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

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

R. Vincent Pohl[‡]
Mathematica
Policy Research

Kyungchul Song[§]
University of
British Columbia

August 2019

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 75% of statistically significant quantile treatment effects become insignificant once corrections for multiple testing are applied.

Keywords: Information; Student Performance; Accountability; Quantile Treatment Effects; Multiple Testing; Bootstrap Tests.

JEL classification: C12; C21; I21; L15.

*We 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 thank SSHRC for research support. All remaining errors are our own. An Online Appendix to this paper is available at https://rvpohl.github.io/files/LehrerPohlSong_Multiple_App.pdf

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

[‡]Email: vincent.pohl@gmail.com.

[§]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 education. 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.

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. 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.

While an increasing number of studies account for possible treatment effect heterogeneity when evaluating programs or other interventions, most conduct statistical inference without allowing for dependence across subgroups. As Fink, McConnell, and Vollmer (2014) point out, a majority of studies based on field experiments published in 10 specific journals estimate separate average causal parameters for different subgroups, but report traditional standard errors and p-values when testing for heterogeneous treatment effects through interaction terms or subgroup analyses. This is inappropriate because each interaction term represents a separate hypothesis beyond the original experimental design and results in a substantially increased type I error.[1] 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), i.e. the probability of rejecting at least one true null hypothesis, in finite samples.

A similar observation can be made for distributional treatment effects. A growing number of studies examine if treatment effects differ across quantiles of the outcome variable, i.e. they estimate quantile treatment effects (QTEs) (e.g., Heckman, Smith, and Clements, 1997;

[1]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, since the probability of a false positive equals 0.05 for each individual hypothesis, the probability of falsely rejecting at least one true null hypothesis may be much larger. Hence, the type I error exceeds the nominal size of the test.

Friedlander and Robins, 1997; Abadie, 2002; Bitler, Gelbach, and Hoynes, 2006; Firpo, 2007). Testing for the presence of positive (or, generally, non-zero) QTEs involves a test of multiple hypotheses, for example 99 hypotheses in the case of percentile treatment effects. Therefore, the 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). As a result, the type I error rates can exceed the desired level of the test, which leads researchers to reject "too many" individual hypotheses.[2] To the best of our knowledge, no published study estimating distributional treatment effects makes such a correction.[3] The absence of these corrections may reflect that multiple testing procedures for QTEs were not previously developed.

We fill this gap in the literature by developing a flexible multiple testing procedure that controls the FWER asymptotically in not only the full sample but also between and within subgroups. One goal of this paper is to provide the practitioner with easy to implement methods that can be applied in settings that empiricists regularly encounter. We formally show the validity of our multiple testing method and in the Online Appendix establish limit results for the bootstrap quantile treatment effect process. While theoretical econometricians have studied the empirical quantile process, we contribute to that literature by providing a general framework for testing quantile treatment effects with estimated propensity scores that does not rely on high level conditions such as requiring an asymptotic linear representation of a semiparametric process.[4] A second goal is to illustrate the importance of making

[2]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.

[3]Among articles published in five high-impact economic journals between 2008 to 2017 that estimate distributional treatment effects none corrects inference for multiple tesing (Allen, Clark, and Houde, 2014; Angrist, Lang, and Oreopoulos, 2009; Bandiera et al., 2017; Banerjee et al., 2015; Behaghel, de Chaisemartin, and Gurgand, 2017; Brown et al., 2014; Crepon et al., 2015; Evans and Garthwaite, 2012; Fack and Landais, 2010; Fairlie and Robinson, 2013; McKenzie, 2017; Meyer and Sullivan, 2008; Muralidharan, Niehaus, and Sukhtankar, 2016).

[4]For example, Koenker and Xiao (2002) and Chernozhukov and Fernandéz-Val (2005) proposed inference on a quantile process, the former based on asymptotic inference using a martingale transform and the latter using subsampling. These studies did not consider quantile treatment effects with estimated propensity scores or bootstrap inference. Similarly, Firpo (2007) proposed efficient estimation of quantile treatment effects but did not give a full limit result for the quantile treatment effect as a stochastic process of quantiles, as we do in our paper. Our work is most closely related to Lee, Song, and Whang (2018) who developed a general framework for testing for functional inequalities which includes inequalities on a nonparametric conditional quantile process. However, their development relies on high level asymptotic linear representation of a semiparametric process unlike ours. In semiparametric econometrics, the demonstration of the root-n consistency and asymptotic normality of an estimator depends on the complexity of the asymptotic linearity representation, which in turn often depends on the complexity of the estimator. See our Online Appendix for details on how the development of our framework does not rely on such conditions.

corrections for multiple testing by reexamining the effectiveness of a program that provided information on student and school test score performance to parents.

Our proposed multiple testing procedure can determine whether a treatment has a (positive) effect for any quantile and detect treatment effect heterogeneity across the outcome distribution and subgroups. Further, it can identify the subgroups and outcome quantiles for which the treatment effect is estimated to be conspicuous beyond sampling variations. Finally, it lets us determine which subgroups exhibit heterogeneous treatment effects. Importantly, we also adjust for dependencies between quantiles within subgroups whereas Bitler, Gelbach, and Hoynes (2017) only allow for treatment effects that vary across subgroups but are constant within subgroup. Thereby, we provide a unified framework to test for treatment effect heterogeneity.

As our results are obtained through a formal multiple testing procedure, they properly take into account the reuse of the same data for different demographic groups or quantile groups and controls the FWER so that it is unaffected by data mining.[5] Controlling the FWER in multiple comparisons across different quantiles is crucial for the validity of the inference procedure, as treatment effects across different quantiles of the outcome distribution are estimated using the same underlying data.[6]

Our use of formal testing procedures for treatment effect heterogeneity is not solely motivated by policy considerations, but also economic theory. The multiple testing approach provides a basis for judging the empirical relevance of treatment effect heterogeneity and sheds light on the pattern of treatment effect heterogeneity across different population groups.[7] Since our multiple testing framework identifies subpopulations with positive responses to the outcome variable, it gives policymakers rich information to more effectively assign different treatments to individuals.[8] For example, policymakers can use the results to modify

[5]More specifically, our procedure involves multiple inequalities of unconditional quantile functions, and draws on a bootstrap method for testing of inequality restrictions. To construct a multiple testing procedure that controls the FWER, we adapt the step-down method proposed by Romano and Wolf (2005) to our context of testing multiple inequalities of unconditional quantiles.

[6]Prior work does not consider conditional quantiles. In contrast to our approach, Lee and Shaikh (2014) do not consider within-subgroup treatment effect heterogeneity. They also require the treatment to be randomly assigned unconditionally whereas our approach permits selection on observables. List, Shaikh, and Xu (forthcoming) additionally consider an experimental settings with multiple treatments, multiple outcomes, and multiple subgroups, but also do not account for within-subgroup treatment effect heterogeneity.

[7]While Crump et al. (2008) focus on heterogeneity of the average treatment effect across subgroups, our focus is on treatment effect heterogeneity across quantiles of the outcome distribution. Specifically, 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.

[8]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

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).[9] 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).

We reexamine data from Andrabi, Das, and Khwaja's (2017) Pakistani school report card field experiment and present evidence that correcting for multiple testing is empirically important and policy relevant. Specifically, 75 percent of the estimated statistically significant QTEs of information provision on children's test scores become insignificant once multiple testing corrections are applied. These findings also demonstrate that the significantly positive effects of providing information to parents reported in Andrabi, Das, and Khwaja (2017) are concentrated in the bottom quintile of the test score distribution. Further, we find clear evidence of treatment effect heterogeneity in the full sample and every subgroup that we consider. Taken together, our results shed new light on the effectiveness of accountability

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.

[9]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.

programs, further indicating how schools and parents respond to the release of information on student performance.

The rest of this paper is organized as follows. In Section 2, we introduce the general testing procedures for treatment effect heterogeneity across quintiles of the outcome distribution and subgroups. In Section 3, we describe the experiment and economic model that underlie the data being investigated. This model predicts heterogenous treatment effects both within and across subgroups. In Section 4, we present results from our empirical application of the methods to the Andrabi, Das, and Khwaja (2017) experimental data. 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 Methodology

2.1 Testing for Treatment Effect Heterogeneity

To develop a multiple testing procedure for various hypotheses of QTEs, we consider the following data generating set-up. Let D_i be a random variable that takes values in $\{0, 1\}$, where $D_i = 1$ indicates 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

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

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 of individual i. The researcher observes a random sample of $(Y_i, D_i, X_i)_{i=1}^n$. We make the following standard assumptions of selection on observables and common support.

Assumption 2.1. *(i)* (Y_{1i}, Y_{0i}) *is conditionally independent of* D_i *given* X_i.

(ii) There exists $\varepsilon > 0$ *such that for all* $x \in \mathcal{X}$ *and* $d \in \{0, 1\}$, $\varepsilon \leqslant p_d(x) \leqslant 1 - \varepsilon$, *where* $p_d(x) = P\{D_i = d | X_i = x\}$.

Further, we assume that X_i can be partitioned $X_i = (X_{1i}, Z_i)$, where Z_i is a discrete random subvector and X_{1i} indicates the vector that is not included in Z_i. The subvector Z_i determines to which subgroup individual i belongs. For each z in the support of Z_i,

$\tau \in (0,1)$, and $d \in \{0,1\}$, we define

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

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. Note that we take τ to run in a continuum. The subgroup QTE at a quantile-subgroup pair (τ, z) is then defined by

$$q^{\Delta}(\tau, z) = q_1(\tau, z) - q_0(\tau, z).$$

Let τ_L, τ_U be such that $\tau_L < \tau_U$ and $\tau_L, \tau_U \in (0,1)$ and let $\mathcal{Z}$ be the support of Z_i. We take $S = [\tau_L, \tau_U] \times \mathcal{Z}$ to be the set of quantile-subgroup pairs (τ, z) that we focus on. We are interested in the hypothesis of the following form: for each $(\tau, z) \in S$,

$$H_0(\tau, z) : \gamma(q^{\Delta}; \tau, z) = 0, \text{ vs } H_1(\tau, z) : \gamma(q^{\Delta}; \tau, z) \neq 0, \tag{1}$$

where $\gamma(q^{\Delta}; \tau, z)$ is a functional of q^{Δ} that depends on (τ, z). Using an appropriate functional $\gamma(q^{\Delta}; \tau, z)$, the hypothesis expressed in (1) also allows for one-sided hypothesis tests. Examples of specific hypothesis testing problems involving QTE are provided in Table 1. The examples illustrate three tests for quantiles and QTEs for individual subgroups as well as three tests for the special case when $Z_i = 1$, which defines quantiles and QTEs for the full sample.

2.1.1 Joint Hypothesis Testing

We can combine the individual hypotheses into a joint hypothesis:

$$H_0 : \Gamma(q^{\Delta}; S) = 0, \text{ vs } H_1 : \Gamma(q^{\Delta}; S) \neq 0, \tag{2}$$

where for each $S' \subset S$, we define

$$\Gamma(q^{\Delta}; S') = \sup_{(\tau,z) \in S'} \gamma(q^{\Delta}; \tau, z). \tag{3}$$

The null hypothesis says that for all $(\tau, z) \in S$, $H_0(\tau, z)$ expressed in (1) is true. As we shall see later, joint hypothesis testing is useful for testing the presence of QTE or QTE heterogeneity.

Table 1: Examples for Joint and Multiple Hypothesis Tests Involving QTE

Hypothesis Being Tested	Indiv. Hypothesis $\gamma(q^\Delta;\tau,z)$	Joint Hypothesis $\Gamma(q^\Delta;S')$	Multiple Hypothesis $\{S_w : w \in W\}$ in (4)
Testing for the presence of QTE in the whole sample			
$H_0(\tau): q^\Delta(\tau) = 0$, vs $H_1(\tau): q^\Delta(\tau) \neq 0$	$\lvert q^\Delta(\tau)\rvert$	$\sup_{\tau\in[\tau_L,\tau_U]} \lvert q^\Delta(\tau)\rvert$	$\{\{\tau\} : \tau \in [\tau_L,\tau_U]\}$
Testing for positive QTE in the whole sample			
$H_0(\tau): q^\Delta(\tau) \leqslant 0$, vs $H_1(\tau): q^\Delta(\tau) > 0$	$\max\{q^\Delta(\tau), 0\}$	$\sup_{\tau\in[\tau_L,\tau_U]} \max\{q^\Delta(\tau), 0\}$ **(H.1)**	$\{\{\tau\} : \tau \in [\tau_L,\tau_U]\}$ **(H.3)**
Testing for QTE heterogeneity in the whole sample			
$H_0(\tau): q^\Delta(\tau) = c$, vs $H_1(\tau): q^\Delta(\tau) \neq c$	$\lvert q^\Delta(\tau) - c\rvert$	$\sup_{\tau\in[\tau_L,\tau_U]} \lvert q^\Delta(\tau) - c\rvert$ **(H.2)**	$\{\{\tau\} : \tau \in [\tau_L,\tau_U]\}$
Testing for the presence of QTE across quantiles and subgroups			
$H_0(\tau,z): q^\Delta(\tau,z) = 0$, vs $H_1(\tau,z): q^\Delta(\tau,z) \neq 0$	$\lvert q^\Delta(\tau,z)\rvert$	$\sup_{(\tau,z)\in S'} \lvert q^\Delta(\tau,z)\rvert$	$\{\{\tau,z\} : (\tau,z) \in [\tau_L,\tau_U]\times\mathcal{Z}\}$
Testing for positive QTE across quantiles and subgroups			
$H_0(\tau,z): q^\Delta(\tau,z) \leqslant 0$, vs $H_1(\tau,z): q^\Delta(\tau,z) > 0$	$\max\{q^\Delta(\tau,z), 0\}$	$\sup_{(\tau,z)\in S'} \max\{q^\Delta(\tau,z), 0\}$	$\{\{\tau,z\} : (\tau,z) \in [\tau_L,\tau_U]\times\mathcal{Z}\}$ **(H.4)**
Testing for QTE heterogeneity in some subgroups			
$H_0(\tau,z): q^\Delta(\tau,z) = c(z)$, vs $H_1(\tau,z): q^\Delta(\tau,z) \neq c(z)$	$\lvert q^\Delta(\tau,z) - c(z)\rvert$	$\sup_{(\tau,z)\in S'} \lvert q^\Delta(\tau,z) - c(z)\rvert$ **(H.5)**	$\{\{\tau,z\} : (\tau,z) \in [\tau_L,\tau_U]\times\mathcal{Z}\}$
Testing for which subgroups QTE are heterogenous			
$H_0(\tau,z): q^\Delta(\tau,z) = c(z)$, vs $H_1(\tau,z): q^\Delta(\tau,z) \neq c(z)$	$\lvert q^\Delta(\tau,z) - c(z)\rvert$	$\sup_{\tau\in[\tau_L,\tau_U]} \lvert q^\Delta(\tau,z) - c(z)\rvert$	$\{\{(\tau,z) : \tau \in [\tau_L,\tau_U]\} : z \in \mathcal{Z}\}$ **(H.6)**

Notes: $q^\Delta(\tau)$ is the same as $q^\Delta(\tau,z)$ except that Z_i is taken to be a constant, say, 1, for all $i = 1,...,n$. We define $c = \frac{1}{\tau_U-\tau_L}\int_{\tau_L}^{\tau_U} q^\Delta(\tau)d\tau$ and $c(z) = \frac{1}{\tau_U-\tau_L}\int_{\tau_L}^{\tau_U} q^\Delta(\tau,z)d\tau$, i.e. the mean QTE for the whole sample and conditional on subgroup. (H.1) through (H.6) denote the hypotheses being tested in the empirical application in Section 4.

2.1.2 Multiple Hypothesis Testing

Often, we are interested in finding out which quantile-subgroup pair (τ, z) is responsible for the rejection of the joint null hypothesis expressed in (2). To address this question, let us consider the following multiple hypothesis testing problem. Suppose that

$$\mathcal{S} = \{S_w \subset S : w \in W\}, \tag{4}$$

where W is an index set, and S_w is a subset of S such that $S_w \cap S_{w'} = \varnothing$ whenever $w \neq w'$. Our focus is to find $w \in W$ such that the violation of the joint null hypothesis expressed in (2) is due to the violation of $H_0(\tau, z)$ for some $(\tau, z) \in S_w$. For this, we first define

$$W_P = \{w \in W : \gamma(q^{\Delta}; \tau, z) \neq 0, \text{ for some } (\tau, z) \in S_w\}.$$

Our goal here is to find a data-dependent set $\hat{W}$ that satisfies

$$\limsup_{n \to \infty} P\{\hat{W} \backslash W_P \neq \varnothing\} \leqslant \alpha. \tag{5}$$

The probability in equation (5) is the probability of mistakenly declaring the violation of $H_0(\tau, z)$ for some $(\tau, z) \in S_w$ for every $w \in \hat{W}$, and is called the FWER in the literature on multiple testing. We aim to construct such a set $\hat{W}$ that controls the FWER under a small number α asymptotically.

2.2 A Bootstrap Step-Down Procedure for Multiple Testing

2.2.1 Estimation of QTE and Bootstrap Joint Testing

The identification and inference on $q^{\Delta}(\tau, z)$ for *each quantile* are established by Firpo (2007). Here we propose joint hypothesis testing and multiple hypothesis testing procedures and provide conditions under which the FWER is asymptotically under control.[10]

To motivate estimation of $q^{\Delta}(\tau, z)$, note that we can identify $q_d(\tau, z)$ by

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

[10]Extending the results to the case of cluster dependence is straightforward, as long as two conditions are satisfied: first, the observations are all identically distributed across the cross-sectional units, and second, the number of the clusters increase to infinity as the number of observations does so. For the bootstrap inference, one can simply use block bootstrap in which one resamples clusters with replacement instead of individual sample units.

where $\omega_{di} = 1\{D_i = d\}/p_d(X_i)$ and $\rho_\tau(x) = x \cdot (\tau - 1\{x \leqslant 0\})$ is the check function. Thus, we estimate $q_d(\tau, z)$ by

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

with $\hat{\omega}_{di} = 1\{D_i = d\}/\hat{p}_d(X_i)$, and $\hat{p}_d(x)$ is the estimated propensity score.[11] Following Firpo (2007), we obtain

$$\hat{q}^\Delta(\tau, z) = \hat{q}_1(\tau, z) - \hat{q}_0(\tau, z).$$

To construct a joint test or a multiple 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 $\hat{Y}_{di}^* = Y_i^* 1\{D_i^* = d\}/\hat{p}_d^*(X_i^*)$, where $\{(Y_i^*, D_i^*, X_i^*)\}_{i=1}^n$ denotes each bootstrap sample and $\hat{p}_1^*(X_i^*)$ the estimated propensity score using the bootstrap sample. Then we construct

$$\hat{q}^{\Delta *}(\tau, z) = \hat{q}_1^*(\tau, z) - \hat{q}_0^*(\tau, z),$$

where $\hat{q}_1^*(\tau, z)$ and $\hat{q}_0^*(\tau, z)$ are the τ-th empirical quantiles of $\{\hat{Y}_{1i}^*\}_{i=1}^n$ and $\{\hat{Y}_{0i}^*\}_{i=1}^n$, respectively, within the samples with $Z_i^* = z$.

For joint hypothesis testing expressed in (2), we construct test statistics

$$T = \Gamma(\hat{q}^\Delta; S), \text{ and } T^* = \Gamma(\hat{q}^{\Delta *} - \hat{q}^\Delta; S),$$

and use as critical value the $(1-\alpha)$-th percentile from the bootstrap distribution of T^*. 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.

2.2.2 Bootstrap Multiple Testing Procedure for QTEs

The multiple testing procedure adapts the step-down method of Romano and Wolf (2005) and Romano and Shaikh (2010) to our set-up. First, let $\hat{q}_b^{\Delta *}$ be the same as $\hat{q}^{\Delta *}$ except that it is made explicit that $\hat{q}_b^{\Delta *}$ is constructed using the b-th bootstrap sample. For each subset $W' \subset W$, we define

$$T_b^*(W') = \sup_{w \in W'} \Gamma(\hat{q}_b^{\Delta *} - \hat{q}^\Delta; S_w).$$

[11] Following Smith and Todd (2005), the propensity score $\hat{p}(x)$ is estimated using data from the full sample.

Setting $\tilde{W}_1 = W$, we take $\hat{c}_{1-\alpha}(\tilde{W}_1)$ to be the smallest c such that

$$\frac{1}{B}\sum_{b=1}^{B} \mathbf{1}\left\{T_b^*(\tilde{W}_1) \leqslant c\right\} \geqslant 1-\alpha.$$

That is, at $\hat{c}_{1-\alpha}(\tilde{W}_1)$, 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 $\hat{c}_{1-\alpha}(\tilde{W}_1)$, i.e., we define

$$\tilde{W}_2 = \left\{w \in W : \Gamma(\hat{q}^{\Delta}; S_w) \leqslant \hat{c}_{1-\alpha}(\tilde{W}_1)\right\},$$

so that $\tilde{W}_2$ is a subset of $\tilde{W}_1$. Now, we take $\hat{c}_{1-\alpha}(\tilde{W}_2)$ to be the smallest c such that

$$\frac{1}{B}\sum_{b=1}^{B} \mathbf{1}\left\{T_b^*(\tilde{W}_2) \leqslant c\right\} \geqslant 1-\alpha.$$

Using this, we define

$$\tilde{W}_3 = \left\{w \in W : \Gamma(\hat{q}^{\Delta}; S_w) \leqslant \hat{c}_{1-\alpha}(\tilde{W}_2)\right\}.$$

This procedure is repeated until at step k, we obtain

$$\tilde{W}_k = \left\{w \in W : \Gamma(\hat{q}^{\Delta}; S_w) \leqslant \hat{c}_{1-\alpha}(\tilde{W}_{k-1})\right\}$$

such that no further element of $\tilde{W}_k$ is eliminated (i.e. $\tilde{W}_k = \tilde{W}_{k-1}$). We take

$$\hat{W} = W \backslash \tilde{W}_k. \tag{6}$$

To incorporate multiple testing for QTE heterogeneity (hypothesis (H.6) in Table 1), we take the definition in (3) and follow the same step-down procedure after replacing $\Gamma(\hat{q}^{\Delta}; S_w)$ by $\Gamma(\hat{q}^{\Delta} - \overline{q}^{\Delta}; S_w)$ and $\Gamma(\hat{q}_b^{\Delta *} - \hat{q}^{\Delta}; S_w)$ by $\Gamma(\hat{q}_b^{\Delta *} - \hat{q}^{\Delta} - (\overline{q}_b^{\Delta *} - \overline{q}^{\Delta}); S_w)$, where

$$\overline{q}^{\Delta}(z) = \frac{1}{\tau_U - \tau_L}\int_{\tau_L}^{\tau_U} \hat{q}^{\Delta}(\tau, z)d\tau, \text{ and } \overline{q}^{\Delta *}(z) = \frac{1}{\tau_U - \tau_L}\int_{\tau_L}^{\tau_U} \hat{q}^{\Delta *}(\tau, z)d\tau.$$

2.2.3 Asymptotic Control of FWER

In this subsection, we provide conditions that ensure the set $\hat{W}$ (in (6)) obtained through the step-down procedure controls the FWER asymptotically. For brevity, we focus on a situation where $Z_i = 1$ for all $i = 1, ..., n$, allowing us to suppress the argument z from

$q_d(\tau, z)$, $q_d^{\Delta}(\tau, z)$, and $\gamma(q^{\Delta}; \tau, z)$, writing them as $q_d(\tau)$, $q_d^{\Delta}(\tau)$, and $\gamma(q^{\Delta}; \tau)$. We also take $W = [\tau_L, \tau_U]$. For each $\tau \in [\tau_L, \tau_U]$, and $d \in \{0, 1\}$, we rewrite

$$\hat{q}_d(\tau) = \arg\min_{q \in \mathbb{R}} \hat{Q}_d(q; \tau),$$

where, for $q \in \mathbb{R}$,

$$\hat{Q}_d(q; \tau) = \sum_{i=1}^{n} \frac{1\{D_i = d\}}{\hat{p}_d(X_i)} \rho_\tau(Y_i - q).$$

We also define its population version:

$$q_d(\tau) = \arg\min_{q \in \mathbb{R}} E[Q_d(q; \tau)],$$

where

$$Q_d(q; \tau) = \sum_{i=1}^{n} \frac{1\{D_i = d\}}{p_d(X_i)} \rho_\tau(Y_i - q).$$

Throughout, we assume that the propensity score is parametrically specified as follows:[12]

$$P\{D_i = 1 | X_i = x\} = G(x; \beta_0),$$

where β_0 is known to lie in a parameter space $\Theta \subset \mathbb{R}^{d_\beta}$. Let $\hat{\beta}$ be the estimator of β_0, so that we take

$$\hat{p}_d(x) = G(x; \hat{\beta})^d (1 - G(x; \hat{\beta}))^{1-d}, d \in \{0, 1\}.$$

We next introduce bootstrap estimator $\hat{\beta}^*$ that is constructed in the same manner as $\hat{\beta}$, with the exception that we use the bootstrap sample $(Y_i^*, X_i^*, D_i^*)_{i=1}^n$ (i.e., the i.i.d. draws from the empirical distribution of $(Y_i, X_i, D_i)_{i=1}^n$) in place of the original sample $(Y_i, X_i, D_i)_{i=1}^n$. Let $\mathcal{F}_n$ be the σ-field generated by $(Y_i, X_i, D_i)_{i=1}^n$. For a matrix A, we define $\|A\| = \sqrt{\text{tr}(A'A)}$. We let

$$V_i = (Y_i, X_i', D_i)', \text{ and } V_i^* = (Y_i^*, X_i^{*\prime}, D_i^*)'.$$

As for the estimators $\hat{\beta}$ and $\hat{\beta}^*$, we make the following assumption.

Assumption 2.2. *There exists a map ψ such that the following two statements hold.*

[12] Extending the results to the situation of nonparametrically specified propensity scores is not difficult. Our focus on parametric specification of the propensity score is motivated by the fact that it is commonly used in empirical applications, despite receiving less attention in the theoretical econometrics literature.

(i)

$$\sqrt{n}(\hat{\beta} - \beta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\psi(V_i) - E\psi(V_i)) + o_P(1),$$

where $\|Var(\psi(V_i))\| < \infty$.

(ii)

$$\sqrt{n}(\hat{\beta}^* - \hat{\beta}) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\psi(V_i^*) - E[\psi(V_i^*)|\mathcal{F}_n]) + o_P(1).$$

This assumption is typically satisfied by most $\sqrt{n}$-consistent and asymptotically normal estimators $\hat{\beta}$.

Let $G_k^{(1)}(x;\beta) = \partial G(x;\beta)/\partial \beta_k$, and for $d \in \{0,1\}$,

$$g_{d,k}(x;\beta) = \left(G_k^{(1)}(x;\beta)\right)^d \left(-G_k^{(1)}(x;\beta)\right)^{1-d},$$

and $g_d(x;\beta) = [g_{d,1}(x;\beta), ..., g_{d,d_\beta}(x;\beta)]'$. Let $g_d^{(1)}(x;\beta) = \partial g_d(x;\beta)/\partial \beta'$. We collect regularity conditions for $g_d(x;\beta)$ and the distribution of Y_{di} below.

Assumption 2.3. *(i) The parameter space* Θ *for* β_0 *is bounded in* $\mathbb{R}^{d_\beta}$ *and*

$$\sup_{x \in \mathcal{X}} \sup_{\beta \in \Theta} \left(\|g_d(x;\beta)\| + \|g_d^{(1)}(x;\beta)\| \right) < \infty.$$

(ii) The set $J_d(\tau_U, \tau_L) \equiv \{q_d(\tau) : \tau \in [\tau_L, \tau_U]\}$ *is bounded for each* $d \in \{0,1\}$.

(iii) The density f_d *of* Y_{di} *is Lipschitz continuous and bounded away from zero on* $J_d(\tau_U, \tau_L)$.

Define $W_P = \left\{\tau \in [\tau_L, \tau_U] : \gamma(q^\Delta(\tau);\tau) \neq 0\right\}$ and let $\hat{W}$ be the set contructed using the step-down procedure explained above. Then let $FWER = P\{\hat{W} \backslash W_P \neq \varnothing\}$.

Theorem 2.1. *Suppose that Assumptions 2.1, 2.2 and 2.3 hold, and that the set of functionals* $\{\gamma(\cdot;\tau) : \tau \in [\tau_L, \tau_U]\}$ *is equicontinuous. Then,*

$$\limsup_{n \to \infty} FWER \leqslant \alpha.$$

The condition on the functionals $\gamma(\cdot;\tau)$ is satisfied by all the examples listed in Table 1. Online Appendix A presents the complete proof of Theorem 2.1 that involves several steps. Briefly, we first obtain the asymptotic linear representation of $\sqrt{n}(\hat{q}^\Delta(\tau) - q^\Delta(\tau))$ that is uniform over $\tau \in [\tau_L, \tau_U]$, using Pollard's convexity lemma; similarly as in Hahn

(1995) and Kato (2009). We next use the maximal inequality in Massart (2007) as in Guerre and Sabbah (2012), to additionally establish the asymptotic equicontinuity of the leading process in the asymptotic linear representation, and its weak convergence to a tight Gaussian process indexed by $\tau \in [\tau_L, \tau_U]$.[13] With these results and using the assumption that γ is a continuous functional, we verify that the conditions of Theorem 2.1 of Romano and Shaikh (2010) are satisfied, thereby obtaining the desired result of asymptotic FWER control.

3 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 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

[13] Note the econometric literature focuses mostly on quantile regression models. Modifications to the standard arguments are needed in our set-up since we estimate a parametric specification of propensity scores in the first step.

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

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 predicting 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.[17]

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

[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).

[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.

[17]In Online Appendix B, 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.

in 2004, approximately 12,000 grade 3 students were surveyed. The follow-up rate was over 96 percent 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.[18]

Table 2 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 incorporates propensity score weighting, which allows us to balance treatment and control group based on these observed variables.

4 Empirical Application

In this section, we obtain new insights extending the findings of Andrabi, Das, and Khwaja (2017) by conducting hypothesis tests based on the framework described in Section 2. Our analysis focuses on the average of standardized test scores across three subjects after random assignment as our outcome variable, and we estimate QTEs of access to report cards for percentiles 1 to 99 using the Firpo (2007) estimator.[19] To balance covariates between the treatment and control groups, we estimate the propensity score $\hat{p}_d(x)$ 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 $B = 9999$.[20]

[18] The data set is available at https://www.aeaweb.org/articles?id=10.1257/aer.20140774.

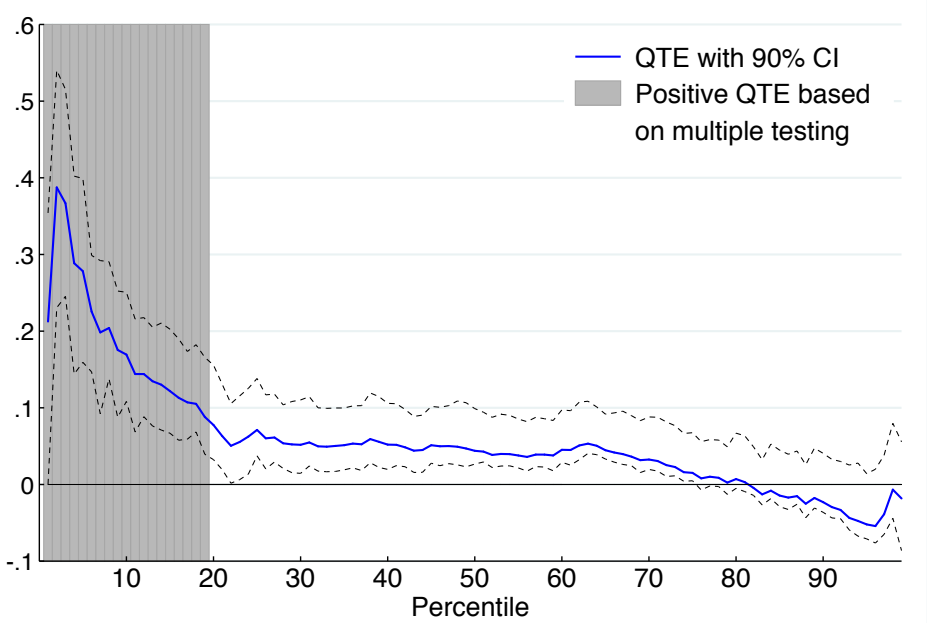
[19] To infer treatment effects for specific individuals from QTEs we have to assume that there are no rank reversals in the test score distribution between the treatment and control groups. While this assumption is likely violated, positive QTEs imply that the treatment has a positive effect for some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

[20] Andrabi, Das, and Khwaja (2017) include district fixed effects treating the data as i.i.d. across districts and do not model inter- and intra-cluster correlation further. We follow this approach to extend their

Table 2: 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.



Note: Multiple testing results show quantiles for which the QTE is positive at a FWER of 5% (see hypothesis (H.3) in Table 1).

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

First, we consider QTEs for the entire sample, i.e. we set $Z_i = 1$ in the notation of Section 2. Figure 1 shows our estimated QTEs for the full sample along with 90 percent pointwise confidence intervals.[21] We find pointwise significant and positive treatment effects extending from the first to the 75th percentile. Starting with the 83th percentile the point estimates for QTEs become negative but the pointwise confidence intervals include zero.

Table 3 summarizes the results for joint hypothesis testing for positive and heterogeneous QTEs. First, we test the null hypothesis of no positive treatment effect at any percentile, i.e. (H.1) in Table 1. As shown in Figure 1, the largest QTE (which occurs at the third percentile) equals 0.387. With the bootstrap critical value of 0.230, we reject the null hypothesis at the 5 percent level. The associated p-value equals 0.003. 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

findings and draw bootstrap samples of individuals instead of villages or districts. When using a block bootstrap that resamples entire villages, the QTE is statistically significant only for one percentile even without adjustments for multiple testing. For completeness, we show the results under block bootstrapping in the Online Appendix.

[21] 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. The pointwise confidence intervals are calculated as the 5th and 95th percentile of the distribution of bootstrapped QTEs.

Table 3: Testing for Presence of Positive QTEs and QTE Heterogeneity Without Subgroups

	Test statistic	Critical value at 5%	p-value
Test for positive QTE (H.1)	0.38747	0.22957	0.0031003
Test for QTE heterogeneity (H.2)	0.32917	0.24014	0.011701

Notes: This table shows test results for hypotheses with $\gamma(q^{\Delta};\tau,z) = \max\{q^{\Delta}(\tau,1),0\}$ and $\gamma(q^{\Delta};\tau,z) = |q^{\Delta}(\tau,1)-c|$, 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, i.e. we test hypotheses (H.1) and (H.2) in Table 1.

across quantiles, i.e. (H.2) in Table 1. The test statistic, which is calculated as the largest deviation from the mean estimated QTE ($c = 0.0583$), equals 0.329. With a bootstrap critical value of 0.240, we also reject this null hypothesis at 5 percent with a p-value of 0.012. This result implies that treatment effects are heterogenous across quantiles, thereby indicating that individuals vary in their response to the report cards.

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 (H.3) in Table 1. The shaded area in Figure 1 corresponds to the set $\hat{W} = W \backslash \tilde{W}_k$. This test accounts for potential dependencies across quantiles of the same outcome variable and the number of individual hypotheses ($|S| = 99$).

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 above the 19th percentile 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 0.088 and 0.387, 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. 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 perfor-

mance. Following Andrabi, Das, and Khwaja (2017), we consider subgroups defined by child characteristics, type of school, and characteristics of the villages.[22]

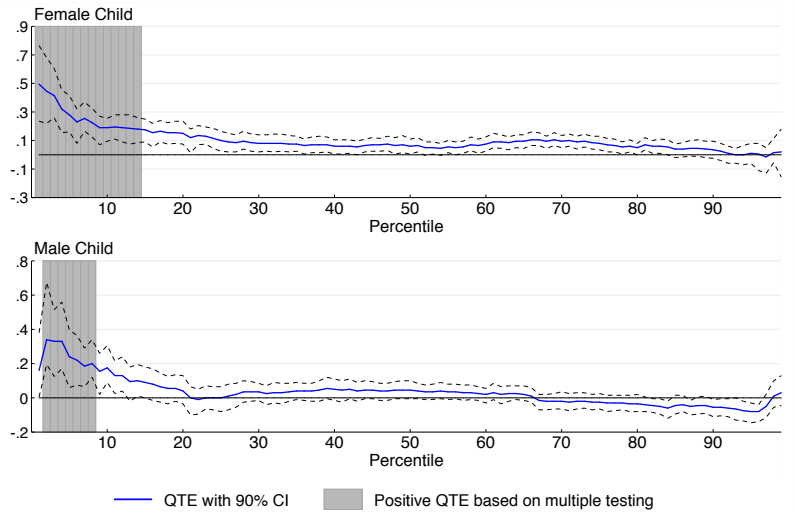
Figure 2 presents QTEs conditional on child gender and child baseline test scores. These figures provide an easy and intuitive way to check which subgroups benefit from being assigned to receive report cards (heterogeneity across subgroups). In addition, we can inspect the figure for each subgroup to 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) in Table 1.

Figure 2a 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 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.[23]

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 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

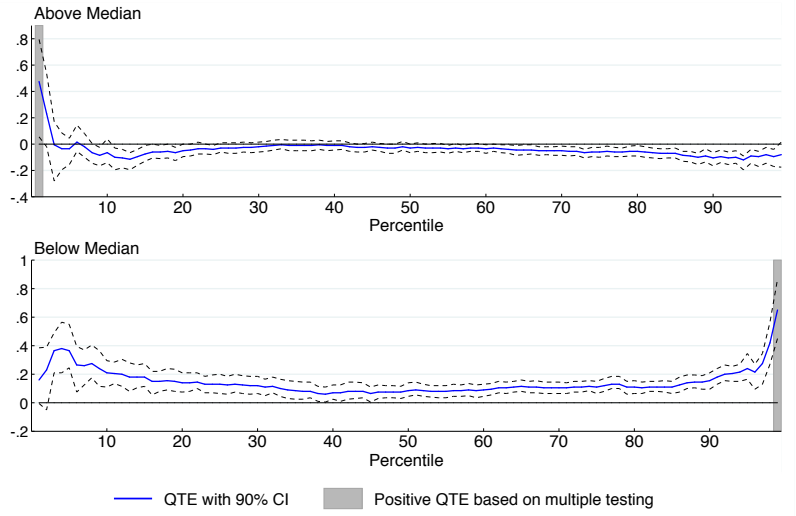
[22]Note that in our application the number of hypotheses being tested is quite small particularly relative to genomic studies from genome wide association studies. If the number of hypotheses were large it is well known that FWER controlling procedures typically have low power, and in response Gu and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

[23]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.



Note: Multiple testing results show quantiles for which the QTE is positive at a FWER of 5 percent (see hypothesis (H.4) in Table 1).

(a) By Child's Gender



Note: Multiple testing results show quantiles for which the QTE is positive at a FWER of 5 percent (see hypothesis (H.4) in Table 1).

(b) By Child's Baseline Test Score

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

literacy rates. In general, these results are important because they can show policymakers which subgroups should be targeted with an accountability program.

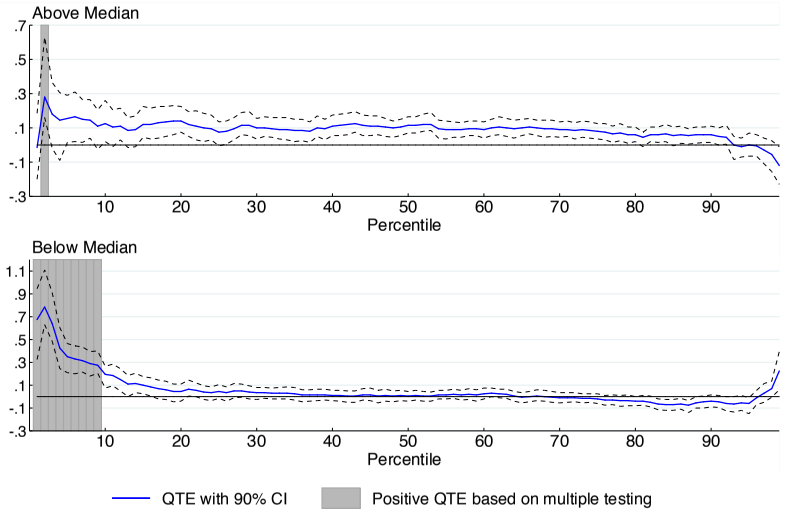
Finally, we consider subgroups defined by the combination of school ownership type (government or private) with one of two different measures of student performance (school level and relative). We first create subgroups by interacting school ownership with school performance in the baseline test to yield four subgroups.[24] 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).[25] 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 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

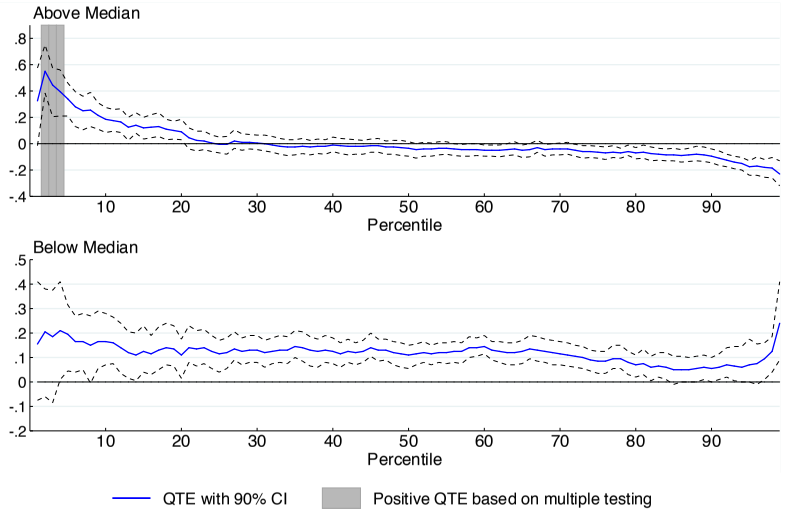
[24]Specifically, following Andrabi, Das, and Khwaja (2017) a school is defined as high-performing if its mean baseline test score exceeds the 60th percentile of all schools' mean scores.

[25]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.



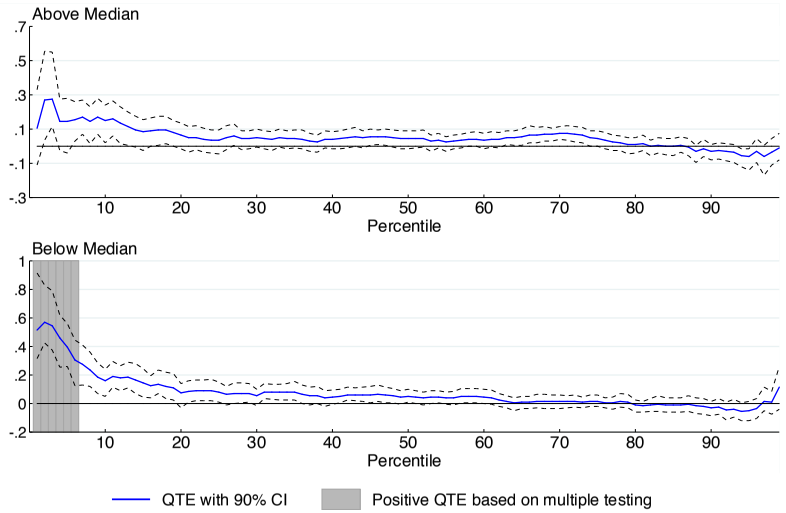
Note: Multiple testing results show quantiles for which the QTE is positive at a FWER of 5 percent (see hypothesis (H.4) in Table 1).

(a) By Village Wealth



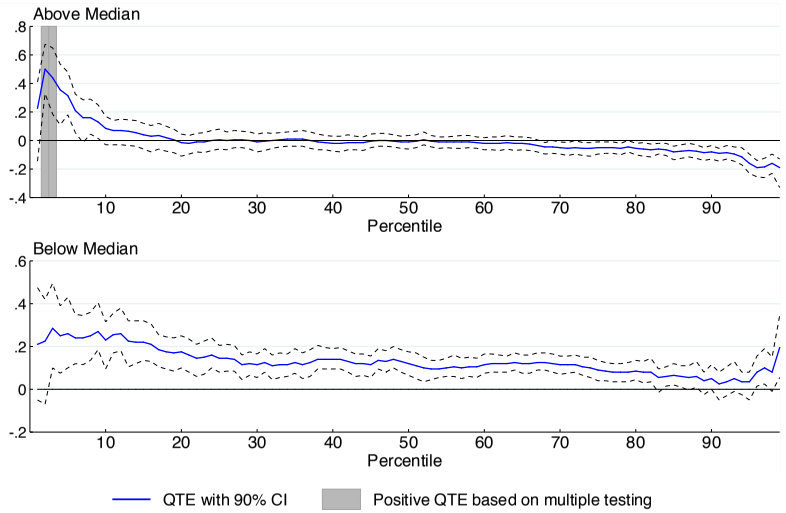
Note: Multiple testing results show quantiles for which the QTE is positive at a FWER of 5 percent (see hypothesis (H.4) in Table 1).

(b) By Village Literacy Rate



Note: Multiple testing results show quantiles for which the QTE is positive at a FWER of 5 percent (see hypothesis (H.4) in Table 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 hypothesis (H.4) in Table 1).

(d) By Village Size

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

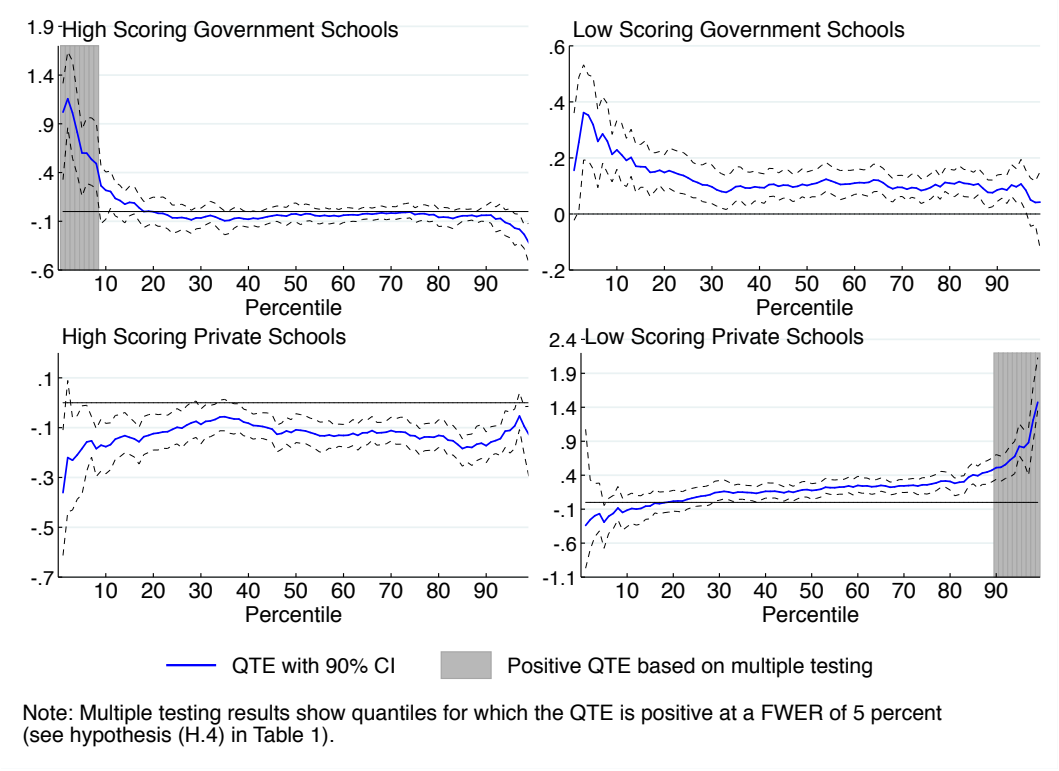
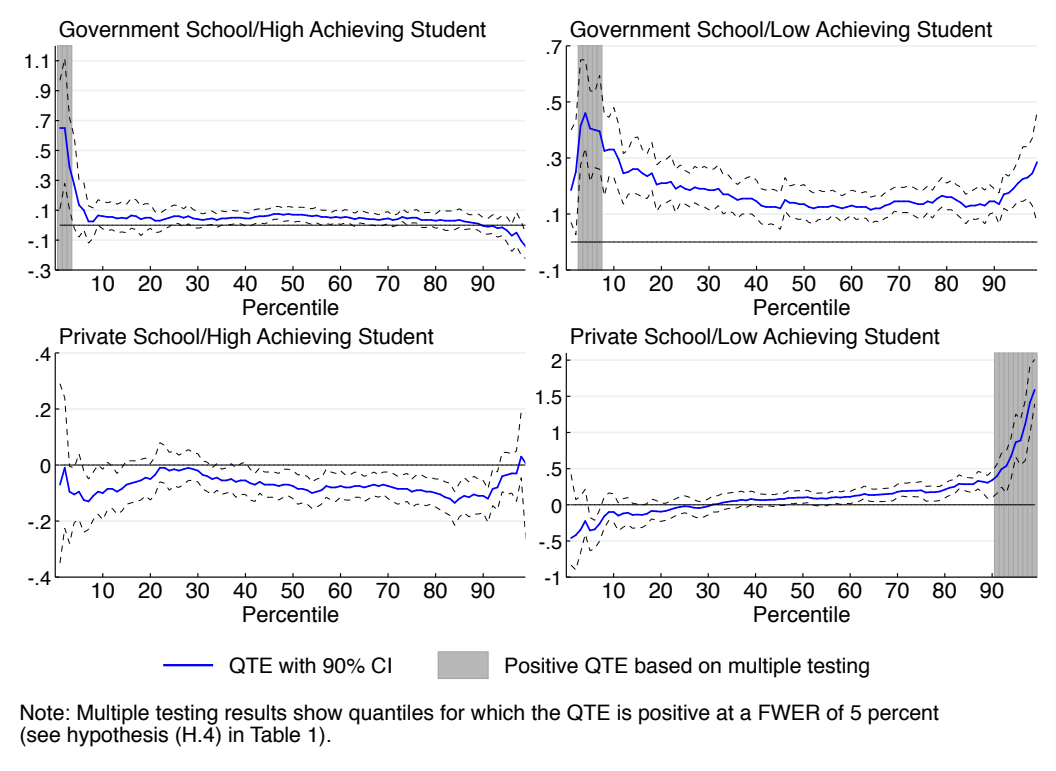


Figure 4: Quantile Treatment Effects and Multiple Testing Results by School Type and Performance

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 4 presents the results for testing hypothesis (H.5) in Table 1. This 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 the hypothesis for all sets of subgroups at a level of 5 percent. We conclude that differences across subgroups do not explain the observed distributional treatment effects in the whole sample. Our test relaxes the assumption of treatment effect homogeneity within subgroups that is implicit in Bitler, Gelbach, and Hoynes (2017).

The test results shown in Table 5 additionally account for potential dependencies within and across subgroups. These test results provide additional insight because they identify the individual subgroups within a class of subgroups that exhibit treatment effect heterogeneity. That is, we test hypothesis (H.6) in Table 1. 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



Note: Multiple testing results show quantiles for which the QTE is positive at a FWER of 5 percent (see hypothesis (H.4) in Table 1).

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

evidence of treatment effect heterogeneity. These results clearly suggest a substantial amount of treatment effect heterogeneity between subgroups and across the student performance distribution within subgroups.

5 Conclusion

In this paper we describe 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

Table 4: Testing for Treatment Effect Heterogeneity Between Subgroups

Subgroup category	Test statistic	Critical value at 5%	p-value
Child's gender	0.45014	0.29377	0.0026003
Child's baseline test score	1.2681	0.41076	0
Village wealth	1.2681	0.33394	0
Village literacy rate	1.2681	0.3618	0
School Herfindahl Index	1.2681	0.26926	0
Village size	1.2681	0.39368	0
School type and school performance	1.2681	0.73724	0.00090009
School type and child's performance relative to school	1.5732	0.6843	0.00020002

Notes: This table shows test results for hypotheses with $\gamma(q^{\Delta};\tau,z) = |q^{\Delta}(\tau,z) - c(z)|$, i.e. these tests show for which subgroups categories we can reject treatment effects that are homogenous within subgroups for all subgroups, i.e. we test hypothesis (H.5) in Table 1.

Table 5: Testing Which Subgroups Exhibit Treatment Effect Heterogeneity

Subgroup category	Test statistic	p-value
Child's gender		
Female	0.450	0
Male	0.304	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 government	1.127	0
Low scoring government	0.237	0
High scoring private	0.0749	0
Low scoring private	1.268	0
School type and child's performance relative to school		
Government/high achieving	0.834	0
Government/low achieving	0.260	0
Private/high achieving	0.0949	0
Private/low achieving	1.573	0

Notes: This table shows results of tests for which subgroups in each subgroup category we can reject homogenous treatment effects, i.e. we test hypothesis (H.6) in Table 1. 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.

a subgroup when determining if the distributional heterogeneity across the full sample is characterized by subgroups.

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.[26]

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, forthcoming) 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.

[26] 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.

References

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

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.

Allen, Jason, Robert Clark, and Jean-François Houde. 2014. "The Effect of Mergers in Search Markets: Evidence from the Canadian Mortgage Industry." *American Economic Review* 104 (10):3365–3396.

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.

Angrist, Joshua, Daniel Lang, and Philip Oreopoulos. 2009. "Incentives and Services for College Achievement: Evidence from a Randomized Trial." *American Economic Journal: Applied Economics* 1 (1):136–163.

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

Bandiera, Oriana, Robin Burgess, Narayan Das, Selim Gulesci, Imran Rasul, and Munshi Sulaiman. 2017. "Labor Markets and Poverty in Village Economies." *Quarterly Journal of Economics* 132 (2):811–870.

Banerjee, Abhijit V., Esther Duflo, Rachel Glennerster, and Cynthia Kinnan. 2015. "The Miracle of Microfinance? Evidence from a Randomized Evaluation." *American Economic Journal: Applied Economics* 7 (1):22–53.

Becker, Gary S. 1995. "Human Capital and Poverty Alleviation." Human Resources Development and Operations Policy Working Paper 52.

Behaghel, Luc, Clément de Chaisemartin, and Marc Gurgand. 2017. "Ready for Boarding? The Effects of a Boarding School for Disadvantaged Students." *American Economic Journal: Applied Economics* 9 (1):140–164.

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.

Brown, Jason, Mark Duggan, Ilyana Kuziemko, and William Woolston. 2014. "How Does Risk Selection Respond to Risk Adjustment? New Evidence from the Medicare Advantage Program." *American Economic Review* 104 (10):3335–3364.

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, V. and I. Fernandéz-Val. 2005. "Subsampling Inference on Quantile Regression Processes." *Sankhya* 67 (2):253–276.

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

Crepon, Bruno, Florencia Devoto, Esther Duflo, and William Pariente. 2015. "Estimating the Impact of Microcredit on Those Who Take It Up: Evidence from a Randomized Experiment in Morocco." *American Economic Journal: Applied Economics* 7 (1):123–150.

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.

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.

Evans, William and Craig Garthwaite. 2012. "Estimating Heterogeneity in the Benefits of Medical Treatment Intensity." *Review of Economics and Statistics* 94 (3):635–649.

Fack, Gabrielle and Camille Landais. 2010. "Are Tax Incentives for Charitable Giving Efficient? Evidence from France." *American Economic Journal: Economic Policy* 2 (2):117–141.

Fairlie, Robert and Jonathan Robinson. 2013. "Experimental Evidence on the Effects of Home Computers on Academic Achievement among Schoolchildren." *American Economic Journal: Applied Economics* 5 (3):211–240.

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.

Guerre, Emmanuel and Camille Sabbah. 2012. "Uniform Bias Study and Bahadur Representation for Local Polynomial Estimators of the Conditional Quantile Function." *Econometric Theory* 28 (01):87–129.

Hahn, Jinyong. 1995. "Bootstrapping Quantile Regression Estimators." *Econometric Theory* 11 (01):105.

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., 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 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.

Kato, Kengo. 2009. "Asymptotics for Argmin Processes: Convexity Arguments." *Journal of Multivariate Analysis* 100 (8):1816–1829.

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

Koenker, R. and Z. Xiao. 2002. "Inference on the Quantile Regression Process." *Econometrica* 70 (4):1583–1612.

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, S., K. Song, and Y.-J. Whang. 2018. "Testing for a General Class of Functional Inequalities." *Econometric Theory* 34:1018–1064.

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. forthcoming. "Multiple Hypothesis Testing in Experimental Economics." *Experimental Economics* .

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.

Massart, Pascal. 2007. *Concentration Inequalities and Model Selection*. Berlin, Heidelberg: Springer-Verlag.

McKenzie, David. 2017. "Identifying and Spurring High-Growth Entrepreneurship: Experimental Evidence from a Business Plan Competition." *American Economic Review* 107 (8):2278–2307.

Meyer, Bruce D and James X Sullivan. 2008. "Changes in the Consumption, Income, and Well-Being of Single Mother Headed Families." *American Economic Review* 98 (5):2221–2241.

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.

Muralidharan, Karthik, Paul Niehaus, and Sandip Sukhtankar. 2016. "Building State Capacity: Evidence from Biometric Smartcards in India." *American Economic Review* 106 (10):2895–2929.

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

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.

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.