Multiple Testing and the Distributional Effects of Accountability Incentives in Education

Lehrer, Steven F. and Pohl, R. Vincent and Song, Kyungchul

Queen's University, University of Georgia, University of British Columbia

24 September 2018

Multiple Testing and the Distributional Effects of Accountability Incentives in Education*

Steven F. Lehrer[†]
Queen's University,
NYU–Shanghai, and NBER

R. Vincent Pohl[‡]
University of Georgia

Kyungchul Song[§]
University of
British Columbia

September 2018

Abstract

Economic theory that underlies many empirical microeconomic applications predicts that treatment responses depend on individuals' characteristics and location on the outcome distribution. Using data from a large-scale Pakistani school report card experiment, we consider tests for treatment effect heterogeneity that make corrections for multiple testing to avoid an overestimation of positive treatment effects. These tests uncover evidence of policy-relevant heterogeneous effects from information provision on child test scores. Further, our analysis reinforces the importance of preventing the inflation of false positive conclusions since over 65% of the estimated statistically significant quantile treatment effects become insignificant once these corrections are applied.

Keywords: information, student performance, quantile treatment effects, multiple testing, bootstrap tests.

JEL classification: C12, C21, I21, L15.

*This paper was previously circulated under the title "Targeting Policies: Multiple Testing and Distributional Treatment Effects." We wish to thank Jonah Gelbach, Pat Kline, Jeff Smith, and seminar and conference participants at the University of Georgia, Hunter College, the University of North Carolina Greensboro, RWI, Sciences Po Paris, Tilburg University, AEA, CLSRN, ESAM, ESNASM, and SOLE/EALE for helpful comments and suggestions. Jacob Schwartz and Thor Watson provided excellent research assistance. Computer code used to generate all of the results in this paper are available in either Stata or Matlab on request. Lehrer and Song respectively thank SSHRC for research support. The usual caveat applies.

[†]School of Policy Studies and Economics Department, email: lehrers@queensu.ca.

[‡]Department of Economics, Terry School of Business, email: pohl@uga.edu.

[§]Vancouver School of Economics, email: kysong@mail.ubc.ca.

1 Introduction

Individuals differ not only in their characteristics but also in how they respond to a particular treatment or intervention. Therefore, treatment effects may vary between subgroups defined by individual characteristics such as gender or race. For example, programs that provide information on schools' performance on standardized tests may lead to a different likelihood that parents "vote with their feet" and move their child to a better school based on parental characteristics such as their education. Similarly, welfare programs that provide work incentives may affect welfare recipients differently according to their demographic and socioeconomic characteristics such as education or number and ages of children. In addition, individuals' response to a particular treatment may vary across quantiles of the unconditional outcome distribution. After all, if a school information provision program improves the odds that a child's performance relative to her peers can be correctly perceived by the parents, parental responses such as switching schools may vary with the child's relative performance.[1]

This diverse and heterogeneous behavior has not only changed how economists think about econometric models and policy evaluation but also has profound consequences for the scientific evaluation of public policy.[2] Although the importance of heterogeneous treatment effects is widely recognized in the causal inference literature, common practice remains to report an average causal effect parameter, even in cases where it is not possible to identify to which subset of individuals this effect applies.[3]

While an increasing number of studies account for possible treatment effect heterogeneity when evaluating either education policies or social programs, most conduct statistical inference without allowing for dependence across subgroups. For example, Fink, McConnell, and Vollmer (2014) report that over 75 percent of studies that analyze data from field experiments published in 10 specific journals estimate separate average causal parameters for different subgroups. Fink, McConnell, and Vollmer (2014) argue that it is inappropriate in those studies to apply traditional standard errors and p-values when testing for heterogeneous treatment effects through interaction terms or subgroup analyses. After all, each interaction term represents a separate hypothesis beyond the original experimental design

[1] Similarly, welfare programs induce kinks in the recipients' budget constraint, so the treatment effect may also vary depending on their pre-treatment earnings.

[2] James Heckman stresses this point in his 2001 Nobel lecture, where he notes that conditional mean impacts including the average treatment effect may provide limited guidance for policy design and implementation (Heckman, 2001).

[3] In particular, a large academic debate (e.g., Deaton, 2010; Heckman and Urzua, 2010; Imbens, 2010) questions whether the local average treatment effect parameter obtained from an IV estimand has policy relevance.

and results in a substantially increased type I error.[4] Lee and Shaikh (2014) address this issue in their study of data from a randomized experiment by adopting a multiple testing procedure for subgroup treatment effects that controls the family-wise error rate (FWER) in finite samples.

A similar observation can be made for distributional treatment effects. A growing number of studies examine if there are different treatment effects across quantiles of the outcome variable, i.e. they estimate quantile treatment effects (QTEs) (e.g., Heckman, Smith, and Clements, 1997; Friedlander and Robins, 1997; Abadie, 2002; Bitler, Gelbach, and Hoynes, 2006; Firpo, 2007). Individual test statistics at different quantiles involve their sample counterparts across different quantiles, which are correlated. A naive approach of comparing individual test results to find quantile groups with positive and statistically significant treatment effects inevitably suffers from the issue of data mining due to the reuse of the same data as emphasized by White (2000).[5] In Online Appendix A, we examine all articles that estimate distributional causal parameters which were published in five high-impact economic journals between 2008–2017, and find that in none of these articles did the authors make corrections for multiple testing.

In this paper, we show that making these corrections is important by reexamining the effectiveness of a market level "soft" accountability policy that consists of reporting information on student and school test score performance to parents.[6] Our contribution is to apply a multiple testing procedure to analyze different dimensions of treatment effect heterogeneity across subgroups and outcome quantiles. Our flexible approach allows us to analyze treatment effect heterogeneity using various hypothesis testing procedures. First, investigating the existence of positive treatment effects for some subgroups or some outcome quantiles is formulated as a hypothesis testing problem. Second, the procedure enables us to identify the subgroups and outcome quantiles for which the treatment effect is estimated to be conspicuous beyond sampling variations. As the result is obtained through a formal multiple testing procedure, it properly takes into account the reuse of the same data for different demographic

[4]The problem when testing multiple hypotheses jointly is the potential over-rejection of the null hypothesis. Intuitively, if the null hypothesis of no treatment effect is true, testing it across 100 subsamples, we expect about five rejections at the 95 percent confidence level. However, if these subsamples depend on each other, more than five rejections may occur. Hence, the type I error would exceed the nominal level of the test. The same issue arises when testing a hypothesis across the percentiles of an outcome variable, as we discuss below.

[5]In part as a response, statistical inference procedures developed in Heckman, Smith, and Clements (1997), Abadie, Angrist, and Imbens (2002), Rothe (2010), and Maier (2011), among others, focus on the whole distribution of potential outcomes to side-step multiple comparisons.

[6]We focus on programs that disclose information on student performance and differs from "hard" accountability programs that tie the pay (and in some case punishments are used in place of rewards) of educators to student test scores.

groups or quantile groups and controls the FWER so that it is unaffected by data mining.[7] Controlling the FWER in multiple comparisons across different quantiles is crucial for the validity of the inference procedure, as estimated treatment effects across different quantiles of the outcome distribution are not independent.[8]

Our use of various formal testing procedures for treatment effect heterogeneity is not solely motivated by policy considerations, but also economic theory.[9] The multiple testing approach provides not only a basis for judging the empirical relevance of treatment effect heterogeneity. It also provides further information on the pattern of treatment effect heterogeneity across different population groups.[10] This information can offer important insights about how scarce social resources are to be distributed by giving policymakers rich information to more effectively assign different treatments to individuals.[11] For example, policy-

[7]More specifically, our procedure involves multiple inequalities of unconditional quantile functions, and draws on a bootstrap method for testing for inequality restrictions. To construct a multiple testing procedure that controls the FWER, i.e. the probability of falsely rejecting at least one true hypothesis, we adapt the step-down method proposed by Romano and Wolf (2005a) to our context of testing multiple inequalities of unconditional quantiles.

[8]Prior work does not consider conditional quantiles and as mentioned in the text, Lee and Shaikh (2014) also adopt a multiple testing procedure to identify subgroups of conspicuous treatment effects. However, there are several notable differences. First, in contrast to our approach, they do not account for within-subgroup treatment effect heterogeneity. Second, Lee and Shaikh (2014) require the treatment to be randomly assigned unconditionally. In contrast, our approach permits selection on observables. Hence it accommodates non-experimental data whenever this assumption is deemed plausible. List, Shaikh, and Xu (2016) extend the multiple testing corrections developed in Lee and Shaikh (2014) to additionally consider an experimental settings with multiple treatments, multiple outcomes, and multiple subgroups, but also do not account for within subgroup treatment effect heterogeneity.

[9]In Online Appendix D, we provide an additional empirical demonstration where we test for treatment effect heterogeneity that is also motivated by an economic model. Specifically, we use a simple static model of labor supply that predicts heterogenous responses to changes in the parameters of a welfare reform policy within and between subgroups. To illustrate the tests we explore the extent of heterogeneity in labor supply responses in the Jobs First welfare experiment across percentiles of the earnings distribution.

[10]Our primary goal is not to develop tests to see if observed behavior is consistent with the quantitative predictions of a theory but rather whether qualitative differences in the pattern of QTEs between subgroups emerge. Our approach differs from Crump et al. (2008) in several aspects. First, Crump et al. (2008) focus on heterogeneity of the average treatment effect across subgroups, while our focus is on treatment effect heterogeneity across quantiles of the outcome distribution, motivated by the findings of Bitler, Gelbach, and Hoynes (2006). Second, Crump et al. (2008) use a joint test for treatment effect heterogeneity covering all the subgroups. In contrast, we use a multiple testing procedure to detect quantiles and/or subgroups for which there is a positive treatment effect. Finally, unlike Crump et al. (2008), we also investigate treatment effect heterogeneity across quantiles *within* each subgroup, so that the focus here is also on whether treatment effect heterogeneity across quantiles is mostly due to subgroup differences or not.

[11]Our interest is not in optimal treatment assignment in the spirit of Manski (2004), Dehejia (2005), and others. Armstrong and Shen (2015) recently extended optimal treatment assignment to additionally consider multiple testing procedures for treatment effects that control for the FWER. In contrast, we do not assume that the researcher ex ante has full knowledge of the distribution of outcomes in the population or of the social planner's welfare function as in the above papers and Kitagawa and Tetenov (2018), among others. Our interest is rather to propose a multiple testing framework for identifying subpopulations with positive responses to the outcome variable.

makers can use the results to modify the design of accountability programs more effectively if they were to know which parents respond to market-level information on school quality. These parents may differ systematically by predetermined characteristics or be characterized by being located between specific percentiles of their child's test score distribution.

The results of this paper contribute to a burgeoning empirical literature surveyed in Figlio and Loeb (2011) that explores how school accountability programs impact education outcomes. Economists have long argued that policies designed to increase competition in markets for education can improve educational outcomes by increasing disadvantaged students' access to high quality schools, and by causing under-performing schools to become more effective or to shrink as families "vote with their feet" (Friedman, 1955; Becker, 1995; Hoxby, 2003). Further, by disclosing information about student and school performance, educators may change their effort since this affects the (implicit) market incentives faced by schools. Indeed, empirical evidence shows that providing information about school-level achievement directly to parents can influence school choice in the United States (Hastings and Weinstein, 2008), Canada (Friesen et al., 2012), the Netherlands (Koning and Van der Wiel, 2012), Brazil (Camargo et al., 2018), and Pakistan (Andrabi, Das, and Khwaja, 2017).[12] However, school performance has also been found to not be the main determinant of choice and that preferences regarding schools are heterogeneous across socioeconomic groups in the United States (Hastings, Kane, and Staiger, 2009), Chile (Schneider, Elacqua, and Buckley, 2006), Pakistan (Carneiro, Das, and Reis, 2013), and the United Kingdom (Gibbons and Machin, 2006).

These findings of heterogeneous responses that appear in many studies are consistent with economic theory and provide a setting to illustrate our methods that examine treatment effect heterogeneity. We reexamine data from Andrabi, Das, and Khwaja (2017) and also make an important methodological contribution to the literature that tests for treatment effect heterogeneity. While Bitler, Gelbach, and Hoynes (2017) allow for multiple tests across subgroups, we also adjust for dependencies between quantiles within subgroups. Thereby, we provide a unified framework to test for treatment effect heterogeneity.

We present evidence that these corrections are empirically important and policy relevant. In our application, slightly over 65 percent of the estimated statistically significant QTEs become insignificant once multiple testing corrections are applied. These findings also demonstrate that the significantly positive effects of providing information to parents

[12]The amount of parental response may depend on the type of information provided. Mizala and Urquiola (2013) provide evidence from Chile that when absolute measures of school achievement are already widely available, there are no changes in enrollment level and socioeconomic composition from receiving an additional highly publicized award.

reported in Andrabi, Das, and Khwaja (2017) are concentrated in the bottom quintile of the test score distribution. We also find that these tests not only generate new insights allowing researchers to better inform policy audiences but are important since recent work by Solon, Haider, and Wooldridge (2015) has shown that researchers who estimate models that do not account for heterogeneous effects may provide inconsistent estimates of average effects, even when unconfoundedness holds (or with experimental data).[13]

The rest of this paper is organized as follows. In section 2, we motivate the tests that we develop by describing the experiment and economic model that underlie the data being investigated. This model predicts heterogenous treatment effects both within and across subgroups. The general testing procedures for treatment effect heterogeneity without and with subgroups are described in Section 3. In Section 4, we present results from our empirical application of the methods to the Andrabi, Das, and Khwaja (2017) experimental data, which yields two main findings. First, while there is clear evidence of treatment effect heterogeneity in the full sample, this is observed in most but not every subgroup. Second, we demonstrate the importance of making corrections for multiple testing since approximately 65 percent of the QTEs become statistically insignificant when we account for potential dependencies. Taken together, our results shed new light on the effectiveness of accountability programs, further indicating how schools and parents respond to the release of information on student performance. The concluding section 5 summarizes the contribution of using these testing approaches in empirical microeconomic research and discusses directions for future methodological work that can aid practitioners.

2 Experimental Design and Data

Andrabi, Das, and Khwaja (2017) conduct an experiment in 112 Pakistani villages to study the impact of providing parents with a detailed two page report card on their child's performance and child's school-level performance on a variety of outcomes. Each report card contained the student's test score and quintile rank (compared to all tested students) in three subject areas, as well as for all of the schools in the village presented information on i) the average score, number of children tested, and iii) quintile rank (across all schools tested in the sample). In accountability systems, such school level report cards are frequently

[13]Under unconfoundedness, it is well known that matching and regression estimators may yield different estimates since they weight observations differently. Intuitively if there are heterogeneous treatment effects across groups in the sample, the OLS estimator gives a weighted average of these effects. The weights depend not only on the frequency of the subgroups, but also upon sample variances within the subgroup. This differs from the sample-weighted average which would be given by the average of each subgroup's partial effect weighted by its frequency in the sample.

postulated to lead to improved parental investment decisions in education. The treatment exogenously increased information in 56 of the 112 villages, and Andrabi, Das, and Khwaja (2017) argue that each village can be viewed as an island economy where private and public schools compete.[14]

The focus of Andrabi, Das, and Khwaja (2017) is to examine the gradient in the estimated causal parameter of providing a report card along both the school type and baseline test score distributions. It is important to stress that the institutional structure of education in Pakistan offers several unique advantages that Andrabi, Das, and Khwaja (2017) exploit to facilitate their study of how competition affects equilibrium school and student outcomes at the market level. Rural villages in Pakistan are typically located at a great distance from each other or are separated by natural barriers. Carneiro, Das, and Reis (2013) find that parents of children in primary school in Pakistan often make enrollment decision that places great weight on the physical distance from home to school. Second, within each village there are multiple affordable private schools, and an estimated 35 percent of all students were enrolled in private schools in 2005. Third, school inputs such as teacher education differ sharply between government and private school and many private schools have a secular orientation. There are very few if any regulations on the private schools that are generally not supported by the government.

The idea that the gradient in the effect of increased information from the report card will differ between public and private schools is consistent with predictions from models of optimal pricing and quality choices in markets with asymmetric information (e.g., Wolinsky, 1983; Shapiro, 1983; Milgrom and Roberts, 1986). These models predict heterogenous responses from improved information. The quality of initially low performing schools as measured by student test scores will increase at a larger rate than responses in initially high-quality schools; and under some assumptions on parental demand for school quality the responses in high quality schools may even be negative. Camargo et al. (2014) develop an alternative model in the spirit of Holmström (1999) of how test score disclosure would lead to heterogenous changes in subsequent student test score performance between public and private schools.[15]

Taken together, these economic models predict students and parents responding to information on school quality and their relative rank within a school, with heterogeneity pre-

[14]These villages are located in one of three selected districts in Pakistan's most populous province, Punjab.

[15]The model they consider is pitched to be a reduced-form version of a dynamic model of managerial effort along the lines of Holmström (1999).

dicting larger behavioral responses to receiving a (more) negative signal.[16] The extent of this heterogeneity can vary across subgroups defined by school type, since administrators in private schools may face greater pressure than public school counterparts and provide a larger response to having negative information being disclosed. Thus, the general shape of treatment effect heterogeneity and the resulting QTEs could be shifted to the left or right, be compressed or stretched, or otherwise be transformed across subgroups without losing their overall shape. In summary, economic theory predicts treatment effect heterogeneity both within and between subgroups, motivating the development of tools to assess its extent in general, as well as in the specific context of the Andrabi, Das, and Khwaja (2017) information provision experiment.

Last, beyond the advantages of the institutional structure, Andrabi, Das, and Khwaja (2017) distinguishes itself from the growing body of work evaluating randomized interventions in developing countries by having collected a rich detailed longitudinal dataset. Beginning in 2004, approximately 12,000 grade 3 students were surveyed. The follow-up rate was over 96% in subsequent years. Schools also completed annual surveys providing rich information on their operations as well as their inputs. A subset of households were also randomly selected for parents to provide additional information on home inputs. In our study, to facilitate comparisons we utilize the same control variables as Andrabi, Das, and Khwaja (2017) and use a standardized grade 4 test score as our primary outcome variable to fully explore treatment effect heterogeneity.

Table 1 shows child-level summary statistics by treatment status for our outcome and subgroup variables. Our outcome variable, "Average test score, round 2," is significantly higher among children in the treated group (whose parents received the school report cards), which is consistent with the findings in Andrabi, Das, and Khwaja (2017). The village-level variables including literacy rate, number of households, school Herfindahl index, and average wealth differ significantly between treatment and control group. Recall that randomization occurred on the village level and not on the child level, and these significant differences disappear in village-level comparisons. We also find significant differences in the fraction of government schools, high-scoring schools, and fathers with above-middle school education by treatment status. Our testing approach incorporate propensity score weighting, which allows us to balance treatment and control group based on these observed variables, see Section 3.1. We consider how the distributional effects of the information experiment vary across and within these subgroups in Section 3.2.

[16] Camargo et al. (2018) present evidence of heterogenous responses in Brazil and Koning and Van der Wiel (2012) also find that test scores increase at a higher rate in schools ranked poorly in national newspapers in the Netherlands.

Table 1: Child-Level Summary statistics

	No report card Mean/Std.dev./N	Report card Mean/Std.dev./N	Difference p-value
Average test score, round 1	−0.0134	−0.0229	0.569
	(0.942)	(0.886)	
	5786	6324	
Average test score, round 2	0.186	0.229	0.012
	(1.004)	(0.943)	
	6266	6538	
Female child	0.425	0.431	0.438
	(0.494)	(0.495)	
	8443	8760	
Child's age	9.680	9.671	0.702
	(1.505)	(1.446)	
	6616	7117	
Village literacy rate	38.46	36.26	0.000
	(12.88)	(10.63)	
	8443	8760	
Number of households in village	708.3	797.3	0.000
	(375.8)	(591.0)	
	8443	8760	
School Herfindahl index	0.181	0.183	0.092
	(0.0680)	(0.0676)	
	8443	8760	
Village wealth	4498.5	4638.6	0.000
(median monthly expenditure)	(1649.4)	(1454.8)	
	8443	8760	
Government school	0.675	0.698	0.003
(excluded category: private)	(0.468)	(0.459)	
	6617	7118	
School size	251.6	248.7	0.386
	(199.5)	(194.9)	
	6617	7118	
High scoring school	0.499	0.486	0.096
(above 60th percentile)	(0.500)	(0.500)	
	8443	8760	
Mother's education above middle school	0.325	0.333	0.498
	(0.469)	(0.471)	
	3097	3278	
Father's education above middle school	0.630	0.590	0.001
	(0.483)	(0.492)	
	3090	3278	

Source: Andrabi, Das, and Khwaja (2017).
Notes: Means, standard deviations (in parentheses), and numbers of observations for children in villages that did not and did receive the information experiment treatment. p-values for the t-test of the null hypothesis that the means do not differ between treatment and control group.

3 Methodology

In this section, we begin by introducing three tests for treatment effect heterogeneity in the full sample. The data generating set-up for the testing problem is as follows. Let D_i be a random variable that takes values in $\{0, 1\}$, where $D_i = 1$ means participation in the program by individual i and $D_i = 0$ being left in the control group. Let Y_i be the observed outcome of individual i defined as

$$Y_i = Y_{0i}D_i + Y_{1i}(1 - D_i),$$

where Y_{1i} denotes the potential outcome of individual i treated in the program and Y_{0i} that of the same individual not treated in the program. Let X_i be a vector of observed covariates pertaining to individual i. The researcher observes a random sample of $(Y_i, D_i, X_i)_{i=1}^n$.

Our parameter of focus is the QTE defined by

$$q_\tau^\Delta = q_{1,\tau} - q_{0,\tau},$$

where for each $\tau \in (0, 1)$ and $d \in \{0, 1\}$,

$$q_{d,\tau} = \inf\{q \in \mathbb{R} : P\{Y_{di} \leqslant q\} \geqslant \tau\}.$$

Throughout this paper, we assume selection on observables (i.e. the conditional independence of (Y_{1i}, Y_{0i}) and D_i given X_i) and the non-overlap condition (i.e., there exists $\eta \in (0, 1/2)$ such that $\eta \leqslant P\{D_i = 1|X_i = x\} \leqslant 1 - \eta$ for all x in the support of X_i.) These assumptions ensure the identification of q_τ^Δ. See Firpo (2007) for details on the efficient estimation of QTE. Motivated by the discussion in the previous section, we propose three additional tests for treatment effect heterogeneity (across τ's) both within and between subgroups.

3.1 Treatment Effect Heterogeneity Without Subgroups

Each test we introduce requires estimates of QTEs that in the full sample are calculated by subtracting the unconditional outcome at quantile τ for the control group from the respective outcome at quantile τ for the treatment group. To control for possible selection on observed variables into treatment and control groups, we weight the outcome variable by inverse propensity score weights (IPSW). We define the IPSW as

$$\hat{\omega}_{1i} = \frac{D_i}{\hat{p}(X_i)} \quad \text{and} \quad \hat{\omega}_{0i} = \frac{1 - D_i}{1 - \hat{p}(X_i)}$$

for treated and control individuals, respectively, where D_i is the treatment indicator, X_i is a vector of observed characteristics, and $\hat{p}(\cdot)$ is the estimated propensity score.[17] We then obtain quantiles of the weighted outcome as follows:

$$\hat{q}_{1,\tau} = \arg\min_q \sum_{i=1}^{n} \hat{\omega}_{1i}\rho_\tau\left(Y_i - q\right) \quad \text{and}$$
$$\hat{q}_{0,\tau} = \arg\min_q \sum_{i=1}^{n} \hat{\omega}_{0i}\rho_\tau\left(Y_i - q\right),$$

where $\rho_\tau(x) = x\cdot(\tau - \mathbf{1}\{x \leqslant 0\})$ is the check function and n is the size of the full sample. That is, $\hat{q}_{\tau,1}$ and $\hat{q}_{\tau,0}$ are the τ-th empirical quantiles of the propensity score weighted outcome variable

$$\left\{\hat{Y}_{1i}\right\}_{i=1}^{n} = \left\{\frac{Y_i D_i}{\hat{p}(X_i)}\right\}_{i=1}^{n} \quad \text{and} \quad \left\{\hat{Y}_{0i}\right\}_{i=1}^{n} = \left\{\frac{Y_i(1-D_i)}{1-\hat{p}(X_i)}\right\}_{i=1}^{n}$$

for treatment and control group, respectively.

Formally, the estimated QTE at τ is then defined as

$$\hat{q}_\tau^{\Delta} = \hat{q}_{1,\tau} - \hat{q}_{0,\tau}.$$

Intuitively, the QTE is equal to the horizontal difference between the graphs of the unconditional outcome distributions of treatment and control group at quantile τ.

3.1.1 Testing for the Presence of Positive Quantile Treatment Effects

The first test is designed to determine whether an intervention had any detectable positive effect on the outcome of interest.[18] We consider the following null and alternative hypotheses:

$$\begin{aligned} H_0 &: q_\tau^{\Delta} \leqslant 0 \text{ for all } \tau \in \mathcal{T} \\ H_1 &: q_\tau^{\Delta} > 0 \text{ for some } \tau \in \mathcal{T}, \end{aligned} \tag{H.1}$$

where $\mathcal{T} \subset [0,1]$ is the finite set of quantiles considered. The alternative hypothesis states that there exists a positive treatment effect for at least one quantile. Therefore, the null hypothesis is rejected if treatment has any positive effect on some range of the outcome distribution, given reasonable power.

[17] Following Smith and Todd (2005) we use the propensity score for the full sample throughout our analysis even when we investigate subgroups.

[18] The idea for this test has policy appeal since, given limited resources, policymakers first need to know if individuals react to a specific policy intervention at all. In contrast, the average treatment effect may conceal positive QTEs if they are entirely offset by negative QTEs in a different range of the outcome distribution.

To develop a bootstrap test of the null hypothesis of no positive treatment effect (H.1), we consider a test statistic of the following form:

$$T_n = \max_{\tau \in \mathcal{T}} \hat{q}_\tau^\Delta. \tag{1}$$

Intuitively, since the null hypothesis states that all QTEs are weakly negative, the largest observed QTE provides the clearest evidence against the null hypothesis (White, 2000). To implement the test, we calculate a critical value using a bootstrap method. Specifically, we first resample with replacement from the original sample B times and construct the propensity score weighted outcomes $\hat{Y}_{1i}^* = Y_i^* D_i^* / \hat{p}^*(X_i^*)$ and $\hat{Y}_{0i}^* = Y_i^*(1-D_i^*)/(1-\hat{p}^*(X_i^*))$, where $\{Y_i^*, D_i^*, X_i^*\}_{i=1}^n$ denotes each bootstrap sample and $\hat{p}^*(X_i^*)$ the estimated propensity score using the bootstrap sample. We construct bootstrap test statistics with bootstrap number B: for $b = 1, \ldots, B$,

$$T_{n,b}^* = \max_{\tau \in \mathcal{T}} \left\{ \hat{q}_\tau^{\Delta *} - \hat{q}_\tau^\Delta \right\}, \tag{2}$$

where $\hat{q}_\tau^{\Delta *} = \hat{q}_{\tau,1}^* - \hat{q}_{\tau,0}^*$ and $\hat{q}_{\tau,1}^*$ and $\hat{q}_{\tau,0}^*$ are the τ-th empirical quantiles of $\{\hat{Y}_{1i}^*\}_{i=1}^n$ and $\{\hat{Y}_{0i}^*\}_{i=1}^n$, respectively. By subtracting $\hat{q}_\tau^\Delta$ we re-center the bootstrap test statistic in order to impose the least favorable configuration under the null hypothesis. We then compare the test statistic (1) to the bootstrap critical value, which is equal to the $(1-\alpha)$-th empirical quantile of the B bootstrap test statistics (2), where α is the nominal level of the test. We reject the null hypothesis if the test statistic exceeds the critical value. Rejection of the null hypothesis (H.1) indicates evidence for positive treatment effects for some range of the outcome distribution.

3.1.2 Testing for General Treatment Effect Heterogeneity

We now test for treatment effect homogeneity, which provides an answer to the policy-relevant question of whether individuals across quantiles differ in their response to a particular intervention. While one may obtain information from a visual inspection of QTEs across quantiles of the outcome distribution, a formal test is necessary to properly account for sampling variations.

We consider the following hypotheses:

$$\begin{aligned} H_0 &: q_\tau^\Delta = c \text{ for all } \tau \in \mathcal{T} \text{ and for some } c \in \mathbb{R} \\ H_1 &: q_\tau^\Delta \neq c \text{ for some } \tau \in \mathcal{T} \text{ and for all } c \in \mathbb{R}. \end{aligned} \tag{H.2}$$

The alternative hypothesis indicates heterogeneity of QTE across quantiles. When the null hypothesis is rejected, it suggests evidence for differential reactions by individuals to the treatment depending on where on the outcome distribution they are located.[19]

To test (H.2) we construct the following test statistic:

$$T_n = \max_{\tau \in \mathcal{T}} \left| \hat{q}_\tau^\Delta - \bar{q}^\Delta \right|, \tag{3}$$

where $\bar{q}^\Delta = \frac{1}{|\mathcal{T}|} \sum_{\tau \in \mathcal{T}} \hat{q}_\tau^\Delta$ is the sample mean of the estimated QTEs. That is, we set the constant c in (H.2) equal to the sample mean, $\bar{q}^\Delta$, and subtract it from the estimated QTEs, so the test statistic will be small if the QTEs are very similar across τ.[20] The max appears in equation (3) to detect the existence of quantiles at which the deviation of the QTE from its mean occurs.

We then follow the same bootstrap approach as in Section 3.1.1 above and calculate the following bootstrap test statistic:

$$T_{n,b}^* = \max_{\tau \in \mathcal{T}} \left| \hat{q}_\tau^{\Delta *} - \hat{q}_\tau^\Delta - \left(\bar{q}^{\Delta *} - \bar{q}^\Delta \right) \right|. \tag{4}$$

The bootstrap test also incorporates re-centering in order to impose the null restriction. To test null hypothesis (H.2), we compare the test statistic (3) to the critical value obtained from the bootstrap test statistics (4), which is equal to the $(1-\alpha)$-th quantile of the bootstrap distribution of $T_{n,b}^*, b = 1, \ldots, B$.

3.1.3 Testing for Which Quantiles the Treatment Effect Is Positive

The next test employs a multiple testing approach to identify the ranges of the outcome distribution that exhibit positive treatment effects. This is important since rejecting null hypothesis (H.1) only informs us that individuals in some range of the outcome distribution exhibit positive QTEs, and rejecting (H.2) just provides evidence that the treatment effect is not constant across the outcome distribution. The identified range can be of considerable

[19] Note that testing for treatment effect heterogeneity has appeared in Appendix E of Heckman, Smith, and Clements (1997) though in a different form. Their null hypothesis posits that the variance of the individual treatment effect $Y_{1i} - Y_{0i}$ is zero, i.e. there is no treatment effect heterogeneity across individuals. On the other hand, our null hypothesis allows for treatment effect heterogeneity across individuals. Our null hypothesis rather states that the QTEs are constant across quantiles. After all, randomness of the individual treatment effect appears to be a less interesting hypothesis to test, given that it is well accepted that individuals have heterogeneous responses to policy interventions including experiments such as PROGRESA (see, e.g., Djebbari and Smith, 2008).

[20] Since the null hypothesis involves an equality, we take the absolute value of the difference between QTE and mean QTE.

interest for policymakers when they wish to define a target group for their policies in a way that carries empirical support. We follow recent developments in the multiple testing literature (see, e.g., Romano and Wolf, 2005a,b) and use a bootstrap based step-down method to identify the quantiles for which positive treatment effects are present.[21] To do so, it is necessary to update the critical value at each step, for example by using a bootstrap method. By combining bootstrap tests of inequality restrictions with multiple testing procedures, we produce a testing procedure suitable for analyzing treatment heterogeneity that controls the FWER at the desired level.

We first define individual hypothesis testing problems as follows: for each τ in a range $\mathcal{T} \subset [0,1]$,

$$\begin{aligned} H_{0,\tau} &: q_\tau^\Delta \leqslant 0 \\ H_{1,\tau} &: q_\tau^\Delta > 0. \end{aligned} \tag{H.3}$$

Then the goal is to find a set of individual hypotheses, for which the null is false, in a way that controls the FWER asymptotically. The FWER here is the probability that we mistakenly declare a positive QTE for at least one $\tau \in \mathcal{T}$.

To implement this approach, we follow Romano and Wolf (2005a) and Romano and Shaikh (2010) by conducting stepwise elimination of quantiles using the bootstrap. More specifically, setting $\mathcal{T}_1 = \mathcal{T}$, we find the smallest $\hat{c}_1$ such that

$$\frac{1}{B}\sum_{b=1}^{B} \mathbf{1}\left\{T_{n,b}^*(\mathcal{T}_1) > \hat{c}_1\right\} \leqslant \alpha,$$

where $T_{n,b}^*(\mathcal{T}_1) = \max_{\tau\in\mathcal{T}_1}\left\{\hat{q}_\tau^{\Delta *} - \hat{q}_\tau^\Delta\right\}$ denotes the bootstrap one-sided test statistic using the b-th bootstrap sample and α is the level of the test. That is, at $\hat{c}_1$, the fraction of test statistics across the B bootstrap samples that exceed that critical value is at most α. Then, we retain those quantiles that do not exceed the critical value $\hat{c}_1$, i.e. we define

$$\mathcal{T}_2 = \left\{\tau \in \mathcal{T}_1 : \hat{q}_\tau^\Delta \leqslant \hat{c}_1\right\},$$

so $\mathcal{T}_2$ is a subset of $\mathcal{T}_1$. Now, we construct $T_{n,b}^*(\mathcal{T}_2) = \max_{\tau\in\mathcal{T}_2}\left\{\hat{q}_\tau^{\Delta *} - \hat{q}_\tau^\Delta\right\}$, find $\hat{c}_2$ such that

$$\frac{1}{B}\sum_{b=1}^{B} \mathbf{1}\left\{T_{n,b}^*(\mathcal{T}_2) > \hat{c}_2\right\} \leqslant \alpha,$$

[21] Romano, Shaikh, and Wolf (2010) provide a recent overview of testing procedures proposed in the literature since the very conservative Bonferroni correction.

and define

$$\mathcal{T}_3 = \left\{\tau \in \mathcal{T}_2 : \hat{q}_\tau^\Delta \leqslant \hat{c}_2\right\}.$$

This procedure is repeated until at step k, we obtain $\mathcal{T}_k = \left\{\tau \in \mathcal{T}_{k-1} : \hat{q}_\tau^\Delta \leqslant \hat{c}_{k-1}\right\}$ such that no further element of $\mathcal{T}_k$ is eliminated (i.e. $\mathcal{T}_k = \mathcal{T}_{k-1}$). Then the resulting set $\mathcal{T}_k$ is the subset of $\mathcal{T}$ such that there is no empirical support for positive and statistically significant treatment effects at quantiles $\tau \in \mathcal{T}_k$. Conversely, this procedure provides evidence for positive treatment effects at quantiles in the set $\mathcal{T}\backslash\mathcal{T}_k$. From the result of Romano and Shaikh (2010), it is not hard to show that this multiple testing procedure asymptotically controls the FWER at α.[22]

3.2 Incorporating Subgroups

The preceding three tests did not consider treatment effect heterogeneity across different subgroups. Tests involving subgroups can be useful when policymakers want to identify subgroups defined by observed variables that exhibit differential treatment effects, or when they are interested in the extent of heterogeneity within subgroups. For example, given limited resources, policymakers may be reluctant to extend programs to groups where a significant fraction does not receive gains. Finally, and consistent with the arguments in Lee and Shaikh (2014) and Fink, McConnell, and Vollmer (2014), it is important to develop tools for statistical inference in this setting that account for dependence both within and across subgroups.

We assume that the subgroup vector Z_i is a subvector of X_i, so we write $X_i = (X_{1i}, Z_i)$, where X_{1i} indicates the vector that is not included in Z_i. Let us define for each z in the support of Z_i, $\tau \in (0,1)$, and $d \in \{0,1\}$,

$$q_{d,\tau}(z) = \inf\{q \in \mathbb{R} : P\{Y_{di} \leqslant q | Z_i = z\} \geqslant \tau\},$$

where Z_i is the subgroup vector taking values from a finite set $\mathcal{Z} = \times_{j=1}^{J} \mathcal{Z}^j$, where $\mathcal{Z}^j$ is the set of values from the j-th category (i.e. each category j corresponds to one observed variable that can take on multiple values). Hence $q_{1,\tau}(z)$ and $q_{0,\tau}(z)$ are the quantiles of the outcome variable in the treatment and control groups conditional on subgroup z. Then the subgroup QTE for subgroup z is defined by

$$q_\tau^\Delta(z) = q_{1,\tau}(z) - q_{0,\tau}(z).$$

[22] See Online Appendix B for a sketch of proofs for the validity of all testing procedures introduced in the paper.

To account for covariates in the analyses, we continue to use inverse propensity score weighting. First, note that we can identify $q_{d,\tau}(z)$ by

$$q_{d,\tau}(z) = \arg\min_{q} E[\omega_{di}\rho_\tau(Y_i - q)\,|Z_i = z], d = 1, 0,$$

where

$$\omega_{1i} = \frac{D_i}{p(X_i)} \quad \text{and} \quad \omega_{0i} = \frac{1 - D_i}{1 - p(X_i)}.$$

Thus, we estimate $q_{d,\tau}(z)$ by

$$\hat{q}_{d,\tau}(z) = \arg\min_{q} \frac{1}{\sum_{i=1}^{n} \mathbf{1}\{Z_i = z\}} \sum_{i=1}^{n} \hat{\omega}_{di}\rho_\tau(Y_i - q)\,\mathbf{1}\{Z_i = z\},$$

with the weights given by

$$\hat{\omega}_{1i} = \frac{D_i}{\hat{p}(X_i)} \quad \text{and} \quad \hat{\omega}_{0i} = \frac{1 - D_i}{1 - \hat{p}(X_i)}.$$

3.2.1 Testing for Which Quantiles and Subgroups the Treatment Effect Is Positive

This test extends the test of hypothesis (H.3) to a setting with subgroups. That is, we identify the quantile-subgroup cells that have statistically significantly positive treatment effects. We consider the following individual hypotheses: for each $\tau \in \mathcal{T}$ and $z \in \mathcal{Z}$,

$$\begin{aligned} H_{0,\tau,z} &: q_\tau^\Delta(z) \leqslant 0 \\ H_{1,\tau,z} &: q_\tau^\Delta(z) > 0. \end{aligned} \tag{H.4}$$

Hence, we test a total number of $|\mathcal{T}| \times |\mathcal{Z}|$ hypotheses. We denote the set of quantile-subgroup cells by $\mathcal{W} = \mathcal{T} \times \mathcal{Z}$.

The test is constructed as follows. First, by resampling with replacement from the original sample, we construct $\hat{Y}_{1i}^* = Y_i^* D_i^* / \hat{p}^*(X_i^*)$ and $\hat{Y}_{0i}^* = Y_i^*(1 - D_i^*)/(1 - \hat{p}^*(X_i^*))$. Then we take our bootstrap one-sided test statistic to be

$$T_{n,b}^*(\mathcal{W}) = \max_{(\tau,z)\in\mathcal{W}} \left\{\hat{q}_\tau^{\Delta *}(z) - \hat{q}_\tau^\Delta(z)\right\}, \tag{5}$$

where $\hat{q}_\tau^{\Delta *}(z) = \hat{q}_{\tau,1}^*(z) - \hat{q}_{\tau,0}^*(z)$, $\hat{q}_{\tau,1}^*(z)$ and $\hat{q}_{\tau,0}^*(z)$ are the empirical quantiles of $\{\hat{Y}_{1i}^*\}_{i=1}^n$ and $\{\hat{Y}_{0i}^*\}_{i=1}^n$, respectively, at quantile τ within the samples with $Z_i^* = z$. To perform multiple

testing, we proceed by eliminating subgroup-quantile cells stepwise. At each step, we retain those (τ, z) cells for which no evidence for positive treatment effect can be found. That is, (τ, z) cells that are eliminated throughout this procedure constitute the subgroup-quantile groups with evidence for positive treatment effects.

Specifically, we take $\mathcal{W}_1 = \mathcal{T} \times \mathcal{Z}$, and find the minimum $\hat{c}_1$ such that

$$\frac{1}{B}\sum_{b=1}^{B}\left\{T_{n,b}^{*}(\mathcal{W}_1) > \hat{c}_1\right\} \leqslant \alpha,$$

where $T_{n,b}^{*}(\mathcal{W}_1)$ is defined in equation (5) and α is the desired FWER. We define

$$\mathcal{W}_2 = \left\{(\tau, z) \in \mathcal{W}_1 : \hat{q}_{\tau}^{\Delta}(z) \leqslant \hat{c}_1\right\}$$

and construct $T_{n,b}^{*}(\mathcal{W}_2) = \max_{(\tau,z)\in\mathcal{W}_2}\left\{\hat{q}_{\tau}^{\Delta *}(z) - \hat{q}_{\tau}^{\Delta}(z)\right\}$ to find the minimum $\hat{c}_2$ such that

$$\frac{1}{B}\sum_{b=1}^{B}\left\{T_{n,b}^{*}(\mathcal{W}_2) > \hat{c}_2\right\} \leqslant \alpha.$$

We then define

$$\mathcal{W}_3 = \left\{(\tau, z) \in \mathcal{W}_2 : \hat{q}_{\tau}^{\Delta}(z) \leqslant \hat{c}_2\right\}.$$

The process is repeated until we obtain $\mathcal{W}_k = \left\{(\tau, z) \in \mathcal{W}_{k-1} : \hat{q}_{\tau}^{\Delta}(z) \leqslant \hat{c}_{k-1}\right\}$ for some k such that no further element of $\mathcal{W}_k$ is eliminated. Then the resulting set $\mathcal{W}_k$ is the subset of $\mathcal{W}$ such that there is no empirical support that the treatment effect at quantile-subgroup pair $(\tau, z) \in \mathcal{W}_k$ is positive. This procedure will yield all the combinations of subgroups and quantiles where positive treatment effects are present; specifically they are given by quantile-subgroup pairs $(\tau, z) \in \mathcal{W}\backslash\mathcal{W}_k$.

3.2.2 Testing for Subgroup-Specific Treatment Effect Heterogeneity Across Quantiles

Here we focus on the question of whether differences across subgroups can explain the observed heterogeneity of QTEs in the full sample. More specifically, we search for evidence that all subgroups exhibit heterogeneity of treatment effects across different quantiles $\tau \in (0, 1)$:

$$\begin{aligned} &H_0 : q_{\tau}^{\Delta}(z) = c_z \text{ for all } \tau \in \mathcal{T}, \text{ for some } c_z \in \mathbb{R}, \text{ and for all } z \in \mathcal{Z} \\ &H_1 : q_{\tau}^{\Delta}(z) \neq c_z \text{ for some } \tau \in \mathcal{T}, \text{ for all } c_z \in \mathbb{R}, \text{ and for some } z \in \mathcal{Z}. \end{aligned} \tag{H.5}$$

The null hypothesis states that the heterogeneity in treatment effects disappears when we condition on Z_i. In other words, it posits that the QTEs are constant across quantiles within all subgroups z. However, the null hypothesis still allows for treatment effect heterogeneity across different subgroups. Rejection of the null hypothesis suggests the presence of QTE heterogeneity across quantiles even after we control for Z_i. Bitler, Gelbach, and Hoynes (2017) ask exactly this question. In contrast to their paper, however, we do not constrain the treatment effect to be constant within subgroups.[23]

We check the validity of the above assumption by testing Hypothesis (H.5) using the following test statistic:

$$T_n = \max_{z \in \mathcal{Z}} \max_{\tau \in \mathcal{T}} \left| \hat{q}_\tau^\Delta(z) - \bar{q}^\Delta(z) \right|, \tag{6}$$

where $\bar{q}^\Delta(z) = \frac{1}{|\mathcal{T}|} \sum_{\tau \in \mathcal{T}} \hat{q}_\tau^\Delta(z)$ is the sample mean of the estimated QTEs for each subgroup. As with test statistic (3), we impose the null hypothesis by subtracting $\bar{q}^\Delta(z)$. For each subgroup, the highest deviation of the estimated QTEs from their sample mean provides the clearest evidence against the null hypothesis. Then to obtain a test statistic that covers all subgroups $z \in \mathcal{Z}$, we take the maximum value over each subgroup's test statistic. Intuitively, we search for evidence that there exists a subgroup that exhibits treatment effect heterogeneity, so we restrict our attention to the subgroups that have the largest degree of heterogeneity.

To construct a bootstrap critical value, we consider the following bootstrap test statistic that is an analog of (4) with subgroups:

$$T_{n,b}^* = \max_{z \in \mathcal{Z}} \max_{\tau \in \mathcal{T}} \left| \hat{q}_\tau^{\Delta *}(z) - \hat{q}_\tau^\Delta(z) - \left(\bar{q}^{\Delta *}(z) - \bar{q}^\Delta(z) \right) \right|. \tag{7}$$

The test statistic in equation (6) is compared to the bootstrap critical value, which equals the $(1-\alpha)$-th quantile of the bootstrap test statistics (7) as described above. This test of hypothesis (H.5) provides a simple and flexible way to see if most treatment effect heterogeneity across quantiles is in fact due to treatment effect heterogeneity across subgroups.

3.2.3 Testing for Which Subgroups Treatment Effects Are Heterogenous

The test of hypothesis (H.5) presented in equation (6) considers whether treatment effect heterogeneity across quantiles exists even after controlling for subgroups. If we reject null hypothesis (H.5), we may also be interested in identifying the subgroups that exhibit QTE heterogeneity. The next test considers this objective by exploring subgroup by subgroup, if

[23] In the test in Section 3.2.3 below we additionally adjust for multiple testing while also relaxing the assumption of constant subgroup-specific treatment effects.

there is treatment effect heterogeneity within subgroups. For each $z \in \mathcal{Z}$, we test

$$\begin{aligned} H_{0,z} &: q^{\Delta}_{\tau,1}(z) = c_z \text{ for all } \tau \in \mathcal{T} \text{ for some } c_z \in \mathbb{R} \\ H_{1,z} &: q^{\Delta}_{\tau,1}(z) \neq c_z \text{ for some } \tau \in \mathcal{T} \text{ for all } c_z \in \mathbb{R}. \end{aligned} \tag{H.6}$$

The null hypothesis (H.6) posits that the QTEs are constant within subgroup. This test can identify the subgroups that exhibit heterogeneity of QTE while accounting for dependencies both between quantiles and between subgroups $z \in \mathcal{Z}$. The test of (H.6) differs from the test of (H.5) by not conditioning on z. The test of (H.6) examines treatment effect heterogeneity separately for each subgroup z and finds subgroups that exhibit treatment effect heterogeneity; whereas the test fo hypothesis (H.5) considers if treatment effect heterogeneity disappears.[24]

We consider the following test statistic:

$$T_n(z) = \max_{\tau \in \mathcal{T}} \left| \hat{q}^{\Delta}_{\tau}(z) - \bar{q}^{\Delta}(z) \right|, \tag{8}$$

which is analog to the test statistic (3) with QTEs calculated by subgroup. As before, we follow Romano and Wolf (2005a) and eliminate the subgroups, for which we cannot reject the null hypothesis (H.6) in a step-down procedure. Then the remaining subgroups (if any) are the ones for which we reject the null hypothesis of no treatment effect heterogeneity. The bootstrap test statistic is defined as

$$T^{*}_{n,b}(\mathcal{Z}_1) = \max_{z \in \mathcal{Z}_1} \max_{\tau \in \mathcal{T}} \left| \hat{q}^{\Delta *}_{\tau}(z) - \hat{q}^{\Delta}_{\tau}(z) - \left(\bar{q}^{\Delta *}(z) - \bar{q}^{\Delta}(z) \right) \right|, \tag{9}$$

where we first take $\mathcal{Z}_1 = \mathcal{Z}$. We then find bootstrap critical values $\hat{c}_{z,1}$ for each subgroup z such that

$$\frac{1}{B} \sum_{b=1}^{B} \mathbf{1} \left\{ T^{*}_{n,b}(\mathcal{Z}_1) > \hat{c}_{z,1} \right\} \leqslant \alpha$$

and define

$$\mathcal{Z}_2 = \left\{ z : T_n(z) \leqslant \hat{c}_{z,1} \right\},$$

i.e. $\mathcal{Z}_2$ is the set of subgroups, for which the test statistic (8) does not exceed the critical value. Hence $z \in \mathcal{Z}_2$ are subgroups that do not exhibit significant treatment effect heterogeneity. We then repeat these steps with $\mathcal{Z}_2$, find a critical values $\hat{c}_{z,1}$ analogously, and so on, until

[24] This test nevertheless differs from testing for treatment effect heterogeneity (hypothesis (H.3) for the entire sample) separately for each subgroup since we use the Romano and Wolf (2005a) approach to identify the subgroup(s) that exhibit heterogeneity in QTEs.

no additional subgroup is eliminated (resulting in the set of subgroups $\mathcal{Z}_k$). Hence, there is evidence for treatment effect heterogeneity for subgroups $z \in \mathcal{Z} \backslash \mathcal{Z}_k$.

4 Empirical Application

In this section, we obtain new insights extending the findings of Andrabi, Das, and Khwaja (2017) by conducting the battery of tests described in the preceding section. Our analysis focuses on the average of test scores across three subjects after random assignment as our outcome variable and estimate QTEs of access to report cards for percentiles 1 to 99 using the Firpo (2007) estimator.[25] To balance covariates between the treatment and control groups, we estimate the propensity score $\hat{p}(x)$ using a parametric logit specification. Specifically, we include district fixed effects, and village wealth, literary rate, school Herfindahl index, and number of households. For the results that follow, we set the level of each test to $\alpha = 0.05$. All test results are based on bootstraps with $B = 999$.

Figure 1 shows our estimated QTEs for the full sample along with 90 percent pointwise confidence intervals.[26] We find pointwise significant and positive treatment effects extending from the first to the 82nd percentile. Starting with the 83th percentile the point estimates for treatment effects become negative but the pointwise confidence intervals include zero.

Table 2 summarizes the test results for hypotheses (H.1) and (H.2) proposed in Section 3.1. First, we test the null hypothesis of no positive treatment effect at any percentile. As shown in Figure 1, the largest QTE (which occurs at the third percentile) equals 0.394, so this value appears in the first row of Table 2. With the bootstrap critical value of 0.223, we reject the null hypothesis at 5 percent. The associated p-value equals 0.0017. Thus, there is clear evidence that the information provision had the desired effect of increasing student performance for at least some individuals. Next, we present results from the test of no treatment effect heterogeneity across quantiles (H.2). The test statistic, which is calculated as the largest deviation from the mean estimated QTE ($\bar{q}^{\Delta} = 0.0583$), equals 0.33. With a bootstrap critical value of 0.236, we also reject this null hypothesis at 5 percent with a p-value of 0.0119. This result implies that treatment effects are heterogenous across quantiles, thereby indicating that individuals vary in their response to the report cards.

[25] To infer treatment effects for specific individuals from QTEs we have to assume that there are no rank reversals in the test score distribution between the treatment and control groups. While this assumption is likely violated, positive QTEs imply that the treatment has a positive effect for some interval of the test score distribution.

[26] We show 90 percent confidence intervals to make them comparable to the multiple testing results, which are obtained from one-sided tests that control the FWER at 5 percent.

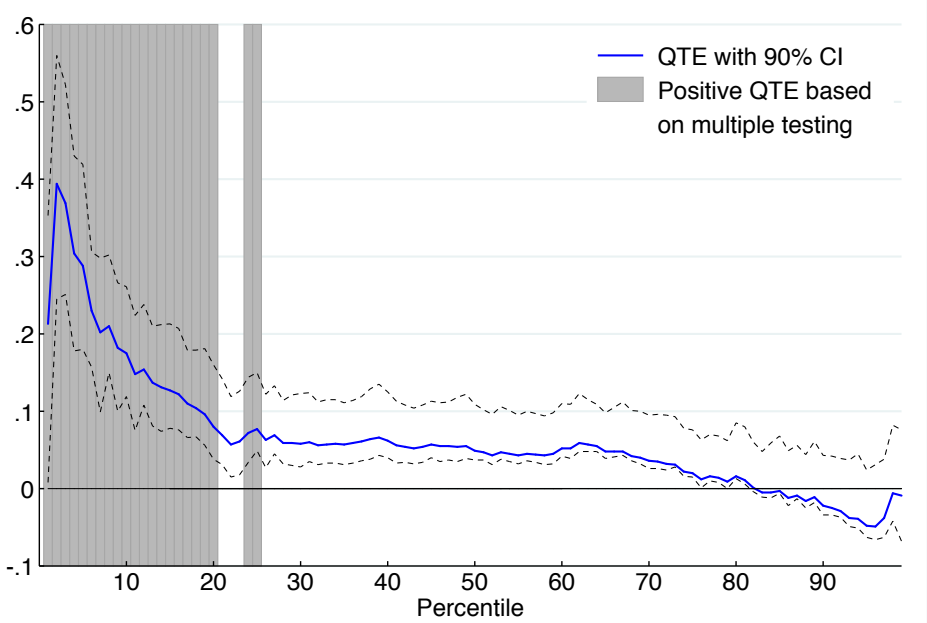


Figure 1: Quantile Treatment Effects and Multiple Testing Results, No Subgroups

Having rejected the null hypothesis of no treatment effect heterogeneity, we now identify the range of the test score distribution where positive treatment effects are located, i.e. we test hypothesis (H.3). As described in Section 3.1.3, this test accounts for potential dependencies across quantiles of the same outcome variable and the number of individual hypotheses ($|\mathcal{T}| = 99$). The shaded area in Figure 1 corresponds to the set $\mathcal{T} \backslash \mathcal{T}_k$, i.e. the percentiles where the treatment effect remains significant using a FWER of $\alpha = 0.05$. Examining the plot we observe that the set of significantly positive QTEs supports the distributional effects predicted by the underlying theory. However, we find that individuals located between the

Table 2: Testing for Presence of Positive QTEs and QTE Heterogeneity Without Subgroups

	Test statistic	Critical value at 5%	p-value
Test of (H.1)	0.394	0.223	0.0017
Test of (H.2)	0.33	0.236	0.0119

Notes: This table shows test results for hypotheses (H.1) and (H.2), i.e. we test that there is no positive treatment effect for all quantiles and that the treatment effect is the same for all quantiles, respectively.

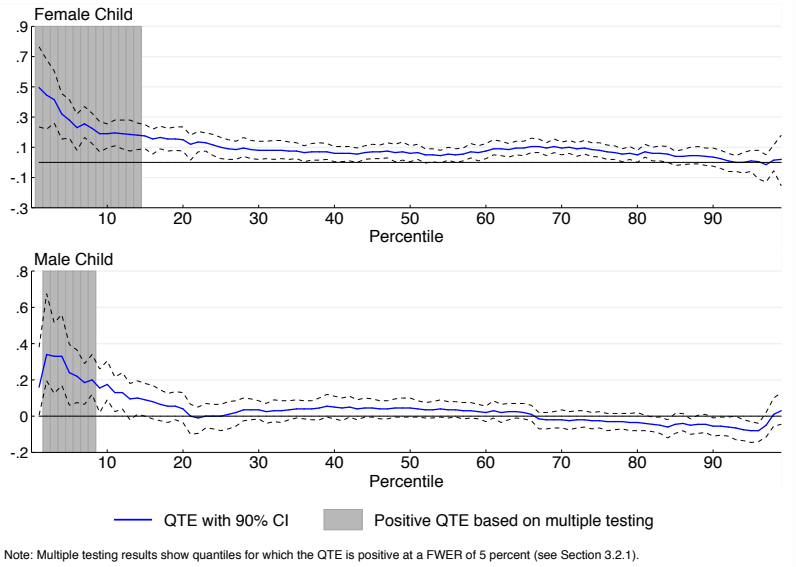
26th and 82nd and the 21st and 23rd percentiles of the test score distribution do not exhibit significant QTEs once we adjust for multiple testing. The smallest and largest quantiles at which QTEs are significantly positive correspond to gains of 0.07 and 0.394, respectively. Hence, we can conclude that the benefits of this particular form of accountability are more confined than one would otherwise find based on traditional statistical inference that ignores potential dependencies and testing at multiple percentiles. We find that there is a more limited range of individuals whose academic outcomes truly increase when assigned to the treatment group.

Next, we present results incorporating subgroups using the tests described in Section 3.2. Economic theory predicts that individuals with different observed characteristics may react differently to the same set of information. In particular, individual and village characteristics may determine for which range of the test score distribution we observe an increase or decrease in test score performance. Following Andrabi, Das, and Khwaja (2017), we consider subgroups defined by child characteristics, education background of their parents, type of school and characteristics of the villages themselves.[27]

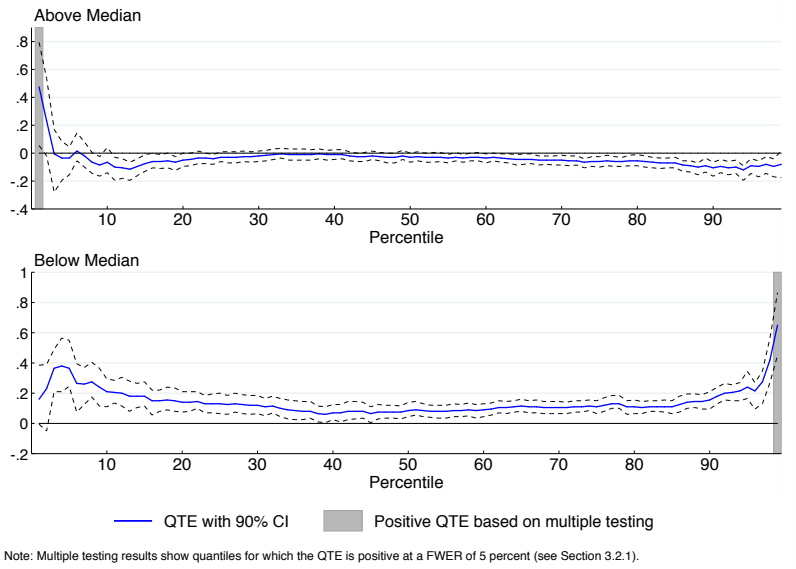
Figure 2 presents QTEs conditional on child gender and child predetermined baseline test scores. These figures provide an easy and intuitive way to check which subgroups benefit from being assigned to receive report cards (heterogeneity across subgroups). In addition, we can inspect the figure for each subgroup to determine the portion of the student test score distribution in which individuals exhibit positive subgroup-specific QTEs (heterogeneity within subgroup). Shaded areas continue to denote significant QTEs based on our multiple testing procedure of testing hypothesis (H.4).

The top panel of Figure 2 presents QTEs for child gender subgroups. The effect of the access to report cards on test scores is larger for girls throughout the test score distribution. For boys, there is no statistically significant positive effect above the 12th percentile (based on the point-wise confidence intervals). When adjusting inference for multiple testing, we find significant effects among girls in the 1st to 14th percentile and boys in the 2nd to 8th percentile. The second panel considers subgroups defined by whether the child's baseline test score was above or below the median. The estimated QTEs and point-wise confidence intervals in Figure 2b show that it is mostly children with a below-median baseline test score who benefit from the report card experiment. When we adjust inference for multiple testing, however, only children in the very top percentile of the post-experiment test score

[27]Note that in our application the number of hypotheses being tested is quite small particularly relative to genomic studies from genome wide association studies. If the number of hypotheses were large it is well known that FWER controlling procedures typically have low power, and in response Gu and Shen (2017) propose an optimal false discovery rate controlling method.



(a) By Child's Gender



(b) By Child's Baseline Test Score

Figure 2: Quantile Treatment Effects and Multiple Testing Results, by Child Characteristics

distribution who scored below the median at baseline exhibit significantly positive QTEs. In addition, children who scored above the median at baseline and whose post-experiment score falls in the first percentile also see a significant effect of information provision.[28]

Next, we construct subgroups based on village characteristics. Figure 3 shows the estimated subgroup-specific QTEs and multiple testing results. We find significant treatment effects are predominantly for children in villages with below-median wealth, above-median literacy rates, below-median school concentration (measured by the school Herfindahl Index), and above-median size. From a policy perspective, it is may be important to know that report cards improve children's test scores in relatively poor villages. At the same time, providing written report cards to parents may not be a successful strategy in villages with low literacy rates. In general, these results are important because they can show policymakers which subgroups should be targeted with an accountability program.

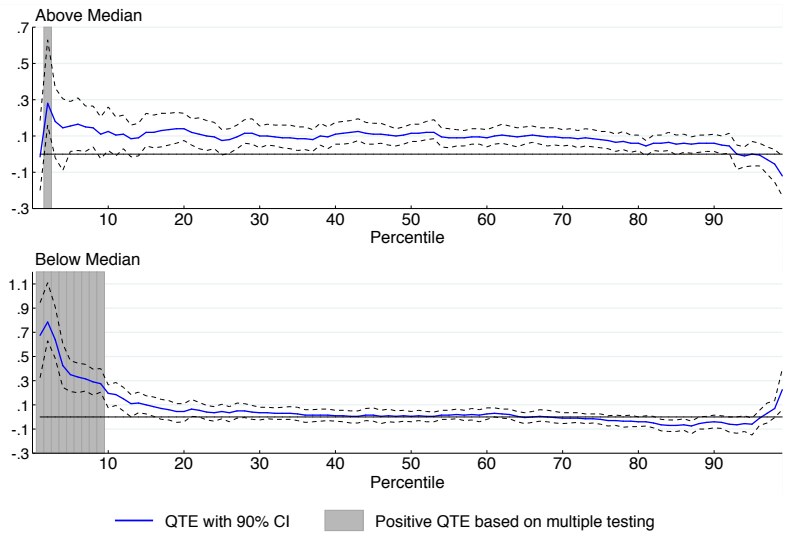
Finally, we consider subgroups defined by the combination of school ownership type (government or private) with one of two different measures of student performance (school level and relative). We first create subgroups by interacting school ownership with school performance in the baseline test to yield four subgroups.[29] Figure 4 illustrates the estimation and multiple testing results. We find that significantly positive QTEs are concentrated among low-scoring children in relatively high-performing government schools and high-scoring children in low-performing private schools. Moreover, consistent with the negative average treatment effect reported in Andrabi, Das, and Khwaja (2017) we do not find any positive effects among children in high-performing private schools.

The second student performance measure we consider pertains to the child's performance at the baseline test relative to his or her school's performance. Specifically, we construct subgroups by dividing the sample into groups defined by the combination of school ownership and whether the child performed above or below the median test score of their respective school at baseline (high and low achieving students, respectively).[30] Figure 5 shows that children in government schools only benefit from the report cards if they are located in the bottom of the test score distribution irrespective of whether they scored above or below

[28]The data also include information on parental educational attainment and monetary and time inputs into the children's human capital. However, the parental survey was only fielded to a third of the sample, and the smaller sample size does not give us enough power to conduct our multiple testing corrections. The results in Table 3 also indicate that we cannot reject the null hypothesis of no treatment effect heterogeneity across subgroups for mother's and father's education.

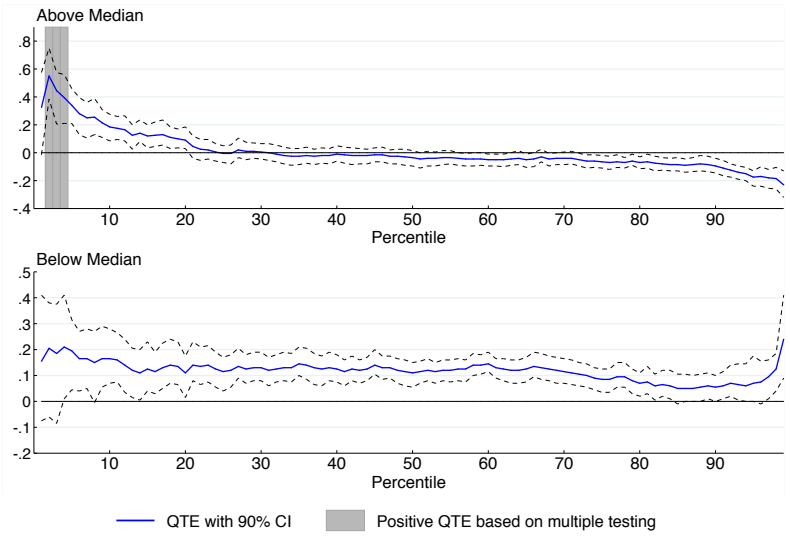
[29]Specifically, following Andrabi, Das, and Khwaja (2017) a school is defined as high-performing if its mean baseline test score exceeds the 60th percentile of all schools' mean scores.

[30]We thank Jishnu Das for pointing out the distinction between these two baseline performance measures. Table VII in Online Appendix III of Andrabi, Das, and Khwaja (2017) shows average treatment effects by children's baseline performance relative to their school.



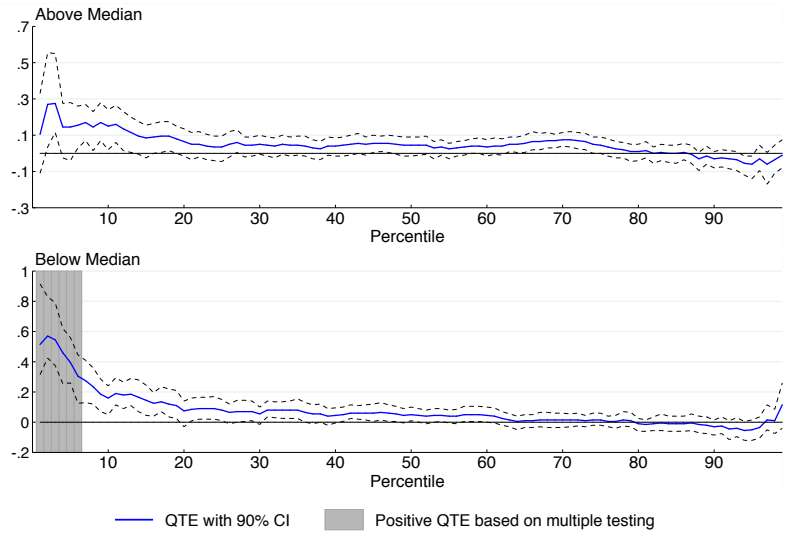
Note: Multiple testing results show quantiles for which the QTE is positive at a FWER of 5 percent (see Section 3.2.1).

(a) By Village Wealth



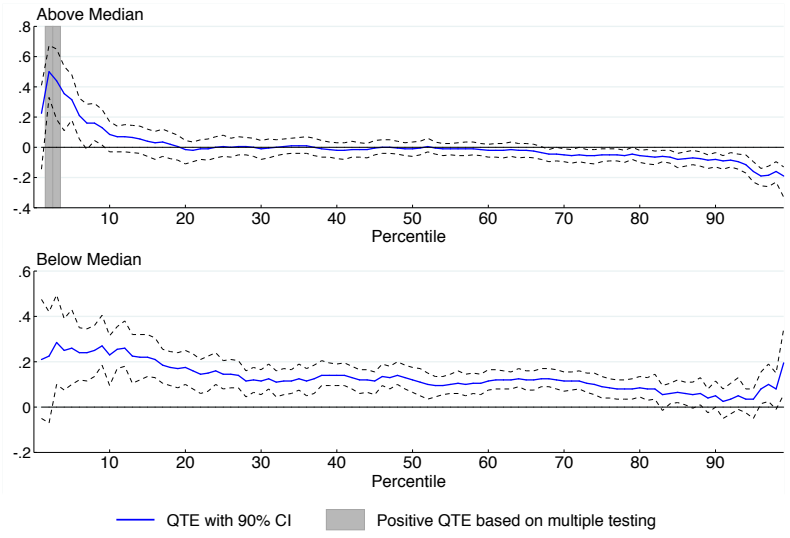
Note: Multiple testing results show quantiles for which the QTE is positive at a FWER of 5 percent (see Section 3.2.1).

(b) By Village Literacy Rate



Note: Multiple testing results show quantiles for which the QTE is positive at a FWER of 5 percent (see Section 3.2.1).

(c) By School Herfindahl Index



Note: Multiple testing results show quantiles for which the QTE is positive at a FWER of 5 percent (see Section 3.2.1).

(d) By Village Size

Figure 3: Quantile Treatment Effects and Multiple Testing Results by Village Characteristics

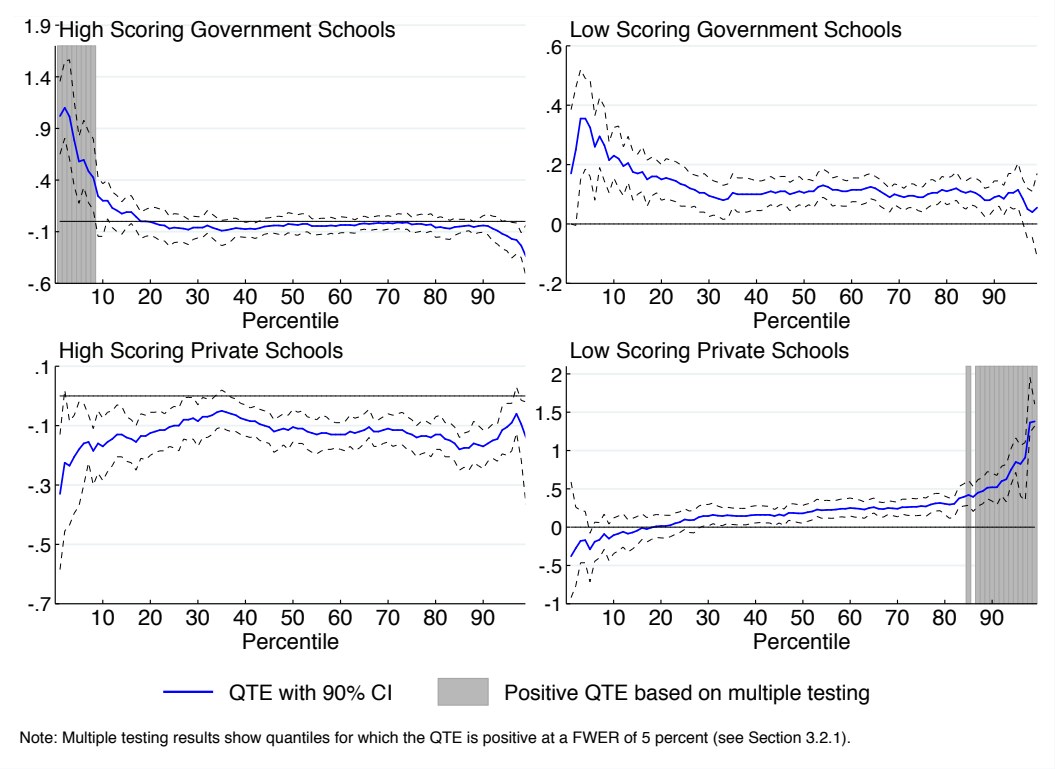


Figure 4: Quantile Treatment Effects and Multiple Testing Results by School Type and Performance

the median of their school's test score at baseline. In addition, the QTEs are significantly positive under corrections for multiple testing among children who score above the 90th percentile and are enrolled in a private school where they scored below the within school median at baseline.

Taken together, our results in Figures 4 and 5 provide additional nuance on the findings of Andrabi, Das, and Khwaja (2017) related to which students in which schools gain from access to report cards. Bitler, Gelbach, and Hoynes (2006) motivate the valuable additional policy insights provided by distributional effects as showing what mean estimates can miss. In Figure 4, our evidence of treatment effect heterogeneity is masked if one estimates average treatment effects even conditional on school type and performance. Further, in Figure 5, while the main result is consistent with Andrabi, Das, and Khwaja (2017) who find that low achieving students benefit from the report card intervention more than high achieving students, we provide additional insights by showing that this benefit is confined to the top decile among low achieving students.

We now formally test for treatment effect heterogeneity between and within subgroups. Table 3 presents the results for hypothesis (H.5) for the same subgroups as above. This

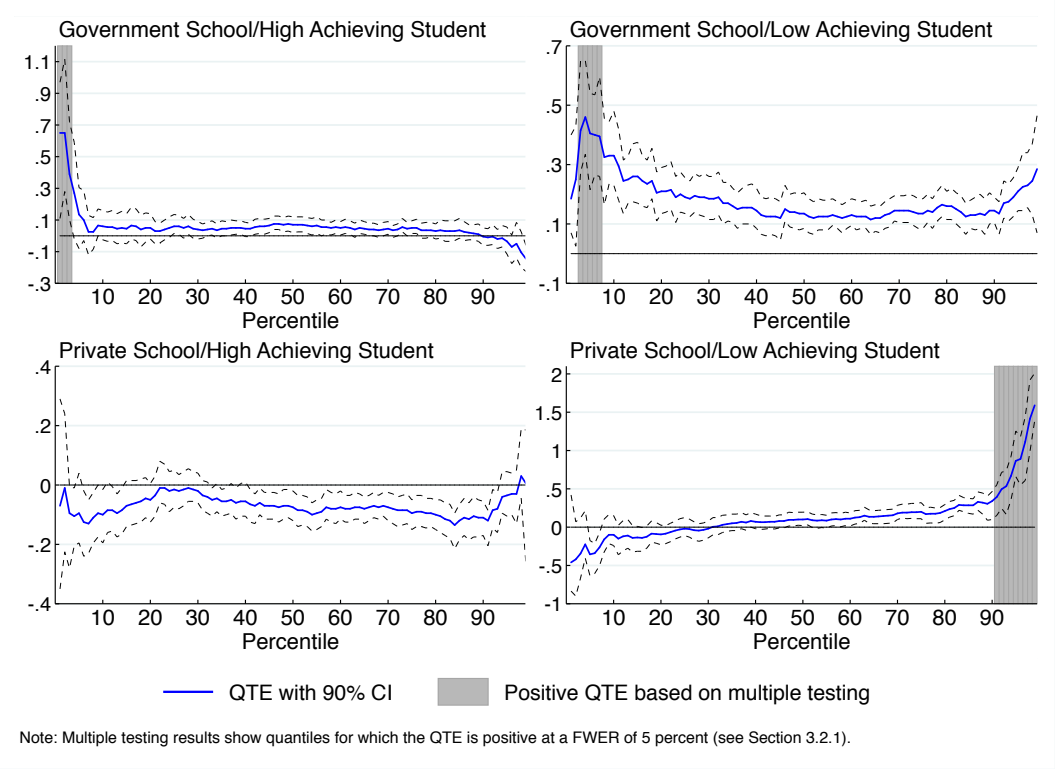


Figure 5: Quantile Treatment Effects and Multiple Testing Results by School Type and Child's Performance Relative to School Performance

null hypothesis posits that there are no differences across subgroups that can explain the observed heterogeneity of QTEs in the full sample. We can reject (H.5) for all but two sets of subgroups at a level of 5 percent. The p-value is largest for subgroups defined by mother's and father's education. Hence, for these two subgroup categories, we cannot reject the null hypothesis that the treatment effect is constant across test score percentiles for all subgroups at the 5 percent level. Overall, however, we conclude that differences across subgroups do not explain the observed distributional treatment effects in the whole sample. Our test relaxes the strong assumption of treatment effect homogeneity within subgroups that is implicit in Bitler, Gelbach, and Hoynes (2017).

The tests of hypothesis (H.6) shown in Table 4 additionally account for potential dependencies within and across subgroups. These test results provide additional insight beyond testing (H.5) because they identify the individual subgroups within a class of subgroups that exhibit treatment effect heterogeneity. In these results, a p-value below 0.05 indicates that the corresponding subgroup exhibits a statistically significant amount of treatment effect heterogeneity across the test score distribution. In each and every subgroup category, we find evidence of treatment effect heterogeneity. These results clearly suggest a substantial

Table 3: Testing for Treatment Effect Heterogeneity Between Subgroups

Subgroup category	Test statistic	Critical value at 5%	p-value
Child's gender	0.394	0.292	0.006
Child's baseline test score	0.516	0.421	0.014
Mother's education	0.292	0.888	0.985
Father's education	0.312	0.963	0.992
Village wealth	0.73	0.343	0
Village literacy rate	0.545	0.358	0
School Herfindahl Index	0.494	0.295	0.002
Village size	0.501	0.389	0.014
School type and school performance	1.166	0.624	0
School type and child's performance relative to school	1.468	0.633	0

Notes: This table shows test results for hypothesis (H.5), i.e. these tests show for which subgroups categories we can reject treatment effects that are homogenous within subgroups for all subgroups.

amount of treatment effect heterogeneity between subgroups and across the student performance distribution within subgroups.

5 Conclusion

In this paper we employ six general tests for treatment effect heterogeneity in settings with selection on observables. These tests allow researchers to provide policymakers with guidance on complex patterns of treatment effect heterogeneity both within and across subgroups. In the present context, the results can guide policymakers in adjusting how information on student performance is provided, for example by introducing more (or different) conditions across villages. In contrast to much of the existing literature, these tests make corrections for multiple testing and therefore provide valid inference under dependence between subgroups and quantiles. Further, our tests generalize the idea of tests considered in Bitler, Gelbach, and Hoynes (2017) by not restricting treatment effects to be constant across quantiles within a subgroup when determining if the distributional heterogeneity across the full sample is characterized by subgroups.

Table 4: Testing Which Subgroups Exhibit Treatment Effect Heterogeneity

Subgroup category	Test statistic	p-value
Child's gender		
Female	0.394	0
Male	0.306	0
Child's baseline test score		
Above median	0.516	0
Below median	0.506	0
Village wealth		
Above Median	0.19	0.01
Below Median	0.73	0
Village literacy rate		
Above median	0.545	0
Below median	0.121	0.01
School Herfindahl Index		
Above median	0.224	0
Below median	0.494	0
Village size		
Above median	0.5	0
Below median	0.155	0.02
School type and school performance		
High scoring govern.	1.073	0
Low scoring govern.	0.226	0
High scoring private	0.0754	0
Low scoring private	1.166	0
School type and child's performance relative to school		
Govern./high achieving	0.593	0
Govern./low achieving	0.277	0
Private/high achieving	0.101	0
Private/low achieving	1.468	0

Notes: This table shows test results for hypothesis (H.6), i.e. these tests show for which subgroups in each subgroup category we can reject homogenous treatment effects. p-values are calculated using a grid with step size 0.005. Hence an entry of zero indicates that the corresponding p-value is below 0.005.

Using data from Andrabi, Das, and Khwaja (2017), we not only present evidence of considerable heterogeneity of the effects of access to report cards on student achievement for most subgroups, but show in which subgroups and which test score quantiles within subgroups the benefits of information provision are highest. In addition, our empirical analysis emphasizes the importance of correcting for multiple testing. Testing across different subgroups is policy relevant, and while Crump et al. (2008) provide an approach to select which subpopulations to study, our tests go further by considering treatment effect heterogeneity conditional on observable characteristics.

Given the considerable attention policymakers pay to developing accountability programs worldwide, our results highlight for which groups targeted information provision would likely yield higher returns. Further, these returns should exceed programs that disclose school quality to parents of all students. That said, education policymakers face additional challenge from incorporating evidence of heterogeneous treatment effects into the design of any policy that may lead to different school choice. While Pareto improvements in welfare can easily be achieved in social and labor policy using ex-post targeted transfers, the effectiveness of redistributing students across schools also depends on how peer groups influence academic outcomes.[31]

We would like to conclude by emphasizing that our multiple testing approach is generally applicable in various other ways beyond what this paper demonstrated. First, the tests can be applied to situations with multiple treatments (e.g., List, Shaikh, and Xu, 2016) or situations where there is selection on unobservables that explore if there is heterogeneity in marginal treatment effects (e.g., Heckman and Vytlacil, 2005; Brinch, Mogstad, and Wiswall, 2017). Second, instead of using inverse propensity score weighting, we may directly use the conditional distribution functions or conditional quantile functions to identify the treatment effects as proposed by Chernozhukov, Fernandez-Val, and Melly (2013). Extending their proposal to multiple testing procedures to test for treatment effect heterogeneity across the distribution or quantile function with or without subgroups has the potential to complement this paper by expanding insights in empirical microeconomics.

References

Abadie, Alberto. 2002. "Bootstrap Tests for Distributional Treatment Effects in Instrumental Variable Models." *Journal of the American Statistical Association* 97 (457):284–292.

[31] These challenges are illustrated in Ding and Lehrer (2007) who use a partial linear model to demonstrate the non-linear shape of the peer effect function changes from convex to concave to convex across the test score distribution.

Abadie, Alberto, Joshua Angrist, and Guido Imbens. 2002. "Instrumental Variables Estimates of the Effect of Subsidized Training on the Quantiles of Trainee Earnings." *Econometrica* 70 (1):91–117.

Andrabi, Tahir, Jishnu Das, and Asim Ijaz Khwaja. 2017. "Report Cards: The Impact of Providing School and Child Test Scores on Educational Markets." *American Economic Review* 107 (6):1535–63.

Armstrong, Timothy B. and Shu Shen. 2015. "Inference on Optimal Treatment Assignments." Cowles Foundation Discussion Paper 1927RR.

Becker, Gary S. 1995. *Human Capital and Poverty Alleviation*. World Bank, Human Resources Development and Operations Policy.

Bitler, Marianne P., Jonah B. Gelbach, and Hilary W. Hoynes. 2006. "What Mean Impacts Miss: Distributional Effects of Welfare Reform Experiments." *American Economic Review* 96 (4):988–1012.

———. 2017. "Can Variation in Subgroups' Average Treatment Effects Explain Treatment Effect Heterogeneity? Evidence from a Social Experiment." *Review of Economics and Statistics* 99 (4):683–697.

Brinch, Christian N, Magne Mogstad, and Matthew Wiswall. 2017. "Beyond LATE with a discrete instrument." *Journal of Political Economy* 125 (4):985–1039.

Camargo, Braz, Rafael Camelo, Sergio Firpo, and Vladimir Ponczek. 2014. "Information, Market Incentives, and Student Performance." IZA Discussion Paper 7941.

———. 2018. "Information, Market Incentives, and Student Performance Evidence from a Regression Discontinuity Design in Brazil." *Journal of Human Resources* 53 (2):414–444.

Carneiro, Pedro, Jishnu Das, and Hugo Reis. 2013. "Parental valuation of school attributes in developing countries: Evidence from Pakistan." Unpublished manuscript.

Chernozhukov, Victor, Ivan Fernandez-Val, and Blaise Melly. 2013. "Inference on Counterfactual Distributions." *Econometrica* 81 (6):2205–2268.

Crump, Richard K., V. Joseph Hotz, Guido W. Imbens, and Oscar A. Mitnik. 2008. "Nonparametric Tests for Treatment Effect Heterogeneity." *Review of Economics and Statistics* 90 (3):389–405.

Deaton, Angus. 2010. "Instruments, Randomization, and Learning about Development." *Journal of Economic Literature* 48 (2):424–455.

Dehejia, Rajeev H. 2005. "Program Evaluation as a Decision Problem." *Journal of Econometrics* 125 (1-2):141–173.

Ding, Weili and Steven F Lehrer. 2007. "Do peers affect student achievement in China's secondary schools?" *The Review of Economics and Statistics* 89 (2):300–312.

Djebbari, Habiba and Jeffrey Smith. 2008. "Heterogeneous Impacts in PROGRESA." *Journal of Econometrics* 145 (1-2):64–80.

Figlio, David and Susanna Loeb. 2011. "School accountability." In *Handbook of the Economics of Education*, vol. 3. Elsevier, 383–421.

Fink, Gunther, Margaret McConnell, and Sebastian Vollmer. 2014. "Testing for Heterogeneous Treatment Effects in Experimental Data: False Discovery Risks and Correction Procedures." *Journal of Development Effectiveness* 6 (1):44–57.

Firpo, Sergio. 2007. "Efficient Semiparametric Estimation of Quantile Treatment Effects." *Econometrica* 75 (1):259–276.

Friedlander, Daniel and Philip K. Robins. 1997. "The Distributional Impacts of Social Programs." *Evaluation Review* 21 (5):531–553.

Friedman, Milton. 1955. *The Role of Government in Education*. Rutgers University Press New Brunswick, NJ.

Friesen, Jane, Mohsen Javdani, Justin Smith, and Simon Woodcock. 2012. "How do school report cards affect school choice decisions?" *Canadian Journal of Economics/Revue canadienne d'économique* 45 (2):784–807.

Gibbons, Stephen and Stephen Machin. 2006. "Paying for primary schools: admission constraints, school popularity or congestion?" *The Economic Journal* 116 (510):C77–C92.

Gu, Jiaying and Shu Shen. 2017. "Oracle and Adaptive False Discovery Rate Controlling Method for One-Sided Testing: Theory and Application in Treatment Effect Evaluation." *The Econometrics Journal* 21 (1):11–35.

Hastings, Justine, Thomas J Kane, and Douglas O Staiger. 2009. "Heterogeneous preferences and the efficacy of public school choice." NBER Working Paper 12145.

Hastings, Justine S and Jeffrey M Weinstein. 2008. "Information, school choice, and academic achievement: Evidence from two experiments." *The Quarterly Journal of Economics* 123 (4):1373–1414.

Heckman, James J. 2001. "Micro Data, Heterogeneity, and the Evaluation of Public Policy: Nobel Lecture." *Journal of Political Economy* 109 (4):673–748.

Heckman, James J., Jeffrey Smith, and Nancy Clements. 1997. "Making The Most Out Of Programme Evaluations and Social Experiments: Accounting For Heterogeneity in Programme Impacts." *Review of Economic Studies* 64 (4):487–535.

Heckman, James J. and Sergio Urzua. 2010. "Comparing IV with Structural Models: What Simple IV Can and Cannot Identify." *Journal of Econometrics* 156 (1):27–37.

Heckman, James J. and Edward Vytlacil. 2005. "Structural Equations, Treatment Effects, and Econometric Policy Evaluation1." *Econometrica* 73 (3):669–738.

Holmström, Bengt. 1999. "Managerial incentive problems: A dynamic perspective." *The Review of Economic Studies* 66 (1):169–182.

Hoxby, Caroline M. 2003. "School choice and school productivity. Could school choice be a tide that lifts all boats?" In *The Economics of School Choice*. University of Chicago Press, 287–342.

Imbens, Guido W. 2010. "Better LATE than Nothing: Some Comments on Deaton (2009) and Heckman and Urzua (2009)." *Journal of Economic Literature* 48 (2):399–423.

Kitagawa, Toru and Aleksey Tetenov. 2018. "Who Should Be Treated? Empirical Welfare Maximization Methods for Treatment Choice." *Econometrica* 86 (2):591–616.

Koning, Pierre and Karen Van der Wiel. 2012. "School responsiveness to quality rankings: An empirical analysis of secondary education in the netherlands." *De Economist* 160 (4):339–355.

Lee, Soohyung and Azeem M. Shaikh. 2014. "Multiple Testing and Heterogeneous Treatment Effects: Re-Evaluating the Effect of PROGRESA on School Enrollment." *Journal of Applied Econometrics* 29 (4):612–626.

List, John A., Azeem M. Shaikh, and Yang Xu. 2016. "Multiple Hypothesis Testing in Experimental Economics." NBER Working Paper 21875.

Maier, Michael. 2011. "Tests For Distributional Treatment Effects Under Unconfoundedness." *Economics Letters* 110 (1):49–51.

Manski, Charles F. 2004. "Statistical Treatment Rules for Heterogeneous Populations." *Econometrica* 72 (4):1221–1246.

Milgrom, Paul and John Roberts. 1986. "Price and advertising signals of product quality." *Journal of Political Economy* 94 (4):796–821.

Mizala, Alejandra and Miguel Urquiola. 2013. "School markets: The impact of information approximating schools' effectiveness." *Journal of Development Economics* 103:313–335.

Romano, Joseph P. and Azeem M. Shaikh. 2010. "Inference for the Identified Set in Partially Identified Econometric Models." *Econometrica* 78 (1):169–211.

Romano, Joseph P., Azeem M. Shaikh, and Michael Wolf. 2010. "Hypothesis Testing in Econometrics." *Annual Review of Economics* 2 (1):75–104.

Romano, Joseph P. and Michael Wolf. 2005a. "Exact and Approximate Stepdown Methods for Multiple Hypothesis Testing." *Journal of the American Statistical Association* 100 (469):94–108.

———. 2005b. "Stepwise Multiple Testing as Formalized Data Snooping." *Econometrica* 73 (4):1237–1282.

Rothe, Christoph. 2010. "Nonparametric Estimation of Distributional Policy Effects." *Journal of Econometrics* 155 (1):56–70.

Schneider, Mark, Gregory Elacqua, and Jack Buckley. 2006. "School choice in Chile: Is it class or the classroom?" *Journal of Policy Analysis and Management* 25 (3):577–601.

Shapiro, Carl. 1983. "Premiums for high quality products as returns to reputations." *The Quarterly Journal of Economics* 98 (4):659–679.

Smith, Jeffrey A. and Petra E. Todd. 2005. "Does Matching Overcome LaLonde's Critique of Nonexperimental Estimators?" *Journal of Econometrics* 125 (1-2):305–353.

Solon, Gary, Steven J. Haider, and Jeffrey M. Wooldridge. 2015. "What Are We Weighting For?" *Journal of Human Resources* 50 (2):301–316.

White, Halbert. 2000. "A Reality Check for Data Snooping." *Econometrica* 68 (5):1097–1126.

Wolinsky, Asher. 1983. "Prices as signals of product quality." *The Review of Economic Studies* 50 (4):647–658.