Testing for Heterogeneous Treatment Effects in Experimental Data: False Discovery Risks and Correction Procedures

Günther Fink
Harvard School of Public Health

Margaret McConnell
Harvard School of Public Health

Sebastian Vollmer
Harvard School of Public Health & University of Hannover

July 2011

Abstract

Randomization has emerged as preferred empirical strategy for researchers in a variety of fields over the past years. While the advantages of RCTs in terms of identification are obvious, the statistical analysis of experimental data is not without challenges. In this paper we focus on multiple hypothesis testing as one statistical issue commonly encountered in economic research. In many cases, researchers are not only interested in the main treatment effect, but also want to investigate the degree to which the impact of a given treatment varies across specific geographic or socio-demographic groups of interest. In order to test for such heterogeneous treatment effects, researchers generally either use subsample analysis or interaction terms. While both approaches have been widely applied in the empirical literature, they are generally not valid statistically, and, as we demonstrate in this paper, lead to an almost linear increase in the likelihood of false discoveries. We show that the likelihood of finding one out of ten interaction terms statistically significant in standard OLS regressions is 42%, and that two thirds of statistically significant interaction terms using PROGRESA data can be presumed to represent false discoveries. We demonstrate that applying correction procedures developed in the statistics literature can fully address this issue, and discuss the implications of multiple testing adjustments for power calculations and experimental design. While multiple testing corrections do require large sample sizes ex-ante, the adjustments necessary to preserve power when corrections are applied appear relatively small.

I. INTRODUCTION

Few things have shaped empirical work on economics as much as the arrival and establishment of randomized controlled trials (RCTs) over the past decade. Initially restricted to a handful of researchers, the number of randomized experiments has grown exponentially across all continents over the past years. The projects listed on the Innovations for Poverty Actions (IPA) and the Abdul Latif Jameel Poverty Action Lab (JPAL) websites suggests that currently more than 100 RCT studies are either planned or in the field (JPAL 2011); John List's "Field Experiments" website currently lists more than 250 field experiments (List 2011).

One of the main advantages of experiments is the relative simplicity of the statistical analysis required to conduct causal inference. With properly done randomization, estimating causal effects corresponds to a simple conditional or unconditional mean comparison between treatment and control groups, with limited need or scope for more sophisticated empirical models. While little econometric work has focused on randomized trials until recently, a few econometric challenges associated with experimental work have emerged over the past few years (Hahn, Hirano and Karlan 2011). This paper focuses on the challenge of estimating (ex-post) heterogeneous treatment effects within experimental settings not initially designed to capture such differences. Most experiments are designed to estimate the average effect of a given treatment of interest. However, researchers may become interested in the interactions of a treatment with some baseline characteristics of interest during or after data collection, and wish to test for heterogeneous treatment effects ex-post. In some cases, researchers may learn during field work that the magnitude of the treatment effect hinges on a variable measured at baseline. In other cases, researchers may find that the theoretical framework applied has clear predictions regarding the expected behavioral changes across different subgroups of interest.

In order to provide some sense of how common ex-post testing for heterogeneous treatment effects is in the literature based on experimental data, we review all articles using field experiment data published in the top 10 journals according to the 2009 Engemann and Wall ranking (2009) as well as the Journal of Development Economics (the top field journal) from 2005 and 2010. Out of 34 articles we classified as field-experiment-based, 26 (76%) estimate

separate treatment effects for subgroups, and 10 articles (29%) estimate treatment effects for ten or more subgroups.

While testing for heterogeneous treatment effects through interaction terms or subgroup analyses is clearly desirable from a research perspective, applying traditional standard errors and p-values is not appropriate. Given that each interaction term represents a separate hypothesis beyond the original experimental design, "trying out" multiple interaction terms corresponds to multiple hypothesis testing, and results in a substantially increased false discovery risk in the empricial analysis.

To illustrate the severity of this issue, we use data from the Programa de Educación, Salud y Alimentación (PROGRESA) in Mexico, and run Monte-Carlo simulations to estimate a large number of heterogeneous treatment effects within the experimental data. We show that any researcher randomly choosing 10 baseline variables as proxy for an underlying characteristic of interest has a 62% chance of finding at least one variable significant at the 5% level. Given that the joint (Bernoulli) distribution for 10 independent binary variables implies a cumulative probabilities of finding at least one irrelevant factor significant with $\alpha = 0.05$ is 40%, this implies that about two thirds of the significant interactions uncovered in our PROGRESA regressions represent false discoveries.

The multiple hypothesis testing highlighted in this paper issue is not new, and has been faced by researchers in several other disciplines such as genomics or brain imaging over the past two decades. Possibly motivated by applications in these quickly evolving fields, recent statistical research has produced a number of powerful methods to correct for multiple hypothesis testing as summarized in Farcomeni (2008). Some of these correction methods can attain high power even when testing millions of hypotheses and are thus suitable for large-scale multiple testing problems in genomics or brain imaging. More importantly, these corrections are relatively easy to implement and very effective in keeping the risk of false discoveries low.

To provide a better sense of the study design implications of multiple testing, we compute ex-ante adjustments needed to sample size if researchers plan to investigate one or multiple

interaction effects ex-post. We show that the required sample size adjustments are relatively small as long as the number of tested heterogeneous treatment effects is reasonably small.

This results presented in this paper contribute to the recent discussion of structure and role of field experiments. Card, DellaVigna and Malmendier (2011) review the literature on field experiments over the last 35 years, finding that the analysis in most field experiments is focused on providing descriptive data instead of testing theories. The paper is also related to the broader literature on heterogeneous treatment effects discussed in Angrist (2004), Green and Kern (2010) as well as Imai and Strauss (2011). While these papers primarily focus on optimal model specification in the presence of heterogeneous treatment effect, we mostly focus on the multiple-testing issue associated with sequential subgroup testing in this paper.

The rest of the paper is organized as follows: We start with a review of the methods used in recently published papers in the field in section 2. We discuss the theoretical and empirical distribution of heterogeneous treatment effects using PROGRESA data in Section 3. We introduce the corrections for multiple testing proposed by the statistics literature in Section 4, and analyze the practical implications of the various correction models in Section 5. In Section 6, we discuss the implications of multiple testing for study design and power calculations; section 7 concludes.

II. LITERATURE REVIEW

Based on the Engemann and Wall (2009) ranking of Journals, we surveyed articles in the top 10 ranked journals as well as the Journal of Development Economics as the most commonly cited field journal. We use the classification proposed by Harrison and List (2004) as a guide to determine how to classify field experiments. We focus exclusively on "natural field experiments," which Harrison and List describe as experiments where a "non-standard[1] subject pool" makes decisions where there is a "field context in either the commodity, task, or information set that the subjects can use" and "the environment is one where the subjects

[1] In the context of experiments in Harrison and List (2004) , a standard subject pool would be undergraduates recruited to perform an experiment in a laboratory setting

naturally undertake these tasks and where the subjects do not know that they are in an experiment." (Harrison and List 2004, p. 1014)

In order to be considered for our literature review, a paper must present evidence from a study where a treatment intervention is randomly assigned by design of the study. Therefore, natural experiments and field experiments where treatments are not randomly assigned are excluded from our review.[2] Furthermore, we exclude papers that focus on econometric methods using data from natural field experiments. We refer to the papers that satisfy this definition as "strict" natural field experiments. While this classification may appear restrictive, it yields a clear decision rule allowing for a consistent review of the existing literature of interest. The main point made in this paper clearly applies to a larger set of empirical papers.

Table 1 shows the journal list, as well as the total number of articles, the number of "strict" natural field experiments.

TABLE 1

"STRICT" NATURAL FIELD EXPERIMENTS PUBLISHED 2005-2009

Journal Rank	Journal Name	Total Articles	"Strict Natural Field Experiments"
1	Quarterly Journal of Economics	283	11
2	Journal of Political Economy	296	1
3	Econometrica	420	5
4	American Economic Review	644	8
5	Review of Economic Studies	292	0
6	Journal of Labor Economics	201	0
7	Journal of Economic Growth	87	0
8	Review of Economics and Statistics	456	1
9	Economic Journal	498	1
10	American Economic Review Papers and Proceedings	592	5
30	Journal of Development Economics	461	2
	Total	*4230*	*34*

[2] Examples include the analysis of data where subjects were quasi-randomly matched as in studies of choices in speed dating (Fisman et al, 2006), the impact of random roommate assignment and random assignment by lottery (Angrist, Bettenger and Kremer 2006).

Of 4,230 articles surveyed in these 11 journals over the period 2005-2009, 34 articles feature evidence from strict natural field experiments. Some of the reviewed journals did not publish any study based on a field experiment (Journal of Labor Economics and Journal of Economic Growth), while the Quarterly Journal of Economics published more than 10 studies based on experiments fitting our "strict" natural field experiment description over the same period.

In Table 2 we take a closer look at the econometric strategy employed by articles featuring evidence from strict natural field experiments. We do not show the full title of each paper, but rather show the paper's index number, and provide the full listing as well as references in the Appendix. We report two measures of heterogeneous treatment effect testing: the number of subgroups for which treatment effects are estimated, and the number of interaction effects between the treatment variable and baseline characteristics tested.

In order to classify a reported result as a "heterogeneous treatment effect test" we followed a series of rules avoiding double-counting of treatment effects as well as incorrect classifications of regressions reflecting a particular study design. First, we only consider tests reported in the main tables of a paper, and exclude all results either shown in an appendix or only mentioned in the text. Second, we do not count reported tests reflecting the original research design. In many instances, researchers test for increasing effects over time, and also for heterogeneous treatment effects across different geographic sites. While one could argue that different time periods and sites reflect distinct sub-groups, we consider them as separate experiments, and thus do not count them as instances of heterogeneous treatment tests. We also do not count heterogeneous treatment effect tests based on baseline characteristics that are measured after the experimental randomization, which raise a whole set of other issues.

In some cases, papers combine subgroup analysis with estimates of interactions with treatments. We count the number of subgroups and interactions separately, and simply report the total number of subgroups and the total number of interactions. Furthermore, we only count each interaction or subgroup analysis once, even if they are considered for more than one dependent variable. Multiple dependent variable testing is indeed very common in the literature

as well, and the associated statistical problem similar to the heterogeneous treatment effects analyzed in this paper (Duflo, Glennester, et al 2008). The complications from considering dependent variables measured multiple times are discussed in recent work by McKenzie (2010).

TABLE 2

STRUCTURE OF EMPIRICAL MODEL IN FIELD EXPERIMENTS

Article Index	Type(s) of Dependent Variable(s): **B**inary, **C**ontinuous or bo**T**h	Number of Subsamples	Number of Interactions Estimated
1	T	2	0
2	B	7	0
3	C	0	4
4	T	11	0
5	T	16	11
6	T	4	0
7	T	0	0
8	C	0	0
9	C	0	0
10	T	0	1
11	T	2	4
12	C	21	10
13	B	0	0
14	B	5	3
15	C	0	3
16	T	15	0
17	B	0	1
18	B	0	0
19	T	6	1
20	C	2	0
21	C	0	0
22	C	7	0
23	C	10	0
24	C	60	2
25	C	3	11
26	C	0	0
27	T	1	1
28	C	7	0
29	B	4	14
30	B	23	0
31	T	2	2

32	C	0	2
33	C	9	6
34	B	0	0

Out of the 34 papers analyzed, 21 articles (62%) estimate separate treatment effects for subgroups, while 16 articles (47%) estimate interaction effects between the treatment and baseline characteristics. Only 8 articles (24%) neither estimate interaction effects nor consider the effect of treatment on subgroups. In some cases, testing for heterogeneous treatment effects is extensive: As Table 2 shows, 10 articles (29%) estimated 10 or more subgroup or interaction effects. Some examples of common interactions or subgroup analyses are sex, age, wealth and education. None of the article corrects (or mentions) multiple hypothesis testing in the empirical analysis

III. TESTING FOR HETEROGENEOUS TREATMENT EFFECTS

While most experiments are designed to investigate the average effects of a specific treatment of interest on a given outcome, researchers often may wish to investigate differences in the impact of the treatment by sex, ethnicity, income level, or other individual or household characteristics. In the simplest case, one may want to simply investigate one particular interaction of interest (possibly reflecting a particular model prediction or anecdotal evidence from the program rollout); in other instances, the researcher may simply be curious to see which factors modify an intervention's impact. Testing for heterogeneous treatment effects without adjusting the estimated standard errors for multiple testing after the fact, however, is highly likely to result in incorrect statistical inference. Given that 95% confidence intervals are constructed to allow for a false discovery probability of 0.05 on each interaction term, the probability of getting k significant p-values with zero true effects is given by the following binomial distribution:

$$f(k,m,\alpha) = \binom{m}{k} \alpha^k (1-\alpha)^{(m-k)}. \quad (1)$$

Plugging in for $\alpha = 0.05$ and $m = 10$, the probabilities of one, two, and more false discoveries with 10 different interaction terms are given by 31.5%, 7.5% and 1.2% as described in in Table 3 below.

TABLE 3

THEORETICAL (BINOMIAL) DISTRIBUTION OF FALSE DISCOVERIES WITH 10 RANDOM INTERACTION TERMS AND ALPHA=0.05.

Event	Probability
No hypothesis significant	0.598
One hypothesis significant	0.315
Two hypothesis significant	0.075
Three or more hypotheses significant	0.012

Heterogeneous Treatment Effects in Practice: PROGRESA

In order to illustrate how the distribution of estimated heterogeneous treatment effects looks in practice, we randomly test for such effects within the experimental data collected as part of the the Programa de Educación, Salud y Alimentación (PROGRESA). PROGRESA, now called OPORTUNIDADES , is a community-level randomized experiment designed to increase school attendance among the poor through a conditional cash transfer program. By providing a cash transfer to poor families large enough to compensate for lost wages from child labor[3] (Skoufias 2005), the conditional cash transfer program was aiming at changing parental schooling decisions.

While PROGRESA's impact on schooling has been well documented (Schultz 2004), it seems natural to ask whether the program impact was contingent on, or mediated by, specific household characteristics of interest at baseline. One may, for example, conjecture that the program impact increases with measures of household poverty or vulnerability. The PROGRESA baseline data from the 1997 includes a large array of measures one could use as potential markers for poverty: size of the household, access to piped water and electricity, asset holdings, characteristics of the dwelling, household size and many more. Given the difficulties associated with correctly

[3] A detailed description of the program as well as links to several evaluation studies are available at http://www.ifpri.org/dataset/mexico-evaluation-progresa.

measuring the income and wealth level of the household, it appears plausible that the interested researcher would consider a larger set of measures, and we shall for simplicity assume that each researcher uses 10 proxies in his analysis. While this may appear high at first sight, 10 interaction terms appear fairly common in the literature: as our review shows, the average paper analyzes 6.4 subgroups and tests for 2.2 interaction effects. The researcher then estimates the following model:

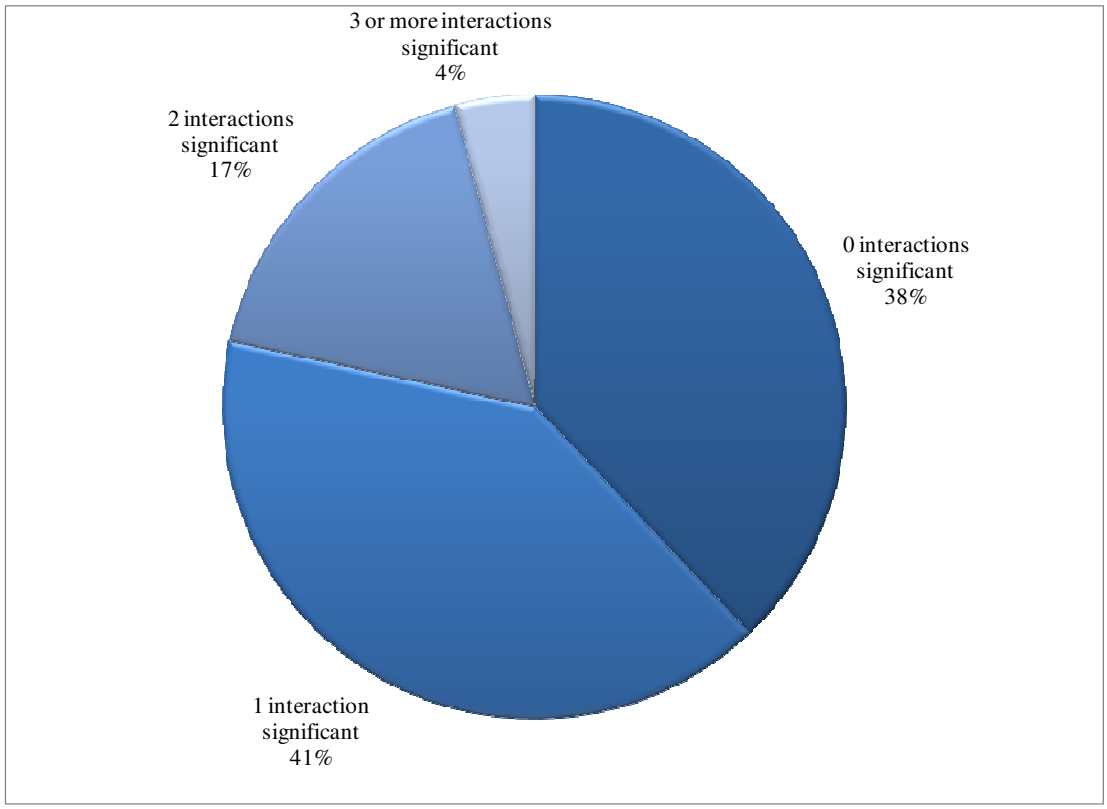
$$y_i = \alpha + \beta T_i + \delta W_i + \lambda (T_i x W_i) + \varepsilon_i, \quad (2)$$

where y_i is the outcome of interest (in this case schooling), T is the treatment indicator (1 if the household was targeted by PROGRESA), WI is one of the 10 poverty indicators coded, and $T_i x W_i$ is the interaction between the poverty indicator and PROGRESA. While we focus on interaction-term-based empirical models in our simulations, it is easy to see that the results will look virtually the same if separate regressions were conducted for each subgroup of interest.[4]
Given that virtually any baseline variable could be interpreted as a proxy for household poverty or vulnerability, we code all available baseline variables with non-zero variation within the treatment and control groups into binary variables. The total list of binary indicators (148 binary variables) is shown in Appendix Table 2. In order to not make any assumptions regarding researchers' choice we assume that each researcher randomly chooses 10 variables out of the set of 148, and runs 10 separate regressions as described in equation (1). We run a Monte-Carlo simulation with 10,000 experiments, where we randomly chose a set of 10 out the 148 variables to interact with the treatment in each round. The Monte-Carlo simulation can thus be viewed as an approximation of a setting where a large number of independent researchers work on a given data set, and each of them subjectively chooses 10 variables as proxies of poverty and vulnerability.

The results of the Monte-Carlo experiment are displayed in Figure 1. With 10 random binary regressors from the PROGRESA baseline data, more than 62% of cases (or independent

[4] Conceptually the main difference between subgroup analysis and interaction terms is that stratified regressions allow all regression coefficients to vary across sub-groups; since most experimental regressions include either few or no control variables, and since these control variables can be presumed to be orthogonal to the treatment, the difference between subgroup analysis and interaction term based regressions is small in practice.

researchers picking 10 interaction terms) find at least one interaction term significant at the 5% level; in 17% of all cases, 2 interaction terms are significant, and in 4% of cases, 3 or more interaction terms are significant.



Notes: Based on 10000 random block of size 11, consisting of the main treatment effect and ten randomly selected and independently tested interaction terms.

FIGURE 1

Empirical Distribution of Significant Coefficients at $\alpha = 0.05$

Given that we expect at least one significant effect in 40.2% of all cases in a completely random setting (Table 3), the likelihood that a single significant interaction term within our PROGRESA thought experiment constitutes a false discovery is about 65% (0.402/0.621).

IV. Statistical Corrections for Multiple Testing

Statisticians have been long aware of the problem of multiple hypothesis testing. The probability of the union of two events A_1 and A_2 is equal to the sum of the two probabilities $P(A_1)$ and $P(A_2)$ minus the probability of the intersect, i.e.

(3) $$P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1 \cap A_2)$$

This means that the sum of $P(A_1)$ and $P(A_2)$ constitutes an upper bound of $P(A_1 \cup A_2)$. If A_1 and A_2 describe very similar events, this upper bound may be distant from the true probability; if A_1 and A_2 are nearly independent $\left(P(A_1 \cap A_2) \approx 0\right)$, this upper bound will be very close to the true probability. In the case of multiple events, equation (1) becomes a bit more complex, but the intuition remains exactly the same as in the two events case.

Based on this basic notion, Boole's inequality states that for a finite set of events $A_1, ..., A_m$ the probability of one event happening can never be greater than the sum of the probabilities of each individual event, i.e.

(4) $$P\left(\bigcup_{i=1}^{m} A_i\right) \leq \sum_{i=1}^{m} P(A_i).$$

Building on this inequality, the Italian mathematician Carlo Emilio Bonferroni proposed a solution to the multiple testing problem. Assume we want to test m (dependent or independent) hypotheses at level α. Boole's inequality[5] implies that at least one of the hypotheses comes out significant with probability less or equal to $m\alpha$. However, in order to keep the chance of false discoveries (Type I errors) low, we would like this upper bound to be α and not $m\alpha$. Bonferroni showed that this can be achieved by testing each single hypothesis at the level $\alpha' = \alpha / m$. This is called the Bonferroni correction, designed to control the so-called familywise error rate (FWER).

[5] Boole's inequality is sometimes also referred to as Bonferroni's inequality.

Duflo et al. (2008) argue that Bonferroni type corrections may not be very useful in the context of economic field experiments, because the control of Type I errors might come at the cost of high Type II errors (less power). At standard 95% confidence intervals, testing for 10 effects simultaneously would require p-values of 0.005 for *each* individual variable. Recent statistical research has produced a number of alternative methods to correct for multiple hypothesis testing, which are much more powerful than the simple Bonferroni method. We focus on frequentist methods here but recent work by Gelman, Hill and Yajima (2010) proposes Bayesian multilevel models which can address the problem of multiple comparisions and increase efficiency.

Among recent developments in frequentist statistical literature is new multiple testing approach introduced in a seminal paper by Benjamini and Hochberg (Benjamini and Hochberg 1995). Rather than focusing on the FWER, the authors define the false discovery rate (FDR) as the expectation of the false discovery proportion (FDP), i.e, the proportion of the rejected null hypothesis which are erroneously rejected. If all null hypotheses are true, the FDR is equivalent to the FWER. Further, if not all null hypotheses are true, it can be shown that any procedure that controls the FWER also controls the FDR. If a procedure controls the FDR only, a gain in power may be expected. The potential for increase is larger when more of the hypotheses are non-true.

Consider testing a set of hypotheses $H_1, H_2, ..., H_m$ based on the corresponding p-values $p_1, p_2, ..., p_m$. Let $p_{(1)} \leq p_{(2)} \leq ... \leq p_{(m)}$ be the ordered p-values, and denote by $H_{(i)}$ the null hypothesis corresponding to $p_{(i)}$. Let k be the largest i for which

$$p_{(i)} \leq \frac{i}{m}\alpha \tag{5}$$

Then reject all $H_{(i)}$ with $i = 1, 2, ..., k$. Benjamini and Hochberg (1995) show that this procedure controls the FDR at α; independence of the test statistics is not needed for the proof.

It is easiest to illustrate the differences between the FWER and the FDR approaches with an example. Consider an experiment with one treatment, but 10 different dependent variables. The ordered p-values on each of the 10 estimated coefficients look as follows:

$$p_{(1)} = 0.001, p_{(2)} = 0.004, p_{(3)} = 0.006, p_{(4)} = 0.008, p_{(5)} = 0.010$$
$$p_{(6)} = 0.040, p_{(7)} = 0.050, p_{(8)} = 0.060, p_{(9)} = 0.100, p_{(10)} = 0.400$$

Without any adjustment for multiple testing we reject seven of the ten hypotheses at a 5 percent level of significance. The Bonferroni-adjustment requires a p-value of $\alpha' = \frac{0.05}{10} = 0.005$, which means that only two out of the ten hypotheses get rejected. With the Benjamini and Hochberg method we check the condition

$$p_{(i)} \leq \frac{i}{10}\alpha. \tag{6}$$

sequentially starting with i = 10. The first p-value to satisfy the condition is $p_{(5)}$ with 0.01 < 0.025; it is straightforward to see that the condition is also satisfied for any $i < 5,$ so that the FDR adjustment leads to a rejection of 5 out of the 10 tested hypotheses.

Genovese and Wasserman (2006) show that the Benjamini and Hochberg method is optimal in the sense that it minimizes the false non-discovery rate (FNR) subject to the constraint that the FDR is controlled at level $\alpha,$ where the FNR is defined as the expectation of the proportion of non-rejections that are incorrect. In other words, the Benjamini and Hochberg method keeps the number of type II errors as small as possible, i.e. the chance of not rejecting a hypothesis when it is false.

V. Correcting for Multiple Hypotheses Tests in Practice

There are two important aspects to consider when it comes to the applicability of these correction procedures: 1) the technical knowledge required for implementation; 2) the statistical and

empirical consequences in terms of type I and type II errors. The first aspect is fortunately straightforward. Thanks to the *multproc* package available in Stata© (Newson 2003), both the FDR and FWER methods are easily implemented in practice. The *multproc* package takes p-values from a set of variables (from single or multiple regressions) as inputs, and calculates corrected critical p-values for a range of correction procedures. The FWER correction is intuitive, as critical p-values are simply divided by the number of hypotheses tested (m). In the case of single control variable and one interaction term, this implies that the critical p-value for significance at the 95% level shifts from 0.05 to 0.05/2 = 0.025. It is easy to see that this adjustment keeps the likelihood of a false discovery at the desired low level independent of the number of hypotheses tested.

The FDR adjustment is slightly more complex since it is taking true discoveries into account, and, as a result, adjusts the p-values to a lesser extent than the FWER method. To see how well these adjustments work, we show the implications of the adjustment with truly independent variables (theoretical binomial) in a first step, and then revisit the PROGRESA results presented in Section 3.

Given that the FDR deviates from the FWER correction only if at least one hypothesis is false, it is easy to see that the two corrections have virtually the same effect if we assume 10 independent interactions without true effect. With $\alpha = 0.05$ and $m = 10$ the adjusted p-value under both correction models can be approximated by[6]

$$p_{adj} = \frac{\alpha}{m} = \frac{0.05}{10} = 0.005 \tag{7}$$

With the adjusted probability of 0.005, we can get the joint distribution within blocks of 10 by plugging the adjusted p-value into the corresponding binomial distribution. As the results displayed in Table 4 show, the chance of false discoveries is indeed reduced to just below5%.

[6] Technically, the FDR calculates separate p-values for each hypothesis. The adjusted FWER p-value corresponds to the p-value for the variable with the lowest p-value. The p-value of the second variable would be is 0.01, the third 0.015 and so on. In practice the first one will already rule out 95% such that the result is very similar to FWER, but it is not the same.

TABLE 4

THEORETICAL (BINOMIAL) DISTRIBUTION OF FALSE DISCOVERIES WITH 10 RANDOM INTERACTION TERMS AND ALPHA=0.005

Event	**Probability**
No hypothesis significant	0.951
One hypothesis significant	0.048
Two or more hypothesis significant	0.010

To provide a better sense of how powerful these corrections are in practice, we show the PROGRESA results displayed in Figure 1 with corrected p-values in Figure 2 below. The assumption underlying the correction is that within each experiment we test 11 hypotheses, the main treatment variable plus 10 randomly selected interaction terms. As Figure 2 shows, neither correction affects the significance of the main treatment effect, which is significant in all cases with both corrections. Large differences emerge, however, with respect to the interaction terms. While we see at least one significant result in 62% of the specifications if no correction is applied, the likelihood of finding a statistically significant results drops by 60% (FDR) and 70% (FWER), respectively, after the correction is applied. This, however, does not mean that researchers applying either correction will never find significant results – as Figure 2 clearly illustrates, the chance of finding one or more significant results in a Table showing 10 interaction terms in the PROGRESA sample is 18% with the Bonferroni FWER correction, and 24% with the Benjamini-Hochberg FDR correction.

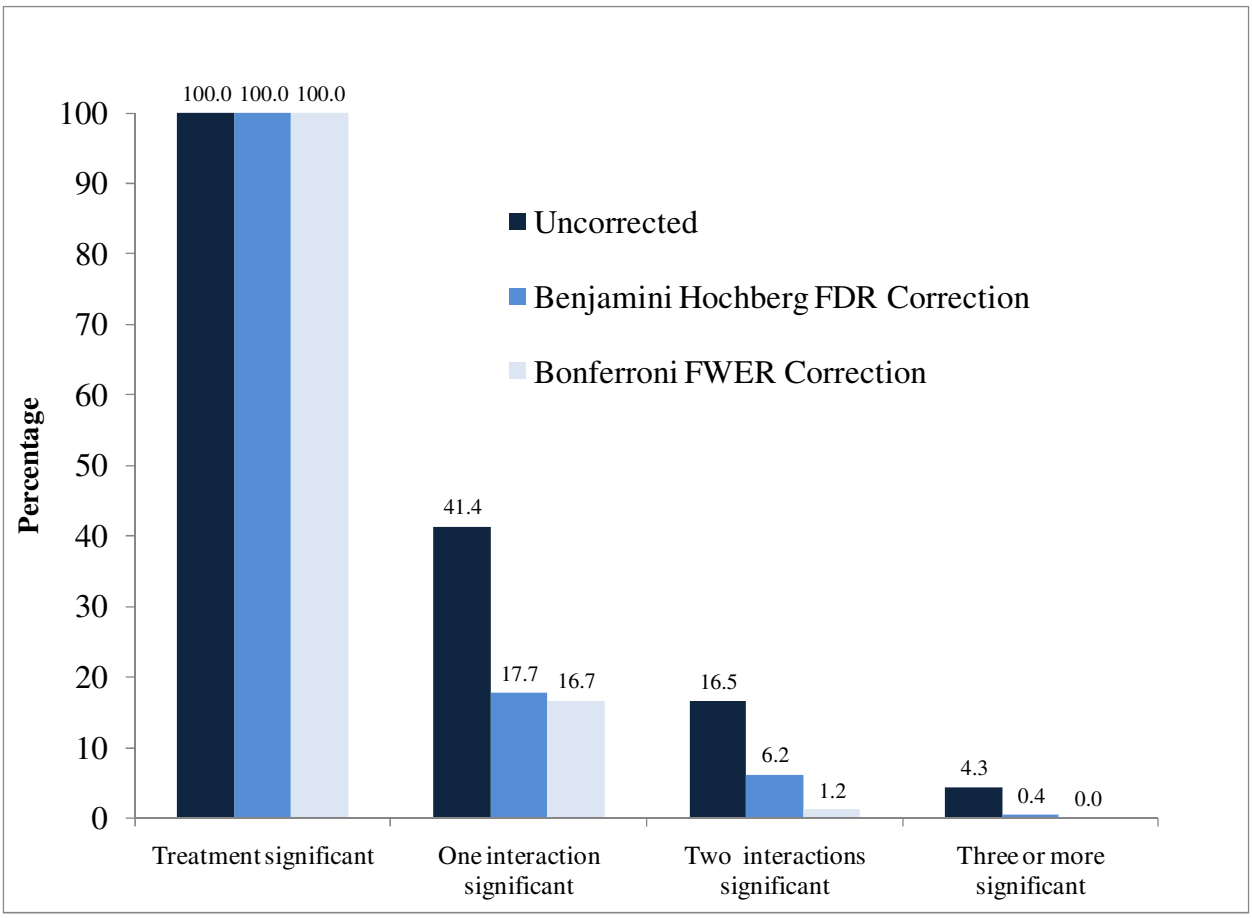


FIGURE 2

Empirical Distribution of Significant Results with and without Corrections

VI. Implications for Study Design and Power Calculations

Given that interaction terms may often be of major importance of researchers designing an experiment, one of the key questions is how much of an adjustment is needed to sample size ex-ante if the researcher plans to test for interaction effects ex-post. To understand what the two corrections imply in terms of power, we show a set of numerical simulations in this section. As a first step, we assume a sample size of 2000, with a corresponding (unadjusted) power of 0.5 and investigate how much power is lost in expected terms once we adjust for multiple hypotheses testing. As Figure 3 illustrates, the power drops in a non-linear fashion from 0.35 to about 0.2 for the FDR approach; as expected the drop is larger for the FWER approach, where the power drops to 0.1 if 10 hypotheses are tested simultaneously.

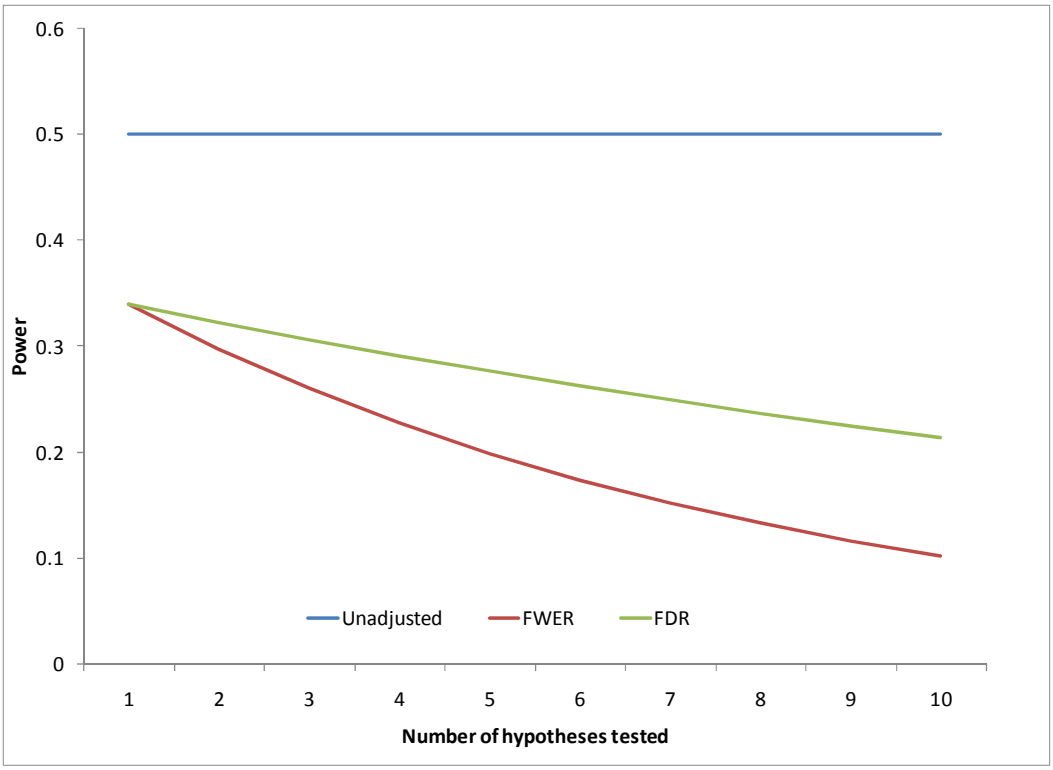


FIGURE 3

Power with Corrections

While these losses in terms of power may dissuade researchers from applying these corrections in practice, the necessary sample adjustment may not be as large as one may think (or fear) as long as the number of interactions the researcher is interested in is reasonably small.

Figures 4 illustrate this point for one and two control variables and their interaction terms, respectively. As pointed out before, standard power calculations do not apply here as two separate treatment groups lower effective group sizes, and increase estimated standard errors. As Figure 6 shows, the power of the study with sample size 5000 and a treatment effect of 0.05 is 0.9 if no interaction term is included. With the interaction term, the power drops to about 0.7. The drop in power due to the multiple testing corrections is comparable in magnitude. With a sample of 5000, the power with one interaction term drops to about 0.5, while the power with two interaction terms is about 0.45.

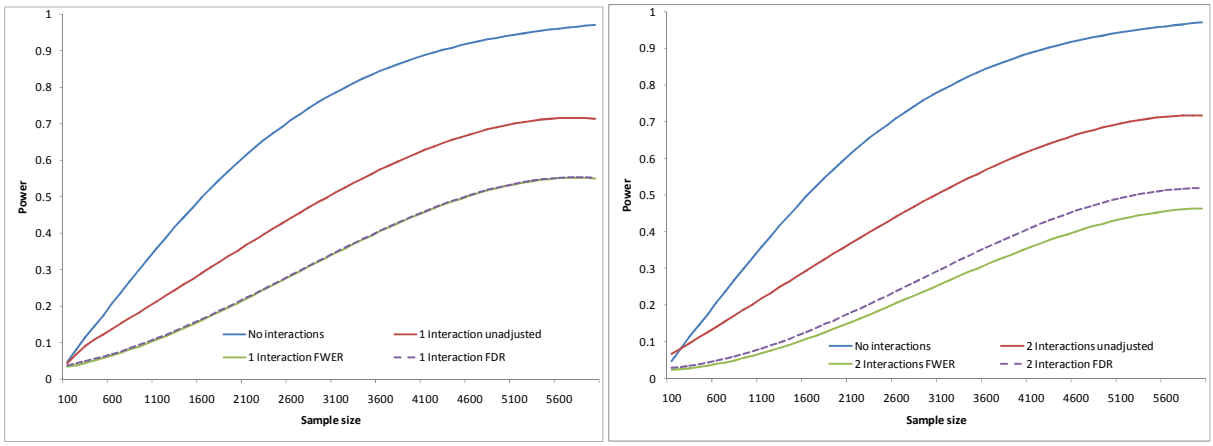


FIGURE 4

Power with one and two Interaction Terms

To see the implications for study design, we show necessary samples with and without corrections for multiple testing in Figure 5 below. As the figure shows, the absolute differences in sample size are rather small for large effect sizes; in relative terms, doing the FWER adjustment implies an average increase in necessary sample size (assumed power is 0.9) of about 28 percent with one interaction, 55% with two interactions, and about 67% with three interactions.[7] These adjustments are not trivial and may appear overly conservative. As discussed earlier, the FWER adjustment reduces the risk of false discoveries under the most conservative assumption of independence across events. Smaller sample size adjustments could in theory be generated by using the FDR approach and by relaxing the assumption regarding event independence; however, this would require imposing a large set of additional distributional assumptions researchers will struggle to make during early stages of field experiments. From a pragmatic perspective, it seems best to base initial sample size calculations on FWER adjusted standard errors. The FWER adjustment will keep the risk of false discoveries at the desired low levels and guarantee sufficient power for either FWER or FDR standard error adjustments ex-post.

[7] We assume that each interaction term is tested separately, so that each regression yields 3 coefficients (covariate, treatment and covariate*treatment) and 3 p-values that need to be adjusted.

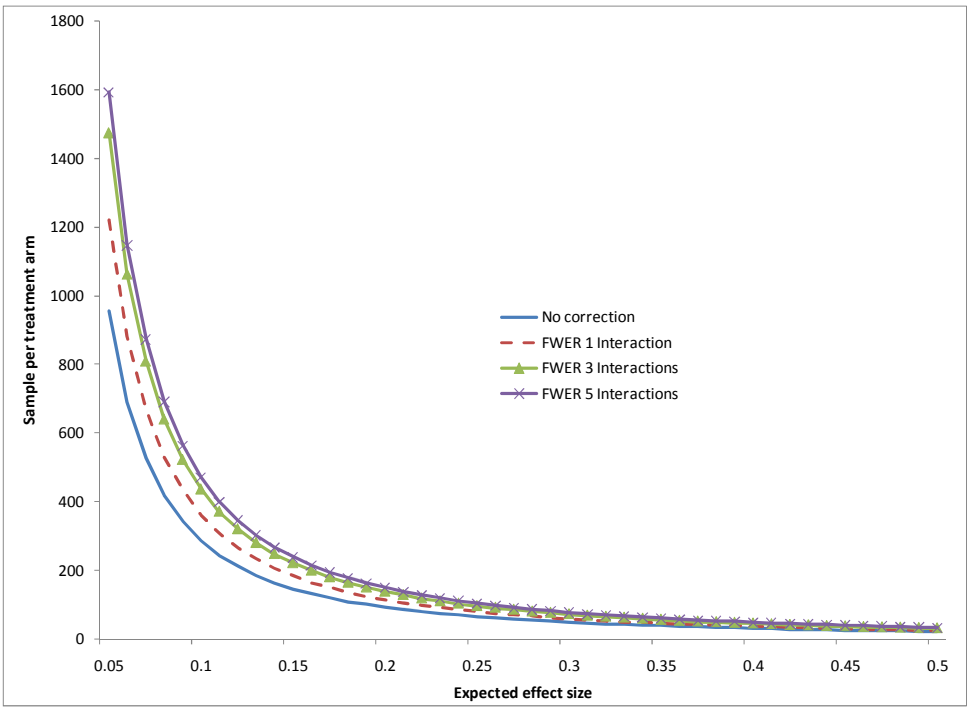


FIGURE 5

Absolute Sample Size with and without FWER Correction

VII. Discussion and Conclusions

Even though the inclusion of multiple interaction terms appears very common in current empirical work based on experimental data, multiple testing corrections are generally not applied in the recemt economics literature. In this paper we demonstrate that standard statistical inference is not valid when heterogeneous treatment effects are tested ex post, and that ignoring this issue is likely to generate a large number of false discoveries. Without any true effect, the likelihood of finding at least one result significant at the 5% level is 40% with 10 interaction effects tested (by one or multiple independent researchers). In the PROGRESA example, we find that about two thirds of the results significant at the 5% level can be assumed to reflect false discoveries. This risk appears high and, more importantly, unnecessary given the readily available correction models developed in the statistics literature.

Three main concerns have been raised regarding the application of correction procedures. The first concern regards the actual reporting of statistical tests conducted. Even if proper adjustments are applied to the final set of interaction terms tested, the underlying variable selection is unobservable ex-post, and may itself be the result of pre-testing. One possible approach to address this issue might be a central registration system similar to the ones used in medical trials. In fact, Duflo et al (2008) suggest that granting agencies create such a database of projects and their ex-ante designs.[8]

The second concern relates to the definition of what counts as a distinct hypothesis. Given that many variables may be used as proxies for a specific factor of interest such as income or human capital, it may be tempting to argue that all interacted variables are related, and thus reflect one single hypothesis. However, given that the correlation between any two proxies of interest is small in most cases empirically[9], treating multiple measures of a specific factor of interest as single hypothesis appears not advisable from a statistical perspective.

Finally and maybe most importantly, there is the concern that applying more stringent standard errors increases the chance of type-II errors, i.e. the chance of not rejecting a hypothesis when it is false. While adjusting p-values clearly comes with some loss of power, we have shown in this paper that the cost in terms of additional sample size required for researchers planning to test for heterogeneous treatment effects ex-post appears well worth the benefit in terms of reduced false discovery risk.

Overall, a wider application of multiple testing procedures in the economics literature appears highly desirable. Testing for heterogeneous treatment effects is of obvious interest to researchers, and neither can, nor should, be avoided in practice. The resulting risk of false discoveries is high, but can be reduced to a minimum if the appropriate correction procedures are applied.

[8] Some researchers have begun to publish their analysis prior to conducting experiments. Another potential alternative is the use of interdisciplinary system where researchers in any field can currently post their designs prior to conducting experiments, such as http://clinicaltrials.gov/.

[9] In the PROGRESA data, the highest correlation between any two indicator variables is 0.15.

Appendix Table 1: Index of Strict Natural Experiments

1	A Field Experiment in Charitable Contribution: The Impact of Social Information on the Voluntary Provision of Public Goods
2	Credit Elasticities in Less-Developed Economies: Implications for Microfinance
3	Do Workers Work More if Wages Are High? Evidence from a Randomized Field Experiment
4	Does Job Corps Work? Impact Findings from the National Job Corps Study
5	Does Price Matter in Charitable Giving? Evidence from a Large-Scale Natural Field Experiment
6	Experimental Analysis of Neighborhood Effects
7	Gift-Exchange in the Field
8	How High are Rates of Return to Fertilizer? Evidence from Field Experiments in Kenya
9	Incentives for Managers and Inequality Among Workers: Evidence From a Firm-Level Experiment
10	Incentives to Exercise
11	Incentives to Learn
12	Information, School Choice, and Academic Achievement: Evidence From Two Experiments
13	Insurance, credit, and technology adoption: Field experimental evidence from Malawi
14	Intra-household allocation of free and purchased mosquito nets
15	Monitoring Corruption: Evidence from a Field Experiment in Indonesia
16	Neighborhood Effects on Crime for Female and Male Youth: Evidence From a Randomized Housing Voucher Experiment
17	Observational Learning: Evidence from a Randomized Natural Field Experiment
18	Obtaining a Driver's License in India: an Experimental Approach to Studying Corruption
19	Power to the People: Evidence From a Randomized Field Experiment on Community-Based Monitoring in Uganda
20	Powerful Women: Does Exposure Reduce Bias
21	Putting Behavioral Economics to Work: Testing for Gift Exchange in Labor Markets Using Field Experiments
22	Remedying Education: Evidence From Two Randomized Experiments in India
23	Requiring a Math Skills Unit: Results of a Randomized Experiment
24	Resource and Peer Impacts on Girls' Academic Achievement: Evidence from a Randomized Experiment
25	Returns to Capital in Microenterprises: Evidence From a Field Experiment
26	Salience and Taxation: Theory and Evidence
27	Saving Incentives for Low- and Middle-Income Families: Evidence From a Field Experiment with H & R Block
28	Social Connections and Incentives in the Workplace: Evidence from Personnel Data
29	The Demand for, and Impact of, Learning HIV Status
30	The Effects of High Stakes High School Achievement Awards: Evidence from a Randomized Trial
31	The importance of being informed: Experimental evidence on demand for environmental quality
32	Toward an Understanding of the Economics of Charity: Evidence From a Field Experiment
33	Tying Odysseus to the Mast: Evidence From a Commitment Savings Product in the Philippines
34	What Matters (and What Does Not) in Households' Decision-Making Regarding Investments in Malaria Prevention?

Appendix Table 2: Binary Variables Used in PROGRESA Regressions

1 bathroom	38 head_nospanish	75 lost_limb	112 rooftype4
2 bathroom_water	39 head_o60	76 mental	113 rooftype5
3 bedrooms1	40 head_perm_unable	77 migrant	114 rooftype6
4 bedrooms2	41 head_preparatoria	78 needs_help_to_move	115 rooftype7
5 bedrooms3	42 head_primary	79 no_waterelectric	116 rooftype8
6 bedrooms4	43 head_primary_income	80 noincome	117 rooftype9
7 bedrooms5	44 head_profesional	81 not_childofhead	118 second_income
8 bedrooms6plus	45 head_retired	82 one_child	119 shared_building
9 blender	46 head_secondary	83 owns_agri_land	120 spouse_away
10 blind	47 head_single	84 owns_animals	121 stove
11 budget_control_head	48 head_single_female	85 owns_cattle	122 three_children
12 budget_control_other	49 head_socialsecurity	86 owns_chicken	123 treatment_dif
13 budget_control_shared	50 head_temp_unable	87 owns_donkey	124 treatment_hospital
14 budget_control_spouse	51 head_u20	88 owns_goatsorsheeps	125 treatment_imss
15 car	52 head_widowed	89 owns_horse	126 treatment_issste
16 cd	53 head_working	90 owns_land	127 treatment_othergov
17 deaf	54 hhsize10plus	91 owns_multiple_pieces	128 treatment_ssa
18 decision_head	55 hhsize2	92 owns_ox	129 truck
19 dumb	56 hhsize3	93 owns_pigs	130 tv
20 electric_lights	57 hhsize4	94 owns_rabbits	131 two_children
21 fan	58 hhsize5	95 piped_inside	132 video
22 father_athome	59 hhsize6	96 piped_water	133 walltype1
23 female	60 hhsize7	97 radio	134 walltype10
24 five_or_more_children	61 hhsize8	98 receives_apoyoINI	135 walltype11
25 floortype1	62 hhsize9	99 receives_becacapicaticaion	136 walltype12
26 floortype2	63 house_paid	100 receives_desayuno_escolar	137 walltype13
27 floortype3	64 house_paying	101 receives_despensa_DIF	138 walltype14
28 floortype4	65 house_provided	102 receives_empleotemporal	139 walltype15
29 four_children	66 house_rented	103 receives_leche	140 walltype2
30 fridge	67 inshool_97	104 receives_ninosdesolid	141 walltype3
31 head_2060	68 kids_medical_head	105 receives_tortilla	142 walltype4
32 head_basica	69 kids_medical_other	106 rooftype1	143 walltype5
33 head_dialect	70 kids_medical_shared	107 rooftype10	144 walltype6
34 head_female	71 kids_medical_spouse	108 rooftype11	145 walltype7
35 head_literate	72 laundry	109 rooftype12	146 walltype8
36 head_married	73 light_meter	110 rooftype2	147 walltype9
37 head_noschool	74 literate	111 rooftype3	148 water_heater

REFERENCES

Angrist, Joshua , and Victor Lavy, "The Effects of High Stakes High School Achievement Awards: Evidence from a Randomized Trial.," American Economic Review, 99((2009), 1384-1414.

Angrist, Joshua D., "Treatment Effect Heterogeneity in Theory and Practice " The Economic Journal, 114 (2004), C52-C83.

Ariely, Dan, Uri Gneezy, George Loewenstein, and Nina Mazar, "Large Stakes and Big Mistakes.," Review of Economic Studies, 76 (2009), 451-469.

Ashraf, Nava, "Spousal Control and Intra-Household Decision Making: An Experimental Study in the Philippines," American Economic Review, , (), pp. . 99 (2009), 1245-1277.

Ashraf, Nava, Dean Karlan, and Wesley Yin, "Tying Odysseus to the Mast: Evidence from a Commitment Savings Product in the Philippines," The Quarterly Journal of Economics, 121 (2006), 635-672.

Bandiera, Oriana, Iwan Barankay, and Imran Rasul, "Incentives for Managers and Inequality among Workers: Evidence from a Firm-Level Experiment," The Quarterly Journal of Economics, 122 (2007), 729-773.

Bandiera, Oriana, Iwan Barankay, and Imran Rasul, "Social Connections and Incentives in the Workplace: Evidence from Personnel Data," Econometrica, 77 (2009), 1047-1094.

Banerjee, Abhijit V., Shawn Cole, Esther Duflo, and Leigh Linden, "Remedying Education: Evidence from Two Randomized Experiments in India," The Quarterly Journal of Economics, 122 (2007), 1235-1264.

Beaman, Lori, Raghabendra Chattopadhyay, Esther Duflo, Rohini Pande, and Petia Topalova, "Powerful Women: Does Exposure Reduce Bias?," The Quarterly Journal of Economics, 124 (2009), 1497-1540.

Benjamini, Y. , and Y. Hochberg, "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing.," Journal of the Royal Statistical Society Series B, 57 (1995), 289--300.

Benjamini, Y. , and Y. Hochberg, "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing.," Journal of the Royal Statistical Society Series, B (1995), 289--300.

Bertrand, Marianne, Simeon Djankov, Rema Hanna, and Sendhil Mullainathan, "Obtaining a Driver's License in India: An Experimental Approach to Studying Corruption.," The Quarterly Journal of Economics, 122 (2007), 1639-1676.

Bitler, Marianne P., Jonah B. Gelbach, and Hilary W. Hoynes, "What Mean Impacts Miss: Distributional Effects of Welfare Reform Experiments.," American Economic Review, 96 (2006), 988-1012.

Björkman, Martina, and Jakob Svensson, "Power to the People: Evidence from a Randomized Field Experiment on Community-Based Monitoring in Uganda," The Quarterly Journal of Economics, 124 (2009), 735-769.

Blumenschein, Karen, Glenn C. Blomquist, Magnus Johannesson, Nancy Horn, and Patricia Freeman, "Eliciting Willingness to Pay without Bias: Evidence from a Field Experiment," Economic Journal, 118 (2008), 114-137.

Boisjoly, Johanne, J. Duncan Greg, Michael Kremer, Dan M. Levy, and Jacque Eccles, "Empathy or Antipathy? The Impact of Diversity," American Economic Review, 96 (2006), 1890-1905.

Card, David, Stefano Della Vigna, and Ulrike Malmendier, "The Role of Theory in Field Experiments," Journal of Economic Perspectives, forthcoming (2011).

Carpenter, Jeffrey, Jessica Holmes, and PeterHans Matthews, "Charity Auctions: A Field Experiment," Economic Journal, 118 (2008), 92-113.

Charness, Gary, and Uri Gneezy, "Incentives to Exercise," Econometrica, 77 (2009), 909-931.

Duflo, Esther, William Gale, Jeffrey Liebman, Peter Orszag, and Emmanuel Saez, "Saving Incentives for Low- and Middle-Income Families: Evidence from a Field Experiment with H&R Block.," The Quarterly Journal of Economics, 121 (2006), 1311-1346.

Duflo, Esther, Rachel Glennerster, and Michael Kremer, "Using Randomization in Development Economics Research: A Toolkit," in *Handbook of Development Economics*, (2008).

Duflo, Esther, Michael Kremer, and Jonathan Robinson, "How High Are Rates of Return to Fertilizer? Evidence from Field Experiments in Kenya.," American Economic Review Papers and Proceedings, 98 (2008), 482-488.
Dupas, Pascaline, "What Matters (and What Does Not) in Households' Decision to Invest in Malaria Prevention?," American Economic Review Papers and Proceedings, 99 (2009), 224-230.
Engemann, Kristie M., and Howard J. Wall, "A Journal Ranking for the Ambitious Economist," Federal Reserve Bank of St. Louis Review., 91 (2009), 127-139.
Falk, Armin, "Gift Exchange in the Field," Econometrica, 75 (2007), 1501-1511.
Farcomeni, A. , "A Review of Modern Multiple Hypothesis Testing, with Particular Attention to the False Discovery Proportion. ," Statistical Methods in Medical Research, 17 (2008), 47--388.
Fearon, James D., Macartan Humphreys, and Jeremy M. Weinstein, "Can Development Aid Contribute to Social Cohesion after Civil War? Evidence from a Field Experiment in Post-Conflict Liberia," American Economic Review Papers and Proceedings, 99 (2009), 287-291.
Fehr, Ernst, and Lorenz Goette, "Do Workers Work More If Wages Are High? Evidence from a Randomized Field Experiment.," American Economic Review, 97 (2007), 298-317.
Fisman, Raymond, Sheena S. Iyengar, Emir Kamenica, and Itamar Simonson, "Gender Differences in Mate Selection: Evidence from a Speed Dating Experiment," The Quarterly Journal of Economics, 121 (2006), 673-697.
Gelman, A., J. Hill, and M.Yajima, "Why we (usually) don't worry about multiple comparisons," Mimeo, (2010).
Genovese, C. , and L. Wasserman, "A Stochastic Process Approach to False Discovery Control," Annals of Statistics, (2004), 1035--1061.
Genovese, C. , and L. Wasserman, "Exceedance Control of the False Discovery Proportion," Journal of the American Statistical Association, 101 (2006), 1408--1417.
Gin, Xavier, and Dean Yang, "Insurance, Credit, and Technology Adoption: Field Experimental Evidencefrom Malawi.," Journal of Development Economics, 89 (2009), 1-11.
Gneezy, Uri, and John A. List, "Putting Behavioral Economics to Work: Testing for Gift Exchange in Labor Markets Using Field Experiments.," Econometrica, (), pp. . 74 (2006), 1365-1384.
Green, Donald P., and Holger L. Kern, "Modeling heterogeneous treatment effects in large-scale experiments using Bayesian Additive Regression Tree," Mimeo, (2010).
Hahn, Jinyong, Keisuke Hirano, and Dean Karlan, "Adaptive Experimental Design Using the Propensity Score," Journal of Business & Economic Statistics, 29 (2011), 96-108.
Harrison, Glenn W., and John A. List, "Field Experiments.," Journal of Economic Literature, 42 (2004), 1009-1055.
Hastings, Justine S., and M. Weinstein Jeffrey, "Information, School Choice, and Academic Achievement: Evidence from Two Experiments.," The Quarterly Journal of Economics, 123 (2008), 1373-1414.
Hoffmann, Vivian, "Intrahousehold Allocation of Free and Purchased Mosquito Nets.," American Economic Review Papers and Proceedings, 99 (2009), 236-241.
Imai, Kosuke , and Aaron Strauss, "Estimation of Heterogeneous Treatment Effects from Randomized Experiments, with Application to the Optimal Planning of the Get-out-the-vote Campaign," Political Analysis, 19 (2011), 1-19.
Abdul Latif Jameel Poverty Action Lab (http://www.povertyactionlab.org/,
Jyotsna, Jalan, and E. Somanathan, "The Importance of Being Informed: Experimental Evidence on Demand for Environmental Quality," Journal of Development Economics, 87 (2008), 14-28.
Karlan, Dean, and John A. List., "Does Price Matter in Charitable Giving? Evidence from a Large-Scale Natural Field Experiment.," American Economic Review, 97 (2007), 1774-1793.
Kling, Jeffrey R., Jeffrey B. Liebman, and Lawrence F. Katz, "Experimental Analysis of Neighborhood Effects.," Econometrica, 75 (2007), 83-119.

Kling, Jeffrey R., Jens Ludwig, and Lawrence F. Katz, "Neighborhood Effects on Crime for Female and Male Youth: Evidence from a Randomized Housing Voucher Experiment," The Quarterly Journal of Economics, 120 (2005), 87-130.
Kremer, Michael, and Edward Miguel, "The Illusion of Sustainability," The Quarterly Journal of Economics, 122 (2007), 1007-1065.
Kremer, Michael, Edward Miguel, and Rebecca Thornton, "Incentives to Learn.," The Review of Economics and Statistics, 91 (2009), 437-456.
Landry, Craig E., Andreas Lange, John A. List, Michael K. Price, and Nicholas G. Rupp., "Toward an Understanding of the Economics of Charity: Evidence from a Field Experiment," The Quarterly Journal of Economics, 121 (2006), 747-782.
Lee, David S., "Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects.," Review of Economic Studies, 76 (2009), 1071-1102.
Field Experiments (http://www.fieldexperiments.com/,
Mel, Suresh de, David McKenzie, and Christopher Woodruff, "Returns to Capital in Microenterprises: Evidence from a Field Experiment," The Quarterly Journal of Economics, 123 (2008), 1329-1372.
Newson, Roger, "Multiple-test Procedures and Smile Plots," The Stata Journal, 3 (2003), 100-132.
Olken, Benjamin A., "Monitoring Corruption: Evidence from a Field Experiment in Indonesia," Journal of Political Economy, 115 (2007), 200-249.
Pozo, Susan, and Charles A. Stull, "Requiring a Math Skills Unit: Results of a Randomized Experiment," American Economic Review Papers and Proceedings, 96 (2006), 437-441.
Schochet, Peter Z., John Burghardt, and Sheena McConnell, "Does Job Corps Work? Impact Findings from the National Job Corps Study," American Economic Review, 98 (2008), 1864-1886.
Schultz, Paul, "School Subsidies for the Poor: Evaluating the Mexican Progresa Poverty Program," Journal of Development Economics, 74 (2004), 199-250.
Shang, Jen, and Rachel Croson, "A Field Experiment in Charitable Contribution: The Impact of Social Information on the Voluntary Provision of Public Goods.," Economic Journal, 119 (2009), 1422-1439.
Skoufias, Emmanual, *PROGRESA and Its Impacts on the Welfare of Rural Households in Mexico* (Washington, D.C.: IFPRI, 2005).
Thornton, Rebecca L., "The Demand for, and Impact of, Learning Hiv Status.," American Economic Review, 98 (2008), 1829-1863.
Todd, Petra E., and Kenneth I. Wolpin, "Assessing the Impact of a School Subsidy Program in Mexico: Using a Social Experiment to Validate a Dynamic Behavioral Model of Child Schooling and Fertility," American Economic Review, 96 (2006), 1384-1417.
van der Laan, M., S. Dudoit, and K. Pollard, "Augmentation Procedures for Control of the Generalized Family-Wise Error Rate and Tail Probabilities for the Proportion of False Positives," Statistical Applications in Genetics and Molecular Biology, 3 (2004), 15.
Whitmore, Diane, "Resource and Peer Impacts on Girls' Academic Achievement: Evidence from a Randomized Experiment.," American Economic Review Papers and Proceedings, 95 (2005), 199-203.