_\theta(X) \pi(\theta) d\theta \tag{9}$$

instead of the prior distribution $f_\theta(X)$. As the posterior distribution of $\theta$ given observation $X$, $\widetilde{\pi}(\theta|X)$, is expressed as

$$\widetilde{\pi}(\theta|X) = f_\theta(X)\pi(\theta)/\widetilde{f}(X), \tag{10}$$

letting the average risk of decision $d$ for given observation $X$ as $\widetilde{L}(d|X)$, we then have

$$\widetilde{L}(d|X) = \int L(\theta, d(X)) \widetilde{\pi}(\theta|X) d\theta, \tag{11}$$

and hence the Bayes risk for a subset $\Omega_0 \subset \Omega$ can be evaluated from the perspective of the sample space and written as

$$R(d|\Omega_0) = \int_{\Omega_0} \widetilde{L}(d|X) \widetilde{f}(X) dX. \tag{12}$$

The above result has the following implications for experiment design:

1. To evaluate this risk through sampling for the whole sample space $\Omega$, experimenters can sample $X_1, \ldots, X_n$ independently and identically (that is, with full randomization) from the posterior distribution $\widetilde{f}(X)$ without knowing the prior distribution. Then the risk can be approximated by

$$R(d|\Omega) \approx R_{\text{FullRand}} := n^{-1} \sum_{i=1}^{n} \widetilde{L}(d|X_i), \quad X_i \sim \widetilde{f}(X). \tag{13}$$

2. The posterior loss function $\widetilde{L}(d|X)$ in the randomization procedure is not the same as the loss function $L(\theta, d(X))$. This is particularly important for the designer of the experiment (or policymaker), since the posterior loss $\widetilde{L}(d|X)$ could be very large (or even unaffordable) for a certain sample $X$ when the posterior loss is not homogeneous over the sample space $\Omega$.

In experimental studies, one usually wants to evaluate the effects of a treatment while keeping the risk of the experiment under control. Assume that the experiment designer has tolerance $\eta$ for the risk of the experiment. The above implications suggest that if the total risk $R(d) \leq \eta$, one could use full randomization (or a randomized controlled trial [RCT]) to make inferences on the effects of the treatment; otherwise, full randomization (or an RCT) should not be used.

Furthermore, this also provides a procedure for a multistage sequential experiment: Suppose the posterior loss $\widetilde{L}(d|X)$ takes a finite set of values $0 \leq v_1 < \cdots < v_K < \infty$ over the sample space. Let $\Omega_k := \{X \in \Omega \mid \widetilde{L}(d|X) = v_k\}$; then $\{\Omega_k, 1 \leq k \leq K\}$ is a disjoint partition of the sample space $\Omega$. The experiment can be carried out in at most $K$ stages. In the $k$th stage, one should sample $\{X_{k,1}, \ldots, X_{k,n_k}\}$ independently and identically from $\widetilde{f}(X)$ with the constraint $X \in \Omega_k$. Since the risk at stage $k$ is approximated by

$$v_k \approx R_{k,\text{Rand}} := \frac{1}{n_k} \sum_{i=1}^{n_k} \widetilde{L}(d|X_{k,i}), \quad X_{k,i} \in \Omega_k. \tag{14}$$

The experiment should be stopped at state $k^*$ such that

$$k^* = \min\{k \le K \mid \sum_{i=1}^{k} v_i \ge \eta\}. \tag{15}$$

Obviously, the above multistage experiment incurs risk that is no larger than that in an experiment with full randomization and is bounded by the experiment designer's risk tolerance.

In general, when the posterior loss $\widetilde{L}(d|X)$ takes continuous values, we might consider the following multistage experiment procedure: Suppose the risk tolerance for the experiment designer is $\eta$. The first stage of the experiment can be done for a subset $\Omega_1$ of the sample space, such that

$$\int_{\Omega_1} \widetilde{L}(d|X)\widetilde{f}(X)dX < \eta. \tag{16}$$

Note that this $\Omega_1$ may or may not be uniquely determined, depending on other constraints of the experiment. In practice, the search of such $\Omega_1$ may not be done in advance, and hence the designer selects a subset $\widetilde{\Omega}$ for the experiment. If the resulting risk $R(d|\widetilde{\Omega}) > \eta$, the experiment should be stopped. Otherwise, the experiment can move on to collect more observations for study in the next stage. If an experiment is finished in at most two stages, the difference of Bayes risk in the one- and two-stage experiments is expressed as

$$\begin{aligned} &R(d|\Omega) - \left[R(d|\Omega_1) + 1_{\{R(d|\Omega_1)\le\eta\}}R(d|\Omega\backslash\Omega_1)\right] \\ &= 1_{\{R(d|\Omega_1)>\eta\}}R(d|\Omega\backslash\Omega_1) \ge 0. \end{aligned} \tag{17}$$

It is apparent that, overall, a sequential experiment procedure incurs a lower Bayes risk than an RCT on the whole sample.

## Examples of Adaptive Sequential Experiment

We discuss here two examples of adaptive sequential experiments, making use of the preceding discussion. The examples here concern two major statistical inference problems, estimation and hypothesis testing, in each step of the experiment design. They further explain how the sampling or randomization should be done for subsets of the sample space when the experiment designer incurs a risk tolerance.

**Example 1. Estimation of the nature of state**. Suppose that $\theta$ is the unknown state of nature governing the outcome of a process, and $\theta$ is an element of the parameter $\Theta$. Given $\theta$, the experiment could generate observations $X_1, \ldots, X_n$ that can be considered as independent and identically distributed samples from the distribution $P_\theta \sim N(\theta, \sigma^2)$ with the density function $f_\theta(x)$. The experiment designer has a prior belief $\pi(\theta) \sim N(\mu, \nu^2)$ on the distribution of $\theta$, that is, $\pi(\theta)$ is the prior distribution of $\theta$. A decision function $d$ here refers to an estimator of $\theta$ using $X_1, \ldots, X_n$ and the prior belief, and the loss function is given by a quadratic function $L(\theta, d) = (\theta - d)^2$. The experiment designer needs to perform an experiment of sampling observations $X_1, \ldots, X_n$ from $P_\theta$, subject to the constraint that the total risk $R(d)$ must have $\eta$ as its upper bound. (To simplify the discussion, we assume the sampling cost is zero.) The questions here include (a) Should the sampling be carried out in the whole sample space $\Omega$ or a subset $\Omega_0 \in \Omega$? and (b) If the sampling can be done in $\Omega_0 \in \Omega$, how should $\Omega_0$ be determined?

Since the experiment designer has a prior belief concerning $\theta$, we shall notice that the posterior distribution of $\theta$, given $X$, is given by

$$\widetilde{\pi}(\theta|X) \sim N\Big(\frac{\frac{\mu}{\nu^2} + \frac{X}{\sigma^2}}{\frac{1}{\nu^2} + \frac{1}{\sigma^2}}, \frac{1}{\frac{1}{\nu^2} + \frac{1}{\sigma^2}}\Big), \tag{18}$$

and the marginal distribution of $X$ on the sample space $\Omega = (-\infty, \infty)$ is

$$\widetilde{f}(X) \sim N(\mu, \nu^2 + \sigma^2). \tag{19}$$

We now discuss the design issues when the designer faces homogeneous and heterogeneous average risks $\widetilde{L}(d|X)$ for different decision (or estimation) functions. We first consider the case based on an optimal decision. The quadratic loss function $L(\theta, d)$ and the Bayesian decision theory imply that the optimal decision or estimator here is the posterior mean of $\theta$, which is expressed as

$$\widehat{\theta}_{\text{Bayes}} = d(X) = \frac{\frac{\mu}{\nu^2} + \frac{X}{\sigma^2}}{\frac{1}{\nu^2} + \frac{1}{\sigma^2}}. \tag{20}$$

Given this decision, its average risk for observation $X$ can be computed as

$$\widetilde{L}(\widehat{\theta}_{\text{Bayes}}|X) = \int (\theta - \widehat{\theta}_{\text{Bayes}})^2 \widetilde{\pi}(\theta|X) d\theta = \frac{\sigma^2\nu^2}{\sigma^2 + \nu^2}. \tag{21}$$

Note that for the quadratic loss function here, the average risk of $\widehat{\theta}_{\text{Bayes}}$ given $X$ is homogeneous over the sample space. Therefore, if $\eta \geq \frac{\sigma^2\nu^2}{\sigma^2+\nu^2}$, we have $\eta \geq B(\widehat{\theta}_{\text{Bayes}})$ and accordingly, $X_1, \ldots, X_n$ can be sampled from the whole sample space $\Omega$. However, if $0 < \eta < \frac{\sigma^2\nu^2}{\sigma^2+\nu^2}$, we shall have that for $X_1, \ldots, X_n \in \Omega_0 \subset \Omega$, and $X_1, \ldots, X_n$ are independent and identically distributed as $\widetilde{f}(X)$:

$$\frac{1}{n}\sum_{i=1}^{n} \widetilde{L}(\widehat{\theta}_{\text{Bayes}}|X_i) \approx \int_{\Omega_0} \widetilde{L}(\widehat{\theta}_{\text{Bayes}}|X) \widetilde{f}(X) dX = \frac{\sigma^2\nu^2}{\sigma^2 + \nu^2} \int_{\Omega_0} \widetilde{f}(X) dX \leq \eta. \tag{22}$$

This implies that $\Omega_0$ satisfies

$$\int_{\Omega_0} \widetilde{f}(X) dX \leq \Big(\frac{1}{\sigma^2} + \frac{1}{\nu^2}\Big)\eta < 1. \tag{23}$$

Furthermore, we should notice that $\Omega_0$ cannot be any set satisfying (23), since the sample mean of $\widehat{\theta}_{\text{Bayes}}$ on $\Omega_0$ needs to match that on the whole sample space. In such a case, $\Omega_0$ should be symmetrical around $\mu$ and uniquely determined, that is,

$$\Omega_0 = \{X \mid \mu - \sqrt{\sigma^2 + \nu^2} z_0 \leq X \leq \mu + \sqrt{\sigma^2 + \nu^2} z_0\}, \tag{24}$$

in which $z_0$ satisfies

$$\int_0^{z_0} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt = \frac{\eta}{2}\Big(\frac{1}{\sigma^2} + \frac{1}{\nu^2}\Big). \tag{25}$$

We now consider another decision (or estimator), the shrinkage estimator, which is not optimal in the Bayes sense but sometimes has desirable properties. The shrinkage estimator originates from the James-Stein estimator (Stein 1956; James and Stein 1961) and has been used to solve various statistical inference problems in recent decades. To explain the idea, we assume that the designer uses the following shrinkage estimator:

$$\widehat{\theta}_{\text{Shrink},\lambda} = \frac{\frac{\mu}{\nu^2} + \lambda\frac{\mu}{\nu^2} + (1-\lambda)\frac{X}{\sigma^2}}{\frac{1}{\nu^2} + \frac{1}{\sigma^2}} = \widehat{\theta}_{\text{Bayes}} + \lambda \frac{\frac{\mu}{\nu^2} - \frac{X}{\sigma^2}}{\frac{1}{\nu^2} + \frac{1}{\sigma^2}}, \tag{26}$$

in which $\lambda$ is a shrinkage parameter specified by the designer. This implies an heterogeneous average risk for observation $X$, that is,

$$\widetilde{L}(\widehat{\theta}_{\text{Shrink},\lambda}|X) = \lambda\left(\frac{\frac{\mu}{\nu^2}-\frac{X}{\sigma^2}}{\frac{1}{\nu^2}+\frac{1}{\sigma^2}}\right)^2 + \frac{1}{\frac{1}{\nu^2}+\frac{1}{\sigma^2}}. \tag{27}$$

This suggests that the although the decision function $\widehat{\theta}_{\text{Shrink},\lambda}$ has the same form for all samples, the designer actually has different risk concerns for different samples, and hence sampling or randomization cannot be simply performed over all samples. In particular, we shall note that if

$$\eta \geq R(\widehat{\theta}_{\text{Shrink},\lambda}|X) = \frac{1}{\frac{1}{\nu^2}+\frac{1}{\sigma^2}} + \frac{\lambda}{\left(\frac{1}{\nu^2}+\frac{1}{\sigma^2}\right)^2}\Big[\big(\frac{1}{\nu^2}-\frac{1}{\sigma^2}\big)\mu^2 + \frac{\sigma^2+\nu^2}{\sigma^4}\Big], \tag{28}$$

the designer can still run randomization on (or sample from) the whole sample space $\Omega$. If $\eta < R(\widehat{\theta}_{\text{Shrink},\lambda}|X)$, the designer should run randomization on a subset $\Omega_1$ of $\Omega$ such that

$$\int_{\Omega_1} \widetilde{L}(\widehat{\theta}_{\text{Shrink},\lambda}|X)\widetilde{f}(X)dX \leq \eta. \tag{29}$$

If the designer still requires the sampling to be symmetrical around the mean of $\widetilde{f}(X)$ or other constraints, then $\Omega_1$ can be uniquely determined.

Example 1 explains how the experiment should be designed when the purpose is to estimate the nature of state. This includes evaluating the effects of a treatment or of a policy intervention. The average risk $\widetilde{L}(d|X)$ describes the loss of decision $d$ for all possible values of the nature of a state on sample $X$, and it can be interpreted as the spillover effect of decision $d$ on sample $X$. Depending on the decision function $d$, the average risk $\widetilde{L}(d|X)$ can be homogeneous or heterogeneous with respect to the sample space. For the heterogeneous case, $\widetilde{L}(d|X)$ is a function of $X$, and the designer has to incorporate this into his risk concern and hence carefully select regions for sampling or randomization. This result explains the case studies in Section 2. The interesting part of Example 1 is that when the average risk $\widetilde{L}(d|X)$ is homogeneous over the sample space, the designer still needs to find sampling regions to keep the risk level within his tolerance.

**Example 2. Hypothesis testing of treatment effects**. Suppose that $\theta$ represents the treatment effect in an experiment. The designer is interested in testing the hypothesis $H_0 : \theta = \theta_0$ versus the alternative, $H_1 : \theta = \theta_1 (\neq \theta_0)$. Assume that $X_1, \ldots, X_n$ are independently and identically sampled from the distribution $P_\theta$ with density function $f_\theta(X)$. Then standard statistical hypothesis testing theory implies that $H_0$ should be rejected if the likelihood ratio $LR = F(\theta_1)/F(\theta_0)$ exceeds some threshold $\gamma$, in which $F(\theta)$ is the likelihood function given by

$$F(\theta) = \prod_{i=1}^{n} f_\theta(X_i). \tag{30}$$

Corresponding to this decision rule, the possible wrong decision includes the cases of accepting $H_0$ when $H_1$ is true and rejecting $H_0$ when $H_0$ is true, which happens with probabilities $\alpha = \text{Prob}(\text{reject } H_0 \mid H_0 \text{ is true})$ and $\beta = \text{Prob}(\text{accept } H_0 \mid H_1 \text{ is true})$, respectively. Because the correct decision incurs no loss, we assume that $w_0(X_1, \ldots, X_n)$ is the loss incurred by rejecting $H_0$ when $H_0$ is true and $w_1(X_1, \ldots, X_n)$ is the loss incurred by accepting $H_0$ when $H_1$ is true. Note that in standard testing theory for fixed samples, the loss functions $w_0$ and $w_1$ are usually constant with respect to the samples. In the case of experiment design, the designer needs to consider the impact of the wrong decision on the samples themselves and the

possible *spillover effects*; hence $w_0$ and $w_1$ should be functions of the sample. Furthermore, we assume that the designer believes that $\theta$ can take on $\theta_0$ and $\theta_1$ with prior probabilities $\pi$ and $1-\pi$, respectively. The designer then obtains the following risk function for this experiment (we still assume that the cost of sampling is zero):

$$R(d|X) = \pi\alpha w_0(X) + (1-\pi)\beta w_1(X). \tag{31}$$

We shall notice that if the spillover effect should not be considered, such that $w_0$ and $w_1$ are constant over $\Omega$, the designer could run full RCT on the sample space. However, since the spillover effect is incorporated into the study, the risk $R(d|X)$ becomes a function of sample $X$, and the designer should consider the problem of minimizing $R(d|X)$ over the sample space. In particular, the equation $R'(d|X) = 0$ implies that the sample $X$ satisfies

$$\pi\alpha w_0'(X) + (1-\pi)\beta w_1'(X) = 0. \tag{32}$$

A simple interpretation of the above constraint is as follows. Suppose the spillover effects of $X$ on $\Omega$ are stratified on two disjoint subsets $\Omega_0$, $\Omega_1 := \Omega\backslash\Omega_0$, that is, $w_i(\Omega_j) = v_{ij}$. If

$$R(d|\Omega_0) = \pi\alpha v_{00} + (1-\pi)\beta v_{10} < R(d|\Omega_1) = \pi\alpha v_{01} + (1-\pi)\beta v_{11}, \tag{33}$$

then the designer should run RCT on the subset $\Omega_0$ first with the minimum risk that he can tolerate.

We might use Example 2 as an interpretation of the case studies in Section 2. For drug development, the sample space $\Omega$ includes all the experimental subjects (animals and humans), the distribution $P_\theta(\cdot)$ represents the results of treatment on different subjects, and $w_i(X)$ $(i = 0, 1)$ represents the spillover effect of a particular treatment on subject $X$. We realize that the animal subject $X \in \Omega_0$ usually has a smaller spillover effect or loss than the human subject $X \in \Omega\backslash\Omega_0$. Hence, in order to develop a safe and effective drug for a particular disease of humans, animal subjects $\Omega_0$ can be "sacrificed" or used in the experiment to determine if the drug contains any toxic components. This view is consistent with the fact that pharmaceutical companies are afraid of adverse effects of their drugs on humans because the FDA may pull the drugs out of the market, evaporating billions of dollars of investment. Similarly, in China's economic reforms, policymakers often place considerable weight on failures of policy experiments. A failure in a big city, for example, Beijing, will be widely known to the whole of China, jeopardizing political stability and politicians' careers. In contrast, if an experiment fails in a remote area, very few people will notice. This is why most Chinese economic reforms have started from a remote location (as with the rural household responsibility system) or in a controlled environment (like the Shenzhen special zone).

## 4. CONCLUSIONS

In the real world, agents often place a great deal of weight on the potential negative effects of a treatment such as a new drug, a policy reform, or a new industrial product. Since randomized controlled trials (RCTs) mainly measure the average treatment effect and do not take into account the potential heterogeneous effect, agents are often reluctant to run RCTs for fear of intolerable adverse consequences for a subgroup of the population from the very beginning. Instead, they prefer to first screen out those options that likely will result in unbearable outcomes through incremental experimentation. Only when they are assured that there are not major risks associated with a treatment do they consider scaling it up.

As a more pragmatic approach, when facing uncertainty about the potential benefits and costs of a treatment, it is better for researchers and policymakers to first conduct experiments in isolated areas. Even if such experiments are not so rigorously conducted as to include control groups, the pilots enable researchers to observe what works and what does not on the ground. Through these sequential experiments, the options with clear negative effects, even if only on some segments of the population, can be eliminated. Where feasible, an RCT can be used to evaluate the average effect of a treatment only after it has been shown to have no major negative side effects in previous stages.

Apart from concerns about potential negative effects, experiment designers also often face general equilibrium and political economy issues in particular with respect to social and economic policies (Acemoglu 2010). The positive effect observed in an RCT on a small scale may dissipate or even reverse itself when the experiment is scaled up. As a matter of fact, the sequential experiment approach proposed in this paper can help mediate this concern by uncovering the true general equilibrium effect step by step. Similarly, the sequential experiment approach can ameliorate concerns about potential political economy risks. Since experiments are conducted sequentially, if political economy risks are revealed at a certain stage of the experiment, the experiment can be immediately called off to avoid increased political backlash.

If evidence based on RCTs is overwhelmingly favored, then there is a tendency for researchers and practitioners to select treatments that obviously have low risk. The selection bias may inherently limit the utility of the chosen treatments for addressing real-world issues in a more relevant way because those treatments with potential big payoffs (and presumably high risks) have likely been screened out (Rodrik 2009).

## REFERENCES


Acemoglu, D. 2010. "Theory, General Equilibrium, and Political Economy in Development Economics." *Journal of Economic Perspective* 24:17-32.

Anthony, S. 2009. "Smart Strategic Experiments: Innovation Inventory #6." *HBR Blog Network, Harvard Business Review*, May 26. http://blogs.hbr.org/anthony/2009/05/seizing_the_silver_lining_chec_5.html.

Anthony, S. 2010. "How P&G Quietly Launched a Disruptive Innovation." *HBR Blog Network, Harvard Business Review*, May 11. http://blogs.hbr.org/anthony/2010/05/how_pg_quietly_launched_a_disruptive_innovation.html.

Arrow, J. K., D. Blackwell, and M. A. Girshick. 1949. "Bayes and Minimax Solutions of Sequential Decision Problems." *Econometrica* 17:213-244.

Arrowsmith, J. 2011. "Trial Watch: Phase III and Submission Failures: 2007-2010." *Nature Reviews: Drug Discovery* 10:87.

Banerjee, A. 2008. *Making Aid Work*. Cambridge, MA, US: MIT Press.

Banerjee, A., and E. Duflo. 2011. *Poor Economics: A Radical Rethinking of the Way to Fight Global Poverty.* New York: Public Affairs.

Barrett, C., and M. R. Carter. 2010. "The Power and Pitfalls of Experiments in Development Economics: Some Non-random Reflections." *Applied Economic Perspective and Policy* 32:515-548.

Besley, T., and A. Case. 1995. "Incumbent Behavior: Vote-Seeking, Tax-Setting, and Yardstick Competition." *American Economic Review* 85:25-45.

Carroll, J. 2010. "Pfizer Writes Off $725M Dimebon Pact after Final Phase III Failure." *FierceBiotech*, January 17. www.fiercebiotech.com/story/pfizer-writes-725m-dimebon-pact-after-final-phase-iii-failure/2012-01-17.

Cohen, J., and P. Dupas. 2010. "Free Distribution or Cost-Sharing? Evidence from a Randomized Malaria Prevention Experiment." *Quarterly Journal of Economics* 125:1-45.

Deaton, A. 2010. "Instruments, Randomization, and Learning about Development." *Journal of Economic Literature* 48:424-455.

Dodge, H. F., and H. G. Romig. 1929. "A Method of Sampling Inspection." *Bell System Technical Journal* 8:613:631.

Du, R. 2010. "The Course of China's Rural Reform." In *Narratives of Chinese Economic Reforms: How Does China Cross the River?*, edited by X. Zhang, S. Fan, and A. De Haan, 15-29. Singapore: World Scientific Publishing.

Duflo, E., R. Glennerster, and M. Kremer. 2008. "Using Randomization in Development Economics Research: A Toolkit." In *Handbook of Development Economics*, edited by T. P. Shultz and J. A. Strauss. 3895-3962. Amsterdam: Elservier.

Fierce Biotech. 2012. "The Top 10 Phase III Failures of 2010." www.fiercebiotech.com/special-reports/top-10-phase-iii-failures-2010. Accessed May 20.

Friedman, L., C. D. Furgerg, and D. L. DeMets. 2010. *Fundamentals of Clinical Trials*. New York: Springer.

Ghosh, B. K. 1991. "A Brief History of Sequential Analysis." In *Handbook of Sequential Analysis*, edited by P. K. Sen and B. K. Ghosh, 1-19. New York: Marcel Dekker.

Govindarajan, V., and C. Trimble. 2004. Strategic Innovation and the Science of Learning." *MIT Sloan Management Review* 45:67-75.

Heilmann, S. 2008. "From Local Experiments to National Policy: The Origins of China's Distinctive Policy Process." *China Journal* 59:1-30.

James, W., and C. Stein. 1961. "Estimation with Quadratic Loss." In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability.* Vol. 1, *Contributions to the Theory of Statistics*, edited by Jerzy Neyman, 361-379. Berkeley, CA, US: University of California Press.

Jennison, C., and B. W. Turnbull. 2000. *Group Sequential Methods with Applications to Clinical Trials.* Boca Raton, FL, US: Chapman & Hall.

Kanbur, R. 2001. "Economic Policy, Distribution, and Poverty: The Nature of the Disagreements." *World Development* 29:1083-1094.

Karlberg, J., and M. A. Speers. 2010. *Reviewing Clinical Trials: A Guide for the Ethics Committee.* Hong Kong: Karlberg, Johan Petter Einar.

Lai, T. L. 2001. "Sequential Analysis: Some Classical Problems and New Challenges." *Statistica Sinica* 11:303-408.

Lai, T. L., and H. Xing. 2008. *Statistical Models and Methods for Financial Markets*. New York: Springer.

Lau, L. J., Y. Qian, and G. Roland. 2000. "Reform without Losers: An Interpretation of China's Dual-Track Approach to Transition." *Journal of Political Economy* 108:120-143.

Lin, J. Y. 1992. "Rural Reforms and Agricultural Growth in China." *American Economic Review* 82:34-51.

Lin, J. Y. 2011. "Watermelons vs Sesame Seeds." *Let's Talk Development: A Blog Hosted by the World Bank's Chief Economist*. June 16. http://blogs.worldbank.org/developmenttalk/watermelons-vs-sesame-seeds.

Lipsky, M. D. and L. K. Sharp. 2001. "From Idea to Market: The Drug Approval Process." *Journal of the American Board of Family Practice* 14:362-367.

Luo, X. 2010. "Collective Learning Capacity and Choice of Reform Path: Theoretical Reflections on the Dual-Track System of Price Reform Process." In *Narratives of Chinese Economic Reforms: How Does China Cross the River?*, edited by X. Zhang, S. Fan, and A. De Haan, 15-29. Singapore: World Scientific Publishing.

Mahalanobis, P. C. 1940. "A Sample Survey of Acreage under Jute in Bengal, with Discussion of Planning of Experiments." In *Proceedings of the Second Session of the Indian Statistical Conference Held in Lahore, 1939*, edited by P. C. Mahalanobis, 73-92. Calcutta, India: Statistical Publishing Society.

Oates, W. E. 1999. "An Essay on Fiscal Federalism." *Journal of Economic Literature* 37:1120-1149.

Qian, Y., G. Roland, and C. Xu. 2006. "Coordination and Experimentation in M-form and U-form Organizations." *Journal of Political Economy* 114:366-402.

Ravallion, M. 2009. "Evaluation in the Practice of Development." *The World Bank Economic Observer* 24:29-53.

Ravallion, M. 2012. "Fighting Poverty One Experiment at a Time: A Review Essay on Abhijit Banerjee and Esther Duflo, Poor Economics." *Journal of Economic Literature* 50:103-114.

Rodrik, D. 2009. "The New Development Economics: We Shall Experiment, but How Shall We Learn?" In *What Works in Development: Thinking Big and Thinking Small*, edited by J. Cohen and W. Easterly, 24-47. Washington, DC: Brookings Institution Press.

Routledge, P. A. 1998. "150 Years of Pharmacovigilance." *Lancet* 351:1200-1201.

Sen, P. K., and B. K. Ghosh. 1991. *Handbook of Sequential Analysis*. London: Marcel Dekker.

Shewhart, W. A. 1931. *Economic Control of Manufactured Products*. New York: Van Nostrand Reinbold.

Siegmund, D. 1985. *Sequential Analysis: Tests and Confidence Intervals.* New York: Springer-Verlag.

Stein, D. 1956. "Inadmissibility of the Usual Estimator for the Mean of a Multivariate Distribution." In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability.* Vol. 1, *Contributions to the Theory of Statistics*, edited by Jerzy Neyman, 197-206. Berkeley, CA, US: University of California Press.

Steinmetz, K., and E. C. Spack. 2009. "The Basics of Preclinical Drug Development for Neurodegenerative Disease Indicators." *BMC Neurology* 9 (Suppl 1): 1-13.

Stratman, H. 2010. "Bad Medicine: When Medical Research Goes Wrong." *Analog Science Fiction and Fact*, September, 20.

Thomke, S. 2003. *Experimentation Matters: Unlocking the Potential of New Technologies for Innovation*. Boston: Harvard Business School Press.

Thompson, W. R. 1933. "On the Likelihood That One Unknown Probability Exceeds Another in View of the Evidence of Two Samples." *Biometrika* 25:285-294.

Wald, A. 1947. *Sequential Analysis*. New York: John Wiley and Sons.

Wald, A., and J. Wolfowitz. 1948. "Optimum Character of the Sequential Probability Ratio Test." *The Annals of Mathematical Statistics* 19:326-339.

Worrall, J. 2007. "Evidence in Medicine and Evidence-Based Medicine." *Philosophy Compass* 2:981-1022.

## RECENT IFPRI DISCUSSION PAPERS

**For earlier discussion papers, please go to www.ifpri.org/pubs/pubs.htm#dp. All discussion papers can be downloaded free of charge.**

1272. *Dynamics of Transformation: Insights from an exploratory review of rice farming in the Kpong Irrigation Project.* Hiroyuki Takeshima, Kipo Jimah, Shashidhara Kolavalli, Xinshen Diao, and Rebecca Lee Funk, 2013.

1271. *Population density, migration, and the returns to human capital and land: insights from Indonesia.* Yanyan Liu and Futoshi Yamauchi, 2013.

1270. *Reverse-share-tenancy and Marshallian inefficiency: Landowners bargaining power and sharecroppers productivity.* Hosaena Ghebru Hagos and Stein T. Holden, 2013.

1269. *The child health implications of privatizing Africas urban water supply.* Katrina Kosec, 2013.

1268. *Group lending with heterogeneous types.* Li Gan, Manuel A. Hernandez, and Yanyan Liu, 2013.

1267. *Typology of farm households and irrigation systems: Some evidence from Nigeria.* Hiroyuki Takeshima and Hyacinth Edeh, 2013.

1266. *Understanding the role of research in the evolution of fertilizer policies in Malawi*. Michael Johnson and Regina Birner, 2013.

1265. *The policy landscape of agricultural water management in Pakistan*. Noora-Lisa Aberman, Benjamin Wielgosz, Fatima Zaidi, Claudia Ringler, Agha Ali Akram, Andrew Bell, and Maikel Issermann, 2013.

1264. *Who talks to whom in African agricultural research information networks?: The Malawi case.* Klaus Droppelmann, Mariam A. T. J. Mapila, John Mazunda, Paul Thangata, and Jason Yauney, 2013.

1263. *Measuring food policy research capacity: Indicators and typologies.* Suresh Chandra Babu and Paul Dorosh, 2013.

1262. *Does freer trade really lead to productivity growth?: Evidence from Africa.* Lauren Bresnahan, Ian Coxhead, Jeremy Foltz, and Tewodaj Mogues, 2013.

1261. *Data needs for gender analysis in agriculture.* Cheryl Doss, 2013.

1260. *Spillover effects of targeted subsidies: An assessment of fertilizer and improved seed use in Nigeria.* Lenis Saweda Liverpool-Tasie and Sheu Salau, 2013.

1259. *The impact of irrigation on nutrition, health, and gender: A review paper with insights for Africa south of the Sahara.* Laia Domenech and Claudia Ringler, 2013.

1258. *Assessing the effectiveness of multistakeholder platforms: Agricultural and rural management councils in the Democratic Republic of the Congo.* Thadde Badibanga, Catherine Ragasa, and John Ulimwengu, 2013.

1257. *The impact of Oportunidades on human capital and income distribution: A top-down/bottom-up approach.* Dario Debowicz and Jennifer Golan, 2013.

1256. *Filling the learning gap in program implementation using participatory monitoring and evaluation: Lessons from farmer field schools in Zanzibar.* Elias Zerfu and Sindu W. Kebede, 2013.

1255. *Agricultural mechanization in Ghana: Is specialization in agricultural mechanization a viable business model?:* Nazaire Houssou, Xinshen Diao, Frances Cossar, Shashidhara Kolavalli, Kipo Jimah, and Patrick Aboagye, 2013.

1254. *A partial equilibrium model of the Malawi maize commodity market.* Mariam A. T. J. Mapila, Johann F. Kirsten, Ferdinand Meyer, and Henry Kankwamba, 2013.

**INTERNATIONAL FOOD POLICY RESEARCH INSTITUTE**

**www.ifpri.org**

**IFPRI HEADQUARTERS**
2033 K Street, NW
Washington, DC 20006-1002 USA
Tel.: +1-202-862-5600
Fax: +1-202-467-4439
Email: ifpri@cgiar.org