JOURNAL OF APPLIED ECONOMETRICS
J. Appl. Econ. **29**: 612–626 (2014)
Published online 28 April 2013 in Wiley Online Library
(wileyonlinelibrary.com) DOI: 10.1002/jae.2327

MULTIPLE TESTING AND HETEROGENEOUS TREATMENT EFFECTS: RE-EVALUATING THE EFFECT OF PROGRESA ON SCHOOL ENROLLMENT

SOOHYUNG LEE[a] AND AZEEM M. SHAIKH[b]*

[a] *Department of Economics, University of Maryland, College Park, MD, USA*
[b] *Department of Economics, University of Chicago, IL, USA*

SUMMARY

The effect of a program or treatment may vary according to observed characteristics. In such a setting, it may not only be of interest to determine whether the program or treatment has an effect on *some* sub-population defined by these observed characteristics, but also to determine for *which* sub-populations, if any, there is an effect. This paper treats this problem as a multiple testing problem in which each null hypothesis in the family of null hypotheses specifies whether the program has an effect on the outcome of interest for a particular sub-population. We develop our methodology in the context of PROGRESA, a large-scale poverty-reduction program in Mexico. In our application, the outcome of interest is the school enrollment rate and the sub-populations are defined by gender and highest grade completed. Under weak assumptions, the testing procedure we construct controls the familywise error rate—the probability of even one false rejection—in finite samples. Similar to earlier studies, we find that the program has a significant effect on the school enrollment rate, but only for a much smaller number of sub-populations when compared to results that do not adjust for multiple testing. Copyright © 2013 John Wiley & Sons, Ltd.

Received 2 July 2012; Revised 27 January 2013

1. INTRODUCTION

The effect of a program or treatment may vary according to observed characteristics, such as gender or age. In such a setting, it may not only be of interest to determine whether the program or treatment has an effect on *some* sub-population defined by these observed characteristics, but also to determine for *which* sub-populations, if any, there is an effect. This paper treats this problem as a multiple testing problem in which each null hypothesis in the family of null hypotheses specifies whether the program has an effect on the outcome of interest for a particular sub-population. For this family of null hypotheses, we construct under weak assumptions a multiple testing procedure that controls the familywise error rate—the probability of even one false rejection—in finite samples.

We require control of the familywise error rate to avoid erroneously finding 'too many' sub-populations for which there is an effect. Indeed, if we were to test each null hypothesis in this family of null hypotheses in a way that controls the usual probability of a Type I error, then the probability of *some* false rejection may be much greater than the nominal level. In other words, the probability of falsely determining that the program or treatment has an effect for some sub-population maybe much greater than the nominal level.

To achieve control of the familywise error rate in finite samples under weak assumptions, we exploit results on stepwise multiple testing procedures developed in Romano and Wolf (2005). The resulting multiple testing procedure differs from classical multiple testing procedures—like Bonferroni and

* Correspondence to: Azeem M. Shaikh, Department of Economics, University of Chicago, 1126 East 59th Street, Chicago, IL 60637, USA. E-mail: amshaikh@uchicago.edu

Holm—in that it incorporates information about the joint dependence structure of the test statistics when determining which null hypotheses to reject. We illustrate the improvement in power by comparing our results with those obtained by Bonferroni and Holm. Similar adjustments for multiple testing have been made by others when making inferences about the effect of a program on multiple outcomes using closely related results on stepwise multiple testing developed in Westfall and Young (1993). See, in particular, Anderson (2008), who analyzes some important early childhood interventions, and Kling *et al*. (2007), who analyze the Moving to Opportunity experiment.

We describe our testing methodology in the context of PROGRESA, a large-scale poverty-reduction program introduced by the Mexican government in 1998. Presently, approximately 2.6 million households in nearly 50,000 rural villages in Mexico are participating in the program. The program is widely credited with decreasing poverty and improving health and educational attainment in regions in which it has been deployed. See, for example, Skoufias (2001), Behrman *et al*. (2005), Djebbari and Smith (2008) and Angelucci and De Giorgi (2009), among others. Similar programs have also been adopted by many other developing countries, including Brazil, Honduras, Jamaica, Chile, Malawi and Zambia.

As described further below, a notable feature of PROGRESA is that treatment status was assigned at the level of the village rather than the individual. For this reason, researchers often use 'clustered standard errors' in their analyses of PROGRESA to allow for dependence across potential outcomes of individuals within villages. It is therefore worth emphasizing that an important feature of our methodology is that it allows for arbitrary dependence across potential outcomes of individuals within villages while controlling the familywise error rate in finite samples—see Remark 3 below. In this way, it accommodates 'spillovers' within villages from treatment.

Among the most commonly analyzed outcomes in previous studies of PROGRESA is school enrollment. See Skoufias *et al*. (2001), Schultz (2004), Todd and Wolpin (2006) and Attanasio *et al*. (2012). We therefore focus on this outcome and analyze the effect of PROGRESA on school enrollment for different sub-populations defined by gender and highest grade completed. Gender varies over two possible values and highest grade completed varies over 10 possible values, so there are 20 sub-populations. Even after adjusting for multiple testing, we find, similar to existing studies, that the program has a significant effect on the school enrollment rate, but in a much smaller number of sub-populations when compared to results that do not adjust for multiple testing.

The remainder of this paper is organized as follows. In Section 2, we provide some limited background information on PROGRESA, including most importantly a description of the way in which treatment status was assigned. In Section 3, we formally describe our set-up and assumptions before developing our testing procedures in Section 4. There we first discuss the problem of testing a single null hypothesis, before turning our attention to the problem of testing multiple null hypotheses. In Section 5, we present the results of applying our methodology to the data from PROGRESA. Section 6 concludes.

2. DESCRIPTION OF PROGRESA

PROGRESA is a large-scale poverty-reduction program introduced by the Mexican government in 1998. The program specifically targeted poverty in rural villages in Mexico by providing cash payments to households in exchange for regular school attendance as well as provisions for free health services, including nutrition supplements and educational seminars on nutrition and hygiene. The program was expanded in May 2000 to the rest of Mexico, after which it was no longer referred to as PROGRESA but as Oportunidades. In this paper, we only use data from the time period before the program was expanded to the rest of Mexico. We now describe the most important aspects of the program for our analysis in further detail.

DOI: 10.1002/jae

2.1. Eligibility

Eligibility criteria for the program were determined according to two baseline surveys conducted in 506 rural villages in Mexico in October 1997 and March 1998. Households that were sufficiently poor according to these baseline surveys were deemed eligible for the program. In May 1998, each of these 506 villages was independently selected into treatment with probability 2/3. All eligible households in 320 of the 506 villages selected in this way were invited to participate in the program. This accounted for approximately 78% of the households in these 320 villages. Nearly all households invited to participate in the program chose to do so.

2.2. Treatment

Eligible households in treated villages received cash transfers every 2 months for each grade-eligible child who attended school regularly. Regular attendance was defined as an attendance rate of at least 85%. Failure to fulfill this requirement would lead to loss of the benefit. The size of the cash transfer varied according to gender and highest grade completed. In particular, the subsidy increased when the child entered 9th grade and the subsidy for girls in 7th through 9th grades was larger than than for boys in 7th through 9th grades. This variation in the size of the cash transfer was intended to offset the opportunity costs of schooling for older children and to promote gender equality in schooling. This variation also makes sub-populations defined by gender and highest grade completed natural for analysis.

2.3. Evaluation

After the program started, three follow-up surveys were administered to all eligible households in the 506 villages in October 1998, May 1999 and November 1999. These surveys included a wide variety of questions, including educational attainment, health, consumption and household structure, and have been used by many researchers to evaluate the impact of the program on various outcomes.

3. SET-UP AND ASSUMPTIONS

3.1. Set-Up

We index villages by $j \in J$, (eligible) individuals in villages by $i \in I_j$ and time periods by $s \in S = S_{\text{base}} \cup S_{\text{follow}-\text{up}}$, where S_{base} and $S_{\text{follow}-\text{up}}$ are disjoint sets of time periods indexing, respectively, the two baseline surveys in October 1997 and March 1998 and the three follow-up surveys in October 1998, May 1999 and November 1999.

Denote by $Y_{i,j,s}(0)$ the outcome of the ith person in the jth village in the sth time period if the jth village were not treated, by $Y_{i,j,s}(1)$ the outcome of the ith person in the jth village in the sth time period if the jth village were treated, by D_j the treatment status of the jth village, and by $Z_{i,j}$ observed characteristics of the ith person in the jth village that will be used to define the different sub-populations of interest. Here,

$$Z_{i,j} = \left(G_{i,j}, A_{i,j}\right)$$

where $G_{i,j}$ is the gender of the ith person in the jth village and $A_{i,j}$ is the highest grade completed by the ith person in the jth village in the second baseline survey in March 1998. In this notation, the observed outcome of the ith person in the jth village in the sth time period is given by

DOI: 10.1002/jae

$$Y_{i,j,s} = \begin{cases} D_j Y_{i,j,s}(1) + \left(1 - D_j\right) Y_{i,j,s}(0) & \text{if } s \in S_{\text{follow-up}} \\ Y_{i,j,s}(0) & \text{if } s \in S_{\text{base}} \end{cases} \tag{1}$$

In order to state our null hypotheses precisely, it is useful to introduce the following shorthand notation. Let

$$D = \left(D_j : j \in J\right)$$

and define

$$\mathcal{D} = \text{supp}(D)$$
$$\mathcal{Z} = \bigcup_{i \in I_j, j \in J} \text{supp}\left(Z_{i,j}\right)$$

i.e. the set of possible values for D and Z. For $z \in \mathcal{Z}$ and $d \in \mathcal{D}$, define

$$Y_z^{\text{base}} = \left(Y_{i,j,s} : i \in I_j, j \in J, s \in S_{\text{base}}, Z_{i,j} = z\right)$$
$$Y_z^{\text{follow-up}}(d) = \left(Y_{i,j,s}\left(d_j\right) : i \in I_j, j \in J, s \in S_{\text{follow-up}}, Z_{i,j} = z\right)$$
$$Y_z = \left(Y_{i,j,s} : i \in I_j, j \in J, s \in S, Z_{i,j} = z\right)$$

In other words, D is the vector of treatment status for the villages, Y_z^{base} is the vector of observed outcomes in the baseline surveys for people in the sub-population corresponding to z, $Y_z^{\text{follow-up}}(d)$ is the vector of potential outcomes in the counterfactual state of the world where treatment status is given by d in the follow-up surveys for people in the sub-population corresponding to z, and Y_z is simply the vectorof observed outcomes for people in the sub-population corresponding to z.

Denote by P the distribution of

$$\left(\left(Y_z^{\text{base}}, Y_z^{\text{follow-up}}(d) : d \in \mathcal{D}, z \in \mathcal{Z}\right), D\right)$$

which is assumed to lie in a large class of possible distributions Ω. The assumptions we impose on Ω are described in Section 3.2 below. For each $z \in \mathcal{Z}$, let

$$\omega_z = \left\{P \in \Omega : Y_z^{\text{follow-up}}(d) \text{ does not depend on } d\right\} \tag{2}$$

In other words, ω_z is the set of distributions for which the program had *no* effect on outcomes in the sense that potential outcomes in the follow-up surveys for people in the sub-population corresponding to z do not depend on the counterfactual state of the world.

In this notation, our goal is to test the family of null hypotheses

$$H_z : P \in \omega_z \text{ for } z \in \mathcal{Z} \tag{3}$$

in a way that controls the familywise error rate—the probability of even one false rejection—in finite samples. More formally, let $\mathcal{Z}_0(P)$ denote the set of true null hypotheses, i.e.

$$\mathcal{Z}_0(P) = \{z \in \mathcal{Z} : P \in \omega_z\}$$

and define

$$\text{FWER}_P = P\{\text{reject} \geq 1 \ H_z \text{ with } z \in \mathcal{Z}_0(P)\}$$

In this notation, our goal is to test the family of null hypotheses (3) in a way that satisfies

$$\text{FWER}_P \leq \alpha \text{ for all } P \in \Omega \tag{4}$$

for some pre-specified value of $\alpha \in (0,1)$ under weak assumptions on Ω.

DOI: 10.1002/jae

Remark 1. By requiring that our testing procedure satisfy equation (4), we ensure that the probability that all of the the null hypotheses rejected by our procedure are false is at least $1-\alpha$. The recent literature on multiple testing has considered error rates less stringent than the familywise error rate. One example is the k-familywise error rate—the probability of k or more false rejections for some $k \geq 1$. Another example is based on the false discovery proportion—the ratio of false rejections to total rejections (defined to be zero when there are no rejections at all). With such error rates, one can only guarantee that the probability that 'most' of the null hypotheses rejected by the procedure are false is at least $1-\alpha$. However, such procedures may have much greater ability to detect false null hypotheses. This feature may be especially valuable when the number of null hypotheses under consideration is very large. See Romano *et al*. (2008) for a discussion of some procedures for control of such error rates. We do not pursue such error rates here because in our application the number of null hypotheses under consideration is relatively small. ■

3.2. Assumptions

In this section, we describe the assumptions we impose upon Ω. The requirements are weak.

Assumption 1. For any $P \in \Omega$,

$$\left(Y_z^{\text{base}}, Y_z^{\text{follow-up}}(d) : d \in \mathcal{D}, z \in \mathcal{Z}\right) \perp\!\!\!\perp D$$

under P.

Our first assumption simply states that the assignment of treatment status was in fact random in the sense that outcomes in the baseline surveys and potential outcomes in the follow-up surveys are independent of treatment status.

Assumption 2. For any $P \in \Omega$,

$$D \sim \text{Bernoulli}(2/3)^{|J|}$$

under P, i.e., $D_j, j \in J$ is an i.i.d. sequence of Bernoulli(2/3) random variables.

Our second assumption simply states the precise way in which treatment status was assigned, i. e. that each village was independently selected for treatment with probability 2/3. It can be weakened considerably. For example, it suffices that the distribution of D is exchangeable in the sense that the distribution of D remains invariant with respect to permutations of its components. See Section 4.1 below for further details.

Remark 2. When treatment status is assigned in more complicated ways, such as stratification, the distribution of D will typically not be exchangeable, but other symmetries in the distribution of treatment status may persist. These symmetries may be exploited in a similar way to exchangeability here to construct tests with finite-sample validity. See Heckman *et al*. (2010, 2011) for examples of this approach. ■

Remark 3. We emphasize that our analysis below will only require Assumptions 1 and 2. In particular, we will make no restrictive assumptions about the dependence structure of outcomes in the baseline surveys and potential outcomes in the follow-up surveys of individuals within the same village. Our methodology therefore accommodates 'spillovers' within villages from treatment. See Barrios *et al*. (2012) for related discussion. ■

DOI: 10.1002/jae

4. METHODOLOGY

In Section 4.1 below, we consider the problem of testing a single null hypothesis of the form

$$H_{\mathcal{Z}'} : P \in \bigcap_{z \in \mathcal{Z}'} \omega_z \tag{5}$$

for some $\mathcal{Z}' \subseteq \mathcal{Z}$ in a way that controls the usual probability of a Type I error at level α. When $\mathcal{Z}'$ is not a singleton, such a null hypothesis is sometimes referred to as a joint null hypothesis. Importantly, rejection of $H_{\mathcal{Z}'}$ allows the researcher to conclude that the program has an effect on potential outcomes for *some* $z \in \mathcal{Z}'$,but does not allow the researcher to conclude for *which* $z \in \mathcal{Z}'$ the program has an effect on potential outcomes. We therefore extend these methods in Section 4.2 to test the family of null hypotheses (3) in a way that satisfies equation (4).

4.1. Testing a Single (Joint) Null Hypothesis

Let $\mathcal{Z}' \subseteq \mathcal{Z}$ be given. In order to describe our test of the single (joint) null hypothesis (5), we first require a test statistic. To this end, define

$$X_{\mathcal{Z}'} = ((Y_z : z \in \mathcal{Z}'), D)$$

and let

$$T_{\mathcal{Z}'} = T_{\mathcal{Z}'}(X_{\mathcal{Z}'})$$

be a test statistic for equation (5). Note that we impose the mild requirement that $T_{\mathcal{Z}'}$ only depend on $X_{\mathcal{Z}'}$. In particular, we assume that it does not depend on Y_z for $z \in \mathcal{Z}'$. We assume further that large values of $T_{\mathcal{Z}'}$ provide evidence against the null hypothesis.

We now describe our construction of a critical value with which to compare $T_{\mathcal{Z}'}$. For this purpose, the following lemma is useful.

Lemma 1. Let $\mathcal{Z}' \subseteq \mathcal{Z}$. If Assumption 1 holds, then

$$(Y_z : z \in \mathcal{Z}') \perp\!\!\!\perp \mathrm{D}$$

under any $P \in \bigcap_{z \in \mathcal{Z}'} \omega_z$.

Proof. From Assumption 1 we have that

$$\left(Y_z^{\text{base}}, Y_z^{\text{follow-up}}(d) : d \in \mathcal{D}, z \in \mathcal{Z}'\right) \perp\!\!\!\perp \mathrm{D}$$

Since $P \in \bigcap_{z \in \mathcal{Z}'} \omega_z$, we have that

$$\left(Y_z^{\text{base}}, Y_z^{\text{follow-up}}(d)\right) = Y_z$$

for all $d \in \mathcal{D}$ and $z \in \mathcal{Z}'$. The desired result thus follows. ■

In order to describe an important implication of Lemma 1, it is useful to introduce the following notation. Denote by $|J|$ the number of elements in the set J. Let $\mathbf{G}$ be the group of permutations on $|J|$ elements and define the action of $g \in \mathbf{G}$ on a $|J|$-dimensional vector v as follows:

DOI: 10.1002/jae

$$gv = \left(v_{g(1)}, \ldots, v_{g(|J|)}\right)$$

Here, the term group is used to describe the mathematical structure of the set of all permutations on $|J|$ elements. Similarly, define the action of $g \in \mathbf{G}$ on $X_{\mathcal{Z}'}$ as follows:

$$gX_{\mathcal{Z}'} = ((Y_z : z \in \mathcal{Z}'), gD) \tag{6}$$

Note that Lemma 1 implies that

$$gX_{\mathcal{Z}'} \overset{d}{=} X_{\mathcal{Z}'}$$

whenever $P \in \cap_{z \in \mathcal{Z}'} \omega_z$ and $g \in \mathbf{G}$. This symmetry in the distribution of the data suggests that we can construct a critical value with which to compare our test statistic by re-evaluating on the data $gX_{\mathcal{Z}'}$ for each $g \in \mathbf{G}$. More specifically, we can use

$$c_{\mathcal{Z}'}(X_{\mathcal{Z}'}, 1-\alpha) = \inf\left\{ t \in \mathbf{R} : \frac{1}{|\mathbf{G}|} \sum_{g \in \mathbf{G}} I\{T_{\mathcal{Z}'}(gX_{\mathcal{Z}'}) \leq t\} \geq 1-\alpha \right\} \tag{7}$$

as our critical value.

The following theorem formalizes the test proposed above.

Theorem 1. Under Assumptions 1 and 2, the test that rejects $H_{\mathcal{Z}'}$ whenever

$$T_{\mathcal{Z}'}(X_{\mathcal{Z}'}) > c_{\mathcal{Z}'}(X_{\mathcal{Z}'}, 1-\alpha)$$

where $c_{\mathcal{Z}'}\left(X_{\mathcal{Z}'}, 1-\alpha\right)$ is defined by equation (7) controls the usual probability of a Type I error at level α, i.e.

$$P\{T_{\mathcal{Z}'}(X_{\mathcal{Z}'}) > c_{\mathcal{Z}'}(X_{\mathcal{Z}'}, 1-\alpha)\} \leq \alpha$$

for all $P \in \bigcap_{z \in \mathcal{Z}'} \omega_z$.

Proof. Consider $P \in \bigcap_{z \in \mathcal{Z}'} \omega_z$. Define

$$\varphi\left(X_{\mathcal{Z}'}\right) = I\left\{T_{\mathcal{Z}'}\left(X_{\mathcal{Z}'}\right) > c_{\mathcal{Z}'}\left(X_{\mathcal{Z}'}, 1-\alpha\right)\right\}$$

Recall that Lemma 1 implies that equation (6) holds under any such P. Hence,

$$\begin{aligned} E_P\left[\sum_{g \in \mathbf{G}} \phi(gX_{\mathcal{Z}'})\right] &= \sum_{g \in \mathbf{G}} E_P[\phi(gX_{\mathcal{Z}'})] \\ &= \sum_{g \in \mathbf{G}} E_P[\phi(X_{\mathcal{Z}'})] \\ &= |\mathbf{G}| E_P[\phi(X_{\mathcal{Z}'})] \end{aligned} \tag{8}$$

On the other hand, because $\mathbf{G}$ is a group,

DOI: 10.1002/jae

$$c_{\mathcal{Z}'}(gX_{\mathcal{Z}'}, 1-\alpha) = c_{\mathcal{Z}'}(X_{\mathcal{Z}'}, 1-\alpha)$$

for any $g \in \mathbf{G}$. We therefore have that

$$E_P\left[\sum_{g\in\mathbf{G}} \phi(gX_{\mathcal{Z}'})\right] \leq |\mathbf{G}|\alpha$$

It follows from equations (8) and (9) that

$$E_P[\phi(X_{\mathcal{Z}'})] \leq \alpha \tag{9}$$

from which the desired conclusion follows immediately. ■

Remark 4. Once equation (6) is established, the proof of Theorem 1 follows from the usual arguments that underlie the validity of 'randomization tests'. See, for example, chapter 15 of Lehmann and Romano (2005) for a textbook discussion of such methods. Nevertheless, we include the details of the argument for completeness. ■

Remark 5. Note that $c_{\mathcal{Z}'}(X_{\mathcal{Z}'}, 1-\alpha)$ defined in equation (7) requires computing $T_{\mathcal{Z}'}(gX_{\mathcal{Z}^e})$ for every $g \in \mathbf{G}$. Since $|\mathbf{G}|$ is large in our case, this is computationally infeasible. We therefore resort to the stochastic approximation to equation (7) defined as follows:

$$\hat{c}_{\mathcal{Z}'}(X_{\mathcal{Z}'}, 1-\alpha) = \inf\left\{t \in \mathbf{R} : \frac{1}{B}\sum_{1\leq i\leq B} I\left\{T_{\mathcal{Z}'}(g_i X_{\mathcal{Z}'}) \leq t\right\} \geq 1-\alpha\right\} \tag{10}$$

where g_1 is the identity element and $g_2, \ldots, g_B$ are i.i.d. Unif($\mathbf{G}$) for some fixed value B. Theorem 1 remains true with $\hat{c}_{\mathcal{Z}'}(X_{\mathcal{Z}'}, 1-\alpha)$ in place of $c_{\mathcal{Z}'}(X_{\mathcal{Z}'}, 1-\alpha)$. See section 15.2 of Lehmann and Romano (2005) for details. In our empirical application below, we compute critical values using such an approximation with $B = 3000$. ■

Remark 6. Note that our analysis only depends on the specific definition of ω_z in equation (2) through Lemma 1. Hence, our analysis is unaffected by changes in the definition of ω_z provided that Lemma 1 continues to hold. For instance, if $\mathcal{Z}'$ were a singleton, then Lemma 1 would continue to hold for

$$\omega_z = \left\{P \in \Omega : \text{the distribution of } \left(Y_z^{\text{base}}, Y_z^{\text{follow-up}}(d)\right) \text{does not depend on } d\right\} \tag{11}$$

As a result, the methodology described above may be used to test the null hypothesis that $P \in \omega_z$ for ω_z defined by equation (11) in a way that controls the usual probability of a Type I error at level α. The need to consider null hypotheses of the form in equation (5) with $\mathcal{Z}'$ not a singleton only arises because of our desire in Section 4.2 below to improve upon multiple testing procedures like Bonferroni and Holm-type corrections. If this were not the case, then it would be possible to define ω_z as in equation (11) throughout our analysis. ■

DOI: 10.1002/jae

4.2. Testing Multiple Null Hypotheses

We now return to the problem of testing the family of null hypotheses (3) in a way that satisfies equation (4). Under Assumptions 1 and 2, it is straightforward to calculate a p-value $\hat{p}_z$ for each H_z using Theorem 1 by simply applying the theorem with $\mathcal{Z}' = \{z\}$ and computing the smallest value of α for which the null hypothesis is rejected. The resulting p-values will satisfy

$$P\{\hat{p}_z \leq u\} \leq u$$

for all $u \in (0,1)$ and $P \in \omega_z$. A crude solution to the multiplicity problem would therefore be to apply a Bonferroni or Holm-type correction. See, for example, Romano *et al*. (2010) for further details. Such an approach would indeed satisfy equation (4), as desired, but implicitly relies upon a 'least favorable' dependence structure among the p-values. To the extent that the true dependence structure differs from this 'least favorable' one, improvements may be possible. For that reason, we use results on stepwise multiple testing procedures developed by Romano and Wolf (2005) for control of the familywise error rate to implicitly incorporate information about the dependence structure when deciding which null hypotheses to reject. Our discussion follows that in Romano and Shaikh (2010), wherein the algorithm is generalized to allow for possibly uncountably many null hypotheses. See also Heckman *et al*. (2010, 2011), where a similar procedure is employed to re-evaluate the High/Scope Perry Preschool program.

In order to describe our testing procedure, we first require a test statistic for each null hypothesis such that large values of the test statistic provide evidence against the null hypothesis. As before, we impose the requirement that the test statistic for H_z depends only on $X_{\{z\}}$. Denote such a test statistic by $T_z(X_{\{z\}})$. Next, for $\mathcal{Z}' \subseteq \mathcal{Z}$, define

$$T_{\mathcal{Z}'}(X_{\mathcal{Z}'}) = \max_{z \in \mathcal{Z}'} T_z\left(X_{\{z\}}\right)$$

Finally, for $\mathcal{Z}' \subseteq \mathcal{Z}$, denote by $c_{\mathcal{Z}'}(X_{\mathcal{Z}'}, 1-\alpha)$ the critical value defined in equation (7) with this choice of$T_{\mathcal{Z}'}(X_{\mathcal{Z}'})$.

Our testing procedure is summarized in the following algorithm.

Algorithm 1.
Step 1. Set $\mathcal{Z}_1 = \mathcal{Z}$. If

$$\max_{z \in \mathcal{Z}_1} T_z\left(X_{\{z\}}\right) \leq c_{\mathcal{Z}_1}(X_{\mathcal{Z}_1}, 1-\alpha)$$

then stop and reject no null hypotheses; otherwise, reject any H_z with

$$T_z\left(X_{\{z\}}\right) > c_{\mathcal{Z}_1}(X_{\mathcal{Z}_1}, 1-\alpha)$$

and go to Step 2.

$$\vdots$$

Step j. Let $\mathcal{Z}_j$ denote the indices of remaining null hypotheses. If

$$\max_{z \in \mathcal{Z}_j} T_z\left(X_{\{z\}}\right) \leq c_{\mathcal{Z}_j}\left(X_{\mathcal{Z}_j}, 1-\alpha\right)$$

then stop and reject no null hypotheses; otherwise, reject any H_z with

DOI: 10.1002/jae

$$T_z\left(X_{\{z\}}\right) > c_{\mathcal{Z}_j}\left(X_{\mathcal{Z}_j}, 1-\alpha\right)$$

and go to Step $j+1$.

⋮

Theorem 2. Under Assumptions 1 and 2, Algorithm 1 satisfies equation (4).

Proof. The claim follows from Theorem 1 and arguments given in Romano and Wolf (2005) or Romano and Shaikh (2010). Since the argument is brief, we include it here for completeness.

Suppose that a false rejection occurs. Let $\hat{j}$ be the *smallest* step at which a false rejection occurs. By the minimality of $\hat{j}$, we must have

$$\mathcal{Z}_{\hat{j}} \supseteq \mathcal{Z}_0(P) \tag{12}$$

It follows that

$$c_{\mathcal{Z}_{\hat{j}}}\left(X_{\mathcal{Z}_{\hat{j}}}, 1-\alpha\right) \geq c_{\mathcal{Z}_0(P)}\left(X_{\mathcal{Z}_0(P)}, 1-\alpha\right) \tag{13}$$

Since a false rejection occurred, we must also have

$$\max_{z\in\mathcal{Z}_0(P)} T_z\left(X_{\{z\}}\right) > c_{\mathcal{Z}_{\hat{j}}}\left(X_{\mathcal{Z}_{\hat{j}}}, 1-\alpha\right)$$

Hence,

$$\max_{z\in\mathcal{Z}_0(P)} T_z\left(X_{\{z\}}\right) > c_{\mathcal{Z}_0(P)}\left(X_{\mathcal{Z}_0(P)}, 1-\alpha\right)$$

and the probability of this event is bounded above by α by Theorem 1. ■

We conclude this section by discussing the choice of $T_z(X_{\{z\}})$ in Algorithm 1. While a researcher may use any choice, we choose to use $T_z\left(X_{\{z\}}\right) = 1-\hat{p}_z$, where $\hat{p}_z$ is a (multiplicity-unadjusted) p-value for testing H_z. As described at the beginning of Section 4.2, such a p-value may be computed as the smallest value of α for which H_z is rejected when applying Theorem 1. We compute p-values for each H_z in this way using two different choices of underlying test statistic. In our first specification, the underlying test statistic is the following 'difference in means':

$$\frac{\sum_{j\in J, i\in I_j, s\in S_{\text{follow-up}}: Z_{i,j}=z} D_j Y_{i,j,s}}{\sum_{j\in J, i\in I_j, s\in S_{\text{follow-up}}: Z_{i,j}=z} D_j} - \frac{\sum_{j\in J, i\in I_j, s\in S_{\text{follow-up}}: Z_{i,j}=z} \left(1-D_j\right) Y_{i,j,s}}{\sum_{j\in J, i\in I_j, s\in S_{\text{follow-up}}: Z_{i,j}=z} \left(1-D_j\right)} \tag{14}$$

In our second specification, the underlying test statistic is the following 'difference in differences':

DOI: 10.1002/jae

$$\left\{\frac{\sum_{j\in J,i\in I_j,s\in S_{\text{follow}-\text{up}}:Z_{i,j}=z} D_j Y_{i,j,s}}{\sum_{j\in J,i\in I_j,s\in S_{\text{follow}-\text{up}}:Z_{i,j}=z} D_j} - \frac{\sum_{j\in J,i\in I_j,s\in S_{\text{follow}-\text{up}}:Z_{i,j}=z} \left(1-D_j\right) Y_{i,j,s}}{\sum_{j\in J,i\in I_j,s\in S_{\text{follow}-\text{up}}:Z_{i,j}=z} \left(1-D_j\right)}\right\} - \left\{\frac{\sum_{j\in J,i\in I_j,s\in S_{\text{base}}:Z_{i,j}=z} D_j Y_{i,j,s}}{\sum_{j\in J,i\in I_j,s\in S_{\text{base}}:Z_{i,j}=z} D_j} - \frac{\sum_{j\in J,i\in I_j,s\in S_{\text{base}}:Z_{i,j}=z} \left(1-D_j\right) Y_{i,j,s}}{\sum_{j\in J,i\in I_j,s\in S_{\text{base}}:Z_{i,j}=z} \left(1-D_j\right)}\right\} \tag{15}$$

It may also be desirable to Studentize equation (14) or (15) in some way, but we do not pursue such modifications here. By using p-values based on these test statistics rather than the test statistics themselves, we ensure that the testing procedure is 'balanced' in the sense that the probability of rejecting any true null hypothesis is roughly equal. See Romano and Wolf (2010) for further details. Note further that this approach remains computationally feasible because the same permutations used to compute $\hat{p}_z$ may be used in Algorithm 1.

Remark 7. Theorem 2 remains true if critical values $\hat{c}_{\mathcal{Z}'}\left(X_{\mathcal{Z}'}, 1-\alpha\right)$ defined in equation (10) are used in place of $c_{\mathcal{Z}'}\left(X_{\mathcal{Z}'}, 1-\alpha\right)$ provided that $g_1, \ldots, g_B$ remain the same throughout the algorithm. If this is not the case, then equation (13) may not hold.

Remark 8. It is straightforward to calculate a multiplicity-adjusted p-value $\hat{p}_z^{\text{adj}}$ for each H_z using Theorem 2 by simply computing the smallest value of α for which each null hypothesis isrejected. The resulting p-values have the property that the procedure that rejects any H_z with $\hat{p}_z^{\text{adj}} \leq \alpha$ satisfies (4).

Remark 9. As an alternative to equation (14) or (15) as the choice of underlying test statistic for testing H_z, a researcher may wish to use the coefficient on D_j from a linear regression involving $X_{\{z\}}$. As an example, one might consider the linear regression of $Y_{i,j,s}$ for some $s \in S_{\text{follow}-\text{up}}$ on D_j using only those observations for which $Z_{i,j} = z$.

Remark 10. The first step of Algorithm 1 may be interpreted as a joint test of the null hypothesis that the program has no effect on *any* of the sub-populations. A similar testing problem is considered in Mitnik *et al*. (2008). The authors there note, as was mentioned previously, that rejection of such a null hypothesis only allows one to conclude that there is *some* sub-population for which the program has an effect. In particular, it does not permit one to answer the more ambitious question of for which sub-populations the program had an effect.

5. EMPIRICAL RESULTS

We use all five of the surveys, i.e. the two baseline and three follow-up surveys, to examine the effect of PROGRESA on school enrollment. Data from each of the five surveys are available from the official website of the program.[1] Following Schultz (2004), we restrict our sample to children whose ages were between 6 and 16 at the time of the first baseline survey in October 1997 and were surveyed in all five of the surveys. See, in particular, Table 3 in Schultz (2004).

[1] http://evaluacion.oportunidades.gob.mx:8010/en/index.php.

DOI: 10.1002/jae

Table I. Summary statistics

	Baseline				Follow-up					
	Oct. 1997		Mar. 1998		Oct. 1998		May 1999		Nov. 1999	
	Treated (1)	Control (2)	Treated (3)	Control (4)	Treated (5)	Control (6)	Treated (7)	Control (8)	Treated (9)	Control (10)
No. children	9,388	5,608	7,718	4,650	9,672	5,783	9,767	5,843	9,763	5,839
Fraction of girls	0.479	0.492 [0.109]	0.480	0.500 [0.029]	0.479	0.493 [0.096]	0.480	0.494 [0.093]	0.479	0.494 [0.082]
Avg. age	10.330	10.353 [0.646]	10.486	10.510 [0.657]	11.281	11.300 [0.725]	11.942	11.923 [0.754]	12.379	12.425 [0.385]
School enrollment										
All	0.857	0.856 [0.858]	0.875	0.873 [0.743]	0.845	0.815 [0.000]	0.829	0.792 [0.000]	0.785	0.744 [0.000]
Boys only	0.869	0.870 [0.841]	0.885	0.883 [0.818]	0.855	0.832 [0.006]	0.836	0.803 [0.000]	0.782	0.759 [0.022]
Girls only	0.844	0.841 [0.722]	0.864	0.862 [0.890]	0.833	0.797 [0.000]	0.821	0.780 [0.000]	0.788	0.728 [0.000]

Note: Bracketed numbers correspond to a *p*-value from an (asymptotic) test of whether the corresponding difference between the treated and untreated children equals zero.

DOI: 10.1002/jae

5.1. Summary Statistics

Table 1 displays for each of the five surveys the total number of treated and untreated children as well as the following five quantities for both the treated and untreated children: (i) fraction of girls; (ii) average age; (iii) average school enrollment rate (for both girls and boys); (iv) average school enrollment rate for boys; and (v) average school enrollment rate for girls. For each of these five quantities, we also report a p-value from an (asymptotic) test of whether the corresponding difference between the treated and untreated children equals zero.

In the baseline surveys, we see little evidence of any differences between the treated and untreated children. This is consistent with the randomization of treatment status. See, for example, Behrman and Todd (1999). On the other hand, in the follow-up surveys, we see evidence that the program had an effect on the school enrollment rate (both for boys and girls together and for boys and girls separately).

5.2. Main Results

Table 2 displays for each of the 20 sub-populations of interest the following eight quantities: column 1 displays a (multiplicity-unadjusted) p-value computed using Theorem 1; column 2 displays a (multiplicity-adjusted) p-value computed using Theorem 2; column 3 displays a (multiplicity-adjusted) p-value obtained by applying a Bonferroni adjustment to the p-values in column 1; column 4 displays a (multiplicity-adjusted) p-value obtained by applying a Holm adjustment to the p-values in column 1; column 5 displays a (multiplicity-unadjusted) p-value computed using Theorem 2; column 6 displays

Table II. Main results

	DI				DID			
		Multiplicity adj.				Multiplicity adj.		
HCG	Unadj. Thm. 4.2 (1)	Thm. 4.2 (2)	Bonf. (3)	Holm (4)	Unadj. Thm. 4.1 (5)	Thm. 4.2 (6)	Bonf. (7)	Holm (8)
Girls								
0	0.908	0.908	1.000	0.908	0.810	0.999	1.000	1.000
1	0.617	0.940	1.000	1.000	0.905	0.905	1.000	0.905
2	0.237	0.939	1.000	1.000	0.032**	0.382	0.633	0.475
3	0.030**	0.344	0.593	0.445	0.709	1.000	1.000	1.000
4	0.037**	0.373	0.746	0.485	0.000***	0.000***	0.007***	0.007***
5	0.001***	0.017**	0.027**	0.023**	0.020**	0.295	0.407	0.346
6	0.001***	0.006***	0.013**	0.012**	0.018**	0.276	0.360	0.324
7	0.549	0.979	1.000	1.000	0.040**	0.401	0.793	0.515
8	0.544	0.988	1.000	1.000	0.396	0.986	1.000	1.000
9+	0.256	0.937	1.000	1.000	0.293	0.967	1.000	1.000
Boys								
0	0.727	0.926	1.000	1.000	0.812	0.992	1.000	1.000
1	0.001***	0.007***	0.013**	0.013**	0.031**	0.394	0.620	0.496
2	0.386	0.985	1.000	1.000	0.862	0.981	1.000	1.000
3	0.001***	0.006***	0.013*	0.013*	0.003***	0.055*	0.067*	0.063*
4	0.230	0.949	1.000	1.000	0.806	1.000	1.000	1.000
5	0.025**	0.311	0.493	0.395	0.188	0.920	1.000	1.000
6	0.034**	0.368	0.686	0.481	0.036**	0.397	0.720	0.504
7	0.598	0.971	1.000	1.000	0.786	1.000	1.000	1.000
8	0.393	0.979	1.000	1.000	0.439	0.990	1.000	1.000
9+	0.414	0.974	1.000	1.000	0.269	0.967	1.000	1.000

Note: HGC, 'highest grade completed'; DI, 'difference in means'; DID, 'difference in differences'. Single, double and triple asterisks indicate p-values less than 10%, 5% and 1%, respectively.

DOI: 10.1002/jae

a (multiplicity-adjusted) *p*-value computed using Theorem 2; column 7 displays a (multiplicity-adjusted) *p*-value obtained by applying a Bonferroni adjustment to the *p*-values in column 5; column 8 displays a (multiplicity-adjusted) *p*-value obtained by applying a Holm adjustment to the *p*-values in column 5.

Regardless of the choice of underlying test statistic, we see in columns 2 and 6 that after accounting for multiple testing we only find that the program had an effect on the school enrollment rate for a limited number of sub-populations. In particular, if, out of concern about initial differences in the school enrollment rate, we were to use a test statistic based on equation (15) ('differences-in-differences'), then we would find that the program only had an effect on two sub-populations, namely girls whose highest grade completed was four and boys whose highest grade completed was three. This finding may merit further investigation, especially since the *p*-values for girls and boys for all other values of highest grade completed are quite large. On the other hand, we see in columns 1 and 5 that by ignoring the multiplicity of comparisons being made one would conclude that the program had an effect on the school enrollment rate for a much larger number of sub-populations. In particular, if we were again to use a test statistic based on equation (15), then we would find that the program had an effect on eight different sub-populations.

As mentioned previously, the *p*-values from Theorem 2 improve upon *p*-values obtained by applying Bonferroni or Holm adjustments by incorporating information about the joint dependence structure of the test statistics when determining which null hypotheses to reject. This feature is evident in Table 2, as the *p*-values from columns 2 and 6 are always smaller (and sometimes by a considerable margin) than the *p*-values in columns 3–4 and 7–8, respectively.

6. CONCLUSION

In this paper, we provide a framework for determining the sub-populations for which a program or treatment has an effect on an outcome of interest. More specifically, we develop under weak assumptions a procedure for testing the family of nullhypotheses in which each null hypothesis specifies whether the program has an effect on the outcome of interest for a particular sub-population in a way that controls the familywise error rate in finite samples. We have applied our methodology to data from PROGRESA and examined the effect of the program on school enrollment and how this effect varies by gender and highest grade completed. Notably, our methodology does not require any restrictions on the dependence structure across potential outcomes of individuals within villages. Similar to earlier studies, we find that the program has a significant effect on the school enrollment rate, but only for a much smaller number of sub-populations when compared to results that do not adjust for multiple testing. We believe our framework will be useful to researchers analyzing similar questions in other empirical settings.

ACKNOWLEDGEMENTS

We thank Aprajit Mahajan, Joseph Romano, Andres Santos, Edward Vytlacil and Joanne Yoong for detailed comments, and Maria Prada for excellent research assistance. The research of the second author has been supported by the National Science Foundation grant DMS-0820310 and the Alfred P. Sloan Foundation.

REFERENCES

Anderson M. 2008. Multiple inference and gender differences in the effects of early intervention: a reevaluation of the Abecedarian, Perry Preschool, and Early Training projects. *Journal of the American Statistical Association* **103**: 1481–1495.

Angelucci M, De Giorgi G. 2009. Indirect effects of an aid program: how do cash transfers affect non-eligibles' consumption? *American Economic Review* **99**: 486–508.

DOI: 10.1002/jae

Attanasio O, Meghir C, Santiago A. 2012. Education choices in Mexico: using a structural model and a randomized experiment to evaluate PROGRESA. *Review of Economic Studies* **79**: 1495–1526.
Barrios T, Diamond R, Imbens G, Kolesar M. 2012. Clustering, spatial correlations and randomization inference. *Journal of the American Statistical Association* **107**: 578–591.
Behrman J, Todd P. 1999. Randomness in the experimental samples of PROGRESA (education, health, and nutrition program). Report submitted to PROGRESA. International Food Policy Research Institute, Washington DC.
Behrman J, Sengupta P, Todd P. 2005. Progressing through PROGRESA: an impact assessment of a school subsidy experiment in Mexico. *Economic Development and Cultural Change* **54**: 237–275.
Djebbari H, Smith J. 2008. Heterogeneous impacts in PROGRESA. *Journal of Econometrics* **145**: 64–80.
Heckman JJ, Moon SH, Pinto R, Savelyev PA, Yavitz A. 2010. Analyzing social experiments as implemented: a reexamination of the evidence from the High/Scope Perry Preschool program. *Quantitative Economics* **1**: 1–46.
Heckman JJ, Pinto R, Shaikh AM, Yavitz A. 2011. Inference with imperfect randomization: the case of the Perry Preschool program. National Bureau of Economic Research Working Paper w16935.
Kling J, Liebman J, Katz L. 2007. Experimental analysis of neighborhood effects. *Econometrica* **75**: 83–119.
Lehmann EL, Romano JP. 2005. Testing Statistical Hypotheses (3rd edn). Springer: New York.
Mitnik O, Imbens G, Hotz V, Crump R. 2008. Nonparametric tests for treatment effect heterogeneity. *The Review of Economics and Statistics* **90**: 389–405.
Romano JP, Shaikh AM. 2010. Inference for the identified set in partially identified econometric models. *Econometrica* **78**: 169–211.
Romano JP, Wolf M. 2005. Exact and approximate stepdown methods for multiple hypothesis testing. *Journal of the American Statistical Association* **100**: 94–108.
Romano JP, Wolf M. 2010. Balanced control of generalized error rates. *The Annals of Statistics* **38**: 598–633.
Romano JP, Shaikh AM, Wolf M. 2008. Formalized data snooping based on generalized error rates. *Econometric Theory* **24**: 404–447.
Romano JP, Shaikh AM, Wolf M. 2010. Hypothesis testing in econometrics. *Annual Review of Economics* **2**: 75–104.
Schultz TP. 2004. School subsidies for the poor: evaluating the Mexican PROGRESA poverty program. *Journal of Development Economics* **74**: 199–250.
Skoufias E. 2001. PROGRESA and its impacts on the human capital and welfare of households in rural Mexico: a synthesis of the results of an evaluation. International Food Policy Research Institute, Washington, DC.
Skoufias E, Parker S, Behrman J, Pessino C. 2001. Conditional cash transfers and their impact on child work and schooling: evidence from the PROGRESA program in Mexico. *Economia* **2**: 45–96.
Todd P, Wolpin K. 2006. 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**: 1384–1417.
Westfall PH, Young SS. 1993. Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment. Wiley: New York.

DOI: 10.1002/jae