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

Steven F. Lehrer† Queen's University and NBER

R. Vincent Pohl‡ Mathematica

Kyungchul Song§ University of British Columbia

March 2021

Abstract

This paper proposes bootstrap-based multiple testing procedures for quantile treatment effect heterogeneity under the assumption of selection on observables, and shows its asymptotic validity. Our procedure can be used to detect the quantiles and subgroups exhibiting treatment effect heterogeneity. We apply the multiple testing procedures to data from a large-scale Pakistani school report card experiment, and 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 because 63 percent of statistically significant quantile treatment effects become insignificant once corrections for multiple testing are applied.

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

JEL classification: C12; C21; I21; L15.

*We thank Jishnu Das, 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 the results in this paper are available in either Stata or MATLAB on request. Lehrer and Song, respectively, 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.

†Department of Economics, email: lehrers@queensu.ca.

‡Division of Health Policy Assessment, 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 treatment or intervention. Therefore, treatment effects may vary between subgroups defined by individual characteristics. For example, providing report cards with information on school test scores may make some parents more likely to move their child to a better school than others based on parental characteristics such as education. In addition, individuals' response to a treatment may vary across quantiles of the unconditional outcome distribution. This type of treatment effect heterogeneity is often modeled via quantile treatment effects (QTE), i.e. the difference in unconditional outcome quantiles between treatment and control groups. For example, if a school information provision program improves the odds that parents correctly perceive their child's performance relative to her peers, parental responses such as switching schools may vary with the child's relative performance, and hence QTE for the effect of the intervention on children's performance would not be constant.

Harnessing the potential policy benefits from treatment effect heterogeneity—ranging from personalized medicine to welfare reform parameters to customized marketing recommendations—requires understanding whether the heterogeneous effects are spurious. A multiple testing approach is useful in this context because it 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. For example, policymakers may be able to modify the design of accountability programs in education more effectively if they know which parents respond to market-level information on school quality. These parents may differ systematically by predetermined characteristics or by the location between specific percentiles of their child's test score distribution. Our proposed testing strategy can be used in many empirical applications to guard against false positives as further outlined in Online Appendix A.

Given the widespread interest in treatment effect heterogeneity and the importance of multiple comparisons corrections, it is somewhat surprising that, to the best of our knowledge, there has been no research that formally establishes the asymptotic validity of a bootstrap multiple testing procedure for functionals of QTE under the assumption of selection on observables. The first contribution of this paper is to fill this gap by providing a formal result of asymptotic validity when the propensity scores that account for selection on observables are parametrically specified. We consider a parametric specification of propensity scores in this paper because, when faced with a large set of covariates, researchers use a parametric specification for the propensity score in their empirical application rather than consider a nonparametric specification. A nonparametric specification has attracted more attention

from from theoretical econometricians despite issues with its practical implementation that include the well-known curse of dimensionality.

The bootstrap-based procedures of multiple testing on distributional treatment effects that we introduce in this paper are motivated by current empirical practice. In many empirical applications, researchers consider a binary treatment effect model under selection on observables. Our testing approach most closely complements three recent papers. Lee and Shaikh (2014) proposed a multiple testing procedure for subgroup treatment effects in randomized experiments that controls the family-wise error rate (FWER), i.e. the probability of rejecting at least one true null hypothesis, in finite samples. Bitler, Gelbach, and Hoynes (2017) adopted a multiple testing procedure based on the Bonferroni correction to test for treatment effect heterogeneity across subgroups and over time. Last, our theoretical work is most closely related to Zhang and Zheng (2020) who formally developed bootstrap inference on quantile treatment effects in a very different setting, by considering a randomized experiment with covariate-adaptive randomization studied earlier in Bugni, Canay, and Shaikh (2018). Below, we further discuss how our study relates to existing contributions in the treatment effect heterogeneity literature.

Our proposed multiple testing procedure controls the FWER in the strong sense and has greater power than Bonferroni-based procedures. It can be used to first 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.

To illustrate our proposed testing strategy, 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 QTE 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 programs, further indicating how schools and parents respond to the release of information on student performance.

1.1 Related literature

In this section, we summarize how this paper contributes to the broad econometrics literature on treatment effect heterogeneity that was recently surveyed in both Athey and Imbens (2017) and Abadie and Cattaneo (2018); as well as adding to the economics of education literature providing evidence on how school accountability programs impact academic outcomes.

Since the publication of Bitler, Gelbach, and Hoynes (2006), empirical researchers in multiple fields, including the economics of education, have increasingly provided evidence on distributional impacts of policies and programs. Bitler, Gelbach, and Hoynes (2006) use data from an experimental evaluation of a welfare reform policy to present evidence of treatment effect heterogeneity that would be predicted by a labor supply model. However, the observed pattern would have been missed if researchers had only reported a mean treatment effect. In a follow-up paper, Bitler, Gelbach, and Hoynes (2017) present evidence that the treatment effect heterogeneity exhibited in this welfare reform evaluation is also not fully characterized by between-subgroup differences in average treatment effects. Thus, although, early empirical investigations, including Heckman, Smith, and Clements (1997) and Friedlander and Robins (1997), document the importance of idiosyncratic impact heterogeneity, the publication of Bitler, Gelbach, and Hoynes (2006) made a compelling case that only estimating average effects may be insufficient.

Our paper first contributes to a growing literature studying inference on the quantile process and distributions of treatment effects. In an early contribution, Koenker and Xiao (2002) propose asymptotic inference on the quantile regression process, which can be subsequently applied to test for QTE heterogeneity. Following a similar spirit, Chernozhukov and Fernandéz-Val (2005) propose subsampling-based testing. In an influential paper, Chernozhukov, Fernandez-Val, and Melly (2013) develop a comprehensive framework of inferences on counterfactual quantities built on the conditional distribution functions of potential outcomes. Recently, Ding, Feller, and Miratrix (2016) and Chung and Olivares (2020) consider permutation tests for the distributional heterogeneity of treatment effects.

Our testing strategy assumes selection on observables and employs inverse probability weighting (IPW) estimators for QTE of a binary treatment initially proposed in Firpo (2007). Similar to our paper, Donald and Hsu (2014) establish weak convergence of inverse-probability weighted quantile processes, yet using a nonparametric series estimator of propensity scores. As noted at the outset of this section, the literature on treatment effect heterogeneity is broad and many contributions consider (1) different settings, (2) a continuous or multivalued treatment, (3) alternative causal parameters, and (4) distributional tests. As an

example, Fan and Wu (2010) consider the distribution of the difference in potential outcomes in a switching regression framework. Zhang and Zheng (2020) and Jiang et al. (2020) employ bootstrap inference of QTE in randomized experiments where the setting involves covariate-adaptive randomization and randomization within matched pairs, respectively. Galvao and Wang (2015) and Cattaneo (2010) provide practical estimation and inference approaches for QTE with a continuous and multivalued treatment, respectively, under the assumption of unconfoundedness. Firpo and Pinto (2016) use inequality measures of the distribution of the potential outcomes to estimate inequality treatment effects. Maier (2011) proposes a nonparametric test of distributional equivalence of potential outcomes under selection on observables. Finally, Goldman and Kaplan (2018) develop a multiple testing procedure for quantiles at different distributions in two-sample Kolmogorov-Smirnov tests.

The second contribution of this paper transpires from the empirical results that contribute to a burgeoning empirical literature surveyed in Figlio and Loeb (2011), which 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 because 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). 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).

1.2 Plan for the paper

The rest of this paper is organized as follows: In Section 2, we introduce the general testing procedures for treatment effect heterogeneity across quantiles of the outcome distribution and subgroups and provide a guide for the practical implementation of these procedures. In Section 3, we illustrate the value of the testing procedure by reexamining the Andrabi, Das, and Khwaja (2017) experimental data. We describe the experiment and economic model that

underlie the data being investigated. This model predicts heterogeneous treatment effects both within and across subgroups. The concluding Section 4 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

In this section, we first introduce joint and multiple hypotheses of QTE that can be used to test for treatment effect heterogeneity within and across subgroups. Then we describe our stepwise bootstrap testing approach for testing multiple hypotheses.

2.1 Testing for treatment effect heterogeneity

To develop a multiple testing procedure for various hypotheses of QTE, 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 placed in the control group. Let Y_i be the observed outcome for 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 by 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 as $X_i = (X_{1i}, Z_i)$, 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. We are interested in the quantile treatment effects at the τ-th percentile, with τ running in a continuum. For each subgroup z in the support of Z_i, we define the unconditional quantile of the potential outcomes at the τ-th percentile as follows: with $\tau \in (0,1)$,

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

The subgroup QTE at a quantile-subgroup pair (τ, z) is then defined by

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

We next introduce individual hypotheses which are specific to a quantile-subgroup pair (τ, z). Later, we build joint and multiple hypothesis testing problems from these individual hypotheses. First, let $\tau_L, \tau_U \in (0, 1)$ be such that $\tau_L < \tau_U$ 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) on which we focus. We are interested in the hypothesis of the following form: for each $(\tau, z) \in S$,

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

where $\gamma(q^{\Delta}; \tau, z)$ is a functional of q^{Δ} that depends on (τ, z). Examples of specific hypothesis testing problems involving QTE are provided in Table 1. This paper's framework applies to a wide range of functionals γ as long as $\gamma(q^{\Delta}; \tau, z)$ is continuous in q^{Δ} uniformly over (τ, z). For example, if we set

$$\gamma(q^{\Delta}; \tau, z) = |q^{\Delta}(\tau, z)|,$$

testing the hypothesis $H_0(\tau, z)$ corresponds to testing for the presence of QTE for (τ, z). This hypothesis can be used to test whether a treatment has a nonzero effect for any quantile-subgroup pair. As another example, if we take

$$\gamma(q^{\Delta}; \tau, z) = \max\{q^{\Delta}(\tau, z), 0\}, \tag{1}$$

testing the hypothesis $H_0(\tau, z)$ amounts to testing the null hypothesis that the QTE for (τ, z) is non-positive.

The individual hypotheses can also be used to test for QTE heterogeneity. For example, suppose that we set

$$\gamma(q^{\Delta}; \tau, z) = |q^{\Delta}(\tau, z) - \bar{q}^{\Delta}(z)|, \tag{2}$$

where

$$\bar{q}^{\Delta}(z) = \frac{1}{\tau_U - \tau_L} \int_{\tau_L}^{\tau_U} q^{\Delta}(\tau, z) d\tau. \tag{3}$$

Table 1: Examples of 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 $\{S_w : w \in W\}$
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 $H_1(\tau,z): q^{\Delta}(\tau,z) > 0$	$\max\{q^{\Delta}(\tau,z),0\}$	$\sup_{(\tau,z)\in S'} \max\{q^{\Delta}(\tau,z),0\}$	$\{\{\tau,z\} : (\tau,z) \in [\tau_L,\tau_U] \times \mathcal{Z}\}$
Testing for QTE heterogeneity in some subgroups			
$H_0(\tau,z): q^{\Delta}(\tau,z) = \bar{q}^{\Delta}(z)$, vs $H_1(\tau,z): q^{\Delta}(\tau,z) \neq \bar{q}^{\Delta}(z)$	$\lvert q^{\Delta}(\tau,z) - \bar{q}^{\Delta}(z)\rvert$	$\sup_{(\tau,z)\in S'} \lvert q^{\Delta}(\tau,z) - \bar{q}^{\Delta}(z)\rvert$	$\{\{\tau,z\} : (\tau,z) \in [\tau_L,\tau_U] \times \mathcal{Z}\}$
Testing for which subgroups QTE are heterogeneous			
$H_0(\tau,z): q^{\Delta}(\tau,z) = \bar{q}^{\Delta}(z)$, vs $H_1(\tau,z): q^{\Delta}(\tau,z) \neq \bar{q}^{\Delta}(z)$	$\lvert q^{\Delta}(\tau,z) - \bar{q}^{\Delta}(z)\rvert$	$\sup_{\tau\in[\tau_L,\tau_U]} \lvert q^{\Delta}(\tau,z) - \bar{q}^{\Delta}(z)\rvert$	$\{\{(\tau,z) : \tau \in [\tau_L,\tau_U]\} : z \in \mathcal{Z}\}$

Notes: We define $\bar{q}^{\Delta}(z) = \frac{1}{\tau_U-\tau_L}\int_{\tau_L}^{\tau_U} q^{\Delta}(\tau,z)d\tau$, i.e. the mean QTE for the whole sample and conditional on subgroup.

Then testing if $H_0(\tau, z)$ is true jointly for all $(\tau, z) \in S$ is tantamount to testing whether there is QTE heterogeneity across quantiles within subgroup z. In the next two subsections, we demonstrate how to combine the individual hypotheses, as outlined above, to construct joint and multiple hypothesis testing problems.

2.1.1 Joint hypothesis testing

In many empirical applications, the researcher is primarily interested in whether $H_0(\tau, z)$ is true for all $(\tau, z) \in S$. To conduct this test, we first combine the individual hypotheses into a joint hypothesis:

$$H_0 : \Gamma(q^{\Delta}; S) = 0, \text{ vs } H_1 : \Gamma(q^{\Delta}; S) \neq 0, \tag{4}$$

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

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

Thus rejecting the null hypothesis in H_0 in (4) means rejecting the hypothesis that $H_0(\tau, z)$ is true for all $(\tau, z) \in S$. For example, testing H_0 against H_1 in (4) with γ as in (1) is equivalent to testing whether QTE is positive at some $(\tau, z) \in S$. Similarly, testing H_0 against H_1 with γ as defined in (2) is equivalent to testing, within subgroup z, whether the QTE are constant across quantiles.

2.1.2 Multiple hypothesis testing

Often, to obtain new policy insights, we are interested in finding out which quantile-subgroup pairs (τ, z) are responsible for the rejection of the joint null hypothesis expressed in (4). To address this question, let us consider the following multiple hypothesis testing problem. First suppose that the set S is partitioned as follows:

$$S = \bigcup_{w\in W} S_w, \tag{5}$$

for some index set W. Our focus is to find $w \in W$ such that the violation of the joint null hypothesis expressed in (4) is due to the violation of $H_0(\tau, z)$ for some $(\tau, z) \in S_w$. For example, if one takes γ as defined in (1) and $S_w = \{(\tau, z)\}$ with $w = (\tau, z)$, and $W = [\tau_L, \tau_U] \times \mathcal{Z}$, our interest is in finding which (τ, z) are responsible for rejecting the null hypothesis that QTE are non-positive for all $(\tau, z) \in [\tau_L, \tau_U] \times \mathcal{Z}$.

We first define

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

W_P is the set of indexes $w \in W$ such that $H_0(\tau, z)$ is violated for some $(\tau, z) \in S_w$. This set depends on the distribution P of data through its dependence on q^{Δ}, and hence we use subscript P in W_P to make this dependence explicit.

Now suppose that one constructs a subset $\hat{W} \subset W$ using observed variables, and proposes the data-dependent set $\hat{W}$ as the collection of indices $w \in W$ such that $H_0(\tau, z)$ is violated at some $(\tau, z) \in S_w$ in the sample at hand. However, if it turns out that

$$\hat{W} \not\subseteq W_P, \tag{6}$$

then the set $\hat{W}$ contains a false positive, i.e., there exists $w \in \hat{W}$ such that the null hypothesis $H_0(\tau, z)$ is mistaken to be violated for some $(\tau, z) \in S_w$ although the hypothesis is in fact true for all $(\tau, z) \in S_w$. The multiple testing literature aims to obtain a data-dependent set $\hat{W}$ such that the probability of (6) is asymptotically controlled under a small number $\alpha > 0$, i.e.,

$$\liminf_{n \to \infty} P\{\hat{W} \subset W_P\} \geqslant 1 - \alpha. \tag{7}$$

The probability of the event in (6) is called the Familywise Error Rate (FWER) in the multiple testing literature. The set $\hat{W}$ satisfying (7) is said to control FWER asymptotically at α. As the asymptotic control holds for all probabilities P, this is called strong control of FWER, see Section 9.1 of Lehmann and Romano (2005). The procedures we introduce in the next subsection construct such a set $\hat{W}$.

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* is established by Firpo (2007). Here we propose joint hypothesis testing and multiple hypothesis testing procedures and provide conditions under which the FWER is controlled asymptotically.

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

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

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

$$\hat{q}_d(\tau, z) = \arg\min_q \frac{1}{\sum_{i=1}^n 1\{Z_i = z\}} \sum_{i=1}^n \hat{\omega}_{di} \rho_\tau (Y_i - q) 1\{Z_i = z\}, \tag{8}$$

with $\hat{\omega}_{di} = 1\{D_i = d\}/\hat{p}_d(X_i)$, and $\hat{p}_d(x)$ is the estimated propensity score. Following Smith and Todd (2005), the propensity score $\hat{p}(x)$ is estimated using data from the full sample. As in 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, using each bootstrap sample, construct

$$\hat{q}_b^{\Delta *}(\tau, z) = \hat{q}_{1,b}^*(\tau, z) - \hat{q}_{0,b}^*(\tau, z),$$

where $\hat{q}_{1,b}^*(\tau, z)$ and $\hat{q}_{0,b}^*(\tau, z)$ are obtained just as $\hat{q}_1(\tau, z)$ and $\hat{q}_0(\tau, z)$ were constructed but using the b-th bootstrap sample.

For joint hypothesis testing expressed in (4), we construct test statistics

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

and use the critical value as the $(1 - \alpha)$-quantile from the bootstrap distribution of T_b^*. By subtracting $\hat{q}^\Delta$ in T_b^*, we re-center the bootstrap test statistic in order to impose the least favorable configuration under the null hypothesis.

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 increases to infinity as the number of observations does so. For bootstrap inference, one can use the block bootstrap in which one resamples clusters with replacement instead of individual sample units.

2.2.2 Bootstrap multiple testing procedure for QTE

The multiple testing procedure adapts the step-down method of Romano and Wolf (2005) and Romano and Shaikh (2010) to our set-up. For each subset $W' \subset W$, we define

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

where the S_ws constitute the partition of S in (5). 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 the above expression, we next 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

$$\hat{W} = W \backslash \tilde{W}_k \tag{9}$$

to be the data-dependent set of indices for which the null hypothesis is violated.

For example, when we perform multiple testing for QTE heterogeneity, i.e. when we identify for which subgroups z the null hypothesis of constant treatment effects within subgroups is violated (the last row in Table 1), we take

$$\begin{aligned}\Gamma(\hat{q}^{\Delta}; S_w) &= \sup_{(\tau,z)\in S_w} \gamma(\hat{q}^{\Delta}; \tau, z) \\ &= \sup_{(\tau,z)\in S_w} \left|\hat{q}^{\Delta}(\tau, z) - \tilde{q}^{\Delta}(z)\right|\end{aligned}$$

and

$$\Gamma(\hat{q}_b^{\Delta*} - \hat{q}^{\Delta}; S_w) = \sup_{(\tau,z)\in S_w} \left|\hat{q}_b^{\Delta*}(\tau, z) - \hat{q}^{\Delta}(\tau, z) - (\tilde{q}_b^{\Delta*}(z) - \tilde{q}^{\Delta}(z))\right|,$$

where

$$\tilde{q}^{\Delta}(z) = \frac{1}{\tau_U - \tau_L}\int_{\tau_L}^{\tau_U} \hat{q}^{\Delta}(\tau, z)d\tau, \text{ and } \tilde{q}_b^{\Delta*}(z) = \frac{1}{\tau_U - \tau_L}\int_{\tau_L}^{\tau_U} \hat{q}_b^{\Delta*}(\tau, z)d\tau. \tag{10}$$

2.2.3 Asymptotic control of FWER

In this subsection, we provide conditions that ensure the set $\hat{W}$ in (9) obtained through the step-down procedure outlined in the prior subsection 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 analogue:

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

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:

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

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

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

Note that the approach of weighting using the inverse propensity score estimated nonparametrically creates what Khan and Tamer (2010) called the issue of irregular identification, and finite sample inference may be unstable without proper trimming combined with assumptions on the tail behavior of the propensity scores. We do not address the issue of trimming because we use a parametric specification of propensity scores, but interested readers can see Ma and Wang (2020) and references therein for proposals on the choice of trimming parameters.

We next introduce the 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.*

(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 list 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 continuous on a closed interval containing $J_d(\tau_U,\tau_L)$ and bounded away from zero on $J_d(\tau_U,\tau_L)$.

Let us introduce a condition for the functional $\gamma(\cdot;\tau)$ as follows.

Assumption 2.4. *$\{\gamma(\cdot;\tau):\tau\in[\tau_L,\tau_U]\}$ are equicontinuous functionals on the space of bounded functions endowed with the sup norm.*

The condition is a mild, technical condition for the functionals γ that are permitted in our framework. All the examples we consider in this paper satisfy this condition.

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 constructed using the step-down procedure explained above. Then let $FWER=P\{\hat{W}\not\subseteq W_P\}$.

Theorem 2.1. *Suppose that Assumptions 2.1–2.4 hold. Then,*

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

The condition on the functionals $\gamma(\cdot;\tau)$ is satisfied by each example listed in Table 1. Online Appendix B 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 $\tau\in[\tau_L,\tau_U]$. While the econometric literature focuses mostly on quantile regression models, modifications to the standard arguments are needed in our set-up because we estimate a parametric specification of propensity scores in the first step. 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. In summary, Theorem 2.1 provides the asymptotic validity of our proposed bootstrap-based multiple testing procedures for quantile treatment effect heterogeneity under the assumption of selection on observables.

2.3 Practical implementation

In this subsection, we provide a brief guide on how to implement our proposed testing procedures. MATLAB and Stata code that implement the tests are available from the authors.

1. Estimate the propensity score of treatment status on observed covariates with the full sample. Following Firpo (2007) and equation (8), estimate the QTE $\hat{q}^{\Delta}(\tau, z)$ for quantiles and subgroups of interest. In researchers bin quantiles, for example, into deciles or percentiles, so here we switch from a continuum of quantiles to a discrete set $\mathcal{T}$. Instead of subgroups, one may include the full sample by setting $Z_i = 1$ if the full sample is the focus.

2. Using each bootstrap sample $(Y_{i,b}^*, D_{i,b}^*, X_{i,b}^*)$, $b = 1, ..., B$, construct $\hat{q}_b^{\Delta *}(\tau, z)$, $b = 1, ..., B$. These quantities are used to calculate bootstrap critical values in the construction of pointwise confidence intervals, uniform confidence intervals and the testing procedures as explained below.

3. To construct a $(1-\alpha) \times 100$ percent point-wise confidence interval for $q^{\Delta}(\tau, z)$, we first obtain the $(1-\alpha)$ percentile of the bootstrap distribution of

$$\left\{ \sqrt{n} \left| \hat{q}_b^{\Delta *}(\tau, z) - \hat{q}^{\Delta}(\tau, z) \right| : b = 1, ..., B \right\},$$

   and denote the percentile to be $c_{1-\alpha}(\tau, z)$. Then the point-wise bootstrap confidence interval is given by

$$\left[\hat{q}^{\Delta}(\tau, z) - \frac{c_{1-\alpha}(\tau, z)}{\sqrt{n}}, \hat{q}^{\Delta}(\tau, z) + \frac{c_{1-\alpha}(\tau, z)}{\sqrt{n}} \right]. \tag{11}$$

4. To test a joint hypothesis, construct a test statistic and a bootstrap critical value as follows. For example, consider the following hypothesis to test for positive QTE, which corresponds to the second row of Table 1:

$$\begin{aligned} &H_0 : q^{\Delta}(\tau, z) \leqslant 0 \text{ for all } (\tau, z) \in \mathcal{T} \times \mathcal{Z}, \text{ vs} \\ &H_1 : q^{\Delta}(\tau, z) > 0 \text{ for some } (\tau, z) \in \mathcal{T} \times \mathcal{Z}. \end{aligned} \tag{H.1}$$

We calculate the test statistic and its bootstrap counterpart as follows:

$$T = \max_{(\tau,z)\in\mathcal{T}\times\mathcal{Z}} \max\{\hat{q}^{\Delta}(\tau,z), 0\}, \text{ and}$$
$$T_b^* = \max_{(\tau,z)\in\mathcal{T}\times\mathcal{Z}} \max\{\hat{q}_b^{\Delta *}(\tau,z) - \hat{q}^{\Delta}(\tau,z), 0\}, b = 1, ..., B.$$

The test statistic T is compared to a bootstrap critical value which is the $1-\alpha$ quantile of $\{T_b^* : b = 1, ..., B\}$. Because a discrete number of quantiles are evaluated in empirical applications, we replace the supremum with the maximum.

As another example that corresponds to the third row of Table 1, consider the following hypothesis to test for treatment effect heterogeneity across quantiles within a subgroup:

$$H_0 : q^{\Delta}(\tau,z) = \bar{q}^{\Delta}(z) \text{ for all } \tau\in\mathcal{T} \text{ and some } z\in\mathcal{Z}, \text{ vs}$$
$$H_1 : q^{\Delta}(\tau,z) \neq \bar{q}^{\Delta}(z) \text{ for some } \tau\in\mathcal{T} \text{ and all } z\in\mathcal{Z}. \tag{H.2}$$

The test statistic and its bootstrap counterpart are calculated as follows:

$$T = \max_{(\tau,z)\in\mathcal{T}\times\mathcal{Z}} |\hat{q}^{\Delta}(\tau,z) - \tilde{q}^{\Delta}(z)|, \text{ and}$$
$$T_b^* = \max_{(\tau,z)\in\mathcal{T}\times\mathcal{Z}} \left|\hat{q}_b^{\Delta *}(\tau,z) - \tilde{q}_b^{\Delta *}(z) - \left(\hat{q}^{\Delta}(\tau,z) - \tilde{q}^{\Delta}(z)\right)\right|, b = 1, ..., B,$$

where $\tilde{q}^{\Delta}(z)$ and $\tilde{q}_b^{\Delta *}(z)$ are as defined in (10). (Note that we replace the integral in $\tilde{q}^{\Delta}(z)$ by the mean of $\{\hat{q}^{\Delta}(\tau,z) : \tau\in\mathcal{T}\}$, and similarly with $\tilde{q}_b^{\Delta *}(z)$.) The bootstrap critical values can be computed from $\{T_b^* : b = 1, ..., B\}$ as before.

5. In the multiple testing procedure, we construct test statistics and bootstrap critical values at each step in the step-down procedure. For example, consider using the procedure to report the quantile-subgroup pairs for which a policy has a positive impact. First, we consider the following individual hypotheses:

$$H_{0,(\tau,z)} : q^{\Delta}(\tau,z) \leqslant 0, \text{ vs}$$
$$H_{1,(\tau,z)} : q^{\Delta}(\tau,z) > 0, \tag{H.3}$$

for each (τ,z). To apply the step-down procedure to find pairs (τ,z) at which the null hypothesis in (H.3) is violated, we use the test statistic and the bootstrap test statistics

for each set $W' \subset \mathcal{T} \times \mathcal{Z}$,

$$T(\tau, z) = \max\{\hat{q}^{\Delta}(\tau, z), 0\}, \text{ and}$$
$$T_b^*(W') = \max_{(\tau,z)\in W'} \max\{\hat{q}_b^{\Delta *}(\tau, z) - \hat{q}^{\Delta}(\tau, z), 0\}, b = 1, ..., B.$$

Let $\hat{c}_{1-\alpha}(W')$ be the bootstrap critical value of $\{T_b^*(W') : b = 1, ..., B\}$. We begin with $\tilde{W}_1 = \mathcal{T} \times \mathcal{Z}$, and following Section 2.2.2, the iterative procedure is repeated until at step $k \geqslant 2$,

$$\tilde{W}_k = \left\{(\tau, z) \in \mathcal{T} \times \mathcal{Z} : T(\tau, z) \leqslant \hat{c}_{1-\alpha}(\tilde{W}_{k-1})\right\}, \tag{12}$$

and eventually $\tilde{W}_k = \tilde{W}_{k-1}$, indicating that no further element of $\tilde{W}_k$ can be eliminated. The set of pairs (τ, z) removed from the set W_1 and not in $\tilde{W}_k$ are those where the null hypothesis $H_{0,(\tau,z)}$ is violated.

Similarly, researchers can use a multiple testing procedure to identify the subgroups in which QTE are heterogeneous across quantiles. As in last row of Table 1, consider

$$H_{0,z} : q^{\Delta}(\tau, z) = \bar{q}^{\Delta}(z) \text{ for all } \tau \in \mathcal{T}, \text{ vs}$$
$$H_{1,z} : q^{\Delta}(\tau, z) \neq \bar{q}^{\Delta}(z) \text{ for some } \tau \in \mathcal{T}. \tag{H.4}$$

To test (H.4), we construct the test statistic and the bootstrap test statistic:

$$T(z) = \max_{\tau\in\mathcal{T}} |\hat{q}^{\Delta}(\tau, z) - \tilde{q}^{\Delta}(z)|, \text{ and}$$
$$T_b^*(W') = \max_{z\in W'} \max_{\tau\in\mathcal{T}} \left|\hat{q}^{\Delta *}(\tau, z) - \tilde{q}^{\Delta *}(z) - \left(\hat{q}^{\Delta}(\tau, z) - \tilde{q}^{\Delta}(z)\right)\right|, b = 1, ..., B,$$

and undertake the step-down procedure as in (12).

When testing (H.3) or (H.4), the step-down procedure yields quantile-subgroup pairs or subgroups at which the hypothesis is rejected at a level α, but researchers may also be interested in p-values. To obtain p-values, we first re-run the step-down procedure for different values of α on a grid. Then, for each quantile-subgroup pair or subgroup, the p-value is the value of α such that we reject the null hypothesis at that value but fail to reject it at the next lower value on the grid of αs.

The above procedures can be used to conduct a wide variety of tests for empirical applications. In the next section, we demonstrate the value of these tests by revisiting Andrabi, Das, and Khwaja (2017).

3 Empirical application

3.1 Experimental design and data

Andrabi, Das, and Khwaja (2017) conduct an experiment in 112 Pakistani villages located in three districts in Pakistan's most populous province, Punjab, to study the impact of providing parents with a detailed two-page report card on their child's own and 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 and, for all of the schools in the village, presented information on the average score, number of children tested, and 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.

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 decisions 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. More recently, Camargo et al. (2014) develop a reduced-form version of a dynamic model of managerial effort along the lines of Holmström (1999) to show how test score disclosure would lead to heterogenous changes in subsequent student test score performance between public and private schools.

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. The extent of this heterogeneity can vary across subgroups defined by school type, because 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 QTE 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 rich longitudinal data. Beginning 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.

Table 2 presents 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). However, the difference is about a third as large as the village-level treatment effect reported by Andrabi, Das, and Khwaja (2017), which is due to the fact that the authors do not weigh by the number of children in each village. 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 at the village level and not

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 (0.942) 5786	−0.0229 (0.886) 6324	0.569
Average test score, round 2	0.186 (1.004) 6266	0.229 (0.943) 6538	0.012
Female child	0.425 (0.494) 8443	0.431 (0.495) 8760	0.438
Child's age	9.680 (1.505) 6616	9.671 (1.446) 7117	0.702
Village literacy rate	38.46 (12.88) 8443	36.26 (10.63) 8760	0.000
Number of households in village	708.3 (375.8) 8443	797.3 (591.0) 8760	0.000
School Herfindahl index	0.181 (0.0680) 8443	0.183 (0.0676) 8760	0.092
Village wealth (median monthly expenditure)	4498.5 (1649.4) 8443	4638.6 (1454.8) 8760	0.000
Government school (excluded category: private)	0.675 (0.468) 6617	0.698 (0.459) 7118	0.003
School size	251.6 (199.5) 6617	248.7 (194.9) 7118	0.386
High scoring school (above 60th percentile)	0.499 (0.500) 8443	0.486 (0.500) 8760	0.096
Mother's education above middle school	0.325 (0.469) 3097	0.333 (0.471) 3278	0.498
Father's education above middle school	0.630 (0.483) 3090	0.590 (0.492) 3278	0.001

Source: Andrabi, Das, and Khwaja (2017).
Note: 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.

at 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. However, our testing approach incorporates propensity score weighting, which allows us to balance treatment and control group based on these observed variables.

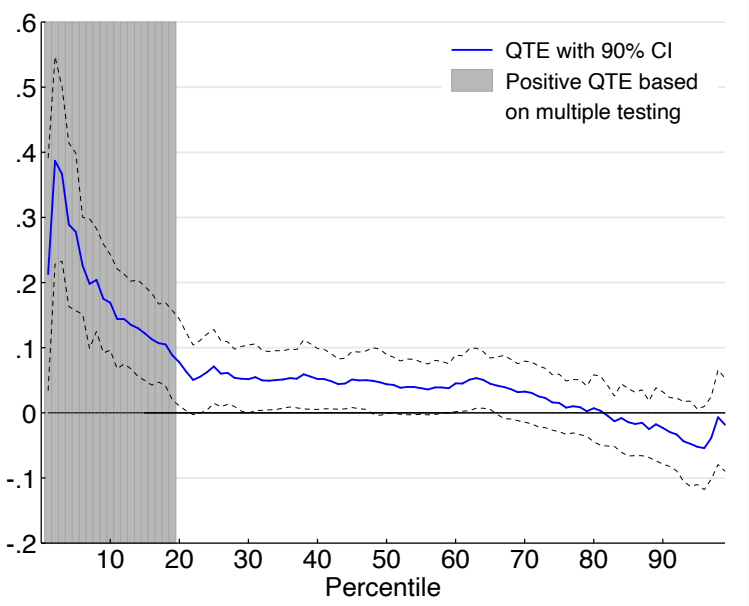
3.2 Results

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. As with any analysis based on a field experiment, the external validity of our findings is limited, and our results are not intended to generalize to distributional treatment effects of school report cards in other contexts. Our analysis focuses on the average of standardized test scores across three subjects after random assignment as our outcome variable, and we estimate QTE of access to report cards for percentiles 1 to 99 using the Firpo (2007) estimator. For the results that follow, we set the level of each test to $\alpha = 0.05$. All test results are based on bootstrapping with $B = 2{,}999$.

To balance covariates between the treatment and control groups, we estimate the propensity score $\hat{p}_d(x)$ using a parametric logit specification, but due to random assignment of the treatment, our results are unlikely to be sensitive to the chosen specification. We include district fixed effects, and village wealth, literary rate, school Herfindahl index, and number of households when estimating the propensity score.

Andrabi, Das, and Khwaja (2017) use stratified randomization of villages (half the villages in each district are treated) and include district fixed effects to account for stratification. For statistical inference, the authors account for clustering within village. In contrast, we draw bootstrap samples of individuals instead of using the block bootstrap to resample villages. Thus, because stratification reduces sampling error, and not accounting for clustering leads to an underestimation of standard errors, it is unclear whether our confidence intervals are too conservative or too wide. To assess the sensitivity of our results to our agnosticism regarding what is the appropriate level at which to cluster and regarding stratification, we additionally present results with $\alpha = 0.1$ instead of 0.05 in Online Appendix C.

To infer treatment effects for specific individuals from QTE we have to assume that there are no rank reversals in the test score distribution between the treatment and control groups. All studies estimating distributional treatment effects have to make this assumption, but it not possible to directly test for rank invariance. Bitler, Gelbach, and Hoynes (2008) develop a test that provides evidence for rank reversals using the distributions of observable



Note: Multiple testing results show quantiles for which the QTE is positive at an FWER of 5 percent (see hypothesis (H.3) in Section 2.3).

Figure 1: QTE and multiple testing results, no subgroups

characteristics of treatment and control group. They find only small deviations from rank preservation for the Self-Sufficiency Project. Djebbari and Smith (2008) and Kottelenberg and Lehrer (2017) use this test in different applications and also find only minor evidence of rank reversals. Irrespective of whether the no-rank-reversal assumption holds, positive QTE imply that the treatment has a positive effect for some interval of the test score distribution.

3.2.1 Results for the full sample

First, we consider QTE for the entire sample, i.e. we set $Z_i = 1$ in the notation of Section 2. Note that QTE relate to the unconditional distribution of round 2 test scores and are unrelated to treatment effects conditional on baseline test scores. Figure 1 shows our estimated QTE for the full sample along with 90 percent point-wise confidence intervals. We present 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 point-wise confidence intervals are calculated according to (11) in Section 2.3. We find point-wise significant and positive QTE extending from the 1st to the 21st, the 24th to 29th, and the 31st to 47th percentiles. From the 82nd to the 99th percentile the point estimates for QTE are negative but the point-wise confidence intervals include zero.

Table 3: Testing for presence of positive QTE and QTE heterogeneity without subgroups

	Test statistic	Critical value at 5%	p-value
Test for positive QTE (H.1)	0.387	0.227	0.00333
Test for QTE heterogeneity (H.2)	0.329	0.238	0.013

Note: This table shows test results for the hypotheses 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 Section 2.3, setting $Z_i = 1$.

Table 3 summarizes the results for joint hypothesis testing for positive and heterogeneous QTE. First, we test the null hypothesis of no positive treatment effect at any percentile, i.e. hypothesis (H.1) in Section 2.3, setting $Z_i = 1$. As shown in Figure 1, the largest QTE (which occurs at the second percentile) equals 0.387. With a bootstrap critical value of 0.227, 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 across quantiles, i.e. hypothesis (H.2), setting $Z_i = 1$. The test statistic, which is calculated as the largest deviation from the mean estimated QTE ($\tilde{q}^{\Delta} = 0.0583$), equals 0.329. With a bootstrap critical value of 0.238, we also reject this null hypothesis at 5 percent with a p-value of 0.013. This result implies 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 hypothesis (H.3), setting $Z_i = 1$. 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 (99 in this case). Examining Figure 1 we observe that the set of significantly positive QTE 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 QTE once we adjust for multiple testing. The smallest and largest quantiles at which QTE are significantly positive correspond to gains of 0.088 and 0.387, respectively, at the 19th and 2nd percentiles. Of the percentiles with point-wise significant QTE, 37 percent remain significant when applying our multiple testing procedure. 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 their parents receive a school report card.

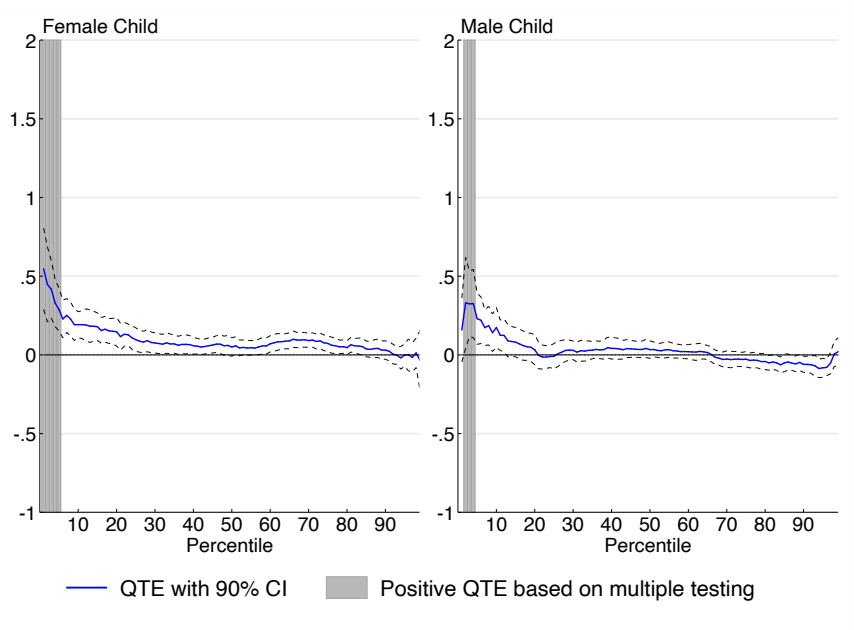
3.2.2 Results by subgroup

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 performance. Following Andrabi, Das, and Khwaja (2017), we consider subgroups defined by child characteristics, type of school, and characteristics of the villages. The research questions we consider are specific to a category of subgroups. For example, we ask whether treatment effects are constant across percentiles of the test score distribution and children's gender. We do not generally conduct multiple testing across different categories of subgroups. However, our methodology would accommodate such a setting, and we do obtain multiple testing results for subgroups defined by school ownership and quality.

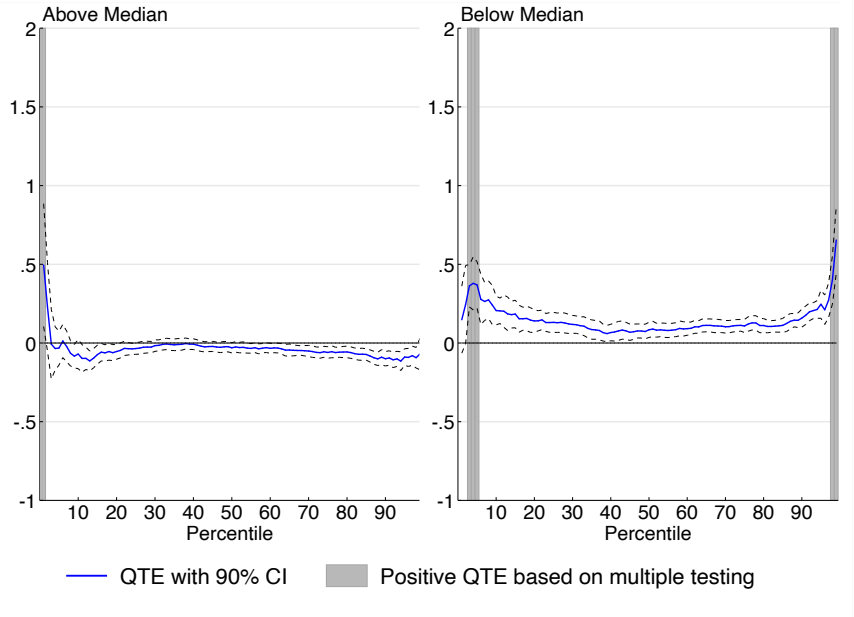
Figure 2 presents QTE conditional on child gender and child baseline test scores. These figures provide an easy and intuitive way to check which subgroups benefit from being assigned to receive report cards (heterogeneity across subgroups). In addition, we can inspect the figure for each subgroup to determine the portion of the student test score distribution in which individuals exhibit positive subgroup-specific QTE (heterogeneity within subgroup). Shaded areas continue to denote significant QTE based on our multiple testing procedure of testing hypothesis (H.3).

Figure 2a presents QTE by child gender. 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 5th percentile and boys in the 2nd to 4th percentile.

The second panel of Figure 2 considers subgroups defined by whether the child's baseline test score was above or below the median. The estimated QTE 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 QTE. 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. This



(a) By child's gender



(b) By child's baseline test score

Note: Multiple testing results show quantiles for which the QTE is positive at an FWER of 5 percent (see hypothesis (H.3) in Section 2.3).

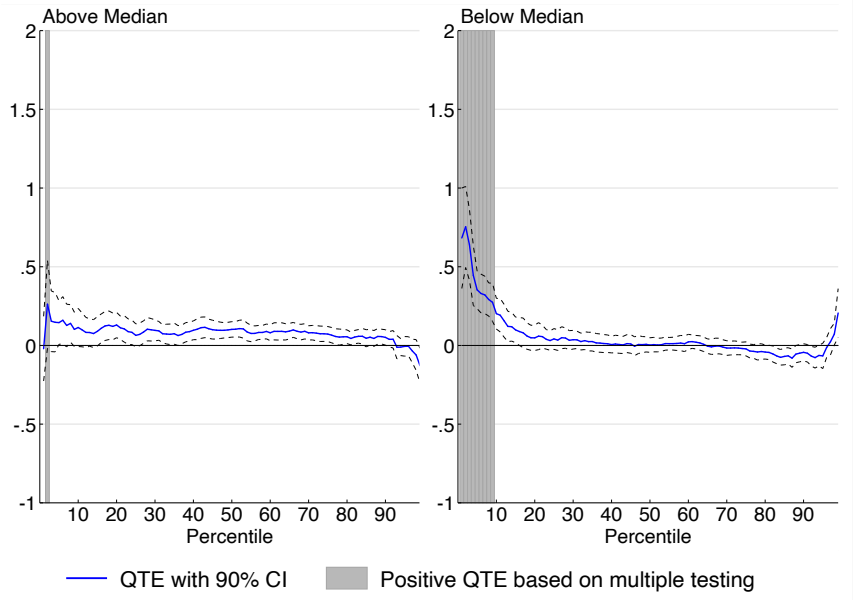
Figure 2: QTE and multiple testing results, by child characteristics

finding suggests that the children who scored below the median at baseline but are in the top percentile at follow-up are those who realized the largest improvement in test scores and thereby benefited the most from the information provision.

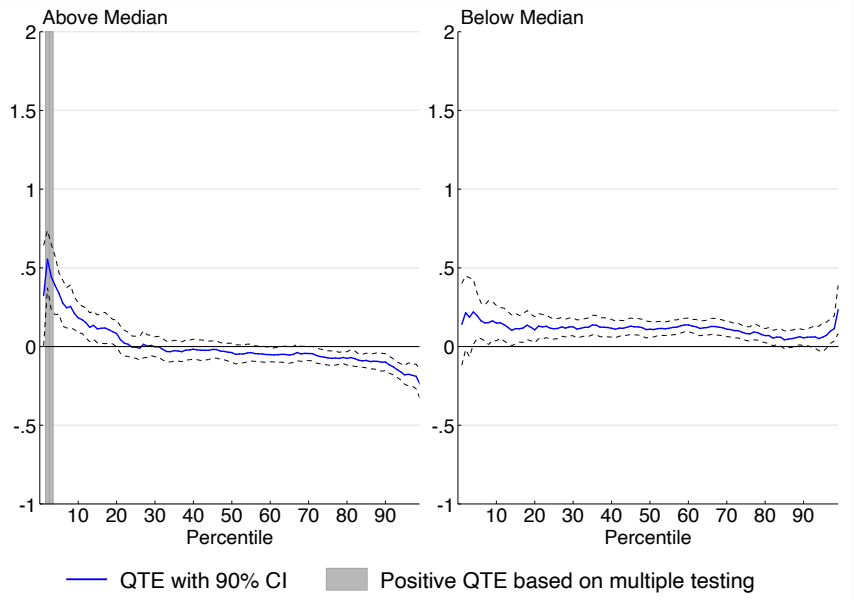
Next, we construct subgroups based on village characteristics. Figure 3 shows the estimated subgroup-specific QTE and multiple testing results. We find significant treatment effects predominantly for children in villages with below-median wealth, above-median literacy rates, below-median school concentration as proxied 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 obviously not be a successful strategy in villages with low literacy rates. In general, these results can show policymakers which subgroups should be targeted with an accountability program. As with the results for the full sample, we only find statistically significant effects in the bottom part of the test score distribution for each of subgroups considered, except for the subgroup of low-performing students shown in Figure 2b. This finding is not surprising because low-performing students tend to benefit from many school interventions that do not change a specific eduction input, including early childhood education interventions (see, e.g., Bitler, Hoynes, and Domina, 2014; Kottelenberg and Lehrer, 2017, for evidence using U.S. and Canadian data respectively).

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. 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. Figure 4 illustrates the estimation and multiple testing results. We find that significantly positive QTE 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). (Table VII in Online Appendix III of Andrabi, Das, and Khwaja (2017) shows average treatment effects by children's

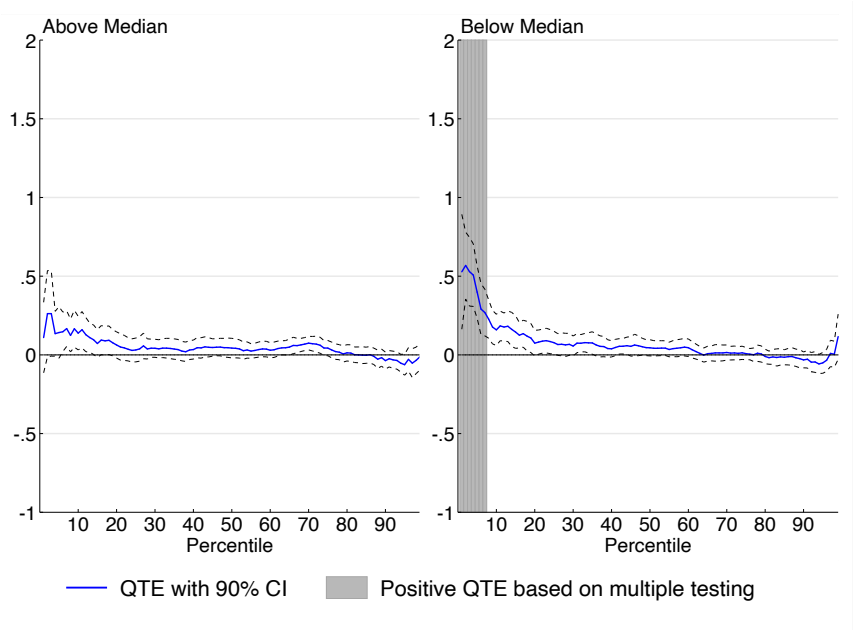


(a) By village wealth

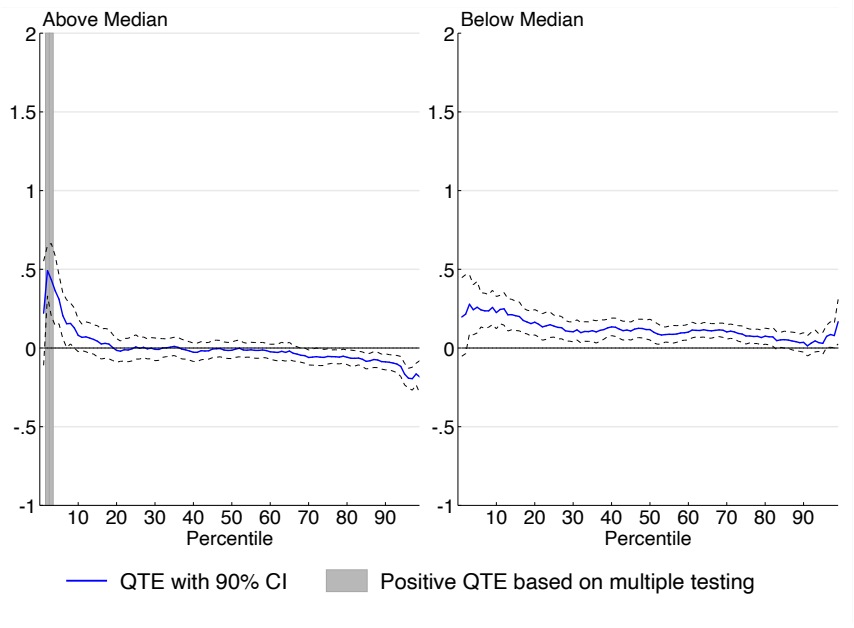


(b) By village literacy rate

Figure 3: QTE and multiple testing results by village characteristics



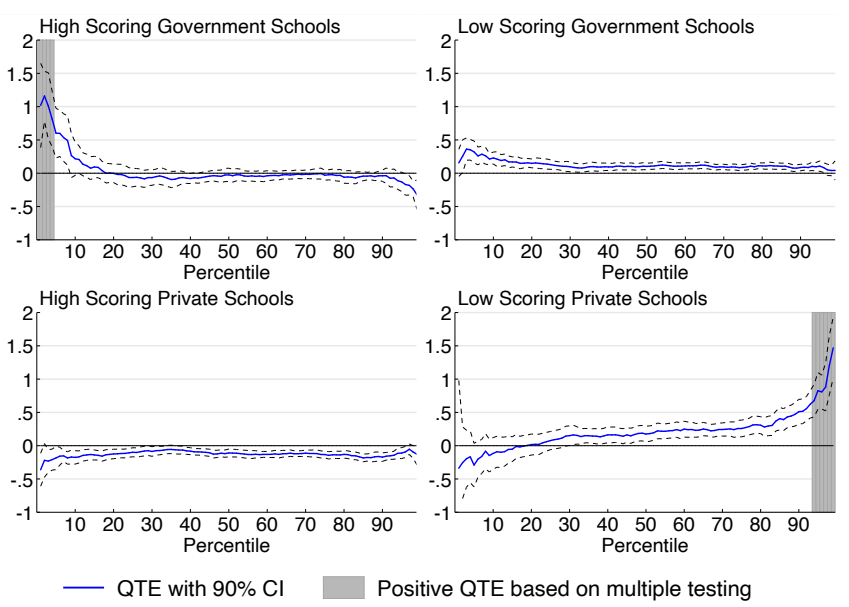
(c) By school Herfindahl index



(d) By village size

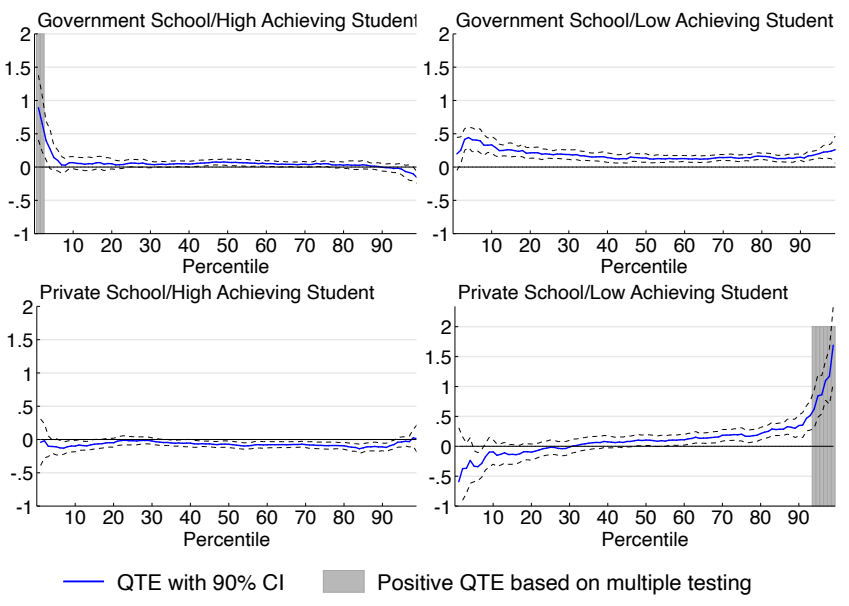
Note: Multiple testing results show quantiles for which the QTE is positive at an FWER of 5 percent (see hypothesis (H.3) in Section 2.3).

Figure 3: QTE and multiple testing results by cillage characteristics (continued)



Note: Multiple testing results show quantiles for which the QTE is positive at an FWER of 5 percent (see hypothesis (H.3) in Section 2.3).

Figure 4: QTE and multiple testing results by school type and performance



Note: Multiple testing results show quantiles for which the QTE is positive at an FWER of 5 percent (see hypothesis (H.3) in Section 2.3).

Figure 5: QTE and multiple testing results by school type and child's performance relative to school performance

Table 4: Testing for treatment effect heterogeneity between subgroups

Subgroup category	Test statistic	p-value
Child's gender	0.45	0.01
Child's baseline test score	0.535	0.019
Village wealth	0.702	0
Village literacy rate	0.555	0
School Herfindahl Index	0.494	0.018
Village size	0.495	0.00701
School type and school performance	1.27	0.0981
School type and child's performance relative to school	1.57	0.002

Note: This table shows test results that indicate for which subgroups categories we can reject treatment effects that are homogenous within subgroups for all subgroups, i.e. we test hypothesis (H.2) in Section 2.3.

baseline performance relative to their school.) 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 QTE 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. The latter finding is consistent with the observed pattern in Figure 2b.

Taken together, our results in Figures 4 and 5 provide additional nuance to 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 4 presents the results for testing hypothesis (H.2). This null hypothesis posits that

there are no differences across subgroups that can explain the observed heterogeneity of QTE in the full sample. We can reject the hypothesis for all sets of subgroups at a level of 5 percent, except for the subgroup category "School type and school performance." We conclude that differences across subgroups do not explain the observed distributional treatment effects in the whole sample.

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.4). 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 most subgroup categories we find evidence of treatment effect heterogeneity for at least one of the subgroups. Exceptions are the subgroup categories "School type and school performance" and "School type and child's performance relative to school" where we cannot reject the null hypothesis for any subgroup at an FWER of 5 percent. These results clearly suggest a substantial amount of treatment effect heterogeneity between subgroups and across the student performance distribution within subgroups.

4 Conclusion

In this paper we introduce general tests for treatment effect heterogeneity in settings with selection on observables. These tests are motivated by empirical practice in settings with many covariates requiring researchers to use a parametric specification for the propensity score. The results of the proposed 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. We establish the asymptotic validity of this bootstrap multiple testing procedure for QTE. In contrast to much of the existing literature on procedures to test for heterogeneous treatment effects, these tests make corrections for multiple testing and therefore provide valid inference under dependence between subgroups and quantiles.

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 demonstrate in which subgroups and which test score quantiles within subgroups the benefits of information provision are largest. In addition, our empir-

Table 5: Testing Which Subgroups Exhibit Treatment Effect Heterogeneity

Subgroup category	Test statistic	p-value
Child's gender		
Female	0.45	0.01
Male	0.304	0.092
Child's baseline test score		
Above median	0.535	0.019
Below median	0.51	0.019
Village wealth		
Above median	0.208	1
Below median	0.702	0
Village literacy rate		
Above median	0.555	0
Below median	0.121	0.81
School Herfindahl Index		
Above median	0.215	1
Below median	0.494	0.018
Village size		
Above median	0.495	0.008
Below median	0.158	0.482
School type and school performance		
High scoring government	1.13	0.128
Low scoring government	0.237	0.307
High scoring private	0.233	0.33
Low scoring private	1.27	0.099
School type and child's performance relative to school		
Government/high achieving	0.834	0.176
Government/low achieving	0.261	0.375
Private/high achieving	0.0949	0.984
Private/low achieving	1.57	0.002

Note: 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.4) in Section 2.3. p-values are calculated using a grid with step size 0.001. Hence an entry of zero indicates that the corresponding p-value is below 0.001.

ical 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 the shape of how peer groups influence academic outcomes (see, e.g., Ding and Lehrer, 2007).

We conclude by emphasizing that our multiple testing approach is generally applicable in various ways beyond what this paper demonstrated. First, the tests can be applied to situations with multiple treatments (e.g., List, Shaikh, and Xu, 2019) or could be extended to situations with 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 microeconomic research.

References

Abadie, Alberto and Matias D. Cattaneo. 2018. "Econometric Methods for Program Evaluation." *Annual Review of Economics* 10 (1):465–503.

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.

Athey, Susan and Guido W. Imbens. 2017. "The Econometrics of Randomized Experiments." In *Handbook of Economic Field Experiments*, vol. 1, edited by Abhijit Banerjee and Esther Duflo, chap. 3. Amsterdam: Elsevier, 73–140.

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

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.

———. 2008. "Distributional impacts of the Self-Sufficiency Project." *Journal of Public Economics* 92 (3):748–765.

———. 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.

Bitler, Marianne P., Hilary W. Hoynes, and Thurston Domina. 2014. "Experimental Evidence on Distributional Effects of Head Start." NBER Working Paper 20434.

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

Bugni, Federico, Ivan A. Canay, and Azeem M. Shaikh. 2018. "Inference Under Covariance-Adaptive Randomization." *Journal of the American Statistical Association* 113 (524):1741–1768.

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." Unbublished manuscript.

Cattaneo, Matias D. 2010. "Efficient semiparametric estimation of multi-valued treatment effects under ignorability." *Journal of Econometrics* 155 (2):138–154.

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.

Chung, EunYi and Mauricio Olivares. 2020. "Permutation Test for Heterogeneous Treatment Effects with a Nuisance Parameter." Unbublished manuscript.

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.

Ding, Peng, Avi Feller, and Luke Miratrix. 2016. "Randomization Inference for Treatment Effect Variation." *Journal of the Royal Statistical Society Series B* 78 (3):655–671.

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):64–80.

Donald, Stephen G. and Yu-Chin Hsu. 2014. "Estimation and Inference for Distribution Functions and Quantile Functions in Treatment Effect Models." *Journal of Applied Econometrics* 178 (3):383–397.

Fan, Yanqin and Jisong Wu. 2010. "Partial Identification of the Distribution of Treatment Effects in Switching Regime Models and Its Confidence Sets." *Review of Economic Studies* 77 (3):1002–1041.

Figlio, David and Susanna Loeb. 2011. "School accountability." In *Handbook of the Economics of Education*, vol. 3, edited by Eric A. Hanushek, Stephen Machin, and Ludger Woessman, chap. 8. Amsterdam: Elsevier, 383–421.

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

Firpo, Sergio and Cristine Pinto. 2016. "Identification and Estimation of Distributional Impacts of Interventions Using Changes in Inequality Measures." *Journal of Applied Econometrics* 31 (3):457–486.

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*. New Brunswick, NJ: Rutgers University Press.

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.

Galvao, Antonio and Liang Wang. 2015. "Uniformly Semiparametric Efficient Estimation of Treatment Effects With a Continuous Treatment." *Journal of the American Statistical Association* 110 (512):1528–1542.

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

Goldman, M. and D. M. Kaplan. 2018. "Comparing Distributions by Multiple Testing Across Quantiles or CDF Values." *Journal of Econometrics* 206 (1):143–166.

Guerre, Emmanuel and Camille Sabbah. 2012. "Uniform Bias Study and Bahadur Representation for Local Polynomial Estimators of the Conditional Quantile Function." *Econometric Theory* 28 (1):87–129.

Hahn, Jinyong. 1995. "Bootstrapping Quantile Regression Estimators." *Econometric Theory* 11 (1):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 Evaluation." *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*, edited by Caroline M. Hoxby, chap. 8. Chicago: University of Chicago Press, 287–342.

Jiang, Liang, Xiaobin Liu, Peter C. B. Phillips, and Yichong Zhang. 2020. "Bootstrap Inference for Quantile Treatment Effects in Randomized Experiments with Matched Pairs." Unbublished manuscript.

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

Khan, Shakeep and Elie Tamer. 2010. "Irregular Identification, Support Conditions, and Inverse Weight Estimation." *Econometrica* 78 (6):2021–2042.

Koenker, Roger. and Zijie 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.

Kottelenberg, Michael J. and Steven F. Lehrer. 2017. "Targeted or Universal Coverage? Assessing Heterogeneity in the Effects of Universal Child Care." *Journal of Labor Economics* 35 (3):609–653.

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.

Lehmann, E. L. and Joseph P. Romano. 2005. *Testing Statistical Hypotheses*. New York: Springer.

List, John A., Azeem M. Shaikh, and Yang Xu. 2019. "Multiple Hypothesis Testing in Experimental Economics." *Experimental Economics* 22:773–793.

Ma, Xinwei. and Jiangshen Wang. 2020. "Robust Inference Using Inverse Probability Weighting." *Journal of the American Statistical Association* 115 (532):1851–1860.

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

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

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

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.

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.

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

Zhang, Yichong and Xin Zheng. 2020. "Quantile Treatment Effects and Bootstrap Inference Under Covariate-Adaptive Randomization." *Quantitative Economics* 11 (3):957–982.