A Beginner's Guide to Randomized Evaluations in Development Economics

E Seul Choi and Booyuel Kim

We present a practical guide to randomized impact evaluation in Development Economics for people who are interested in starting a field experiment for the first time. We provide an overview of the important steps in planning and implementing a randomized evaluation from beginning to end. The steps include issues such as research question development, local partner institution, institutional review board, pre-analysis plan, sample-size calculation, random sampling, randomization, survey implementation, data entry, and data analysis.

Keywords: Impact evaluation, Randomized Controlled Trial (RCT), Potential outcome model

JEL Classification: C93

I. Introduction

Randomized evaluations in development economics are receiving considerable attention; there has been a dramatic increase in randomized evaluations carried out not only by academia but also by governments, international agencies, and non-governmental organizations to test the effectiveness of a specific program or policy (Banerjee, and Duflo 2009).[1] The rise in randomized evaluations is

E Seul Choi, Graduate Student, Department of Economics, Seoul National University, Seoul 08826, Republic of Korea. (E-mail): codnsznzn@gmail.com; Booyuel Kim, Corresponding author, Assistant Professor, KDI School of Public Policy and Management, Sejong 30149, Republic of Korea. (E-mail): bkim@kdischool.ac.kr, (Tel): +82-44-550-1023, respectively.

[1] For one example, Jameel Poverty Action Lab (J-PAL), one of the leading institutions in rigorous impact evaluations has conducted 784 randomized

in line with the credibility revolution in empirical economics which emphasizes the identification of causal effects (Angrist, and Pischke 2010). Previously, sensitivity analysis, where researchers presented the rigorousness of their results with different specifications or functional forms had been considered a salutary econometric practice (Leamer 1985). However, the sensitivity analysis was limited in identifying the causal effect of a program because it is difficult to know the difference between the outcome a participant experienced under the program and the experienced outcome if the participant did not take part in the program. Although the causal effect of a program for each person is not identifiable (this is the fundamental problem of causal inference), a randomized evaluation makes it possible to create group-level counterfactual which is free from selection bias, and to provide internally valid estimates of the average treatment effect (Rubin 1974).

In this paper we provide a brief overview of the important steps in planning and implementing a randomized evaluation from start to finish including issues such as research question development, local partner institution, IRB (Institutional Review Board), PAP (Pre-Analysis Plan), sample size calculation, random sampling, randomization, survey implementation, data entry, and data analysis.[2]

II. Research Question Development

Not all research questions need a randomized evaluation to be answered, thus, it is important to ask the right research question that requires a randomized evaluation. If we are interested to know what the needs are in a specific context, then descriptive need assessment can be enough. However, if we want to know whether a program or policy works or not, then we need to conduct a randomized evaluation to answer this question. Furthermore, randomized evaluations should be performed if we aim to answer one of the following research questions. 1) Which elements of the program matter the most? 2) Which of the two different programs produce a better outcome? 3) Are there complementarities among the programs? 4) Can the results from one

evaluations in 69 countries since 2003.

[2] More detailed discussions on these topics can be found in Glennerster and Takavarasha (2013) and Duflo *et al.* (2007).

context be replicated in another context?

A good place to start developing research questions is with existing literature reviews. We can quickly summarize the information in existing research regarding issues of interest. The information can also help identify the most important knowledge gaps, where new research questions begin. High quality literature review on randomized evaluations can be found within institutions such as Jameel Poverty Action Lab (J-PAL) at MIT, Innovations for Poverty Action (IPA), the International Initiative for Impact Evaluation (3ie), and the Development Impact Evaluation Initiative (DIME) in the Research Group of the World Bank.[3] Many academic journals in economics, which are all reliable resources, also publish up-to-date literature reviews. These journals include The Handbook of Economics series, the Annual Review of Economics, the Journal of Economic Perspectives, and the Journal of Economic Literature.

After developing research questions for a randomized evaluation, the next step is to incorporate the questions into the survey questionnaire to analyze the research hypotheses and its possible mechanisms. It is useful to review the existing survey questionnaire of related research, which is often available on its author's personal webpage or on the publication journal's website.[4] Another good source for questionnaire development is the government's or multilateral organization's survey questionnaire such as Demographic and Health Surveys (DHS).[5] Demographic and Health Surveys (DHS) are nationally-representative household surveys that provide data for a wide range of indicators in the areas of population, health, and nutrition. Also, the surveys have been administered in more than 90 developing countries on a regular

[3] Research institutions dedicated to randomized evaluations have literature reviews of the evidence in specific areas available on their websites. J-PAL: https://www.povertyactionlab.org/; IPA: http://www.poverty-action.org/; 3ie: http://www.3ieimpact.org/; DIME: http://www.worldbank.org/en/research/dime.

[4] For example, we can download survey questionnaires of selected research from Esther Duflo on her webpage. We can also find many survey questionnaires of published papers under the "Additional Material Section" of academic journals' websites such as American Economic Reviews (AER) and American Economic Journal (AEJ).

[5] DHS model questionnaires can be downloaded from http://dhsprogram.com/What-We-Do/Survey-Types/DHS-Questionnaires.cfm.

basis. By reviewing the DHS questionnaire of a specific country, we can easily understand the localized questionnaires in the context.

For example, KOICA (Korea International Cooperation Agency) has been implementing community-driven development (CDD) projects in 100 rural villages of nine different regions in Myanmar; KOICA would like to measure the short-term effect of CDD projects on social capital, and the survey questionnaire was developed from the DHS questionnaire and World Bank's Integrated Questionnaire for the Measurement of Social Capital (Grootaert *et al.* 2004).

III. Local Partner Institution

As most randomized evaluations in Development Economics take place in developing countries, it is important to find a local partner institution that implements the research project. A local partner institution provides legal and logistical bases for a randomized evaluation project. The government's approval for the research project often comes through the local partner institution which hires local staff and implements the intervention.

When Kim *et al.* (2016) started a randomized evaluation project for male circumcision and HIV/AIDS prevention in Malawi, they worked with Daeyang Luke Hospital in Lilongwe, Malawi. The project and Institutional Review Board (IRB) approvals were obtained through Daeyang Luke Hospital. Meanwhile, the project intervention (free male circumcision provision to male students at secondary schools) was implemented in the catchment areas of Daeyang Luke Hospital. Daeyang Luke Hospital was an ideal partner for this medical intervention.

When creating a research partnership with a local institution, it is important that a local partner institution understands why randomized evaluations are useful. Sometimes, local non-governmental organizations (NGOs) have negative perceptions on a research-based project (especially, randomized evaluation project). The NGOs assume that a project intervention is designed not for the benefit of the targeted people but only for the researchers. In this case, it is difficult to build long-term partnership with the local institutions. Thus, building a mutual consensus among the researchers and a local partner institution based on the importance of randomized evaluations is one of the critical prerequisites for the success of a field research.

IV. Project Approval and Institutional Review Board (IRB)

Before implementing a randomized evaluation, we need to get approvals for the project intervention and Institutional Review Board (IRB) in advance. Kim *et al.* (2016) in partnership with the Daeyang Luke Hospital, got the approval for male circumcision project from the Malawi government (Ministry of Health). On top of this official project approval from the Ministry, they got the project approval from traditional authorities in the catchment areas as well.

Meanwhile, it is common for researchers to have IRB approvals both from the country where they plan to conduct research and from their home institution. Kim *et al.* (2016) through the local partner institution first obtained the IRB approval from the National Health Sciences Research Committee under the Ministry of Health, Malawi and then, later applied for the IRB approval from the home institution, Columbia University.

V. Pre-Analysis Plan (PAP)

When comparing the treatment and control groups on a very large number of outcome variables, we would probably be able to find one or more significant differences among two groups by chance. We could also analyze outcomes for many different sub-groups until we find significant heterogeneous treatment effects. To avoid the danger of these data mining and cherry picking practices, it has become common to register a Pre-Analysis Plan (PAP) of a randomized evaluation in the social sciences, like in clinical drug trials.[6]

Registering a PAP requires more detailed information than registering the existence of an evaluation. Finkelstein *et al.* (2010) provided one of the early PAPs in economics presenting analysis plan on the effect of extending public health insurance to low income adults (the Oregon Medicaid Experiment). This PAP includes 1) randomization design and data sources, 2) estimating equations and analytical framework, 3) verification of randomization and baseline results, 4) primary and supporting analyses, 5) explanatory analyses, and 6) interpretations

[6] In 2013, the American Economic Association started a registry for randomized evaluations in the social sciences (www.socialscienceregistry.org). We can download PAPs of selected randomized evaluations in this website.

and caveats.

A PAP is useful to register when there are many alternative ways to measure outcomes or to specify the estimating equation. Also, when we have a strong ex-ante theory to think that a project or program will have heterogeneous effects on sub-groups of the population of interest, a registered PAP prevents us from being accused of data mining. Also, a PAP allows us to show that these sub-groups were not chosen randomly but were well-planned. However, some researchers do not agree with a PAP registration because it may reduce the flexibility of analysis.

When a PAP is written before the baseline data are collected, and the intervention and evaluation has started, this may be the best time to avoid any accusations of data mining. However, writing a PAP earlier on signifies that we do not have enough information to improve the analysis plan. Thus, the most effective time to write a PAP is after the endline data have been collected, but before they have been analyzed. Finkelstein *et al.* (2010) used this approach when they wrote the pre-analysis plan.

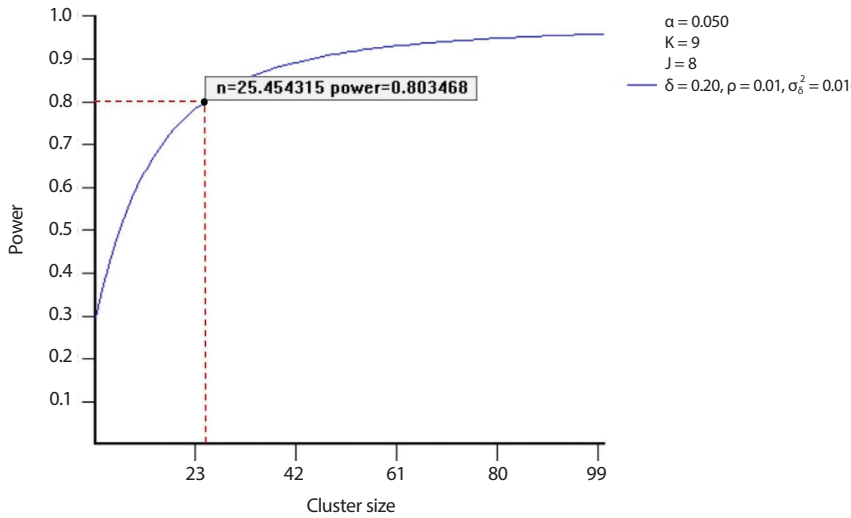
VI. Sample Size Calculation

The right sample size of a randomized evaluation in advance is important to determine. With a budget constraint, we need to calculate the optimal sample size to make a proper statistical inference. When we make a statistical inference, statistical power plays an important role in choosing the sample size.[7] Statistical power is determined by 1) the significance level, 2) the minimum detectable effect size (MDE), 3) the variance of the outcome of interest, 4) the allocation fractions among the treatment and control groups, and 5) the sample size. After this, we can calculate the optimal sample size using the determinants of statistical power.[8]

Although most statistical packages, such as Stata, have sample size functions, it is much more convenient to use a software package, Optimal Design, which is specifically designed to calculate statistical

[7] If we define false positive in a statistical test with probability α and false negative with probability κ, then significance level is α and power, the probability that detects a treatment effect when there is true one is $(1 - \kappa)$.

[8] For more details on statistical power and sample size calculation, see Glennerster and Takavarasha (2013) Chapter Six.



Source: Optimal Design.

FIGURE 1
SAMPLE SIZE AND POWER CALCULATION FOR KOICA'S PROJECT IN MYANMAR

power and sample size.[9] Optimal Design can show graphically how statistical power changes with sample size at different levels of its determinants such as MDE.

Figure 1 shows an example of calculating a sample size with the Optimal Design software. For KOICA (Korea International Cooperation Agency)'s community-driven development project in rural Myanmar, the team covered nine regions (K, the total number of stratum, is nine) and the project was designed to have at least eight villages per each region (J, the number of clusters per stratum, is eight). We set significance level to five percent (α = 0.05) and standardized MDE to 0.2 SD (δ = 0.2). In this setting, we find that we need at least 26 households (n, the cluster of size, is 25.45 households) per village to have 80 percent power.

VII. Random Sampling

Having a representative sample from the population of interest is

[9] Optimal Design can be freely downloaded from http://hlmsoft.net/od/.

Table 1
Assessment of Sample Representativeness

	Number of household			Occupation	Land	Household Asset			
	Total	Female	Single (dummy)	Non-farming (dummy)	Sown area (2013-14, acre)	Live-stock	Agricultural machine	Car or Motorcycle	Tailor machine
POPULATION (N = 18,456 households)									
Mean	4.01	2.07	0.04	0.50	2.90	12.78	0.19	0.01	0.002
(SD)	(1.67)	(1.12)	(0.20)	(0.50)	(5.85)	(107.56)	(0.55)	(0.14)	(0.05)
SAMPLE (n = 5,515 households)									
Mean	3.98	2.06	0.04	0.49	2.94	12.49	0.19	0.01	0.002
(SD)	(1.67)	(1.12)	(0.20)	(0.50)	(5.85)	(107.56)	(0.55)	(0.14)	(0.04)
t	-1.40	-0.90	-0.75	-0.93	0.41	-0.20	0.54	0.42	-0.53
(p-value)	(0.16)	(0.37)	(0.45)	(0.35)	(0.68)	(0.84)	(0.59)	(0.67)	(0.60)

Notes: Standard deviations and p-values are reported in parentheses.
Source: 2015 preliminary feasibility study data (census) and Baseline Data from KOICA Saemaul Undong (SMU) Project in Myanmar

one of the important issues for consideration before implementing a randomized evaluation. In the 1936 U.S. presidency election, the most influential poll in America predicted that the Republican candidate Mr. Landon, would have a landslide win against the Democratic incumbent, Franklin Roosevelt. However, the result was completely the opposite because the poll was based on the sample of magazine subscribers and owners of cars and telephones during the Great Depression. The majority of these people were affluent and Republican. This example shows the importance of having a representative sample when we use sample statistics to estimate population parameters.

If the census data for population of interest are available, then we can have a random sampling from the census, and compare sample statistics to population characteristics. When KOICA implemented CDD projects in 100 rural villages in Myanmar, the team collected census data for the entire 18,456 households in 100 villages of 9 regions, in partnership with local governments. Then, they randomly selected 5,515 households out of 18,456 households. Table 1 compares the population characteristics to the sample averages on selected outcomes such as number of household, occupation, land size, and household assets. Results show that the population and sample means are very similar, and none of them are significantly different from one another.

When a long-term field research is implemented, a demographic census surveillance system is ideal to establish in the project catchment area. Ghana Navrongo Health Research Center (NHRC) can be considered as one of the best examples in this issue. After the successful implementation of Ghana Vitamin A Supplementation Trials (VAST) project in 1980s, the research team set up a demographic census surveillance system for entire households in their catchment areas and implemented the census for approximately 150,000 households every three months. Because of this excellent foundation for field research, academic institutions, multilateral organizations, and aid agencies came to Navrongo Health Research Center to launch various research projects based on this census surveillance system.[10]

[10] For more information on Navrongo Health Research Center (NHRC), see http://www.navrongo-hrc.org/

VIII. Randomization

As Rubin (1974) pointed out in his potential outcomes approach, randomization ensures that there are no systematic differences among the treatment and control groups, which create a valid counterfactual on group-level average.[11]

Three different ways can introduce randomization into a program. First, we can randomly assign access to a program or policy, which is the most common practice. When Kim (2016) provided a one-year tuition and monthly education stipends to female students in 33 public secondary schools in Malawi, Table 2 shows that 62 out of 124 classrooms were randomly selected as the treatment group. With a budget constraint, it is not possible to provide a program to everyone, and thus, it is necessary to determine who would or would not receive the treatment. In many cases, lottery (randomization) is preferred because it is considered as an impartial method. After much discussion with 33 public secondary school headmasters on how to choose the scholarship beneficiaries, they all agreed that it was very difficult to identify the students who most needed the scholarship within a school and among schools; thus, they decided to randomly select the treatment group.[12]

Table 2
Experimental Design for Girls' Education Support Program

Group	Assignment	Classrooms	Students
G1	Treatment	62	2,102
G2	No treatment (Control)	62	1,895
Total		124	3,997

Note: The level of randomization is at the classroom level.
Source: Kim (2016).

[11] Duflo, Glennerster, and Kremer (2007) explained in detail how randomization solved the selection bias.

[12] Kim (2016) used random number generating function in Excel. Division Education Officer (DEO) clicked the randomization button in Excel to generate random numbers in front of 33 secondary school headmasters. Once the random assignment was completed, the results were printed out, and all 33 headmasters put his/her sign on the randomization results.

Second, we can randomly assign the timing of access to a program, designating who receives the treatment first, and who gets it later. Miguel and Kremer (2004) used this phase-in randomization design in their deworming program for 75 primary schools in Kenya, over three years. These 75 schools were randomly divided into three groups of 25 schools. In 1998, Group 1 schools received deworming treatment, whereas the other 50 schools stayed untreated. Group 1 and 2 schools received the treatment in 1999, and finally, Group 3 schools got the treatment in 2001. Thus, in 1998, Group 1 schools were in the treatment group, whereas Group 2 and Group 3 schools were in the control group. In 1999, Group 1 and Group 2 were the treatment groups and Group 3 was the control group. When Kim *et al.* (2016) provided free male circumcision to male secondary schools, the partner hospital's operational capacity was limited, only allowing 10 students to get circumcised per day. Thus, the researchers also used this second randomization approach.

This second approach is commonly used when we face administrative constraints and cannot provide a program at the same time. Compared to the first randomization approach, this approach is relatively free from ethical criticisms on a randomized evaluation because everyone would eventually receive a program. However, one of the drawbacks in this approach is that we cannot evaluate the long-term effect of a program as the control group no longer exists at the end of the program.

Third, when we are unable to randomly assign access to a program, we can randomly assign encouragement methods for participants to take up the program. This third approach is useful when a program or a policy is universally open to everyone, but the take-up rate is sub-optimal. For example, in Ethiopia, family planning service is free and thus, it is impossible to randomly assign the family planning service to participants. However, the take-up rate of free family planning service is relatively low because of many reasons such as transportation cost. In this case, we can randomly provide transportation reimbursement only to the treatment group.

Three aspects of programs that can be randomly assigned (access, timing of access, and encouragement) clearly show that there are opportunities to randomize in almost every circumstance for a field research. When the type of randomization is determined, we need to choose the level of randomization. Usually, the level of randomization is influenced by the level at which the program or data is implemented

or collected, respectively.[13] However, we do not always randomize at the level of implementation or data collection when spillovers among treatment and control groups are expected to occur. Spillovers can be physical, behavioral, or informational and they can be positive or negative. In this case, we choose the level of randomization to limit spillovers to the control group. Miguel and Kremer (2004) chose school-level randomization, which limits spillovers within a school, because students treated by a deworming program reduce disease transmission in their school.

If measuring spillovers is one of the main research questions, it may be useful to conduct a two-level randomization, both at the individual level and the group level. Table 3 shows Kim *et al.* (2016)'s two-level randomization design for the male circumcision program. First, they randomly selected three different groups, 100% treatment classrooms, 50% treatment classrooms, and No Treatment classrooms. Then, all male students in the 100% treatment classrooms (Group 1) received a free male circumcision offer, whereas none of the male students in no treatment classrooms (Group 4) got the treatment. Then, within the 50% treatment classrooms, they randomly picked half of the male students for the treatment (Group 2). The spillover group includes those who were in 50% treatment classrooms and did not receive the treatment (Group 3). In this two-level randomization design, the within-class spillovers can be measured by comparing the spillover group (Group 3) with the control group (Group 4).

Moreover, randomization within a group sometimes leads to resentment. People in the control group may be less likely to cooperate with the program implementation or survey participation when they see others in the same group receiving the treatment and they do not. It might cause systematic attrition bias if the control group is less likely to participate in the endline survey because of resentment. Randomizing at a higher level can reduce resentment because people in the same group are treated similarly. Kim (2016) considered possible resentment within a classroom if some of the students received the scholarship, whereas others in the same classroom did not receive anything. Instead of having randomization at the individual level, class-level randomization

[13] The most common levels at which to randomize are the individual and the community level.

Table 3
Experimental Design for Male Circumcision Program

	Group	Assignment	Classrooms	Students
100% Treatment	G1	Treatment	41	1,293
50% Treatment	G2	Treatment	41	679
	G3	No treatment		679
No Treatment	G4	No treatment (Control)	42	1,323
Total			124	3,974

Notes: The randomization was done in two levels. First, classrooms for each grade across 33 schools were randomly assigned to the 100% treatment, 50% treatment, and no treatment group. Then, within 50% treatment classrooms, only half of the students were randomly assigned for treatment at the individual level.
Source: Kim *et al.* (2016).

was chosen to avoid within-classroom resentment.

Although randomization at a higher level can effectively deal with issues like spillovers or resentment, there is a trade-off between level of randomization and statistical power. When we choose to randomize groups rather than individuals, we have fewer units to randomize; thus, the statistical power is likely to be reduced. Therefore, power calculation should be carefully considered when we choose the level for a randomized evaluation.

IX. Survey Implementation

Implementing surveys in field experiments is one of the most important processes for a randomized evaluation research. A baseline survey needs to be implemented before a program starts. As a baseline survey is the crucial foundation for any field research, the preparation of the survey should be carefully planned. Hiring and training enumerators can be the first step in preparing a baseline survey. It would be ideal to hire an enumerator who has experience working in similar field surveys. When Kim *et al.* (2016) recruited enumerators for the baseline survey, they contacted National Statistical Office (NSO) of Malawi; they received the list of enumerators who participated in 2010 Malawi Demographic and Health Survey and live in Lilongwe area.

After hiring survey enumerators, we need to have an intensive training session for the enumerators. It is important to make sure that enumer-

Table 4
Enumerator's Training Through Pilot Surveys

	Survey Duration	Total Errors	Skip Pattern	Missing Value	Outlier	Others	Data Entry
PILOT 1	1h 49min	25.92	2.17	14.53	0.33	2.77	6.13
PILOT 2	1h 35min	1.65	0.06	1.06	0.00	0.16	0.37

Notes: The unit for total errors is the number of errors counted. The total number of variables increased from 873 at the first pilot survey to 880 at the second pilot, because some questions were adjusted. The observed errors are categorized into five types. The first type is the skip pattern of question that has an instruction of "skip the next question if answer of this question is the following." The second type of error is the missing value. The outlier was also checked when values are four standard deviations from the mean. Other types of errors were placed in "others' category. Finally, the data entry mistake is checked.

ators completely understand not only survey questionnaire but also the survey logistics. To check the survey readiness of enumerators, a few small-scale pilot surveys are recommended. Table 4 shows two rounds of the pilot survey for KOICA's CDD project in Myanmar. During the first pilot survey, the enumerators made 26 errors on average, while the number of total errors was less than two on the second pilot survey. After confirming that the error rates have significantly decreased between the first and second pilot surveys, KOICA decided to start with the baseline survey.

When a program intervention is completed, we need to implement the endline survey. One of the key issues during the endline survey is attrition. If we have a high attrition rate or a systematically different attrition rate between the treatment and control group, this can be a threat to the internal validity of a randomized evaluation. To minimize the attrition of the endline, we need to collect detailed tracking information from the baseline survey. It is a common practice that we implement a tracking survey for the attrition of the endline survey. Kim (2016) conducted the endline survey and was able to reach 68.4% of the baseline students (2,733 students), 31.6% (1,264 students) were lost because of absence, transfer, and dropout. Then, the researcher randomly chose 15% of the lost students (187 out of 1,264 students) for the tracking survey. Out of 187 students, 128 students were reached for

the tracking survey at their homes. This decision resulted to an effective survey follow-up rate of 90.4%.[14]

After implementing a survey, data entry is the last process before we start data analysis. It is very common to make mistakes during the data entry process, thus, we should pay special attention to data entry to increase accuracy of data. Double data entry is recommended though it substantially increases costs and processing time compared to single data entry.[15] It is also recommended to use software such as EpiData, which is specifically designed for data entry process.[16] In the EpiData software, we can set an eligible range of legal values for each variable in advance to initially and effectively prevent erroneous data entry. Moreover, EpiData provides a double entry verification process. Figure 2 presents double entry verification results for five different variables (fields). A total of two out of five variables (v3a and v4) have discrepancies between the first and second data entry. With this information, a data entry clerk can detect the data entry errors and make corrections based on actual survey results.

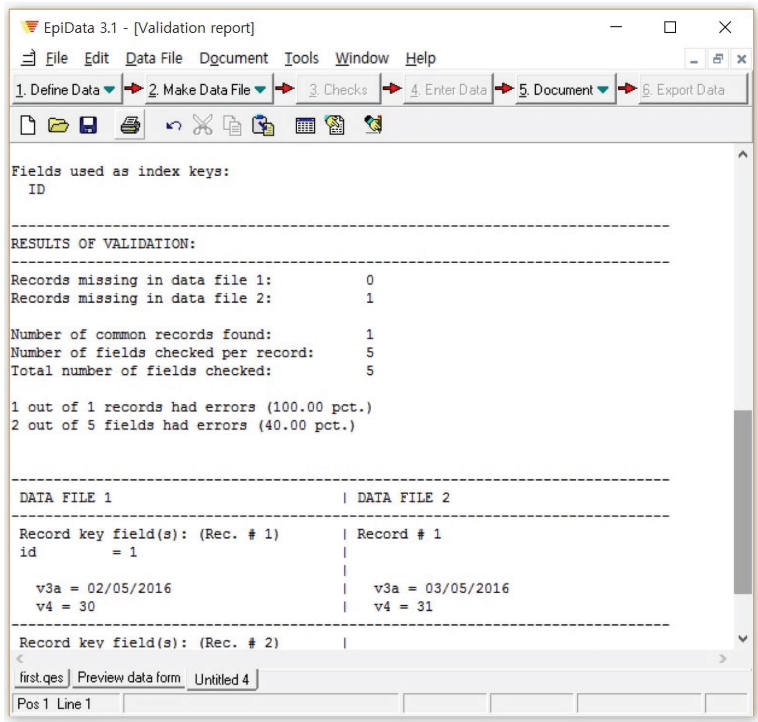
Recently, mobile-based survey has received much attention as an alternative to paper-based survey. One of the biggest advantages of mobile-based survey is it does not need a data entry process because data entry is being processed simultaneously as the survey is administered. Commcare is the most widely adopted mobile survey platform for low-resource settings. Various academic institutions, multilateral organizations, and donor agencies have been using Commcare for their research on health, education, and environment, in more than 50 countries.[17]

[14] The effective survey rate (ESR) is a function of the regular endline rate (RER) and home-visit tracking rate (HTR) as follows: ESR = RER + (1 – RER) * HTR (Baird *et al.* 2012). Overall, ESR is 90.4% (68.4% + 31.6% * 69.6%). Weight for home-visit tracking survey in Kim (2016) is 6.67 because he conducted a 15% random sampling from the sample attrition. It is desirable to achieve ESR greater than 90%. For one example, the ESR of Thomas *et al.* (2012) was 95%.

[15] Double data entry process is as follows. First, all data is entered into a data file. Then, the data is entered again and compared with the first data during the second data entry process. Discrepancies are brought to the attention of the data entry person.

[16] EpiData software can be freely downloaded from http://epidata.dk/.

[17] There are five levels for Commcare platform (Scope, Launch, Boost, Growth, and Scale). The basic level, Scope can be freely downloaded from https://www.commcarehq.org/home/.



Source: EpiData.

FIGURE 2
DOUBLE ENTRY VERIFICATION PROCESS

X. Analysis

When the survey data is ready to use, the first analysis we need to perform for a randomized evaluation is to check the baseline covariate balance. Random assignment should, in expectation, create treatment and control groups that have similar baseline characteristics. Kim (2016) conducted the OLS regression analysis to check randomization balance for the girls' education support program. Table 5, column 2 shows that none of the demographic characteristics, except for father's education, can predict the likelihood that a girl is assigned to the education support program. The F-tests for the joint significance of all the predetermined demographic variables on the girls scholarship is

TABLE 5
RANDOMIZATION BALANCE FOR GIRLS EDUCATION SUPPORT PROGRAM

Dependent Variable	Avg. (S.D) (1)	Girls Scholarship (2)
Age (year)	16.16	0.003
	(1.856)	(0.011)
Orphan	0.054	-0.037
	(0.227)	(0.032)
Father's tertiary education	0.198	0.047**
	(0.399)	(0.022)
Mother's tertiary education	0.082	-0.024
	(0.274)	(0.030)
Father's white-collar job	0.256	-0.031
	(0.436)	(0.023)
Mother's white-collar job	0.106	-0.017
	(0.307)	(0.027)
Household Assets (0-16)	7.59	0.003
	(3.455)	(0.007)
Conventional School	0.245	0.086
	(0.430)	(0.102)
p-value of joint F-test		0.325
Observations		3,993
R-squared		0.055

Notes: Orphan equals one when both parents died. Parent's tertiary education equals one when they graduated from a two-year college or four-year university. Parent's while-collar job equals one when they have a professional or government job. Household Assets are defined the total number of assets they own from the list of 16 asset questions. Conventional school equals one when a student is enrolled in a conventional secondary school. Column 2 shows randomization balance for girls' scholarship intervention. Robust standard errors clustered by classroom are reported in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Source: Kim (2016).

insignificant (p = 0.325), and does not reject that all baseline coefficients are jointly equal to zero. These results show that randomization for the intervention was well-balanced across predetermined baseline characteristics.

After checking the baseline balance among the treatment and control groups, we need to investigate whether or not the likelihood

Table 6
Relationship between Survey Attrition and Baseline Characteristics

Dependent variable	= 1 if surveyed in follow-up or home-visit surveys			
	Treatment (1)	Adjusted (2)	Main effect (3)	Interaction (4)
Girls Education Support	0.040*	0.038*		0.068
	(0.022)	(0.022)		(0.207)
Age		-0.015**	-0.012	-0.006
		(0.006)	(0.008)	(0.011)
Orphan		-0.081**	-0.071	-0.025
		(0.039)	(0.057)	(0.079)
Father's tertiary education		-0.003	-0.009	0.001
		(0.025)	(0.038)	(0.051)
Mother's tertiary education		-0.070**	-0.069	-0.000
		(0.034)	(0.050)	(0.069)
Father's white-collar job		-0.024	-0.041	0.035
		(0.017)	(0.027)	(0.034)
Mother's white-collar job		-0.028	-0.004	-0.049
		(0.027)	(0.038)	(0.053)
Household Assets		-0.001	-0.005	0.006
		(0.003)	(0.004)	(0.007)
Conventional School		0.049**	-0.008	0.102**
		(0.023)	(0.035)	(0.048)
Observations	3,997	3,993	3,993	
R-squared	0.014	0.024	0.027	

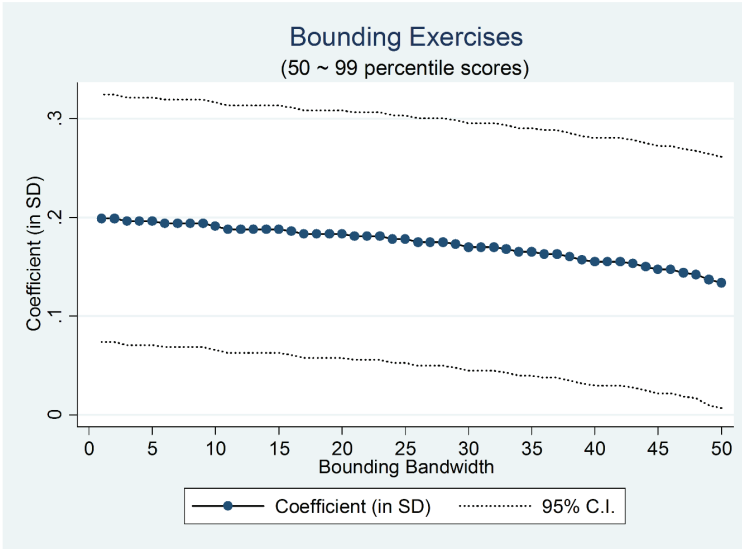
Notes: Regressions are OLS models with grade fixed effects. Robust standard errors clustered by classroom are reported in parentheses. The weight of 6.67 is given to home-visit survey sample. Columns 3 and 4 present results from one regression with main effects (Column 3) and all covariates interacted with treatment effect (Column 4). Baseline values of the following variables are included as controls: age, orphan status, parents' tertiary education, parents' white-collar job, household asset ownership, and school type. *** p<0.01, ** p<0.05, * p<0.10

Source: Kim (2016).

of remaining in the endline survey varies by the assignment groups and baseline characteristics. Attrition balance between two research groups is also important because it threaten the internal validity of a randomized evaluation. Kim (2016) conducted OLS regression

analyses to check attrition balance for the girls' education support program. Table 6 Column 2 shows that when students who receive the scholarship intervention after controlling the full set of demographic characteristics, they were 3.8 percentage points more likely to stay in the sample. This result causes a negative attrition (or retention) bias. In this case, we can check whether or not this attrition differs by baseline characteristics (Thomas *et al.* 2001); Columns 3 and 4 present no evidence that the survey attrition of the girls' education support program is systematically related to baseline characteristics.

When we have attrition bias, we can place bounds on the treatment effect and the largest and smallest average estimated treatment effects that would be obtained, if the missing attrition were filled in with extremely high or low outcomes (Lee 2009). For example, Kim (2014) performed bounding exercises for cognitive ability outcome. Initially, 50 percentile cognitive test score is assigned to the attrition sample. Then, the percentile score assigned to attrition students of the control group increases by 1 percentile, whereas the percentile score of the attrition students of the treatment group decreases by 1 percentile. Therefore,



Source: Kim (2016).

FIGURE 3
BOUNDING EXERCISES FOR THE COGNITIVE ABILITY OUTCOME

the second bounding practice is that attrition students in the control group is assigned 51 percentile score, and those in the treatment group is given a 49 percentile score. Finally, a 99 percentile score is assigned to the control group attrition, and 1 percentile score for the treatment group attrition. Figure 3 presents the cognitive ability outcomes under bounding exercises. The first dot at the zero bandwidth shows that the effect of girls' education support program on cognitive outcome is 0.2 standard deviations, when a 50 percentile score is assigned to the attrition. The last dot at the 50^{th} bandwidth shows that the effect decreases to 0.16 standard deviations, when the highest 99 percentile score and the lowest 1 percentile score are assigned to the attrition. As the dashed lines show upper and lower intervals of the 95% confidence level, all the bounding exercises are still statistically significant.

After checking the baseline and attrition balances among the treatment and control groups, we first perform the most basic analysis to calculate the intention-to-treat (ITT) estimate of average treatment effects on the outcomes of interest. This ITT analysis compares the mean outcomes of all those who are randomly assigned to the treatment with the people in the control group. The ITT estimates what happens to the average person given access to the program, and does not measure the effect of the actual participation in the program, because not all of those who are randomly assigned to the treatment group participated in the program. We can also calculate the treatment of treated (TOT) estimate, which is the average treatment effect among people who are treated. However, the TOT estimate can be biased when there are systematic differences among those who and who are not treated. In this case, the balance between the subgroup of treated people in the treatment group and the control group may not hold. Therefore, the ITT estimation is generally preferred over the TOT estimation, because the former provides unbiased estimates.

When we conduct ITT analysis, it is a common practice to report the estimated program impact both with and without the baseline covariates. When we have perfect balance on baseline covariates among treatment and control groups, adding covariates to the regression does not affect the main effects. However, the addition can give us a more precise estimate of the effect of a program when covariates reduce the unexplained variance of outcome variable. For example, Kim (2016) conducted the ITT analysis of girls' education support program on school attendance in Malawi. Table 7 Column (1) shows the ITT

TABLE 7
ITT ANALYSIS ON SCHOOL ATTENDANCE

Dependent variable	Self-reported absence			
	(1)	(2)	(3)	(4)
Girls Education Support	-1.707***	-1.645***	-2.187***	-2.187***
	(0.345)	(0.275)	(0.437)	(0.420)
Mean in the control group	3.794		0.672	
Baseline Covariates	No	Yes	No	Yes
Observations	2,715	2,704	2,700	2,689
R-squared	0.027	0.047	0.027	0.042

Notes: The dependent variable in Columns (1) and (2) are based on follow-up survey, whereas Columns (3) and (4) are based on the differences among the baseline and follow-up surveys. Regressions are OLS models with grade fixed effects. Robust standard errors clustered by classroom are reported in parentheses. The weight of 6.67 is given to home-visit survey sample. Baseline values of the following variables are included as controls: age, orphan status, parents' tertiary education, parents' white-collar job, household asset ownership, and school type. *** p<0.01, ** p<0.05, * p<0.10
Source: Kim (2016).

estimates without baseline covariates, whereas Column (2) indicates the ITT estimates with baseline covariates. As we can see, the magnitudes of the estimated effects are very similar (girls in the treatment group are 1.7 or 1.6 days per semester less likely to be absent), whereas the standard errors decrease (from 0.345 to 0.275) when we add baseline covariates.

When a program expects to have heterogeneous treatment effects for different subgroups in a population, we can conduct heterogeneity analysis by creating interaction terms among the treatment dummy and subgroup variables. We can estimate heterogeneous treatment effects on a subgroup by dropping all the other samples which do not belong to the subgroup. However, this subgroup analysis has smaller observations, and so may not have as much statistical power as when we estimate the average treatment effects on the whole sample. Thus, in general, using interaction terms is preferred to subgroup analysis for heterogeneity analysis. For example, Kim *et al.* (2016) tried to examine the heterogeneous effects of prior beliefs on male circumcision take-up in Malawi. Table 8 Column 3 shows that students, who think that male circumcision is very painful, are 3.8 percentage points less likely to take male circumcision surgery when they receive a free male circumcision offer.

Table 8
Heterogeneous Effects by Prior Beliefs

Dependent Var.	Circumcision Take-up			
	(1)	(2)	(3)	(4)
MC offer	0.146***	0.144***	0.162***	0.151***
	(0.021)	(0.023)	(0.023)	(0.023)
Knowing MC benefit		0.007		
		(0.012)		
MC offer and knowing MC benefit		0.003		
		(0.023)		
Thinks that MC is very painful			-0.023**	
			(0.010)	
MC offer and thinks that MC is very painful			-0.038*	
			(0.021)	
Thinks that MC is only for Muslims				-0.014
				(0.016)
MC offer and thinks that MC is only for Muslims				-0.032
				(0.029)
Observations	3,952	3,949	3,945	3,942

Notes: This table shows the heterogeneous effects on male circumcision take-up. MC offer variable equals one when students get a free male circumcision offer. All columns use grade fixed effects and robust standard errors clustered at the classroom level in parentheses. Control variables include age, circumcising ethnicity, circumcising religion (Muslim), orphan of both parents, father's good education, mother's good education, father's good job, mother's good job, household assets, and school type. *** p<0.01, ** p<0.05, * p<0.10

Source: Kim *et al.* (2016).

Finally, when we have randomization at the group level rather than the individual level because of spillovers or resentment issues, we can collapse the data to group aggregates and analyze it at the group level. However, when there are many clusters, it is more common to analyze at the individual level, and correct the regression by clustering standard errors at the level of randomization. This process allows outcomes within a group to be correlated to one other. Most statistical packages have commands that allow the estimation of clustered standard error.

XI. Conclusion

A randomized evaluation has been proven to be a powerful research method to identify the causal effects of a program or a policy. Also, there has been rapidly increasing trend in the academic community of all different disciplines which implements the randomized evaluation approach in empirical research. In this paper, we try to present, by using practical examples, a beginner's guide to a randomized evaluation, for people who are interested to launch a field experiment for the first time. We provide a brief overview of the important steps in planning and implementing a randomized evaluation from start to finish. However, this paper is limited in covering the very basics of a randomized evaluation, and there are various important topics not discussed in this paper. For researchers who are interested to investigate randomized impact evaluations further, the study of Glennerster and Takavarasha (2013), which is based on the research experiences of J-PAL, can provide a detailed and comprehensive picture of this new trend.

(Received 4 October 2016; Revised 12 October 2016; Accepted 13 October 2016)

References

Angrist, J. D., and J. S. Pischke. "The Credibility Revolution in Empirical Economics: How Better Research Design Is Taking the Con out of Econometrics." *The Journal of Economic Perspectives* 24 (No. 2 2010): 3-30.

Baird, S., J. H. Hicks, M. Kremer, and E. Miguel. "Worms at Work: Public Finance Implications of a Child Health Investment." Working Paper, 2012, Unpublished.

Banerjee, A. V., and E. Duflo. "The Experimental Approach to Development Economics." *Annual Review of Economics* (No. 1 2009): 151-78.

Duflo, E., R. Glennerster, and M. Kremer. Using Randomization in Development Economics Research: A Toolkit. Center for Economic Policy Research (CEPR) Discussion Paper No. 6059, 2007.

Finkelstein, A., S. Taubman, H. Allen, J. Gruber, J. Newhouse, B. Wright, K. Baicker, and the Oregon Health Study Group. The short-

run impact of extending public health insurance to low income adults: evidence from the first year of The Oregon Medicaid Experiment – Analysis Plan. 2010. Available at http://www.nber.org/oregon/documents/analysis-plan/analysis-plan-one-year-2010-12-01.pdf.

Glennerster, R., and K. Takavarasha. *Running Randomized Evaluations: a Practical Guide*. Princeton: Princeton University Press, 2013.

Grootaert, C., D. Narayan, and V. N. Jones, and M. Woolcock. Measuring Social Capital: An Integrated Questionnaire. World Bank Working Paper No. 18, 2004.

Kim, B. "Short-Term Impacts of a Cash Transfer Program for Girls' Education on Academic Outcomes: Evidence from a Randomized Evaluation in Malawian Secondary Schools." *Seoul Journal of Economics* 29 (No. 4 2016): 553-72.

Kim, B. Essays on Education and Health in Developing Countries. Columbia University Academic Commons, 2014. Available at: http://dx.doi.org/10.7916/D8C53J55.

Kim, B., H. Kim, and C. Pop-Eleches. Peer Effects in the Demand for Male Circumcision: Evidence from Secondary Schools in Malawi. AEA RCT Registry, 2016, Unpublished.

Leamer, E. "Sensitivity Analyses Would Help." *American Economic Review* 75 (No. 3 1985): 308-13.

Lee, D. "Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects." *Review of Economic Studies* 76 (No. 3 2009): 1071–102.

Miguel, E., and M. Kremer. "Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities." *Econometrica* 72 (No. 1 2004): 159-217.

Rubin, D. "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies." *Journal of Educational Psychology* 66 (No. 5 1974): 688-701.

Thomas, D., E. Frankenberg, and J. P. Smith. "Lost but Not Forgotten: Attrition and follow-up in the Indonesia Family Life Survey." *Journal of Human Resources* 36 (No. 3 2001): 556–92.

Thomas, D., F. Witoelar, E. Frankenberg, B. Sikoki, J. Strauss, C. Sumantri, and W. Suriastini. "Cutting the Costs of Attrition: Results from the Indonesia Family Life Survey." *Journal of Development Economics* 98 (No. 1 2012): 108–23.