ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Statistical Methodology
Series B

J. R. Statist. Soc. B (2016)
78, *Part* 3, *pp.* 655–671

Randomization inference for treatment effect variation

Peng Ding, Avi Feller and Luke Miratrix

Harvard University, Cambridge, USA

[Received May 2014. Revised March 2015]

Summary. Applied researchers are increasingly interested in whether and how treatment effects vary in randomized evaluations, especially variation that is not explained by observed covariates. We propose a model-free approach for testing for the presence of such unexplained variation. To use this randomization-based approach, we must address the fact that the average treatment effect, which is generally the object of interest in randomized experiments, actually acts as a nuisance parameter in this setting. We explore potential solutions and advocate for a method that guarantees valid tests in finite samples despite this nuisance. We also show how this method readily extends to testing for heterogeneity beyond a given model, which can be useful for assessing the sufficiency of a given scientific theory. We finally apply our method to the National Head Start impact study, which is a large-scale randomized evaluation of a Federal preschool programme, finding that there is indeed significant unexplained treatment effect variation.

Keywords: Causal inference; Head Start; Heterogeneous treatment effect; Randomization test

1. Introduction

Researchers and practitioners are increasingly interested in whether and how treatment effects vary in randomized evaluations. For example, we might be interested in assessing the effect of scaling up a promising intervention evaluated on a limited subpopulation (O'Muircheartaigh and Hedges, 2014). If we use only observed characteristics to predict the programme's effectiveness on the new population, we might wonder whether we are missing critical unexplained variation, which could undermine our generalization. Similarly, we might want to determine whether different theoretical models are sufficiently rich to explain observed behaviour in a randomized experiment. For instance, is a simple model of constant treatment effects within subgroups sufficient to explain observed labour supply behaviour in welfare reform experiments? Or is there meaningful unexplained variation, as predicted by labour supply theory (Bitler *et al.*, 2010)? The goal of this paper is to build a framework to assess treatment effect variation that is not explained by observed covariates, which is also known as *idiosyncratic variation* (e.g. Heckman *et al.* (1997) and Djebbari and Smith (2008)).

Unfortunately, assessing such variation is difficult—to paraphrase *Anna Karenina*: 'constant treatment effects are all alike; every varying treatment effect varies in its own way'. In general, researchers investigating specific types of idiosyncratic variation must therefore rely on strong modelling assumptions to draw meaningful conclusions from the data (Cox, 1984; Heckman *et al.*, 1997; Gelman, 2004). The key contribution of our paper is an approach that tests for

Address for correspondence: Avi Feller, Department of Statistics, Harvard University, 1 Oxford Street, Cambridge, MA 02139, USA.
E-mail: avifeller@fas.harvard.edu

 1369–7412/16/78655

the presence of unexplained treatment effect variation without requiring any such modelling assumptions. In the simplest case, the method proposed is a test of the null hypothesis that the treatment effect is constant across all units. More generally, the approach tests whether there is significant unexplained variation beyond a specified model of treatment effect.

Of course, all treatment effects vary in practice, especially in the social sciences, which are our area of application. The key question is whether the unexplained variation is sufficiently large to be of substantive importance. As with all omnibus-type testing procedures, rejecting this null hypothesis does not provide any indication of the source of the unexplained variation. Rather, we view this procedure as a non-parametric first step in characterizing the results of a randomized experiment.

In the simplest no-covariate case, the goal of this approach is to test whether the treatment outcome distribution is the same as the control outcome distribution shifted by the average treatment effect (ATE), a constant. Such testing would be straightforward if this shift were known—we could simply apply standard Kolmogorov–Smirnov- (KS) type tests. However, since the shift is not known, it is a nuisance parameter that we must estimate. In this case, otherwise sensible methods, such as 'plug-in' (PI) approaches, can fail, even asymptotically (e.g. Babu and Rao (2004)). Incorporating covariates only compounds this problem.

Testing features of distributions in the presence of nuisance parameters has a long history in statistics and econometrics, where in the latter it is known as the *Durbin problem* (Durbin, 1973). Several references tackle this issue in the context of comparing treatment and control outcome distributions, appealing to various asymptotic justifications to bypass the nuisance parameter problem. These include a martingale transformation (Koenker and Xiao, 2002) and subsampling (Chernozhukov and Fernández-Val, 2005).

We take a different approach, exploiting the act of randomization as the 'reasoned basis for inference' (Fisher, 1935). The corresponding Fisher randomization test (FRT) does not rely on further model assumptions, asymptotics or regularity conditions (for a review, see Rosenbaum 2002a). For the constant treatment effect case, when the ATE is assumed known, the FRT procedure yields an exact p-value for the sharp null hypothesis of a constant treatment effect (for one generalization, see Abadie (2002)). When the ATE is unknown, the null hypothesis is no longer sharp. To correct for this, we first construct a confidence interval (CI) for the ATE, repeat the FRT procedure pointwise over that interval and then take the maximum p-value. As Berger and Boos (1994) showed, this procedure guarantees a valid test, despite the presence of the nuisance parameter. This process readily generalizes for testing treatment effects beyond a hypothesized model.

Our FRT-based approach has several key advantages. First, since the FRT approach is justified by the physical randomization alone, it yields valid inference in finite samples without relying on asymptotics or requiring absolutely continuous outcomes. Second, the FRT automatically accounts for complex experimental designs, such as stratified and matched pair randomizations or even re-randomization (Morgan and Rubin, 2012). Third, this procedure is valid for any test statistic, though some statistics will be more powerful in certain settings. With this flexibility, researchers can easily extend the FRT approach, tailoring the specific test statistic to their particular problem of interest.

Using this framework, we assess treatment effect variation in the National Head Start impact study (HSIS), which is a large-scale randomized evaluation of Head Start, a Federal preschool programme (Puma *et al.*, 2010). After evaluating a range of null models, we find that there is substantial unexplained treatment effect variation, even when considering heterogeneity across age of student, dual language learner status and baseline academic skill level, suggesting that policy makers should not base key decisions on the topline results alone.

The paper proceeds as follows. Section 2 describes treatment effect variation by using the potential outcomes framework as well as how variation depends on the chosen outcome scale. Section 3 gives an overview of various measures of treatment effect variation. Section 4 outlines the FRT method that we propose, and Section 5 generalizes this approach to incorporate covariates. Sections 6–8 provide some simulation studies, apply this approach to Head Start and discuss next steps. The on-line supplementary material contains all proofs as well as additional details.

The programs that were used to analyse the data can be obtained from

`http://wileyonlinelibrary.com/journal/rss-datasets`

2. Defining treatment effect variation

Following the causal inference literature, we describe our approach by using the potential outcomes framework (Neyman, 1990; Rubin, 1974). We focus on the case of a randomized experiment with a binary treatment Z_i and continuous outcome Y_i. Let N be the number of subjects in the study, with N_1 of them randomly assigned to treatment and N_0 of them assigned to control. As usual, we invoke the stable unit treatment value assumption, which states that there is only one version of the potential outcomes and that there is no interference between subjects (Rubin, 1980).

With this set-up, the potential outcomes for subject i under treatment and control are $Y_i(1)$ and $Y_i(0)$. The *science table* is the $N \times 2$ table containing the potential outcomes for all N units (Rubin, 2005). Each individual's observed outcome is a function of the treatment assignment and the potential outcomes:

$$Y_i^{\text{obs}} = Z_i Y_i(1) + (1 - Z_i) Y_i(0),$$

where the randomness comes only from the random treatment assignment. Let $\mathbf{Z}$ and $\mathbf{Y}^{\text{obs}}$ denote the treatment assignment and observed outcome vectors respectively. We define the individual treatment effect in the usual way as $\tau_i = Y_i(1) - Y_i(0)$, but note that other contrasts are also possible. Finally, we define the finite sample ATE as

$$\tau = \frac{1}{N} \sum_{i=1}^{N} \{Y_i(1) - Y_i(0)\}.$$

This is a statement about the N units that we observe. In other words, we condition on the sample at hand.

The treatment effect is constant if $\tau_i = \tau$ for all $i = 1, \ldots, N$. Otherwise, we say that the treatment effect varies across experimental units. In the language of hypothesis testing, we can define the constant treatment effect null as

$$H_0^{\text{C}} : Y_i(1) - Y_i(0) = \tau \qquad \forall\, i \quad \text{for some } \tau. \tag{1}$$

If τ were known to be $\tau = \tau_0$, this hypothesis becomes sharp.

2.1. Constant shift

We cannot, however, directly observe any individual level treatment effects τ_i, since we only ever observe one potential outcome for each unit. Instead, we observe the marginal distributions of the treatment and control groups. Because of this, much of the literature (see, for example, Cox (1984)) defines a 'constant treatment effect' as a statement that the marginal cumulative

distribution functions (CDFs) of the potential outcomes of the experimental and control unit distributions $F_0(y)$ and $F_1(y)$ are a constant shift apart:

$$H_0 : F_1(y) = F_0(y - \tau) \qquad \text{for some } \tau. \tag{2}$$

Rejecting H_0 implies rejecting the more restrictive null hypothesis H_0^{C} that $\tau_i = \tau$ for all i, but rejecting H_0^{C} does not necessarily imply rejecting H_0. That said, it is difficult to imagine the practical situation in which there is a substantial, varying treatment effect that nonetheless yields parallel CDFs. Even more interestingly, the two nulls appear to be indistinguishable given observed data. Therefore, although not formally correct, we generally view tests for H_0^{C} as tests for H_0. Simulation studies, which are not shown, suggest that this practice generally leads to valid, if somewhat conservative, tests. Understanding this relationship is an important area of future work and is closely related to the interplay between Neyman and Fisher style tests. See, for example, Ding (2014).

2.2. Treatment effect variation and scaling

Whether a given treatment effect is constant critically depends on the scale of the outcomes. For example, a job training programme that has a constant effect in earnings does not have a constant effect in log-earnings. This scaling issue is a particularly salient issue if the outcome is, say, test scores in an educational context where scale is not necessarily well defined.

Cox (1984) demonstrated the importance of scaling in a special case first explored by G. E. H. Reuter: if the marginal CDFs of $Y(1)$ and $Y(0)$ do not cross, there is a monotone transformation such that the distributions of the transformed treatment and control outcomes are a constant shift apart. Unfortunately, G. E. H. Reuter has since died and his proof is lost to the literature; we provide a proof of this theorem in the on-line supplementary material.

Theorem 1. Assume that $F_1(\cdot)$ and $F_0(\cdot)$ are both continuous and strictly increasing CDFs of the marginal distributions of $Y(1)$ and $Y(0)$ respectively, with strict stochastic dominance $F_1(y) < F_0(y)$ for all y on $[F_0^{-1}(0), F_1^{-1}(1)]$. There is an increasing monotone transformation g such that the CDFs of $g\{Y(1)\}$ and $g\{Y(0)\}$ are parallel.

Although the applicability of this result is limited to non-crossing CDFs, it nonetheless emphasizes the importance of scale and of understanding the problem at hand. In general, whether a given transformation is substantively reasonable depends on the context: a cube root transformation might be very sensible if the outcome is in cubic centimetres but not if the outcome is in dollars (Berrington de González and Cox, 2007).

3. Measures of treatment effect variation

There are many approaches to measuring treatment effect variation, dating back to early work on non-additivity in randomized experiments (see Berrington de González and Cox (2007)). We briefly highlight three basic measures: comparing marginal variances, comparing marginal CDFs and comparing marginal quantiles. The usual testing procedures with the measures that are discussed here typically yield reasonable inference only when particular conditions, such as normality or asymptotic regularity, are met. In the next section, we show how the FRT can yield exact p-values with any of these test statistics, regardless of whether these conditions are met.

To fix notation, assume that the potential outcome for treatment z is drawn from the distribution of $Y(z)$, with marginal probability density function $f_z(y)$, CDF $F_z(y)$, quantile function $F_z^{-1}(q)$, mean μ_z and variance σ_z^2. Sample analogues are denoted with circumflexes.

3.1. Comparing variances

Following Cox (1984), we can assess treatment effect heterogeneity by examining the marginal variances of the treatment and control outcomes. In particular, if the treatment effect is constant, $Y_i(1) = Y_i(0) + \tau$ and $\text{var}\{Y_i(1)\} = \text{var}\{Y_i(0)\}$. Therefore, unequal sample variances imply treatment effect heterogeneity, although the converse is not necessarily true. This makes the variance ratio

$$t_{\text{var}} = \hat{\sigma}_1^2 / \hat{\sigma}_0^2$$

an attractive statistic, especially if the researcher believes that the treatment plausibly induces greater variance (e.g. Gelman (2004)).

Furthermore, if the marginal distributions of potential outcomes are normal, then t_{var} follows an F-distribution and the corresponding test is the uniformly most powerful test of equal variance. However, as we show in the on-line supplementary material, the F-test is highly sensitive to departures from normality, even asymptotically. We also provide a test that uses higher order moments, such as kurtosis, to improve inference in this case.

3.2. Comparing cumulative distribution functions

In general, second-order moments might not capture some important features of heterogeneity, especially when τ_i varies with $Y_i(0)$. For example, a classroom intervention might have the largest effect on the lowest performing students. An alternative approach compares marginal CDFs rather than higher order moments, suggesting the use of a KS-like test to compare the treatment and control groups. The classic KS statistic, which measures the maximum pointwise distance between two curves, is $t_{\text{KS}} = \max_y |\hat{F}_1(y) - \hat{F}_0(y)|$. This test, however, could reject if the treatment effect is constant but non-zero, since it is an omnibus test for any difference in distribution.

To focus on heterogeneous treatment effects, we want to shift one of the CDFs by the ATE, and then to compare the resulting distributions. In particular, if τ were known, we could calculate

$$t_{\text{KS}}(\tau) = \max_y |\hat{F}_0(y) - \hat{F}_1(y + \tau)|.$$

Under the null hypothesis, the two aligned CDFs should be the same and we can directly compare the observed test statistic with the null distribution for the classic, non-parametric distribution-free KS test.

In practice, τ is unknown and is therefore a nuisance parameter. One natural approach is to plug in the difference-in-means estimate, $\hat{\tau} = \hat{\mu}_1 - \hat{\mu}_0$, yielding the 'shifted' KS (SKS) statistic:

$$t_{\text{SKS}} = \max_y |\hat{F}_0(y) - \hat{F}_1(y + \hat{\tau})|.$$

As we prove in the on-line supplementary material, however, comparing this test statistic with the usual null KS distribution yields invalid p-values. In particular, t_{SKS} converges to an asymptotic distribution that depends on the underlying distributions of the outcomes.

3.3. Comparing quantiles

A third approach focuses on quantiles rather than on CDFs. In this formulation,

$$F_1^{-1}(q) = F_0^{-1}(q) + \tau(q),$$

where $\tau(q)$ is the *quantile process* for the treatment effect. If the effect is constant, then $\tau(q)$ is constant across q. On the basis of this, Chernozhukov and Fernández-Val (2005) proposed a

class of test statistics based on the estimated quantile process,

$$t_{\mathrm{QP}} = \|\hat{\tau}(q) - \hat{\tau}\|,$$

where $\hat{\tau}(q)$ is an estimate of the treatment effect at the qth quantile, and $\|\cdot\|$ is some norm, such as sup.

As in the CDF case, τ remains a nuisance parameter, so the Durbin problem remains. Chernozhukov and Fernández-Val (2005) solved this via subsampling. Their main argument is that, under some regularity conditions, a particular form of recentred subsampling can yield asymptotically valid tests for whether $\tau(q)$ is constant, despite dependence on $\hat{\tau}$. Chernozhukov and Fernández-Val (2005) and Linton *et al*. (2005) also proposed a bootstrap variant of the subsampling procedure, though the bootstrap does not have the same general theoretical guarantees as subsampling. For other approaches for inference on quantiles, see Doksum and Sievers (1976), Rosenbaum (1999) and Koenker and Xiao (2002).

4. A randomization test for treatment effect variation

Our analytic approach is based on the FRT. To perform an FRT, a researcher needs three main ingredients: a randomized treatment assignment mechanism, a sharp null hypothesis and a test statistic $t(\mathbf{Z}, \mathbf{Y}^{\mathrm{obs}})$, such as those in the previous section. Under the sharp null, all missing potential outcomes can be imputed and are thus known. Given all the potential outcomes, a researcher can then enumerate the possible values of a specified test statistic under all possible randomizations. This enumeration forms the exact null distribution, called the reference distribution, of that statistic.

4.1. Fisher randomization test with known τ

First consider a sharp null hypothesis of no heterogeneity for a known τ:

$$H_0^{\tau} : Y_i(1) = Y_i(0) + \tau \qquad \forall\, i.$$

Given this null, we can immediately impute all missing potential outcomes from the observed data. For a unit with $Z_i = 1$, the potential outcome under treatment is Y_i^{obs} and the potential outcome under control is $Y_i^{\mathrm{obs}} - \tau$. For a unit with $Z_i = 0$, the potential outcome under treatment is $Y_i^{\mathrm{obs}} - \tau$ and the potential outcome under control is Y_i^{obs}.

The steps of the FRT are then as follows:

(a) Calculate the test statistic for the observed data, $t = t(\mathbf{Z}, \mathbf{Y}^{\mathrm{obs}})$.
(b) Given the observed outcomes Y_i^{obs}, the treatment assignment Z_i and the sharp null H_0^{τ}, generate the corresponding science table.
(c) Enumerate all possible treatment assignments $\tilde{\mathbf{Z}}$, under the given treatment assignment mechanism. These are all possible treatment assignments that we could have observed for a given experiment. Typically, there are too many possible enumerations so we instead take a random sample from the set of all possible assignment vectors.
(d) For each possible assignment $\tilde{\mathbf{Z}}$, compute
    (i) the observed outcomes $\tilde{\mathbf{Y}}^{\mathrm{obs}}$ given $\tilde{\mathbf{Z}}$ and the science table, and
    (ii) the test statistic $\tilde{t} = t(\tilde{\mathbf{Z}}, \tilde{\mathbf{Y}}^{\mathrm{obs}})$.
    The resulting distribution of $\tilde{t}$ across all randomizations is the exact distribution of the test statistic given the units in the sample and the null hypothesis.
(e) Compare the observed statistic t with its null distribution and obtain the p-value

$$p(\tau) \equiv \Pr(t \geqslant \tilde{t}).$$

This procedure yields an exact test for any test statistic assuming that τ is known. For instance, we can use as test statistics any of the measures of treatment effect heterogeneity that were discussed in the previous section, such as $t_{\text{var}}, t_{\text{SKS}}$ and t_{QP}.

4.2. Fisher randomization test with unknown τ

When τ is unknown, the null hypothesis is no longer 'sharp' in the sense that we can no longer impute all the missing potential outcomes. We provide two options.

4.2.1. Fisher randomization test plug-in method

One option is to impute the science table with the estimated $\hat{\tau}$ instead of τ, and to run the FRT to obtain the distribution of t for that table. Ideally, if $\hat{\tau}$ is close to τ, the resulting science tables will be close in that the exact reference distribution for the imputed science table should look similar to the true reference distribution for our sample. If this is so, then inference from this PI procedure should be close to the case where τ is known, i.e. $p(\hat{\tau}) \approx p(\tau)$. Nonetheless, as Berger and Boos (1994) discussed, there are no general theoretical guarantees from such a procedure. In fact, as we show in the simulation studies, this approach can lead to invalid results when $\hat{\tau}$ is highly variable, such as for skewed distributions, though it does appear to have sensible size for approximately normal outcomes.

Nevertheless, this approach is distinct from appealing to the asymptotic distribution of a given test statistic. Instead, this attempts to generate a reference distribution based on the data at hand, which may make the Durbin problem far less severe. Even so, as we show next, we can guarantee validity with a mild extension of this approach.

4.2.2. Fisher randomization test confidence interval method

An alternative approach is to find the maximum p-value across all values of the nuisance parameter, $\tau' \in (-\infty, \infty)$:

$$p_{\text{sup}} = \sup_{\tau'} p(\tau')$$

where $p(\tau')$ is obtained by performing an FRT under the sharp null hypothesis $H_0^{\tau'}$. Although p_{sup} is conservative, it is still valid since $\Pr(p_{\text{sup}} \leqslant \alpha) \leqslant \Pr\{p(\tau) \leqslant \alpha\} \leqslant \alpha$. This approach, however, leads to two complications in practice:

(a) maximizing a quantity over the entire real line is computationally intractable and
(b) doing so can lead to a dramatic loss in statistical power.

Berger and Boos (1994) proposed a convenient fix to these issues—rather than maximize over the entire real line, they instead maximize over a $(1-\gamma)$-level CI for τ, CI_γ:

$$p_\gamma = \sup_{\tau' \in \text{CI}_\gamma} p(\tau') + \gamma.$$

Following Rosenbaum (2002a), we could obtain an exact (under H_0^C) CI, CI_γ, by inverting FRTs for a sequence of sharp null hypotheses, $Y_i(1) - Y_i(0) = \tau'$. In practice, we approximate this CI on the basis of the Neyman variance estimator (Neyman, 1990). The following proposition guarantees the validity of the resulting p-value.

Proposition 1. Given that CI_γ is a $(1-\gamma)$-level CI for τ, p_γ is a valid p-value, in the sense that $\Pr(p_\gamma \leqslant \alpha) \leqslant \alpha$ under the null hypothesis.

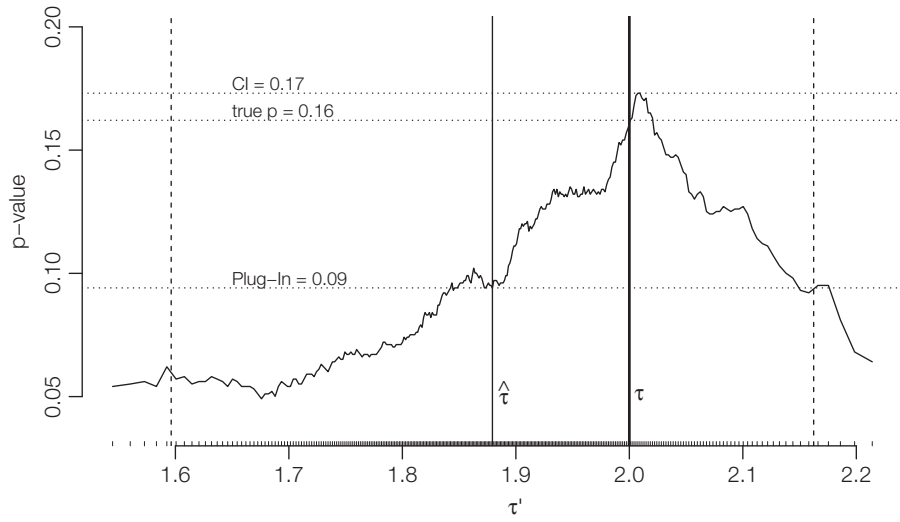

Fig. 1. *p*-values over the range of the nuisance parameter (the rug plot indicates the grid of sampled τ): ¦, bounds of 99.9% CI_γ on τ; ·····, *p*-values from, bottom to top, plugging in $\hat{\tau}$, using the known τ and maximizing the *p*-value over CI_γ

As Berger and Boos (1994) noted, the behaviour of the *p*-values at the tails of the nuisance parameter interval can be complex and, unsurprisingly, depends on both the specific test statistic and the value of the nuisance parameter. For example, they might climb or be driven to zero. Although we cannot provide theoretical guarantees, in our experience, the *p*-values for our chosen test statistics tend towards 0 or remain flat for values of τ' that are moderately far from τ, which suggests that our method does not sacrifice much in terms of power.

To illustrate this procedure, we simulate a balanced randomized experiment with a constant treatment effect, $N = 200$, $Y_i(0) \sim^{\text{IID}} \text{exponential}(1)$, and $Y_i(1) = Y_i(0) + 2$. Fig. 1 shows *p*-values from FRTs for a fixed data set under $H_0^{\tau'}$ for τ' in a 99.9% CI, following the procedure described above using t_{SKS} as the test statistic. If the true τ were known, we could obtain the exact *p*-value of $p = 0.16$. The *p*-value at the observed value of $\hat{\tau}$ is too low, around $p = 0.09$, demonstrating why a simple PI approach may yield incorrect size. Finally, taking the maximum *p*-value over the 99.9% CI yields a *p*-value of $p = 0.17$, which is only slightly larger than the true value of 0.16. This figure, with the 'mountain' shape, is typical for this test statistic under many data generation processes.

5. Incorporating covariates

In practice, we typically observe a vector of individual level pretreatment covariates $\mathbf{X}$ that are possibly related to the outcome. This can help to increase the power of our test and also enable exploration of variation beyond that which can be explained by $\mathbf{X}$.

5.1. Using covariates to improve power

To improve power we allow the chosen test statistic to account for the relationship between the covariates and outcome, such as via a linear regression of outcome on covariates and treatment with no interaction. In the linear regression case, for example, we can generate a 'regression-adjusted KS statistic'. This statistic compares the CDFs of the residuals of a regression of Y on $\mathbf{X}$ and Z (with no interaction of $\mathbf{X}$ and Z). Let $\hat{e}_i = Y_i^{\text{obs}} - \hat{Y}_i$ be the residuals of a prespecified

regression, with $\hat{Y}_i$ being the associated predicted values. Then define our test statistic as

$$t_{\mathrm{RKS}} = \max_y |\hat{F}_{\mathrm{e1}}(y) - \hat{F}_{\mathrm{e0}}(y)| \tag{3}$$

where $\hat{F}_{\mathrm{e1}}(y)$ and $\hat{F}_{\mathrm{e0}}(y)$ are the empirical CDFs of the residuals $\hat{e}_i$ for the treatment and control groups respectively.

To motivate this, consider the simple regression of Y on Z. The residuals of this regression are $\hat{e}_i = Y_i^{\mathrm{obs}} - \hat{\mu}_1$ for the treated units and $\hat{e}_i = Y_i^{\mathrm{obs}} - \hat{\mu}_0$ for the control units. Since $\hat{\tau} = \hat{\mu}_1 - \hat{\mu}_0$ and $\hat{F}_0(y) - \hat{F}_1(y + \hat{\tau}) = \hat{F}_0(y - \hat{\mu}_0) - \hat{F}_1(y - \hat{\mu}_1)$, the regression-adjusted KS statistic for the simple regression of Y on Z is equivalent to the SKS statistic: without covariates, $t_{\mathrm{SKS}} = t_{\mathrm{RKS}}$.

Now, by including covariates we hope to remove outcome variation due to covariates, making variation in treatment effect more readily apparent. As long as $\mathbf{X}$ is predictive of Y, regression adjustment reduces residual variation in the marginal outcomes but cannot directly reduce variation in the treatment effect. In general, covariate adjustment will therefore yield more powerful test statistics. Importantly, since the validity of the approach is justified by randomization alone, this adjustment does not require any underlying model assumptions. This approach is analogous to classical, model-assisted covariate adjustment in randomized experiments (Rosenbaum, 2002b; Lin, 2013).

We can easily repeat this approach with t_{var}, redefining the test statistic via the residual variance after a regression of Y on $\mathbf{X}$. However, accounting for covariates with quantile-based statistics is more complicated, with two basic approaches in the literature. In the conditional approach, we redefine t_{QP} via the estimate of $\tau(q)$ in a quantile regression of Y on both $\mathbf{X}$ and Z, as in Koenker and Xiao (2002). In the unconditional approach, we redefine t_{QP} via the estimate of $\tau(q)$ in a weighted quantile regression of Y on Z, with weights defined as a given function of $\mathbf{X}$, as in Firpo (2007) or Firpo *et al.* (2009).

5.2. Treatment effect variation beyond covariates

In many applications, the constant treatment effect null hypothesis may be of limited scientific interest. Instead, we wish to investigate whether there is significant treatment effect variation beyond a particular model for the treatment. For example, Bitler *et al.* (2010) proposed the constant treatment effect within subgroups model, which assumes that the ATE differs across observable subgroups (e.g. by education or age group) but is otherwise constant within those subgroups.

To make this more precise, let $\mathbf{W}$ be an $n \times (k+1)$ matrix of the unit vector and k pretreatment covariates. The unit vector corresponds to the overall ATE and the covariates allow for modelled treatment effect heterogeneity. We then replace the null hypothesis of a constant treatment effect with the assumption that the individual level treatment effects are a particular function of $\mathbf{W}$:

$$H_0^{\mathbf{W}}: \quad Y_i(1) - Y_i(0) = \boldsymbol{\beta}^{\mathrm{T}} \mathbf{W}_i \qquad \forall\, i \quad \text{for some } \boldsymbol{\beta}, \tag{4}$$

where $\boldsymbol{\beta}$ is some (unknown) vector of coefficients for $\mathbf{W}$. Under the null, there is some β such that the set of $Y_i(1) - \boldsymbol{\beta}^{\mathrm{T}} \mathbf{W}_i$ yields the same CDF as the set of $Y_i(0)$.

We can easily test this hypothesis by using the regression-adjusted KS statistic t_{RKS} constructed via the residuals of a regression of Y on $\mathbf{W}$, Z and $\mathbf{W} \times Z$. This regression yields point estimates $\hat{\boldsymbol{\beta}}$ with a corresponding $(k+1)$-dimensional $(100 - \gamma)\%$ confidence region. To obtain the FRT PI p-value, we simply use the science table based on $\hat{\boldsymbol{\beta}}$. To obtain the FRT CI p-value, we must repeat the FRT procedure for each point in a potentially high dimensional grid. We defer a detailed discussion of the estimation issues in this setting to Ding *et al.* (2015).

We can also extend this regression approach to account for covariates that are not assumed to interact with the treatment (i.e. those in $\mathbf{X}$ but not $\mathbf{W}$). Furthermore, we can allow the treatment effect model to be arbitrarily flexible, including series expansions on the covariates, such as splines or higher order polynomials. See Crump *et al*. (2008) for a discussion of non-parametric estimation in this context.

5.3. Subgroup variation

We briefly turn to the special case in which the treatment effect is assumed to vary across discrete groups. Let Y_{ik}^{obs} be the observed outcome of unit i in group k, for $i = 1, \ldots, n_k$ and $k = 1, \ldots, K$, with n_{1k} the number of treated units in group k.

For example, consider a stratified experiment, where both n_k and n_{1k} are fixed. Of course, we can always analyse a stratified experiment as if it were K separate, completely randomized experiments. However, we can also test whether variation across strata explains the full variation in treatment effects. This corresponds to the following joint null hypothesis of stratum-specific treatment effects $\mathcal{T} \equiv (\tau_1, \ldots, \tau_K)$:

$$H_0^{\text{joint}} : Y_{ik}(1) = Y_{ik}(0) + \tau_k \qquad \forall\, i, \quad \forall\, k, \quad \text{for some } \mathcal{T}.$$

Under this null, the pooled CDF of the recentred-by-stratum outcomes of all the units under treatment (i.e. the residuals from outcome regressed on strata) would be the same as for the control.

To test the null, we then need a measure of discrepancy between the estimates of the two CDFs as our test statistic. Several choices are possible. First, we can use t_{RKS}, the regression-based test statistic above, letting $\mathbf{W}$ be a matrix of indicators for stratum membership and $\boldsymbol{\beta}$ be $\mathcal{T}$ (with no intercept). However, if the proportions of treated units differ across strata or if homoscedasticity is implausible, pooling may not be appropriate. Instead, we can post-stratify by weighting each group-by-stratum empirical CDF with weight proportional to the stratum size. The revised $\hat{F}_{\text{ez}}$ is then

$$\hat{F}_{\text{ez}}(y) = \sum_{k=1}^{K} \frac{n_k}{n} \hat{F}_{\text{ekz}}(y)$$

where $\hat{F}_{\text{ekz}}(y)$ is the empirical CDFs of the $Y_{ik}^{\text{obs}} - \hat{\mu}_{kz}$ for those units in stratum k with $Z_i = z$.

Similarly, we might instead take a weighted average of individual stratum level test statistics as

$$t_{\text{WSKS}} = \sum_{k=1}^{K} \frac{n_k}{n} t_{\text{SKS},k}.$$

Fig. 2 demonstrates this last approach by extending the results from Fig. 1 into two dimensions. Here we have two distinct subgroups, one of 75 units and one of 375 units, and simulate a balanced randomized experiment with a treatment effect that is constant within each subgroup, but not constant overall. The baseline distributions are exponential. We then test for treatment effect heterogeneity beyond these discrete subgroups. To do this we search over a confidence set, which is depicted in Fig. 2, for a maximum p-value. We again see the 'mountain shape' and end up with a final p-value of 0.46, *versus* $p = 0.43$ for known τ and $p = 0.39$ for the plug-in $\hat{\tau}$. As in the one-dimensional case, the plug-in p-value is lower than the true p-value. Moreover, the maximum p-value is only modestly higher than the truth, as the p-values fall away at moderate distances from the true τ. This plot is typical over several simulation settings. Finally, as expected, if we do an omnibus test for heterogeneity beyond a single ATE, we reject

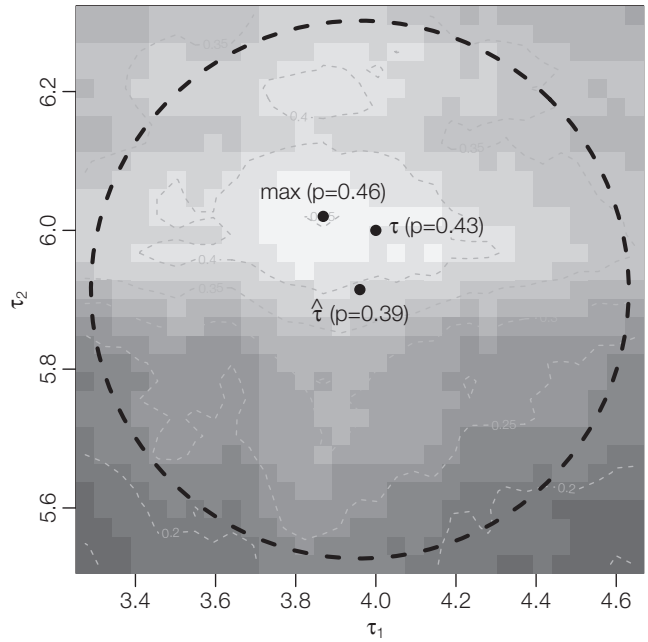

Fig. 2. *p*-values over the region of the nuisance parameters (the *p*-values due to the maximum, the PI and the oracle truth are all marked on the plot): ◌, $1-\gamma$ confidence region for the nuisance parameters

with $p < 0.005$. Our model of constant treatment effect within groups is thus significantly better than a single average, and we have no evidence for needing a more complex model.

The constant treatment effect within subgroups model of Bitler *et al*. (2010) is equivalent to H_0^{joint} except that the number of treated units in each group, n_{k1}, is possibly random rather than fixed. Here simply conditioning on the observed n_{k1} for each group (i.e. considering only randomizations that maintain the n_k) and performing the analysis as above yields valid inference. This is a conditional randomization test; the analogue of post-stratification for testing rather than estimation (see, for example, Holt and Smith (1979) and Miratrix *et al*. (2013)).

6. Simulation studies

We now turn to a series of simulation studies that confirm the validity of the FRT approach and assess power under a range of plausible scenarios.

6.1. Validity results

First, we examine the various methods under the null hypothesis of a constant treatment effect. To assess validity, we repeat the following steps 5000 times each for a given test statistic and underlying distribution:

(a) generate a sample from the underlying distribution, assuming a constant treatment effect;
(b) randomly assign treatment and obtain observed outcomes;
(c) calculate our test statistic t_{SKS}; finally
(d) calculate a *p*-value by using each of several different approaches described below.

We assess five methods.

(i) *Naive PI*: this method calculates the usual KS p-value, assuming that the estimated treatment effect is in fact the true treatment effect.
(ii) *FRT PI and FRT CI*: these methods are the two FRT-based approaches discussed above. For this simulation, we use a 99.9% CI_γ for $\hat{\tau}$ (i.e. $\gamma = 0.001$).
(iii) *Subsampling*: this method is the subsampling approach of Chernozhukov and Fernández-Val (2005), with their recommended subsampling size of $b = 20 + n^{1/4}$.
(iv) *Bootstrap*: this method is based on the bootstrap that was proposed by Chernozhukov and Fernández-Val (2005) and Linton *et al.* (2005), using the t_{SKS} test statistics. To generate the bootstrap distribution, we de-mean the treatment and control groups and sample with replacement from the pooled vector of residuals, keeping the number of treatment and control units fixed.

We assess these five methods for the following distributions: the standard normal, t_5, standard exponential and log-normal, each with a constant treatment effect of 1 unit.

Table 1 shows the rejection rates for a test of size $\alpha = 0.05$ for each method and data-generating process. As expected, the naive PI approach fails dramatically, either yielding hyperconservative or highly invalid size. The FRT PI approach appears to work well for the symmetric normal and t_5-distributions but leads to invalid size for the skewed exponential and log-normal distributions. The FRT CI approach corrects for this, yielding exact or conservative size for all data generation processes that are assessed here, where conservative indicates lower than nominal rejection rates. It is encouraging that, even when the FRT CI is conservative, it is not dramatically so, suggesting that we are not giving up too much power due to the maximization procedure. Subsampling yields correct, if slightly conservative, rejection rates overall. Finally, the bootstrap approach is invalid for the normal, t_5- and exponential distributions.

The bootstrap approach that we used seemed the most promising choice. Other alternatives to the bootstrap exist, but they seem to perform even more poorly. For example, one seemingly obvious bootstrap is to sample repeatedly, with replacement, N_1 treatment cases and N_0 control cases from their respective original samples, calculating the resulting test statistic. Ideally, this would capture the variability of the entire process, giving a valid p-value for the actual observed test statistic. Unfortunately, even if the null hypothesis were true, the bootstrap null would generally not be in this context, and so we would end up simulating our distribution under a 'near alternative' which gives poor size. We confirmed this intuition with simulations, which are not shown in this paper, that indeed show that this approach can fail catastrophically.

Table 1. Size of $\alpha = 0.05$ tests, in percentage points, for various methods under H_0^{C}†

Method	*Results for the following models and values of n:*							
	Normal		t_5		*Exponential*		*Log-normal*	
	100	*1000*	*100*	*1000*	*100*	*1000*	*100*	*1000*
FRT PI	4.5	5.1	5.4	5.2	11.3	7.7	15.1	7.0
FRT CI	1.9	3.8	2.1	3.7	4.1	4.9	4.5	5.0
Subsampling	2.3	4.6	1.5	1.6	2.1	2.3	1.1	0.7
Bootstrap	9.3	8.8	7.4	8.3	6.3	6.1	5.0	6.5
Naive PI	0.0	0.0	0.0	0.0	12.3	21.4	36.2	44.7

†Estimates are based on 5000 replications, which implies a simulation standard error of approximately 0.3 percentage points.

6.2. General power simulations

To assess the power of these methods under select alternatives we mirror a set of simulation studies that were conducted by both Koenker and Xiao (2002) and Chernozhukov and Fernández-Val (2005). For these simulations, we repeatedly generate data with different levels of treatment effect heterogeneity, denoted by σ_τ, and we estimate the probability that a method would reject the null hypothesis of constant treatment effect (at $\alpha = 0.05$) given draws of data and random treatment assignment. Since the bootstrap, FRT PI, and naive PI are invalid tests, we do not include them here.

We use a binary version of the data generation process from Chernozhukov and Fernández-Val (2005):

$$Y_i(0) = \varepsilon_i,$$
$$\tau_i = 1 + \sigma_\tau Y_i(0).$$

with $\varepsilon_i \sim N(0, 1)$. This model can also be expressed as the classic additive treatment effect model under normality, $Y_i(1) = Y_i(0) + \tau_i$, where $Y_i(0) \sim N(0, 1)$; $\tau_i \sim N(1, \sigma_\tau^2)$ on the margin, and $\sigma_\tau = 0$ corresponds to a constant treatment effect. Note that, as Cox (1984) observed, the F-test is the uniformly most powerful test in this setting.

We then extend the simulations from Chernozhukov and Fernández-Val (2005) by imposing log-normality rather than normality. In particular, we assume a treatment effect of the form

$$\log\{Y_i(0)\} = \varepsilon_i,$$
$$\tau_i = 1 + \sigma_\tau Y_i(0).$$

Then marginally $Y_i(1) \sim$ log-normal $\{\log(\sigma_\tau + 1), 1\} + 1$. In either case, for $\sigma_\tau > 0$, the treatment effect increases with $Y_i(0)$, which is non-negative. Rosenbaum (1999) called this kind of treatment effect variation a *dilated effect*.

Table 2 shows the main power results. For normal outcomes, both the FRT CI and the subsampling methods have correct size when $\sigma_\tau = 0$. However, subsampling appears to be more powerful for $\sigma_\tau > 0$, perhaps because the asymptotics 'kick in' quickly under normality. For log-normal outcomes, however, the situation is reversed, with much greater rejection rates under the FRT CI method than under subsampling.

7. Application to the Head Start impact study

Initially launched in 1965, Head Start is the largest Federal preschool programme today, serving around 900 000 children each year at a cost of roughly $8 billion. The National HSIS is the first major randomized evaluation of the programme (Puma *et al.*, 2010). The published report found that, on average, providing children and their families with the opportunity to enrol in Head Start improved children's key cognitive and social–emotional outcomes. The report also included ATE estimates for a variety of subgroups of interest, though there is only significant impact variation across a small number of the reported pretreatment covariates.

After these findings were released, many researchers argued that the reported topline results masked critical variation in programme impacts. For example, Bitler *et al.* (2013) showed that the treatment is differentially effective across quantiles of the test score distribution; Bloom and Weiland (2015) explored variation in programme impacts across select subgroups and across the 351 Head Start centres in the study; and Feller *et al.* (2014) investigated differential effects based on the setting of care that each child would have received in the alternative treatment condition.

Table 2. Rejection rates for $\alpha = 0.05$ tests, in percentage points, under select alternative hypotheses with different levels of treatment effect variation σ_τ and data generation processes†

N	*Results for FRT CI*			*Results for subsampling*		
	$\sigma_\tau = 0$	$\sigma_\tau = 0.2$	$\sigma_\tau = 0.5$	$\sigma_\tau = 0$	$\sigma_\tau = 0.2$	$\sigma_\tau = 0.5$
Normal outcomes						
100	2.3	5.5	23.1	2.4	8.4	39.2
400	3.5	24.8	93.0	4.0	40.1	98.4
800	3.5	52.3	100.0	4.6	72.9	100.0
Log-normal outcomes						
100	4.7	7.3	19.3	1.2	2.2	5.9
400	4.7	19.8	70.5	0.6	3.4	32.9
800	4.6	35.1	94.1	0.8	9.1	70.7

†Estimates are based on 5000 replications, which imply a simulation standard error of approximately 0.3 percentage points.

All these approaches, however, estimate treatment effect variation by relying on a specific set of models, such as quantile or hierarchical regression. Given the breadth of research in this area, a natural question is whether the topline and subgroup ATEs for the HSIS are indeed sufficient summaries of the programme's effect. We investigate this question by focusing on the Peabody picture vocabulary test, which is a widely used measure of cognitive ability in early childhood. We also utilize a rich set of pretreatment covariates, including pretest score, child's age, child's race, mother's education level and mother's marital status. In addition, we follow the experimental design and ensure that the randomizations that are used in the FRT procedure are stratified by Head Start centre. For the sake of exposition, we restrict our analysis to a complete-case subset of the HSIS, with $N_1 = 2238$ in the treatment group and $N_0 = 1348$ in the control group. Note that this restriction could lead to a range of inferential issues which we do not explore here; see Feller *et al.* (2014) for a detailed discussion.

As shown in Table 3, we apply the FRT procedure to a set of increasingly flexible null hypotheses. The least flexible models, models 1 and 2, assesses the null hypothesis of constant treatment effect across all units without and with covariate adjustment, using the t_{SKS}-statistic and the t_{RKS}-statistic respectively. Model 3 adjusts for pretreatment covariates and allows the treatment effect to vary by child's age (3 *versus* 4 years old). The most flexible model, model 4, allows the treatment effect to vary by child's age, child's dual language learner status and an indicator for whether the child was in the bottom quartile on an assessment of pre-academic skills before the study. The resulting p-values are roughly $p = 0.03$ for the model without covariates and $p < 0.01$ across all three models that adjust for covariates, clearly demonstrating significant unexplained variation regardless of the exact specification. This provides evidence that there is indeed substantial treatment effect variation beyond that explained by these subgroups.

8. Discussion

Researchers are increasingly interested in assessing treatment effect heterogeneity. We propose a framework to unify and generalize some existing statistical procedures for inference about such variation, using randomization as the 'reasoned basis for inference' for the testing procedure. As

Table 3. FRT *p*-values for the HSIS, based on 2000 repetitions†

Model	*p-value*	*Treatment effect varies by*	*Control for covariates*
1	0.033	—	—
2	0.005	—	✓
3	0.005	Age	✓
4	0.003	Age Dual language learner Pre-academic skills	✓

†Models 1 and 2 correspond to a null hypothesis of constant treatment effect. Models 3 and 4 allow the treatment effect to vary across given covariates.

a result, the method does not rely on any further model assumptions, asymptotics or regularity conditions. We use simulation studies to confirm that this approach yields valid results in finite samples and that its power is competitive with some existing approaches, especially subsampling. Finally, we apply this method to the National HSIS, a large-scale randomized evaluation, and find that there is indeed significant unexplained treatment variation.

Other randomization-based approaches to heterogeneity also exist. These methods typically specify a model for heterogeneity and test based on that model. For example, Rosenbaum (2011) provided randomization tests for rare but substantial effects, Rosenbaum (2001) constructed a randomization-based interval estimate for the attributable effect of a treatment on a binary outcome and Rosenbaum (1999) proposed a randomization-based procedure for non-negative and non-decreasing quantile treatment effects under the assumption of rank preservation; see section 2.4.4 of Rosenbaum (2010) for discussion of testing general null hypotheses of non-zero treatment effects. By contrast, we attempt to test for heterogeneity in an unstructured way, though the choice of test statistic is motivated by the problem at hand. As additional assumptions on the structure of the heterogeneity will increase statistical power, using these approaches may be more appropriate than our omnibus method when such assumptions are met.

There is one important complication that we do not directly address here: the case of discrete outcomes. Even though the FRT procedure still yields valid inference in this setting, the constant treatment effect hypothesis may no longer be of scientific interest. This is a fundamental issue and is not specific to any particular testing procedure. For example, consider a semicontinuous outcome distribution, with a large point mass at zero, such as in the Connecticut Jobs First evaluation, where roughly half the sample has no earnings (Bitler *et al*., 2006). Here, the constant effect null hypothesis implies that welfare reform has the same dollar impact regardless of whether the individual starts with zero earnings, which is nonsensical. In future work, we hope to explore different approaches for this setting, including latent variable formulations and principal stratification (Nolen and Hudgens, 2011).

In the end, our approach offers a flexible framework for assessing treatment effect variation in randomized experiments, allowing researchers to incorporate a broad range of test statistics and to accommodate complex experimental designs. Most of all, our goal is to give applied researchers a set of tools so that inference about treatment effect variation can becomes a standard step in the analysis of randomized experiments. Next steps are to explore the role of covariates in treatment effect variation and, in particular, the interplay between systematic and idiosyncratic treatment effect variation.

Acknowledgements

The authors thank Alberto Abadie, Marianne Bitler, Paul Rosenbaum, Don Rubin, Tyler VanderWeele and participants at the Atlantic Causal Inference Conference, the Joint Statistical Meetings and the Harvard–Massachusetts Institute of Technology econometrics workshop for helpful comments. We especially thank Sir David Cox for his insights and for bringing G. E. H. Reuter's lost proof to our attention. We also thank the Joint Editor and two reviewers for their very helpful feedback. The research that is reported here was partially funded under co-operative agreement #90YR0049/02 with the Agency for Children and Families of the US Department of Health and Human Services. The opinions expressed are those of the authors and do not represent these institutions.

References

Abadie, A. (2002) Bootstrap tests for distributional treatment effects in instrumental variable models. *J. Am. Statist. Ass.*, **97**, 284–292.

Babu, J. G. and Rao, C. R. (2004) Goodness-of-fit tests when parameters are estimated. *Sankhya* A, **66**, 63–74.

Berger, R. L. and Boos, D. D. (1994) P values maximized over a confidence set for the nuisance parameter. *J. Am. Statist. Ass.*, **89**, 1012–1016.

Berrington de González, A. and Cox, D. R. (2007) Interpretation of interaction: a review. *Ann. Appl. Statist.*, **1**, 371–385.

Bitler, M. P., Domina, T. and Hoynes, H. W. (2013) Experimental evidence on distributional effects of Head Start. *Working Paper*. University of California in Irvine, Irvine. (Available from `http://www.socsci.uci.edu/~mbitler/papers/bdh-hsis-paper.pdf`.)

Bitler, M. P., Gelbach, J. B. and Hoynes, H. W. (2006) What mean impacts miss: distributional effects of welfare reform experiments. *Am. Econ. Rev.*, **96**, 988–1012.

Bitler, M. P., Gelbach, J. B. and Hoynes, H. W. (2010) Can variation in subgroups' average treatment effects explain treatment effect heterogeneity?: evidence from a social experiment. *Working Paper*. University of California in Irvine, Irvine. (Available from `http://www.socsci.uci.edu/~mbitler/papers/bgh-subgroups-paper.pdf`.)

Bloom, H. S. and Weiland, C. (2015) Quantifying variation in Head Start effects on young children's cognitive and socio-emotional skills using data from the Head Start Impact Study. *Working Paper*. Manpower Demonstration Research Corporation, New York. (Available from `http://www.mdrc.org/sites/default/files/quantifying_variation_in_head_start.pdf`.)

Chernozhukov, V. and Fernández-Val, I. (2005) Subsampling inference on quantile regression processes. *Sankhya*, **67**, 253–276.

Cox, D. R. (1984) Interaction. *Int. Statist. Rev.*, **52**, 1–24.

Crump, R. K., Hotz, V. J., Imbens, G. W. and Mitnik, O. A. (2008) Nonparametric tests for treatment effect heterogeneity. *Rev. Econ. Statist.*, **90**, 389–405.

Ding, P. (2014) A paradox from randomization-based causal inference. *Working Paper*. Harvard University, Cambridge. (Available from `http://arxiv.org/abs/1402.0142`.)

Ding, P., Feller, A. and Miratrix, L. W. (2015) Decomposing treatment effect heterogeneity. *Technical Report*. Department of Statistics, Harvard University, Cambridge.

Djebbari, H. and Smith, J. (2008) Heterogeneous impacts in PROGRESA. *J. Econmetr.*, **145**, 64–80.

Doksum, K. A. and Sievers, G. L. (1976) Plotting with confidence: graphical comparisons of two populations. *Biometrika*, **63**, 421–434.

Durbin, J. (1973) *Distribution Theory for Tests Based on the Sample Distribution Function*. Philadelphia: Society for Industrial and Applied Mathematics.

Feller, A., Grindal, T., Miratrix, L. and Page, L. (2014) Compared to what?: variations in the impacts of Head Start by alternative care-type settings. *Working Paper*. Harvard University, Cambridge. (Available from `http://scholar.harvard.edu/files/feller/files/feller_grindal_miratrix_page_12_6_14.pdf`.)

Firpo, S. (2007) Efficient semiparametric estimation of quantile treatment effects. *Econometrica*, **75**, 259–276.

Firpo, S., Fortin, N. M. and Lemieux, T. (2009) Unconditional quantile regressions. *Econometrica*, **77**, 953–973.

Fisher, R. A. (1935) *The Design of Experiments*, 1st edn. Edinburgh: Oliver and Boyd.

Gelman, A. (2004) Treatment effects in before-after data. In *Applied Bayesian Modeling and Causal Inference from an Incomplete Data Perspective*, pp. 195–202. London: Wiley.

Heckman, J. J., Smith, J. and Clements, N. (1997) Making the most out of programme evaluations and social experiments: accounting for heterogeneity in programme impacts. *Rev. Econ. Stud.*, **64**, 487–535.

Holt, D. and Smith, T. M. F. (1979) Post stratification. *J. R. Statist. Soc.* A, **142**, 33–46.

Koenker, R. and Xiao, Z. (2002) Inference on the quantile regression process. *Econometrica*, **70**, 1583–1612.

Lin, W. (2013) Agnostic notes on regression adjustments to experimental data: reexamining Freedman's critique. *Ann. Appl. Statist.*, **7**, 295–318.
Linton, O., Maasoumi, E. and Whang, Y.-J. (2005) Consistent testing for stochastic dominance under general sampling schemes. *Rev. Econ. Stud.*, **72**, 735–765.
Miratrix, L. W., Sekhon, J. S. and Yu, B. (2013) Adjusting treatment effect estimates by post-stratification in randomized experiments. *J. R. Statist. Soc.* B, **75**, 369–396.
Morgan, K. L. and Rubin, D. B. (2012) Rerandomization to improve covariate balance in experiments. *Ann. Statist.*, **40**, 1263–1282.
Neyman, J. (1990) On the application of probability theory to agricultural experiments. *Statist. Sci.*, **5**, 465–472.
Nolen, T. L. and Hudgens, M. G. (2011) Randomization-based inference within principal strata. *J. Am. Statist. Ass.*, **106**, 581–593.
O'Muircheartaigh, C. and Hedges, L. V. (2014) Generalizing from unrepresentative experiments: a stratified propensity score approach. *Appl. Statist.*, **63**, 195–210.
Puma, M., Bell, S. H., Cook, R., Heid, C. and Shapiro, G. (2010) *Head Start Impact Study: Final Report*. Washington DC: Department of Health and Human Services, Administration for Children and Families.
Rosenbaum, P. R. (1999) Reduced sensitivity to hidden bias at upper quantiles in observational studies with dilated treatment effects. *Biometrics*, **55**, 560–564.
Rosenbaum, P. R. (2001) Effects attributable to treatment: inference in experiments and observational studies with a discrete pivot. *Biometrika*, **88**, 219–231.
Rosenbaum, P. R. (2002a) *Observational Studies*. New York: Springer.
Rosenbaum, P. R. (2002b) Covariance adjustment in randomized experiments and observational studies. *Statist. Sci.*, **17**, 286–327.
Rosenbaum, P. R. (2010) *Design of Observational Studies*. New York: Springer.
Rosenbaum, P. R. (2011) A new u-statistic with superior design sensitivity in matched observational studies. *Biometrics*, **67**, 1017–1027.
Rubin, D. B. (1974) Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.*, **66**, 688–701.
Rubin, D. B. (1980) Comment on "Randomization analysis of experimental data: the Fisher Randomization Test". *J. Am. Statist. Ass.*, **75**, 591–593.
Rubin, D. B. (2005) Causal inference using potential outcomes: design, modeling, decisions. *J. Am. Statist. Ass.*, **100**, 322–331.

Supporting information

Additional 'supporting information' may be found in the on-line version of this article:

'Supplementary materials for "Randomization inference for treatment effect variation"'.