Can Variation in Subgroups's Average Treatment Effects Explain Treatment Effect Heterogeneity? Evidence from a Social Experiment: A Comment*

EunYi Chung[†] Mauricio Olivares[‡]

University of Illinois at Urbana-Champaign

April 4, 2020

Abstract

A common practice in empirical research to evaluate heterogeneity in the treatment effect involves calculating mean impacts that are constant within subgroups but vary across them. In a recent paper, Bitler, Gelbach, and Hoynes (2017) propose a new approach to challenge the validity of this practice, and find evidence of poor explanatory performance when analyzing heterogeneity of Connecticut's Jobs First welfare reform. In this comment, we argue that while their conclusion is supported by the data, their proposed permutation test suffers from an estimated-parameter problem, raising issues that need to be addressed to make inference valid. We provide numerical and theoretical evidence to show the relevance of our comment as well as an alternative test if one wants to conduct asymptotically valid permutation inference. While we also find evidence against the conventional approach to heterogeneity, our empirical evaluation of Connecticut's Jobs First sheds additional insights for a wide range of subgroups.

Keywords: Permutation Test, Khmaladze transformation, Heterogeneous treatment effects, Connecticut's Jobs First.

JEL Classification: C12, C14, C46.

*We are grateful to MDRC for granting access to the experimental data used in this paper. We would also like to thank Marianne Bitler, Jonah Gelbach, and Hilary Hoynes for their useful comments and generous feedback that led to the improvement of the paper as well as for making their code for replication publicly available. Lastly, we thank Dan Bernhardt for helpful discussion and encouragement during this project. All errors are our own. Address correspondence to EunYi Chung (eunyi@illinois.edu).

[†] EunYi Chung is an Assistant Professor in the Department of Economics, University of Illinois at Urbana-Champaign. Email: eunyi@illinois.edu

[‡] Mauricio Olivares is a Ph.D. student in the Department of Economics, University of Illinois at Urbana-Champaign. Email: lvrsgnz2@illinois.edu

1 Introduction

Detecting variation in subgroup's average treatment effects is a popular way to determine treatment effect heterogeneity. In its simplest form, this approach characterizes the variation in the treatment effect by investigating mean impacts that (a) differ across subgroups—typically defined by demographic or pre-intervention characteristics—but (b) are the same for each individual within the group. If the average treatment effect differs significantly across subgroups, the variation provides evidence in favor of treatment effect heterogeneity.

This approach is widely used due to its simplicity—once the subgroups are defined, the researcher can proceed straightforwardly. However, in a recent paper Bitler, Gelbach, and Hoynes (2017)—henceforth, BGH—use data from a randomized experiment, Connecticut's Jobs First welfare reform, to challenge the adequacy of this conventional approach. BGH argue that if the conventional approach were correct, then one might shift the observations in the control group by adding these subgroup-specific treatment effects to the actual outcome. This simple transformation yields an auxiliary outcomes distribution that they call the *simulated outcomes under treatment*. If the standard approach is a good representation of the heterogeneity in the treatment effect, then *i)* the distribution of the simulated outcomes should be "close," in some sense, to the distribution of the actual observed outcomes under treatment, and *ii)* the quantile treatment effects (QTE) obtained from this simulated outcomes distribution should agree with QTE based on the actual observed outcomes distribution.

BGH show that the conventional approach does a poor job of capturing the heterogeneity in the treatment effect for Connecticut's Jobs First program. On the one hand, QTE point estimates based on the simulated outcomes distribution fail to replicate the patterns predicted by labor theory, both in sign and magnitude.[1] On the other hand, BGH's proposed permutation test rejects the null hypothesis of equality of distributions between the actual outcomes and the simulated outcomes across a wide range of subgroups.[2]

The main challenge for BGH's testing procedure is that, when constructing the simulated outcomes distribution, the treatment effect is unknown and therefore needs to be estimated. The error involved in the estimation of the treatment effects, however, leaves a mark in the asymptotic distribution of the test statistic—the statistic is no longer asymptotically pivotal, meaning its asymptotic distribution depends on the underlying data generating process (DGP).[3]

[1] In previous work on the same randomized experiment, however, BGH confirmed these patterns are consistent with the QTE estimates based on the actual observed outcomes distribution. See Bitler, Gelbach, and Hoynes (2006) for a more thorough discussion.

[2] Their permutation test is based on the Kolmogorov–Smirnov statistic which in their application is the supremum distance between the cumulative distribution functions of the simulated outcomes under treatment and the actual observed outcomes.

[3] The problem of nonstandard distributions for sup-norm tests, or procedures based on sup-norm functionals like the permutation tests discussed here, falls into the classical goodness-of-fit problem with estimated nuisance parameter, and is also commonly referred in the literature as *the Durbin problem* (Durbin, 1973).

The practical consequence of this dependency is to make it difficult, if not impossible, to obtain critical values. Permutation-based inference is not immune to the effects of estimated parameters—the permutation distribution is no longer able to mimic the true unconditional limiting distribution.[4] This means that the resulting permutation test fails to control the type 1 error, even in large samples. BGH argue that, while their approach is subject to the estimated-parameter problem, their proposed permutation test satisfies a condition that allows it to bypass this complication and that their test is asymptotically valid.

In this comment we reexamine BGH's proposed testing procedure, raising issues that need to be addressed to make inference valid. In particular, we present theoretical, empirical, and numerical evidence showing that their permutation test is generally unable to control the type 1 error and, consequently, the estimated-parameter problem cannot be considered strictly solved. To motivate our analysis, and to illustrate the relevance of our comment, we first conduct a simple Monte Carlo experiment with no subgroups. Our exercise emphasizes how the dependency on the underlying DGP may jeopardize the usefulness of a permutation test in their context—for some DGPs, BGH's permutation test may either under-reject or over-reject under the null hypothesis of equality of distributions, with the latter being more problematic.

To explain these findings, we rely on an equivalent description of the permutation test, namely the permutation distribution. Our results show that the permutation distribution under BGH's setup behaves asymptotically like the distribution of the statistic as if the treatment effect were known, that is, as the supremum of a Brownian bridge. Heuristically, this behavior arises because when shuffling the pooled data using the simulated outcomes and the actual outcomes, the construction of the permutation test treats the observations *as if* they were i.i.d. However, this is not the true sampling distribution of the statistic, which is now (the supremum of) a Brownian bridge plus a drift component that emerges as a result of the estimation of the treatment effect. This means that the permutation distribution under BGH's approach is mimicking a Brownian bridge as opposed to a Brownian bridge with a drift.

These results caution against BGH's approach to testing the adequacy of the conventional approach. We propose an alternative procedure to conduct permutation-based inference that guarantees type 1 error control in large samples. This permutation test, introduced in Chung and Olivares (2019), exploits the martingale transformation of the empirical process. In a nutshell, the martingale transformation clears the empirical process out from the estimated treatment effect by decomposing it into two parts—a martingale with a standard Brownian motion limiting behavior, and a second part that vanishes in large samples. This strategy delivers a statistic that does not depend on the underlying DGP. Therefore, the permutation distribution based on the martingale-transformed statistic behaves asymptotically like the true unconditional limit distribution under the null. Consequently, the permutation test based on the martingale-transformed statistic controls the limiting rejection probability under weak assumptions, all of which are satisfied in a wide variety of applications, including the present

[4] See Online Appendix, Section 1.3 for a formal definition of the permutation distribution.

one. The implementation of this test is straightforward using the companion `RATest` R package, freely available on CRAN.

We confirm many of BGH's results for Connecticut's Jobs First experiment. Most notably, we reject the null hypothesis of equality of distributions between the actual earnings and the simulated earnings across a wide range of subgroups. In contrast to BGH's analysis, our permutation test rejects the null hypothesis of equality of distributions for a variety of subgroups. For example, for subgroups defined by the interaction of levels of education and number of quarters before random assignment with positive earnings, BGH's test provides evidence in favor of the conventional approach at the 5% level. Meanwhile, our proposed permutation test provides evidence against such conventional approach.

2 Permutation Test with Estimated Parameters

To illustrate the estimated-parameter problem and to show how it can lead to flawed inferences in the context of permutation tests when ignored, we investigate the performance of BGH's permutation test in a simplified Monte Carlo study without subgroups.

BGH's strategy boils down to testing for equality of distributions of the simulated outcomes and the actual outcomes under treatment based on a random sample of N individuals, where m of them receive treatment and $n = M - m$ belong to the control group. Let $Y_i(1)$ and $Y_i(0)$ be the potential outcomes for the ith unit with and without treatment, respectively, and denote their corresponding cumulative distribution functions by $F_1(\cdot)$ and $F_0(\cdot)$.

Under the conventional approach, an estimate of the simulated outcome under treatment is readily available and given by $\hat{Y}_i^* = Y_i(0) + \hat{\delta}$, where $\hat{\delta}$ is the difference in sample means from both groups. Collect the data into $Z = \left(\hat{Y}_1^*, \ldots, \hat{Y}_n^*, Y_1(1), \ldots, Y_m(1)\right) = (Z_1, \ldots, Z_N)$.

A natural statistic is the two-sample Kolmogorov–Smirnov (2SKS) statistic:

$$K_{m,n,\hat{\delta}}(Z) = \sup_y \left| V_{m,n}(y, \hat{\delta}) \right| \ , \tag{1}$$

where

$$\begin{aligned} V_{m,n}(y, \hat{\delta}) &= \sqrt{\frac{mn}{N}} \left(\frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{\hat{Y}_i^* \leq y\}} - \frac{1}{m} \sum_{i=1}^{m} \mathbb{1}_{\{Y_i(1) \leq y\}} \right) \\ &= \sqrt{\frac{mn}{N}} \left(\hat{F}_1^*(y) - \hat{F}_1(y) \right) \ . \end{aligned} \tag{2}$$

Under the null hypothesis of equal distributions between simulated and actual outcomes under treatment, each permutation of the N values in Z is equally likely to lead back to the original observations. Thus, the construction of a permutation test based on (1)—like the one proposed by BGH—consists of evaluating the 2SKS statistic for all permutations of the pooled data Z. One then rejects the null hypothesis of equal distributions at α level if the 2SKS statistic based on the original sample exceeds the $1-\alpha$ quantile of the empirical distribution that follows from recalculating the 2SKS statistics, also known as permutation distribution.[5]

For our Monte Carlo exercise, we generate data consistent with the conventional approach, which in this simplified version reduces to imposing that $Y_i(0)$ and $Y_i(1)$ are a constant shift apart. Let $Y_i(0) = \varepsilon_i$, and $Y_i(1) = \delta + Y_i(0)$ for some δ. In each of the following specifications, ε_i are i.i.d. according to one of the following probability distributions: standard normal, lognormal, Student's t distribution with 5 degrees of freedom, and exponential. To illustrate the effect that the estimated parameter has on the permutation test, we include a permutation test in the infeasible case when δ is known so that we do not have the estimated-parameter problem. We call this case the classical 2SKS statistic.

Table 1 reports the rejection probabilities under the null hypothesis of equality of distributions and a constant unitary shift, $\delta = 1$. This simple Monte Carlo experiment illustrates how the estimated parameter affects the validity of the permutation test—BGH's test either under-rejects (normal and t cases) or severely fails to control the rejection probability (lognormal and exponential cases). It follows that BGH's proposed permutation test does not control the type 1 error in the presence of an estimated parameter in the sense that the rejection probability is far from the nominal level α. Table 1 also reveals, however, that when the treatment effect is known, the permutation test based on the classical 2SKS delivers rejection probabilities under the null hypothesis that are very close to the nominal level for all specifications and sample sizes that we consider.[6]

This numerical example underscores the potential severity of the issues caused by the estimation step needed for the construction of the simulated outcomes distribution, offsetting the validity for permutation-based inference under BGH's paradigm. In the following section, we elaborate on the theoretical properties of the permutation test considered by BGH, shedding further light on these findings.

2.1 Discussion

Whether a permutation test is valid for inference in the present context depends on the asymptotic properties of the permutation distribution.[7] More specifically, we require the permutation distribution to behave, asymptotically, like the true unconditional limiting distribution of the

[5] See the online Appendix, Section 1.3, for a more detailed and formal construction of the permutation test based on (1).

[6] For additional results under a similar Monte Carlo design, including Fisher's randomization (Ding, Feller, and Miratrix, 2015), bootstrap, and subsampling tests (Chernozhukov and Fernández-Val, 2005), and the martingale-based permutation test, see Chung and Olivares (2019, Section 5).

[7] For a more technical discussion, see Sections 2.3, and Theorems A.1 and A.2 in Chung and Olivares (2019).

Table 1: Size of $\alpha = 0.05$ tests H_0 : Equality of simulated and actual outcomes distributions.

		Distributions			
N	Method	Normal	Lognormal	Student's t	Exponential
$N = 13$	Classical 2SKS	0.0494	0.0482	0.0522	0.0522
$n = 8,\ m = 5$	BGH	0.0000	0.0298	0.0002	0.0014
$N = 50$	Classical 2SKS	0.0528	0.0506	0.0460	0.0460
$n = 30,\ m = 20$	BGH	0.0002	0.3116	0.0014	0.0764
$N = 80$	Classic 2SKS	0.0452	0.0516	0.0510	0.0510
$n = 50,\ m = 30$	BGH	0.0000	0.3244	0.0016	0.0872
$N = 200$	Classic 2SKS	0.0472	0.0548	0.0486	0.0486
$n = 120,\ m = 80$	BGH	0.0004	0.3912	0.0032	0.1532

Rejection probabilities are computed using 5000 replications across Monte Carlo Experiments. We set $\delta = 1$ in all specifications.

2SKS statistic. We show that this is *not* the case when δ is unknown, and that the permutation test considered in BGH's can under- or overreject under the null hypothesis.

We begin the discussion by considering the classical case (δ is *known*) as a stepping stone to the more challenging case with estimated $\hat{\delta}$. In the classical case when δ is known, the 2SKS statistic weakly converges to (the supremum of) an F_0-Brownian bridge (Van der Vaart, 2000, Theorem 19.3 and Corollary 19.21).[8] Moreover, when δ is known, the permutation distribution behaves asymptotically like the true unconditional limiting distribution of the 2SKS statistic. This means that the permutation test has rejection probability equal to the nominal level α in large samples under the sharp null hypothesis (Chung and Olivares, 2019, Theorem A.2).[9]

What happens to the permutation test if we replace δ by the estimate $\hat{\delta}$, as in BGH's proposed permutation test? It turns out that the permutation distribution based on (1) behaves asymptotically like the limiting distribution of the 2SKS statistic as if δ were known, i.e., like an F_0-Brownian bridge (Chung and Olivares, 2019, Theorem 2). Intuitively, this is so because in both cases, the permutation distribution is treating the observations as if they were i.i.d. Specifically, the simulated outcomes are generated the same way—by adding δ, or its estimate, to the permutation-generated control group.

BGH reach this conclusion in a slightly different way, relying on the verification of technical conditions due to Præstgaard (1995, eq (2.14)-(2.15)). In a nutshell, the theory in Præstgaard (1995) allows us to determine the weak convergence of the 2SKS statistic to an F_0-Brownian

[8] The relevance of this result is that the limit distribution is the same for every continuous distribution function F_0, implying there is no dependency on the underlying DGP.

[9] As a matter of fact, when δ is known the permutation test is exact in finite samples under the sharp null (Chung and Olivares, 2019, see Section 2.2).

bridge under the null hypothesis when the supremum distance is taken over a more general class of indexing functions, possibly dependent on the sample size (Præstgaard, 1995, Corollary 1).[10] BGH argue that the testing problem at hand falls into this class, implying that the permutation distribution based on (1) asymptotically behaves, under the null hypothesis, like the limit distribution of the 2SKS statistic *as if δ were known*. Therefore, the permutation-based critical values for the 2SKS statistic are asymptotically valid.

However, as previously noted, the asymptotic behavior of the 2SKS statistic (1) differs from the classical case in important ways. Its asymptotic distribution is not (the supremum of) a Brownian bridge but also contains a non-negligible "drift," due to $\hat{\delta}$. Loosely speaking, if a smoothness condition about F_0 and its density f_0 is satisfied (Chung and Olivares, 2019, Assumption A.2), then

$$V_{m,n}(y,\hat{\delta}) \approx \underbrace{V_{m,n}(y,\delta)}_{\text{as if } \delta \text{ is known}} + \underbrace{\sqrt{\frac{mn}{N}}\left(f_0(y)(\hat{\delta}-\delta)\right)}_{\text{drift}} .$$

This decomposition is key to understanding the asymptotic properties of the 2SKS statistic with estimated parameters. Observe that the first summand corresponds with the classical case, whose limit distribution is the Brownian bridge. However, the introduction of the drift implies that the 2SKS statistic converges in distribution to (the supremum of) a certain Gaussian process whose limit distribution may depend on the model or the estimator (see Chung and Olivares (2019, Theorem 1) or Ding, Feller, and Miratrix (2015, Theorem 4)).

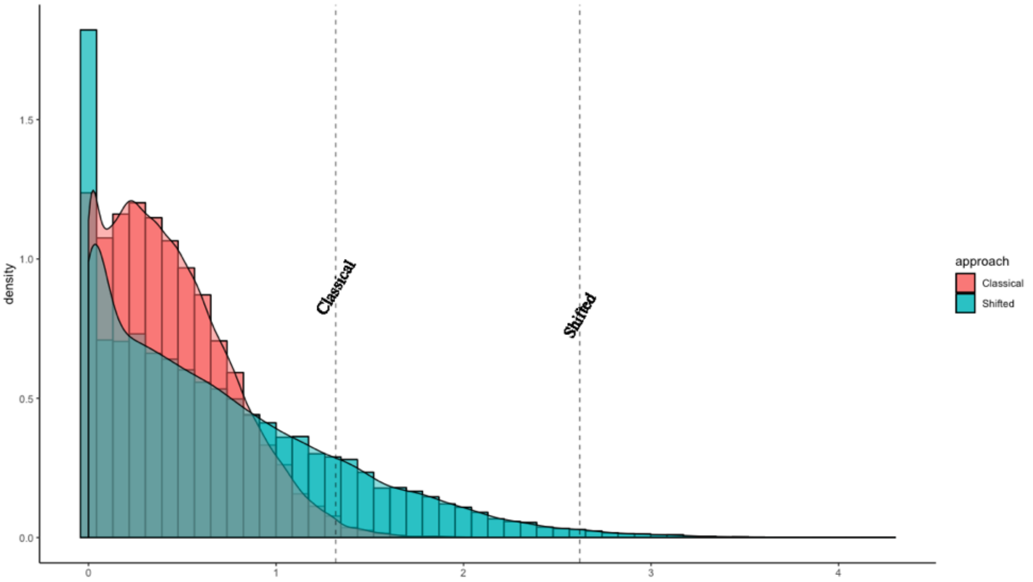
This discrepancy between the permutation distribution and the true unconditional limiting sampling distribution illustrates that the result from Præstgaard (1995) is not appropriate for this problem,[11] offsetting the validity of BGH's procedure—under their setup, the limiting rejection probability of their testing procedure tends to a value that differs from the nominal level α. This means that one may have underrejection or overrejection under H_0, with the latter being more problematic. Figure 1 plots the sampling distributions of the 2SKS, as well as their $1-\alpha$ quantiles, when δ is known (classical) and unknown (shifted) based on numerical simulation (see the online Appendix, Section 2, for a formal description of the covariance structures). This difference in the quantiles of the corresponding sampling distributions highlights the gravity of the estimated-parameter problem.

In summary, BGH's approach to inference should not be viewed as a formal device to validate the adequacy of the conventional approach, but rather as a heuristic tool to motivate the use of permutation-based inference. However, there also exists a different permutation test for this testing problem that is asymptotically valid in the presence of estimated nuisance parame-

[10] See also Van der Vaart and Wellner (1996, Theorems 3.7.1 and 3.7.2) for further details.

[11] While the case when δ is known is interesting from a theoretical point of view—it yields finite- and large-sample validity—it is often dismissed as implausible for we rarely know what δ is in practice. See Caughey, Dafoe, and Miratrix (2017) for an in-depth discussion about this.

Figure 1: Simulated sampling distributions of the 2SKS statistic



Histograms calculated after simulating 20,000 sample paths, each one of them formed based on 1000 observations. Covariance structures calculated assuming F_0 follows the (standard) lognormal distribution, and $\delta = 1$ in the classical case. Vertical lines represent the 99th quantiles of the respective sampling distributions.

ters. This alternative permutation test, introduced by Chung and Olivares (2019), exploits the martingale transformation of the empirical process introduced by Khmaladze (1981).[12]

The idea behind the martingale transformation is to modify (2) so that the limit distribution of statistics based on it, like the Kolmogorov–Smirnov statistic, does not depend on the underlying DGP. Roughly speaking, this transformation clears (2) out from the effect of the estimated parameters by decomposing it into two parts—a martingale with a standard Brownian motion limiting behavior, and a second part that becomes negligible in large samples. The relevance of this decomposition is twofold. On the one hand, the limiting distribution of a statistic based on it does *not* depend on the underlying DGP. On the second hand, and more importantly, the permutation distribution based on this statistic behaves like the true unconditional limiting distribution in large samples (Chung and Olivares, 2019, Theorem 4). Hence, a permutation test based on this martingale-transformed statistic yields a permutation test whose limiting rejection probability equals the nominal level α under the null hypothesis. The implementation

[12] A closely related but different permutation test is due to Ding, Feller, and Miratrix (2015). While their permutation test can also be applied to the current testing problem, we stick to the martingale-transformed permutation test of Chung and Olivares (2019) since the former test is more conservative. See the discussion in Chung and Olivares (2019) for a more detailed comparison.

of the martingale-based permutation test is straightforward using the `RATest` `R` package.[13]

3 Revisiting Connecticut's Jobs First

We now return to the case of multiple subgroups defined by covariates. We apply our martingale-based permutation test to the empirical analysis of the Connecticut's Jobs First welfare reform, and consider exactly the same sample, subgroups, and multiple testing Bonferroni adjustment as BGH. We stick to this multiple testing adjustment to help us compare our results with BGH's procedure without prejudicing against the latter.[14]

Table 2 shows the results from BGH as well as the martingale based permutation test for the joint null hypothesis, for several sets of subgroups.[15] As in BGH, each row represents the results for a particular set of subgroups, and the total number of subgroups equals the number of tests contained in Column 2. Columns 3-4 are the empirical results directly from BGH's Table 2, which we include as they are for a fair comparison. Columns 5-6 correspond to the martingale based permutation test.

We confirm many of BGH's results for the Connecticut's Jobs First—we reject the null hypothesis of equality of distributions between the actual earnings and the simulated earnings across a wide range of subgroups. In other words, the conventional approach does a poor job of explaining the heterogeneity in the treatment effect of the welfare reform on earnings.[16] However, we find that compared to BGH's permutation test, our method rejects the null hypothesis of equality of distributions for a variety of subgroups. For example, for subgroups defined by the number of pre-random assignment quarters with positive earnings, BGH's permutation test fails to reject the null hypothesis that the simulated earnings distribution and the actual earnings distribution under treatment are equal at 5% level. We reach the same conclusion if we interact this with education levels, welfare receipt seven quarters pre-random assignment, or earnings level in the seventh quarter pre-random assignment. This is not the case when applying the martingale based permutation test. That is, in all the aforementioned

[13] See the CRAN repository for documentation. For theoretical and computational aspects with regards the martingale transformation, see Chung and Olivares (2019, Section), and references therein.

[14] We have also implemented our test with alternative, asymptotically more efficient multiple testing adjustments considered in Chung and Olivares (2019, Section 4), leading to similar conclusions to those we show here. The formal statement of the testing procedure considered in BGH, as well as the multiple testing procedures are relegated to the online Appendix, Section 1.

[15] See BGH's Section V.A. We have also considered the same exercise using Ding, Feller, and Miratrix (2015), reaching similar results as the ones reported here. See the online Appendix, Section 3.

[16] This conclusion–the inadequacy of the conventional approach at explaining heterogeneity in the treatment effect– follows if one conducts inference using different methods too. For example, BGH argue that this is the case if one performs subsampling-based inference as in Chernozhukov and Fernández-Val (2005) or bootstrap-based inference, as in Linton, Maasoumi, and Whang (2005), adjusting for multiple testing as above in both cases.

Table 2: Testing for Heterogeneity in the Treatment Effect by Subgroups, Time-varying mean treatment effects by subgroup with participation adjustment

		BGH's Permutation Test		Martingale-based Permutation Test	
Subgroup	Number of Tests	Number of Reject at 10%	Number of Reject at 5%	Number of Reject at 10%	Number of Reject at 5%
Full Sample	7	4	4	7	7
Education	21	3	1	9	9
Age of youngest child	21	3	1	11	10
Marital status	21	2	1	14	14
Earnings level seventh Q pre-RA	21	2	1	17	16
Number of pre-RA Q with earnings	21	1	0	17	16
Welfare receipt seventh Q pre-RA	14	3	3	14	14
Education subgroups interacted with					
Age of youngest child	49	1	0	14	14
Marital status	35	3	3	18	17
Earnings level seventh Q pre-RA	63	1	0	15	14
Number of pre-RA Q with earnings	63	0	0	13	11
Welfare receipt seventh Q pre-RA	42	1	0	15	14
Age of youngest child interacted with					
Marital status	35	1	1	17	15
Earnings level seventh Q pre-RA	63	0	0	17	14
Number of pre-RA Q with earnings	49	1	1	14	12
Welfare receipt seventh Q pre-RA	42	1	0	14	13
Marital status subgroup interacted with					
Earnings level seventh Q pre-RA	63	2	1	14	11
Number of pre-RA Q with earnings	63	0	0	15	11
Welfare receipt seventh Q pre-RA	42	1	0	14	13
Earnings level seventh Q pre-RA subgroups interacted with					
Number of pre-RA Q with earnings	49	0	0	16	15
Welfare receipt seventh Q pre-RA	42	1	1	17	15
Number of quarters any earnings pre-RA subgroup interacted with					
Welfare receipt seventh Q pre-RA	42	0	0	14	13

All reported results account for multiple testing using Bonferroni adjustment. The martingale based permutation test was calculated based on 1000 permutations.

cases, BGH's provides evidence in favor of the traditional approach at the 5% level, whereas the martingale-based permutation test rejects in all cases. The same is true if we consider the interaction between age of youngest child with earnings level in the seventh quarter before random assignment—BGH's test provides evidence in favor of the conventional approach at the 10% level, whereas our test provides evidence against the conventional approach.

References

Bitler, M. P., Gelbach, J. B., and Hoynes, H. W. (2006). What mean impacts miss: Distributional effects of welfare reform experiments. *American Economic Review*, 96(4):988–1012.

Bitler, M. P., Gelbach, J. B., and Hoynes, H. W. (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.

Caughey, D., Dafoe, A., and Miratrix, L. (2017). Beyond the sharp null: Randomization inference, bounded null hypotheses, and confidence intervals for maximum effects. *arXiv preprint arXiv:1709.07339*.

Chernozhukov, V. and Fernández-Val, I. (2005). Subsampling inference on quantile regression processes. *Sankhyā: The Indian Journal of Statistics*, pages 253–276.

Chung, E. and Olivares, M. (2019). Permutation test for heterogeneous treatment effects with a nuisance parameter. *Working Paper*, pages 1–50. Latest version.

Ding, P., Feller, A., and Miratrix, L. (2015). Randomization inference for treatment effect variation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*.

Durbin, J. (1973). Weak convergence of the sample distribution function when parameters are estimated. *The Annals of Statistics*, pages 279–290.

Khmaladze, E. V. (1981). Martingale approach in the theory of goodness-of-fit tests. *Theory of Probability & Its Applications*, 26(2):240–257.

Linton, O., Maasoumi, E., and Whang, Y.-J. (2005). Consistent testing for stochastic dominance under general sampling schemes. *The Review of Economic Studies*, 72(3):735–765.

Præstgaard, J. T. (1995). Permutation and bootstrap kolmogorov-smirnov tests for the equality of two distributions. *Scandinavian Journal of Statistics*, pages 305–322.

Van der Vaart, A. W. (2000). *Asymptotic statistics*, volume 3. Cambridge university press.

Van der Vaart, A. W. and Wellner, J. (1996). *Weak convergence and empirical processes: with applications to statistics*. Springer Science & Business Media.