Electronic Journal of Statistics
Vol. 18 (2024) 465–489
ISSN: 1935-7524
https://doi.org/10.1214/23-EJS2209

Wilcoxon-Mann-Whitney statistics in randomized trials with non-compliance

Lu Mao

*Department of Biostatistics and Medical Informatics, 207A WARF Office Building,
610 Walnut St., University of Wisconsin-Madison, Madison, Wisconsin 53726, U.S.A.*
e-mail: lmao@biostat.wisc.edu

Abstract: The Mann–Whitney-type stochastic shift $\mathbb{P}(Y > X)$ has long been used as a scale-free alternative to the mean difference in measuring the distance between two populations. It has recently been recast as a causal estimand, but only in standard settings where confounders are fully captured. We study the Mann–Whitney treatment effect (MWTE) in randomized trials with non-ignorable non-compliance, where the treatment received is confounded by unknown factors. First, we define and estimate a local MWTE on the compliers via the standard principal-stratification approach with randomization status as an instrumental variable (IV). Then, we derive sensitivity bounds for the local effect estimand when key IV assumptions such as exclusion restriction and monotonicity are violated. Finally, we study the asymptotic operating characteristics of the local MWTE estimator in testing the treatment effect. Analytic bounds on the asymptotic relative efficiencies show that this IV-based test is likely superior to standard intent-to-treat tests under location-shift alternatives. The proposed methodology is applied to the famous National Job Training Partnership Act Study as an illustration.

MSC2020 subject classifications: Primary 62D20; secondary 62G20.
Keywords and phrases: Instrumental variable, asymptotic relative efficiency, causal estimand, exclusion restriction, rank tests, intent-to-treat analysis.

Received April 2022.

1. Introduction

The past three quarters of a century has seen an enduring popularity of the rank-based Wilcoxon–Mann–Whitney (WMW) test, initially proposed by Wilcoxon (1945) with its statistical properties explained later by Mann and Whitney (1947). Compared with the standard t-test, the WMW test makes no distributional assumptions, is robust against outliers, and may even gain efficiency under heavy-tailed distributions. Besides testing, the statistic has also been used to quantify treatment effect as measured by the stochastic shift $\mathbb{P}(Y_1 > Y_0)$, where Y_k is a (continuous) outcome from group k $(k = 1, 0)$ (Halperin, Gilbert and Lachin, 1987; McGraw and Wong, 1992; Cliff, 1993; Grissom, 1994; Hauck, Hyslop and Anderson, 2000; Kraemer and Kupfer, 2006; Acion et al., 2006; Brumback, Pepe and Alonzo, 2006; Newcombe, 2006a,b; Zhou, 2008; Kieser, Friede and Gondan, 2013). We refer to this parameter as the Mann–Whitney treatment effect (MWTE), as it is the estimand of Mann and Whitney (1947)'s version of the statistic.

The causal meaning of the MWTE is made precise in a series of recent papers that recast it under the counterfactual framework of causal inference (Rubin, 1978). Let $Y(k)$ denote the potential outcome under treatment arm k, where $k = 1$ indicates the active treatment and $k = 0$ indicates the control. Then, the causal MWTE is defined by

$$\tau_g = \mathbb{P}\big\{Y_i(1) \geq Y_j(0)\big\}, \tag{1}$$

where i and j index independent units. It is oftentimes transformed into the Mann–Whitney odds $\tau_g/(1 - \tau_g)$, i.e., the odds of having a greater outcome under the treatment against the control, for better interpretation. As a rank-based estimand, the MWTE is unaffected by the scale of the outcome as it is invariant under monotone transformation. As a result, it is more robust against outliers than is the traditional average treatment effect (ATE).

Studies on the causal MWTE have only just begun to emerge. For randomized trials, Fay et al. (2018) investigated the properties of τ_g and its relationship with the unidentifiable, subject-level $\mathbb{P}\{Y(1) \geq Y(0)\}$. For non-randomized studies with measured confounders, Wu et al. (2014) used a functional response model to estimate τ_g. Viewing the estimation of τ_g as a semiparametric problem, Mao (2018) studied both the inverse probability weighted and doubly robust methods. More recently, Zhang et al. (2019) extended the MWTE to censored outcomes and proposed various inferential strategies. To our knowledge, inference on the MWTE with unknown confounding has not yet been considered in the literature.

A common case of unknown confounding is non-compliance in randomized controlled trials (RCTs). Indeed, participants in an RCT, whether in a sociological or medical setting, often violate their randomization status and choose their own treatment. A standard approach to addressing the self-selection bias is to follow the intent-to-treat (ITT) principle, by analyzing data according to the randomization status rather than the treatment received. The ITT analysis produces valid estimates of the causal effect of treatment *assignment* (or policy), as well as unbiased tests of the sharp null hypothesis on the treatment effect itself (Robins and Greenland, 1996). To quantify the latter, a complementary approach is to use the randomization status as an "instrumental variable" (IV) to the treatment received. The IV (or Wald) estimator, studied in detailed in the seminal paper by Angrist, Imbens and Rubin (1996), provides a valid estimate of the local ATE on the compliers, i.e., those who would comply with the assignment no matter which group they are assigned to.

To accommodate non-compliance in WMW analysis, the ITT approach would be the easiest, but not necessarily the best. Like the case with ATE, it would only allow us to estimate the stochastic shift caused by randomization, not by the treatment. Even in testing, the ITT may be suboptimal as it places equal weights on compliers and non-compliers, who may show different treatment effects. For better interpretation and possible improvement in efficiency, there is an interest in an alternate IV approach to WMW analysis, one that emulates Angrist, Imbens and Rubin (1996)'s approach to the ATE.

We aim to develop this approach. In Section 2, we define the local MWTE, re-express it as a contrast of marginal distributions on the compliers, and construct a simple nonparametric plug-in estimator using the standard results of Imbens and Rubin (1997). We further develop estimable sensitivity bounds for the local MWTE when randomization has a direct effect on the outcome or when defiers exist in the population, both violating key assumptions needed for point identification. We also explore the identification of the MWTE on the overall population, with or without extra assumptions. In Section 3, the asymptotic power of the IV-based WMW test is derived explicitly under additive treatment effects and compared with the ITT-based WMW test and t-test. It is shown that the IV-based WMW test tends to outperform the standard tests in realistic scenarios. The empirical performance of the estimation and testing procedures is evaluated by simulations in Section 4. Section 5 demonstrates our methods on real data from the National Job Training Partnership Act Study. Concluding remarks in Section 6 summarize the present work and discuss future research topics.

2. Estimation and sensitivity analysis

2.1. Definition and estimation

Let $Z = 1, 0$ denote the randomized treatment assignment. Use $A(z)$ to denote the potential treatment under assignment z ($z = 1, 0$). Under the Stable Unit Treatment Value Assumption (SUTVA) (Rubin, 1978), the observed treatment and outcome are $A = ZA(1)+(1-Z)A(0)$ and $Y = AY(1)+(1-A)Y(0)$, respectively. Throughout, we assume that the distributions of the $Y(k)$ are absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}$.

According to the compliance status of each subject, the population is divided into four sub-populations, or four principal strata (Frangakis and Rubin, 2002). These are compliers, with $A(z) = z$; always-takers, with $A(z) = 1$; never-takers, with $A(z) = 0$; and defiers, with $A(z) = 1-z$ ($z = 1, 0$). Unless noted otherwise (such as in Sections 2.2 and 2.3), we proceed under the following standard assumptions (Angrist, Imbens and Rubin, 1996).

(A1) (Exclusion Restriction) Treatment assignment has a causal effect on the outcome only through the treatment received. In other words, if $Y(z, k)$ denotes the potential outcome under assigned treatment z and received treatment k ($z = 1, 0; k = 1, 0$), then $Y(z, k) = Y(k)$ almost surely.

(A2) (Randomization) $\{Y(1), Y(0), A(1), A(0)\} \perp\!\!\!\perp Z$.

(A3) (Relevance of Instrument) Treatment assignment has a non-trivial effect on the treatment received, i.e., $\mathbb{P}\{A(1) = 1\} \neq \mathbb{P}\{A(0) = 1\}$.

(A4) (Monotonicity) There are no defiers, i.e., $A(1) \geq A(0)$ almost surely.

Assumption (A4) leaves us with only three compliance classes. Use $\mathcal{C} = 0, 1$, or 2 to indicate the always-taker, complier, or never-taker, respectively. Then, we can define the local MWTE by

$$\tau_{\mathrm{c}} = \mathbb{P}\big\{Y_i(1) \geq Y_j(0) \mid \mathcal{C}_i = \mathcal{C}_j = 1\big\}, \tag{2}$$

where i and j index independent units. Similarly to τ_g defined in (1), τ_{c} quantifies the treatment effect by stochastic shift, only restricted to the sub-population of compliers. That is, τ_{c} is the probability of a complier under treatment having a greater outcome than another under control.

It is clear from (2) that τ_{c} is a functional only of the marginal (potential) outcome distributions on the compliers. Let $\nu_{k\mathrm{c}}$ denote the cumulative distribution function of $Y(k)$ given $\mathcal{C} = 1$, i.e., $\nu_{k\mathrm{c}} = [Y(k) \mid \mathcal{C} = 1]$ $(k = 1, 0)$. Here and in the sequel, we use $[X \mid C]$ to denote the conditional distribution of a random variable X given an event C. Then, the local MWTE can be expressed as $\tau_{\mathrm{c}} = \int \nu_{0\mathrm{c}}(y)\nu_{1\mathrm{c}}(\mathrm{d}y)$.

The observed data consist of a random n-sample (Z_i, A_i, Y_i) $(i = 1, \ldots, n)$ of (Z, A, Y). Based on the observed data, we can construct a nonparametric plug-in estimator for τ_{c} by estimating the $\nu_{k\mathrm{c}}$ along the lines of Imbens and Rubin (1997). The idea is intuitively illustrated by the diagram in Fig. 1. Clearly, both randomized groups are mixtures of the principal strata in the form of

$$\begin{aligned} \omega_1(\cdot) &= p_{\mathrm{c}}\nu_{1\mathrm{c}}(\cdot) + p_{\mathrm{a}}\nu_{1\mathrm{a}}(\cdot) + p_{\mathrm{n}}\nu_{0\mathrm{n}}(\cdot), \\ \omega_0(\cdot) &= p_{\mathrm{c}}\nu_{0\mathrm{c}}(\cdot) + p_{\mathrm{a}}\nu_{1\mathrm{a}}(\cdot) + p_{\mathrm{n}}\nu_{0\mathrm{n}}(\cdot), \end{aligned} \tag{3}$$

where $p_{\mathrm{a}} = \mathbb{P}(\mathcal{C} = 0)$, $p_{\mathrm{c}} = \mathbb{P}(\mathcal{C} = 1)$, $p_{\mathrm{n}} = \mathbb{P}(\mathcal{C} = 2)$, $\nu_{1\mathrm{a}} = [Y(1) \mid \mathcal{C} = 0]$, and $\nu_{0\mathrm{n}} = [Y(0) \mid \mathcal{C} = 2]$. Because of the randomization assumption (A2), $(p_{\mathrm{a}}, \nu_{1\mathrm{a}})$ and $(p_{\mathrm{n}}, \nu_{0\mathrm{n}})$ can be identified and estimated empirically from the randomized control and treatment groups, respectively (and so can we estimate $p_{\mathrm{c}} = 1 - p_{\mathrm{a}} - p_{\mathrm{n}}$ given (A4)). Meanwhile, we can also identify and empirically estimate the distributions of the "per-protocol" groups $\mu_z = [Y \mid A = Z = z]$ $(z = 1, 0)$. Although the per-protocol groups consists of subjects with unobserved compliance status, we can nevertheless tease out the complier distributions from the mixtures. Indeed, since $(p_{\mathrm{a}} + p_{\mathrm{c}})\mu_1 = p_{\mathrm{a}}\nu_{1\mathrm{a}} + p_{\mathrm{c}}\nu_{1\mathrm{c}}$ and $(p_{\mathrm{n}} + p_{\mathrm{c}})\mu_0 = p_{\mathrm{n}}\nu_{0\mathrm{n}} + p_{\mathrm{c}}\nu_{0\mathrm{c}}$, we have that $\nu_{1\mathrm{c}}(\cdot) = (1 + \rho_{\mathrm{a}})\mu_1(\cdot) - \rho_{\mathrm{a}}\nu_{1\mathrm{a}}(\cdot)$ and that $\nu_{0\mathrm{c}}(\cdot) = (1 + \rho_{\mathrm{n}})\mu_0(\cdot) - \rho_{\mathrm{n}}\nu_{0\mathrm{n}}(\cdot)$, where $\rho_{\mathrm{a}} = p_{\mathrm{a}}/p_{\mathrm{c}}$ and $\rho_{\mathrm{n}} = p_{\mathrm{n}}/p_{\mathrm{c}}$. We can thus construct estimators $\widehat{\nu}_{k\mathrm{c}}$ for $\nu_{k\mathrm{c}}$ $(k = 1, 0)$ by replacing the unknown quantities with their empirical analogs. Then the local MWTE is estimated by the plug-in $\widehat{\tau}_{\mathrm{c}} = \mathcal{M}(\widehat{\nu}_{1\mathrm{c}}, \widehat{\nu}_{0\mathrm{c}})$, where $\mathcal{M}(\eta_1, \eta_2) = \int \eta_2(y)\eta_1(\mathrm{d}y)$. We call $\mathcal{M}$ the MW functional, which will play a bigger part in the sensitivity analysis. The asymptotic normality of $\widehat{\tau}_{\mathrm{c}}$ and its variance can be derived via the functional delta method. The details can be found in the Supplementary Materials.

2.2. Bounds under exclusion restriction

When the Exclusion Restriction (ER) assumption (A1) is violated, randomization itself has a direct effect on the outcome. In that case, the notation $Y(k)$ is no longer adequate to capture the variety of the potential outcomes. We

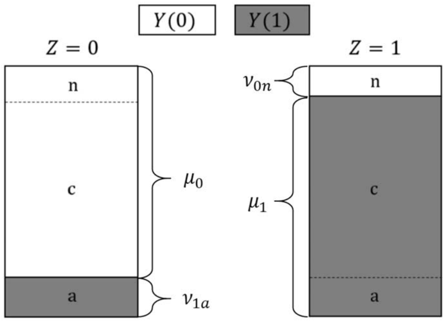

FIG 1. *A schematic illustration of the composition of the randomized groups by compliance class and potential outcome type (c: compliers; a: always takers; n: never takers). Dotted lines represent latent, unobservable boundaries.*

instead revert to the notation $Y(z,k)$ for $z,k \in \{1,0\}$ defined in the statement of (A1). Let $\nu_{zz,\mathrm{c}} = [Y(z,z) \mid \mathcal{C} = 1]$, $\nu_{z1,\mathrm{a}} = [Y(z,1) \mid \mathcal{C} = 0]$, and $\nu_{z0,\mathrm{n}} = [Y(z,0) \mid \mathcal{C} = 2]$ $(z = 1,0)$. Due to randomization direct effects, we now have that $\nu_{11,\mathrm{a}} \neq \nu_{01,\mathrm{a}}$ and $\nu_{10,\mathrm{n}} \neq \nu_{00,\mathrm{n}}$; See Figure S1 in the Supplementary Materials for a diagrammatic illustration similar to Fig. 1.

For the compliers, it is impossible, without unrealistic assumptions, to disentangle the treatment effect from that of randomization. We thus re-define the local MWTE in this case by $\tau_{\mathrm{c}}^{\mathrm{ER}} = \mathcal{M}(\nu_{11,\mathrm{c}}, \nu_{00,\mathrm{c}})$, which is the combined stochastic shift by the treatment and randomization. Let $\nu_{z\mathrm{c}}^{*\mathrm{ER}}$ denote the estimand (i.e., asymptotic limit) of $\widehat{\nu}_{z\mathrm{c}}$ $(z = 1,0)$ and let $\tau_{\mathrm{c}}^{*\mathrm{ER}}$ denote that of $\widehat{\tau}_{\mathrm{c}}$. We first quantify the biases of the $\nu_{z\mathrm{c}}^{*\mathrm{ER}}$ with regard to $\nu_{zz,\mathrm{c}}$, and then use them to bound the bias of $\tau_{\mathrm{c}}^{*\mathrm{ER}}$ with regard to $\tau_{\mathrm{c}}^{\mathrm{ER}}$. In order to precisely characterize the conditions for attainment of the bounds, we need the following definition on a rank order for measures on the real line.

Definition 2.1. *Let η_1 and η_2 be two Borel measures on $\mathbb{R}$. We say that η_1 is ranked lower than η_2, denoted $\eta_1 \lhd \eta_2$ or $\eta_2 \rhd \eta_1$, if the support of η_1 lies entirely to the left of that of η_2.*

When applied to probability measures, the relation $\eta_1 \lhd \eta_2$ can viewed as a strengthened stochastic order in the sense that $\mathbb{P}(X_1 < X_2) = 1$ with $X_k \sim \eta_k$ $(k = 1,2)$. In fact, $\eta_1 \lhd \eta_2$ is equivalent to $\mathcal{M}(\eta_1, \eta_2) = 0$ or $\mathcal{M}(\eta_2, \eta_1) = 1$.

Theorem 2.1 (Simple bounds for $\tau_{\mathrm{c}}^{\mathrm{ER}}$ without ER)**.** *Under (A2)–(A4), we have that*

$$\nu_{1\mathrm{c}}^{*\mathrm{ER}} = \nu_{11,\mathrm{c}} + \rho_{\mathrm{a}}(\nu_{11,\mathrm{a}} - \nu_{01,\mathrm{a}}), \textit{ and } \nu_{0\mathrm{c}}^{*\mathrm{ER}} = \nu_{00,\mathrm{c}} + \rho_{\mathrm{n}}(\nu_{00,\mathrm{n}} - \nu_{10,\mathrm{n}}).$$

Therefore,

$$-(\rho_{\mathrm{a}} + \rho_{\mathrm{n}} + \rho_a\rho_n) \leq \tau_{\mathrm{c}}^{*\mathrm{ER}} - \tau_{\mathrm{c}}^{\mathrm{ER}} \leq \rho_{\mathrm{a}} + \rho_{\mathrm{n}} + \rho_a\rho_n. \tag{4}$$

The upper bound in (4) is attained if and only if $\nu_{01,\mathrm{a}} \lhd \nu_{00,\mathrm{c}} \lhd \nu_{11,\mathrm{a}}$, $\nu_{00,\mathrm{n}} \lhd \nu_{11,\mathrm{c}} \lhd$

$\nu_{10,\mathrm{n}}$*, and* $\nu_{01,\mathrm{a}} \lhd \nu_{00,\mathrm{n}} \lhd \nu_{11,\mathrm{a}} \lhd \nu_{10,\mathrm{n}}$. *The lower bound is attained if and only if* $\lhd$ *in the above rank orders is replaced by* $\rhd$.

By (4), the maximum bias of $\widehat{\tau}_{\mathrm{c}}$ increases with the non-compliance rates. Under fixed non-compliance rates, maximum bias is achieved when $\nu_{11,\mathrm{a}}$ is well separated from $\nu_{01,\mathrm{a}}$, and $\nu_{10,\mathrm{n}}$ from $\nu_{00,\mathrm{n}}$, as a result of randomization direct effects on the non-compliers. We can easily estimate the bounds in (4) and derive their asymptotic distributions, based on which confidence intervals for $\tau_{\mathrm{c}}^{\mathrm{ER}}$ can be constructed. See the Supplementary Materials for details.

Angrist, Imbens and Rubin (1996) established a similar result to (4) on the bias of the local ATE estimator in terms of the non-compliance rates and the average randomization direct effect on the non-compliers. In their case, however, it is not easy to construct bounds similar to (4) in terms of only the non-compliance rates except when the range of the outcome can be bounded (Horowitz and Manski, 2000; Blanco, Flores and Flores-Lagunes, 2013).

The bounding interval in (4) has a radius greater than the overall non-compliance rate and may thus be too wide to be useful in practice. For sharper bounds, we seek instead to characterize the identification region (Manski, 2003) for $\tau_{\mathrm{c}}^{\mathrm{ER}}$. To that end, we first consider the identification regions for the $\nu_{zz,\mathrm{c}}$ $(z = 1, 0)$. It is clear that the only constraint on $\nu_{zz,\mathrm{c}}$ is that it is mixed within μ_z as a component probability measure with a known (i.e., identifiable) mixing proportion. More formally,

$$\nu_{11,\mathrm{c}} \in \mathcal{V}\big\{\mu_1, (1+\rho_{\mathrm{a}})^{-1}\big\} \text{ and } \nu_{00,\mathrm{c}} \in \mathcal{V}\big\{\mu_0, (1+\rho_{\mathrm{n}})^{-1}\big\}, \tag{5}$$

where $\mathcal{V}(\mu, r) = \{\nu : \nu \text{ is mixed in } \mu \text{ with proportion } r\}$.

For two distribution functions η_1 and η_2, recall that η_2 is stochastically greater than η_1, denoted by $\eta_1 \preceq \eta_2$, if $\eta_1(y) \geq \eta_2(y)$ for all $y \in \mathbb{R}$. Equipped with the partial order $\preceq$, $\mathcal{V}(\mu, r)$ as a space of distributions always has a greatest and a least element. These are the upper and lower truncated distributions of μ cut off at its $(1-r)$th and rth quantiles, respectively.

Lemma 2.1. *Given* $\mathcal{V}(\mu, r)$*, let* $\overline{\mathcal{V}}(\mu, r)(\mathrm{d}y) = r^{-1}I\{y > \mu^{-1}(1-r)\}\mu(\mathrm{d}y)$ *and* $\underline{\mathcal{V}}(\mu, r)(\mathrm{d}y) = r^{-1}I\{y \leq \mu^{-1}(r)\}\mu(\mathrm{d}y)$*. Then, we have that* $\overline{\mathcal{V}}(\mu, r), \underline{\mathcal{V}}(\mu, r) \in \mathcal{V}(\mu, r)$ *and that* $\underline{\mathcal{V}}(\mu, r) \preceq \nu \preceq \overline{\mathcal{V}}(\mu, r)$ *for all* $\nu \in \mathcal{V}$.

The stochastic-order bounds on the identification regions (5) for the $\nu_{zz,\mathrm{c}}$ can then be utilized to characterize the identification region for $\tau_{\mathrm{c}}^{\mathrm{ER}}$. This is possible because the MW functional $\mathcal{M}$ is monotonic with respect to the partial order of its arguments in the sense that

$$\eta_1 \preceq \eta_2 \Rightarrow \mathcal{M}(\eta_1, \eta) \leq \mathcal{M}(\eta_2, \eta) \text{ and } \mathcal{M}(\eta, \eta_1) \geq \mathcal{M}(\eta, \eta_2)$$

for probability measures η, η_1, and η_2.

The following theorem summarizes the results on the identification region for $\tau_{\mathrm{c}}^{\mathrm{ER}}$. The fact that every point within the bounds in (2.2) is attainable can be justified through the bilinearity of $\mathcal{M}$ and the convexity of the identification regions for the $\nu_{zz,\mathrm{c}}$. The bounds themselves can be estimated by replacing the proportions and distribution functions with their sample analogs.

Theorem 2.2 (Identification region for $\tau_{\mathrm{c}}^{\mathrm{ER}}$ without ER)**.** *Let* $\overline{\nu}_{11,\mathrm{c}} = \overline{\mathcal{V}}\{\mu_1, (1+\rho_{\mathrm{a}})^{-1}\}$*,* $\underline{\nu}_{11,\mathrm{c}} = \underline{\mathcal{V}}\{\mu_1, (1+\rho_{\mathrm{a}})^{-1}\}$*,* $\overline{\nu}_{00,\mathrm{c}} = \overline{\mathcal{V}}\{\mu_0, (1+\rho_{\mathrm{n}})^{-1}\}$*, and* $\underline{\nu}_{00,\mathrm{c}} = \underline{\mathcal{V}}\{\mu_0, (1+\rho_{\mathrm{n}})^{-1}\}$*. Then, the identification region for* $\tau_{\mathrm{c}}^{\mathrm{ER}}$ *is*

$$\mathcal{M}(\underline{\nu}_{11,\mathrm{c}}, \overline{\nu}_{00,\mathrm{c}}) \leq \tau_{\mathrm{c}}^{\mathrm{ER}} \leq \mathcal{M}(\overline{\nu}_{11,\mathrm{c}}, \underline{\nu}_{00,\mathrm{c}}).$$

More generally, the identification regions for the $\nu_{zz,\mathrm{c}}$ in (5) and Lemma 2.1 shed light on a testable implication of the ER assumption.

Proposition 2.1 (A testable implication of ER)**.** *Under (A2)–(A4), a testable implication of (A1) is*

$$\underline{\nu}_{zz,\mathrm{c}} \preceq \nu_{z\mathrm{c}} \preceq \overline{\nu}_{zz,\mathrm{c}} \quad (z = 1, 0). \tag{6}$$

Given these stochastic orders, (A1) can be informally checked by plotting $\widehat{\nu}_{z\mathrm{c}}(y)$ against the estimated $\underline{\nu}_{zz,\mathrm{c}}(y)$ and $\overline{\nu}_{zz,\mathrm{c}}(y)$. Under a sound ER assumption, $\widehat{\nu}_{z\mathrm{c}}(y)$ should be uniformly sandwiched between its bounding functions. Otherwise, the assumption may be violated. Huber and Mellace (2015) took a similar approach to testing the ER, but they focused on the implied inequalities on the mean without exploiting the full strength of (6).

2.3. Bounds under violation of monotonicity

When the Monotonicity assumption (A4) is violated, there are defiers in the population. As a result, the observed non-compliers in the randomized groups are no longer pure samples of always-takers or never- takers but are rather contaminated with defiers (see Figure S2 in the Supplementary Materials for a schematic illustration). Use $\mathcal{C} = -1$ to indicate a defier and write $p_{\mathrm{d}} = \mathbb{P}(\mathcal{C} = -1)$. In this case, the true compliance class probabilities $p_{\mathrm{a}}, p_{\mathrm{n}}, p_{\mathrm{c}}$, and p_{d} are no longer fully identified. Nonetheless, they are constrained by the identifiable proportions $p_{\mathrm{a}}^* = \mathbb{P}(\mathcal{C} = -1, 0)$, $p_{\mathrm{n}}^* = \mathbb{P}(\mathcal{C} = -1, 2)$, and $p_{\mathrm{c}}^* = 1 - p_{\mathrm{a}}^* - p_{\mathrm{n}}^*$ (i.e., estimands of the estimators for $p_{\mathrm{a}}, p_{\mathrm{n}}$, and p_{c} in Section 2.1, respectively) through

$$p_{\mathrm{a}}^* = p_{\mathrm{a}} + p_{\mathrm{d}}, \quad p_{\mathrm{n}}^* = p_{\mathrm{n}} + p_{\mathrm{d}}, \quad \text{and } p_{\mathrm{c}}^* = p_{\mathrm{c}} - p_{\mathrm{d}}. \tag{7}$$

Let $\nu_{k\mathrm{c}}^{*\mathrm{D}}$ denote the estimand of $\widehat{\nu}_{k\mathrm{c}}$ $(k = 1, 0)$ and $\tau_{\mathrm{c}}^{*\mathrm{D}}$ the estimand of $\widehat{\tau}_{\mathrm{c}}$. Write $\nu_{k\mathrm{d}} = [Y(k) \mid \mathcal{C} = -1]$ $(k = 1, 0)$. The following theorem gives the biases of the $\widehat{\nu}_{k\mathrm{c}}$, based on which the bias of $\widehat{\tau}_{\mathrm{c}}$ can be bounded.

Theorem 2.3. *Under (A1)–(A3), we have that*

$$\nu_{k\mathrm{c}}^{*\mathrm{D}} = \nu_{k\mathrm{c}} + \lambda(\nu_{k\mathrm{c}} - \nu_{k\mathrm{d}}), \tag{8}$$

where $\lambda = p_{\mathrm{d}}/(p_{\mathrm{c}} - p_{\mathrm{d}})$*. Furthermore, if* $p_{\mathrm{c}} > p_{\mathrm{d}}$*, then*

$$(1+\lambda)^{-2}\big(\tau_{\mathrm{c}}^{*\mathrm{D}} - \lambda^2\big) \leq \tau_{\mathrm{c}} \leq (1+\lambda)^{-2}\big(\tau_{\mathrm{c}}^{*\mathrm{D}} + 2\lambda + 2\lambda^2\big). \tag{9}$$

The lower bound in (9) *is attained if and only if* $\nu_{1\mathrm{c}} \lhd \nu_{0\mathrm{d}} \lhd \nu_{1\mathrm{d}} \lhd \nu_{0\mathrm{c}}$*; the upper bound is attained if and only if the reverse order is true.*

The identity in (8) strengthens Proposition 3 of Angrist, Imbens and Rubin (1996), which expresses the bias of the local ATE in terms of λ in a similar form. According to (9), the length of the bounding interval tends to zero with $\tau_{\mathrm{c}}^{*\mathrm{D}} \to \tau_{\mathrm{c}}$ when $\lambda \downarrow 0$ (i.e., $p_{\mathrm{d}} \downarrow 0$). To obtain estimable bounds for τ_{c}, we can use the identifiable constraints in (7) to bound λ and exploit the monotonicity of the bounds in (9) as functions of λ. Write $\rho_{\mathrm{a}}^* = p_{\mathrm{a}}^*/p_{\mathrm{c}}^*$ and $\rho_{\mathrm{n}}^* = p_{\mathrm{n}}^*/p_{\mathrm{c}}^*$.

Corollary 2.1 (Simple estimable bounds for τ_{c} without Monotonicity)**.** *Suppose that $p_{\mathrm{c}} > p_{\mathrm{d}}$, we have that*

$$(1+\lambda_{\max})^{-2}\left(\tau_{\mathrm{c}}^{*\mathrm{D}} - \lambda_{\max}^2\right) \le \tau_{\mathrm{c}} \le (1+\lambda_{\max})^{-2}\left(\tau_{\mathrm{c}}^{*\mathrm{D}} + 2\lambda_{\max} + 2\lambda_{\max}^2\right), \quad (10)$$

where $\lambda_{\max} = \rho_{\mathrm{a}}^* \wedge \rho_{\mathrm{n}}^*$. The upper and lower bounds in (10) are attained if and only if $p_{\mathrm{a}}p_{\mathrm{n}} = 0$ and the rank-order conditions for the attainment of the corresponding bounds in Theorem 2.3 are satisfied.

Additional discussions about the global MWTE can be found in the Supplementary Materials.

3. Hypothesis testing

In this section, we study the operating characteristics of the hypothesis test based on $\widehat{\tau}_{\mathrm{c}}$ in comparison with standard testing procedures. Specifically, we consider the following tests on the treatment effect based on (a) the ITT or IV estimator for the ATE (A-ITT or A-IV, respectively, which turn out to be asymptotically equivalent); (b) the ITT estimator for the MWTE (WMW-ITT) (i.e., the standard WMW test under the ITT principle); and (c) our estimator $\widehat{\tau}_{\mathrm{c}}$ for the local MWTE (WMW-IV). Throughout the section, we assume that (A1)–(A4) hold.

3.1. Null and alternative hypotheses

The ITT tests are based on contrasts between ω_1 and ω_0, while the IV tests on those between $\nu_{1\mathrm{c}}$ and $\nu_{0\mathrm{c}}$. Under the sharp null hypothesis

$$H_0 : Y(1) = Y(0) \text{ almost surely},$$

we have that $\nu_{1\mathrm{c}} = \nu_{0\mathrm{c}}$ regardless of the mechanism underlying compliance. By (3), equality of the complier distributions also implies $\omega_1 = \omega_0$ for the ITT tests. As a result, all tests are valid under H_0. In fact, sharpness of the null is also in a sense necessary for their validity. This is because any scenario outside H_0 can lead to a non-zero effect size in the presence of non-compliance, whether it be ATE or MWTE. To see this, suppose that $\mathbb{P}\{Y(1) \neq Y(0)\} > 0$. Without loss of generality, assume that $\mathbb{P}\{Y(1) > Y(0)\} > 0$. If the compliers are precisely those for whom $Y(1) > Y(0)$, then we must have that $\nu_{0\mathrm{c}} \prec \nu_{1\mathrm{c}}$, that is, $\nu_{0\mathrm{c}} \preceq \nu_{1\mathrm{c}}$ but $\nu_{0\mathrm{c}} \neq \nu_{1\mathrm{c}}$. This yields a strictly positive non-centrality parameter, and thus incorrect type I error, for each of the four tests. By the same token, it is clear that each test is consistent against the alternative hypothesis

$$H_A : Y(1) > Y(0) \text{ almost surely.}$$

Instead of a generic subject in the general population, we could also formulate H_0 and H_A on a generic complier. This change of perspective, however, would have no substantive impact on our subsequent discussion, as treatment effects manifest themselves only on the compliers (those on the always- and never-takers are unobservable given their unmovable treatments; see Fig. 1). As a result, an overall effect affects the test behavior only through the implied effect on the compliers.

3.2. Pittman efficiency under location-shift models

To compare the power of the tests, we evaluate their asymptotic efficiency under a sequence of "contiguous" alternatives $H_{A,n} \subset H_A$ that approaches H_0 as n increases to infinity (see, e.g., Ch. 12 of van der Vaart, 1998). To recapitulate the idea of contiguity, suppose in a general context we are interested in testing $H_0 : \theta = 0$ against $H_A : \theta > 0$ for some parameter θ. Moreover, suppose that T_n is a regular estimator for some transformation $f(\theta)$ of θ such that $\sqrt{n}\{T_n - f(\theta)\}/\widehat{\sigma}$ converges weakly to the standard normal distribution under every θ, where $\widehat{\sigma}^2$ is a consistent variance estimator. Then, a two-sided asymptotic level-α $(0 < \alpha < 1)$ test can be constructed by rejecting H_0 if $|\sqrt{n}\{T_n - f(0)\}/\widehat{\sigma}| > z_{1-\alpha/2}$, where $z_{1-\alpha/2} = \Phi^{-1}(1-\alpha/2)$ and $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution. To evaluate the asymptotic power of the test, consider a sequence of continuous alternatives $H_{A,n} : \theta = n^{-1/2}h$ with $h > 0$. It can be shown that the power function under $H_{A,n}$ converges to $\Phi(\zeta h - z_{1-\alpha/2})$, where $\zeta = \sigma_0^{-1}|\dot{f}(0)|$ and σ_0^2 is the limit of $\widehat{\sigma}^2$ under H_0. Here and in the sequel, $\dot{g}(x) = \mathrm{d}g(x)/\mathrm{d}x$ for a generic function g. The quantity ζ^2, called the Pittman efficiency, is inversely proportional to the sample size needed to achieve a certain power under a fixed $\theta > 0$ and is an intrinsic characteristic of the test. As a result, the ratio of ζ^2 between two tests is commonly used to measure their asymptotic relative efficiency (ARE).

For A-ITT, A-IV, WMW-ITT, and WMW-IV, the test statistics are $T_n = \int y\{\widehat{\omega}_1(\mathrm{d}y) - \widehat{\omega}_0(\mathrm{d}y)\}, \widehat{p}_{\mathrm{c}}^{-1}\int y\{\widehat{\omega}_1(\mathrm{d}y) - \widehat{\omega}_0(\mathrm{d}y)\}, \mathcal{M}(\widehat{\omega}_1, \widehat{\omega}_0)$, and $\mathcal{M}(\widehat{\nu}_{1\mathrm{c}}, \widehat{\nu}_{0\mathrm{c}})$, respectively, where $\widehat{\omega}_k(y)$ $(k = 1, 0)$ and $\widehat{p}_{\mathrm{c}}$ are empirical estimates of $\omega_k(y)$ and p_{c}, respectively. Standard variance estimators are used for the $\widehat{\sigma}^2$. In particular, that for $\mathcal{M}(\widehat{\nu}_{1\mathrm{c}}, \widehat{\nu}_{0\mathrm{c}})$ is provided in Section S1.1 of the Supplementary Materials. Consider the contiguous alternatives under a location-shift treatment effect model

$$H_{A,n} : Y(1) = Y(0) + n^{-1/2}h, \;\; h > 0. \tag{11}$$

Meanwhile, assume that the joint distribution of $\{\mathcal{C}, Y(0)\}$ does not change with n. Under this set-up, denote $\nu = p_{\mathrm{a}}\nu_{\mathrm{a}} + p_{\mathrm{c}}\nu_{\mathrm{c}} + p_{\mathrm{n}}\nu_{\mathrm{n}}$, where ν is the distribution of $Y(0)$, $\nu_{\mathrm{a}} = [Y(0) \mid \mathcal{C} = 0]$, $\nu_{\mathrm{c}} = [Y(0) \mid \mathcal{C} = 1]$, and $\nu_{\mathrm{n}} = [Y(0) \mid \mathcal{C} = 2]$.

Lemma 3.1. *Under the contiguous alternatives $H_{A,n}$ in (11), the Pittman efficiencies of the four tests under consideration are*

$$\zeta^2_{\text{A-ITT}} = \zeta^2_{\text{A-IV}} = \frac{q(1-q)p_{\mathrm{c}}^2}{\mathrm{var}\{Y(0)\}},$$
$$\zeta^2_{\text{WMW-ITT}} = 12q(1-q)p_{\mathrm{c}}^2\left\{\int \dot\nu_{\mathrm{c}}(y)\dot\nu(y)\mathrm{d}y\right\}^2,$$
$$\zeta^2_{\text{WMW-IV}} = \frac{q(1-q)p_{\mathrm{c}}^2\{\int \dot\nu_{\mathrm{c}}(y)^2\mathrm{d}y\}^2}{\mathrm{var}[\nu_{\mathrm{c}}\{Y(0)\}]},$$

where $q = \mathbb{P}(Z=1)$, $\dot\nu(y) = \mathrm{d}\nu(y)/\mathrm{d}y$ and $\dot\nu_{\mathrm{c}}(y) = \mathrm{d}\nu_{\mathrm{c}}(y)/\mathrm{d}y$ are the densities of the overall and complier outcome distributions, respectively.

Remark 1. The three Pittman efficiencies share a common factor $q(1-q)$, derived from the treatment-control randomization ratio (maximized under 1:1). Between the three, WMW-IV will be more efficient if $\dot\nu_{\mathrm{c}}(y)$ has heavy tails, giving rise to a large integral of the squared density. The WMW-ITT test will benefit similarly from heavy-tailed distributions, but will also depend on the overlap between $\dot\nu_{\mathrm{c}}(y)$ and $\dot\nu(y)$. These qualitative arguments will be made more precise in Sections 3.3 and 3.4.

Remark 2. A major challenge in deriving the Pittman efficiency for WMW-IV lies in the asymptotic null variance of $\widehat\tau_{\mathrm{c}}$, which, unlike the standard WMW test statistic, is not distribution-free. In fact, uncertainty in $\widehat\tau_{\mathrm{c}}$ comes from both the random observations within each principal stratum and the random sizes of the strata (see Fig. 1). The two sources of variation are quantified and combined using weak convergence theory for empirical processes indexed by random sample sizes (see, e.g., Ch. 3.5 of van der Vaart and Wellner, 1996). The details of the proof can be found in Section S1.9 of the Supplementary Materials.

Because the A-ITT and A-IV tests are asymptotically equivalent by Lemma 3.1, we shall refer to them indiscriminately using the generic term "t-test" and use ζ_t^2 to denote their common Pittman efficiency. To compare the relative efficiency of WMW-IV against WMW-ITT and the t-test, it is crucial to bound $\mathrm{var}[\nu_{\mathrm{c}}\{Y(0)\}]$, the denominator of $\zeta^2_{\text{WMW-IV}}$. For precise statement of the attainment conditions for the bounds, we need the following definition.

Definition 3.1. *For two probability measures η_1 and η_2 on the real line, we say that η_1 halves η_2, denoted by $\eta_1 \vdash \eta_2$, if there exist two sub-probability measures η_* and η^* such that $\eta_2 = \eta_* + \eta^*$, $\eta_*(\mathbb{R}) = \eta^*(\mathbb{R}) = 2^{-1}$, and $\eta_* \lhd \eta_1 \lhd \eta^*$.*

Lemma 3.2. *We have that*

$$12^{-1}p_{\mathrm{c}} \leq \mathrm{var}\big[\nu_{\mathrm{c}}\big\{Y(0)\big\}\big] \leq 12^{-1}(3-2p_{\mathrm{c}}),$$

with upper bound attained if and only if $\nu_{\mathrm{c}} \vdash \nu_{\bar{\mathrm{c}}}$ and lower bound attained if and only if $\nu_{\bar{\mathrm{c}}} \vdash \nu_{\mathrm{c}}$, where $\nu_{\bar{\mathrm{c}}} = (p_{\mathrm{a}}+p_{\mathrm{n}})^{-1}(p_{\mathrm{a}}\nu_{\mathrm{a}}+p_{\mathrm{n}}\nu_{\mathrm{n}})$ is the pooled null distribution for the non-compliers.

3.3. Comparing WMW-IV versus WMW-ITT

We consider the ARE between WMW-ITT and WMW-IV in two scenarios. In the first, non-compliance is at random (i.e., uninformative) so that $\nu_{\mathrm{c}} = \nu_{\bar{\mathrm{c}}}$; in the second, non-compliance is so strongly informative that ν_{c} and $\nu_{\bar{\mathrm{c}}}$ are completely separated. We also give a universal lower bound for the ARE.

Theorem 3.1 (ARE between WMW-ITT vs WMW-IV)**.** *Under the contiguous alternatives $H_{A,n}$ in (11), the following results hold.*

(a) If $\nu_{\mathrm{c}} = \nu_{\bar{\mathrm{c}}}$, then

$$\frac{\zeta^2_{\text{WMW-ITT}}}{\zeta^2_{\text{WMW-IV}}} \equiv 1.$$

(b) If ν_{c} and $\nu_{\bar{\mathrm{c}}}$ are disjointly supported, then for all $p_{\mathrm{c}} \in (0,1)$,

$$p_{\mathrm{c}}^3 \le \frac{\zeta^2_{\text{WMW-ITT}}}{\zeta^2_{\text{WMW-IV}}} \le p_{\mathrm{c}}^2(3 - 2p_{\mathrm{c}}) < 1. \tag{12}$$

The first equality is attained if and only if $\nu_{\bar{\mathrm{c}}} \vdash \nu_{\mathrm{c}}$ and the second equality is attained if and only if $\nu_{\mathrm{c}} \vdash \nu_{\bar{\mathrm{c}}}$.

(c) The lower bound in (12), along with its necessary and sufficient attainment conditions, is universal (i.e., holds without the disjointness condition).

Hence, WMW-IV is as efficient as WMW-ITT when the non-compliers are identically distributed with the compliers, and is strictly more efficient when the two distributions are disjointly supported. Furthermore, in the latter scenario, the efficiency gain of WMW-IV increases with the non-compliance rate by (12). The ranges of the ARE in the two scenarios are graphed in Figure S3 of the Supplementary Materials. For "intermediate" cases, simulation studies suggest that WMW-IV still tends to outperform WMW-ITT (see Section 4).

Remark 3. For completeness we have provided the Pittman efficiency and ARE values for all p_{c} over $(0,1)$. It is implicitly understood, however, that the asymptotic formulas are accurate in finite samples only when p_{c} is not too small, say $p_{\mathrm{c}} \ge 0.5$. Their accuracy becomes suspect under weak instruments with very small p_{c} (see, e.g., Burgess, Small and Thompson, 2017; Zhao et al., 2020). In that case, non-asymptotic evaluations may instead be preferable (Fieller, 1954; Nelson and Startz, 1990).

3.4. Comparing WMW-IV versus t-test

The relative efficiency between the rank-based WMW-IV and the scale-based t-test depends on two factors. One is the shape of the outcome distribution ν, in particular the heaviness of its tails. The other is the structure and rate of non-compliance. The former has been studied thoroughly for various choices of ν under perfect compliance. For instance, the ARE ranges from $3/\pi$ for the light-tailed normal distribution to $3/2$ for the heavier-tailed Laplace (i.e., double exponential) distribution (see, e.g., Ch. 14 of van der Vaart, 1998). By Lemma 3.1, it is easy to see that these results carry over to the case when non-compliance is at random.

Proposition 3.1. *Under contiguous alternatives $H_{A,n}$ in (11), if $\nu_{\mathrm{c}} = \nu_{\bar{\mathrm{c}}}$, then, regardless of p_{c},*

$$\frac{\zeta_t^2}{\zeta_{\text{WMW-IV}}^2} \equiv \big[12\mathrm{var}\big\{Y(0)\big\}\big]^{-1}\bigg\{\int \dot{\nu}(y)^2\mathrm{d}y\bigg\}^{-2},$$

which is identical to the ARE under perfect compliance.

In the case with informative non-compliance, we consider a simple symmetric non-compliance model (SNM) in order to obtain explicit results regarding the dependence of the ARE on the compliance rate. To be specific, let

$$\mathcal{C} = \begin{cases} 0, & \text{if } Y(0) < -c(\nu, p_{\mathrm{c}}) \\ 1, & \text{if } |Y(0)| \leq c(\nu, p_{\mathrm{c}}) \\ 2, & \text{if } Y(0) > c(\nu, p_{\mathrm{c}}) \end{cases} \tag{13}$$

where $c(\nu, p_{\mathrm{c}}) = \nu^{-1}\{2^{-1}(1 + p_{\mathrm{c}})\}$ and ν is a symmetric distribution about zero. This model is plausible when, for example, the sickest patients always take the treatment while the healthiest never take it (possibly due to unpleasant side effects). This exemplifies a common situation where the compliers are the "typical" ones in the population with the non-compliers tending to be more extreme.

Under the SNM, we can use Lemma 3.1 to express the ARE comparing the t-test against WMW-IV in terms of ν and p_{c}. To study the impact of non-compliance in particular, we consider the standardized ARE, namely, the ratio between the AREs with and without non-compliance under the same ν.

Theorem 3.2. *Under the contiguous alternatives $H_{A,n}$ in (11) and the symmetric non-compliance structure (13), we have that*

$$\frac{\zeta_t^2}{\zeta_{\text{WMW-IV}}^2}(\nu, p_{\mathrm{c}}) = 12^{-1}p_{\mathrm{c}}^4(3 - 2p_{\mathrm{c}})\big[\mathrm{var}\big\{Y(0)\big\}\big]^{-1}\bigg\{\int_{-c(\nu,p_{\mathrm{c}})}^{c(\nu,p_{\mathrm{c}})} \dot{\nu}(y)^2\mathrm{d}y\bigg\}^{-2}, \tag{14}$$

where $(\zeta_t^2/\zeta_{\text{WMW-IV}}^2)(\nu, p_{\mathrm{c}})$ is the ARE under outcome distribution ν and compliance rate p_{c}. In addition, if the outcome distribution is unimodal, i.e., $\dot{\nu}(\cdot)$ is non-increasing on $[0, \infty)$, then the standardized ARE, denoted by

$$\mathcal{R}(\nu, p_{\mathrm{c}}) = \frac{(\zeta_t^2/\zeta_{\text{WMW-IV}}^2)(\nu, p_{\mathrm{c}})}{(\zeta_t^2/\zeta_{\text{WMW-IV}}^2)(\nu, 1)},$$

is strictly increasing in p_{c} and satisfies

$$p_{\mathrm{c}}^4(3 - 2p_{\mathrm{c}}) < \mathcal{R}(\nu, p_{\mathrm{c}}) \leq p_{\mathrm{c}}^2(3 - 2p_{\mathrm{c}}) < 1 \tag{15}$$

for all $p_{\mathrm{c}} \in (0, 1)$.

The following proposition further elucidates the attainment conditions for the upper bound in (15).

Proposition 3.2. *Under the conditions of Theorem 3.2 with a unimodal ν, the following statements are equivalent:*

(a) $\mathcal{R}(\nu, p_c^*) = {p_c^*}^2(3 - 2p_c^*)$ *for some* $p_c^* \in (0,1)$*;*
(b) $\mathcal{R}(\nu, p_c) = p_c^2(3 - 2p_c)$ *for all* $p_c \in (0,1)$*;*
(c) ν *is the a uniform distribution, i.e.,*

$$\dot{\nu}(y) = (2c)^{-1} I(-c \leq y \leq c) \quad \textit{almost everywhere}$$

for some $c > 0$*.*

According to Theorem 3.2 and Proposition 3.2, under the SNM with a unimodal outcome distribution, non-compliance always plays to the advantage of WMW-IV against the t-test. The least favorable case for WMW-IV is the uniform distribution, for which the relative efficiency gain under compliance rate p_c compared to perfect compliance is $p_c^{-2}(3 - 2p_c)^{-1}$. In all other scenarios, the relative gain is strictly greater than that for every $0 < p_c < 1$. A graphical illustration is presented in Figure S4 of the Supplementary Materials.

Using formula (14), we can derive the raw (i.e., non-standardized) ARE $(\zeta_t^2/\zeta_{\text{WMW-IV}}^2)(\nu, p_c)$ as a function of p_c under various choices of ν. Table 1 lists the derived functions under the distributions considered in Ch. 14 of van der Vaart (1998). The derivations are straightforward, if somewhat tedious, and are relegated to the Supplementary Materials. When $p_c = 1$, the table reduces to Table 14.2 of van der Vaart (1998) in the standard setting. Some example functions are plotted in Fig. 2 to better visualize the different trends across different ν.

4. Simulation studies

We first assessed the estimation of the local MWTE τ_c. We generated data by

$$Y(1) = Y(0) + \theta, \text{ where } Y(0) \sim N(0,1), \tag{16}$$

under the symmetric non-compliance structure described in Section 3.4. Let ν be the standard normal distribution and $p_c = 68\%$. Under this set up, a subject is an always-taker if $Y(0) < -1$, a never-taker if $Y(0) > 1$, and a complier if $|Y(0)| \leq 1$. For $\theta = 0, 0.2, 0.5$, the true values of the local MWTE are $\tau_c = 0.500, 0.599, 0.732$, respectively. We evaluated the performance of $\widehat{\tau}_c$ in terms of bias, standard error, and confidence interval estimation. The results are summarized in Table 2. It can be seen that the estimator is virtually unbiased, even for sample size as small as $n = 200$. The standard error estimator described in Section S1.1 of the Supplementary Materials accurately reflects the variations in the estimator. Finally, the empirical coverage probability of the 95% confidence interval closely approximates the nominal rate.

Table 1
Asymptotic relative efficiency between WMW-IV and the t-test under the symmetric non-compliance model with different outcome distributions ν.

Distribution	Efficiency (WMW-IV/t-test)
Logistic	$\dfrac{\pi^2(3-p_c^2)^2}{36p_c^2(3-2p_c)}$
Normal	$\dfrac{3[2\Phi\{\sqrt{2}\Phi^{-1}(\frac{1+p_c}{2})\}-1]^2}{\pi p_c^4(3-2p_c)}$
Laplace	$\dfrac{3(2-p_c)^2}{2p_c^2(3-2p_c)}$
Uniform	$p_c^{-2}(3-2p_c)^{-1}$
t_d $(d>2)$	$\dfrac{12B(1/2,d+1/2)^2[2F_{2d+1}\{\sqrt{2+d^{-1}}F_d^{-1}(\frac{1+p_c}{2})\}-1]^2}{(d-2)B(1/2,d/2)^4p_c^4(3-2p_c)}$
$\frac{3}{4}(1-y^2)\vee 0$	$\dfrac{27g(p_c)^2\{3g(p_c)^4-10g(p_c)^2+15\}^2}{2000p_c^4(3-2p_c)}$

$F_d(\cdot)$ denotes the cumulative distribution function of t_d (t-distribution with d degrees of freedom); $B(\cdot,\cdot)$ is the beta function; $g(p_c)$ denotes the unique root of $(2-y)(1+y)^2=2+2p_c$ in $[0,1]$.

Table 2
Simulation results for estimation of the local MWTE.

n	τ_c	Bias	SE	SEE	CP
200	0.500	0.007	0.080	0.078	0.952
	0.599	0.008	0.084	0.085	0.961
	0.732	0.008	0.094	0.089	0.948
500	0.500	0.001	0.048	0.049	0.972
	0.599	0.003	0.053	0.053	0.970
	0.732	0.002	0.057	0.056	0.966
1000	0.500	0.002	0.035	0.034	0.966
	0.599	0.000	0.037	0.037	0.972
	0.732	0.001	0.041	0.040	0.970
2000	0.500	−0.003	0.024	0.024	0.953
	0.599	−0.001	0.025	0.026	0.946
	0.732	0.000	0.028	0.029	0.948

SE, empirical standard error of the estimator; SEE, empirical average of the standard error estimator; CP, empirical coverage rate of the 95% confidence interval. Each entry is based on 2,000 replicates.

Next, we assessed the empirical power of WMW-IV, WMW-ITT, and the t-test compared theoretically in Section 3. We first generated data under the location-shift Gaussian symmetric non-compliance model. The set-up is similar to that of the first set of simulations, except that we vary the compliance rate over 90%, 80%, 70%, 60%. The empirical power under a range of location-shift treatment effect θ are scatter-plotted in Fig. 3, overlaid with the theoretical asymptotic power functions. It can be seen that the empirical power agrees with the theoretical values fairly well. As suggested by Theorems 3.1 and 3.2,

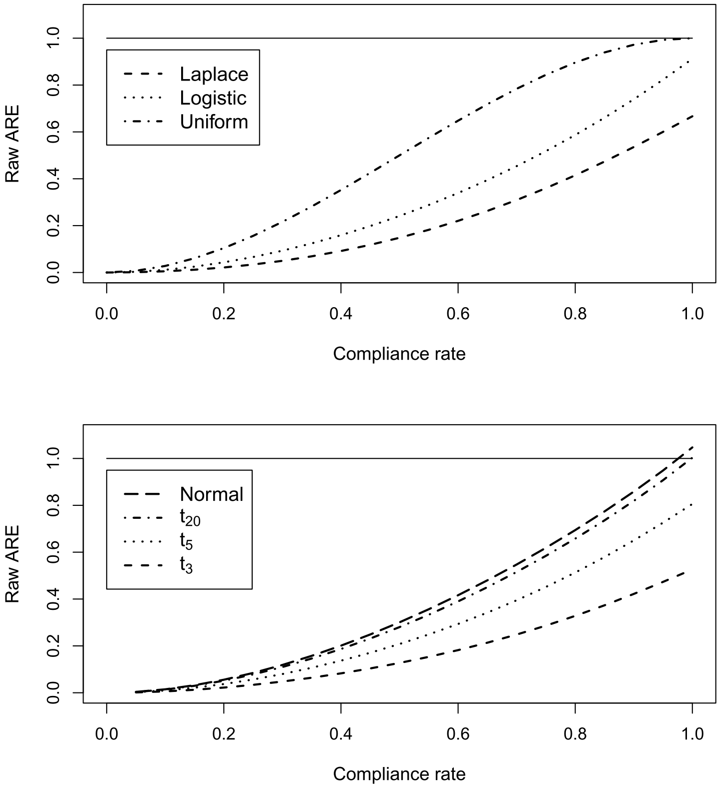

FIG 2. *The raw (i.e., non-standardized) ARE comparing the t-test versus WMW-IV as a function of* p_c *under the symmetric non-compliance model with different outcome distributions.*

the relative efficiencies of WMW-IV over WMW-ITT and t-test increase with the non-compliance rate. At 60% compliance rate, for example, the WMW-IV enjoys substantially higher power than WMW-ITT, followed by the t-test. The lead widens as p_c reduces further to $50\%, 40\%, 30\%$ and 20% (see Figure S5 of the Supplementary Materials), though the asymptotic power function becomes less accurate at the lowest of compliance rates (see Remark 3).

We also conducted simulations to evaluate the power of the three tests outside the symmetric non-compliance model. Specifically, we generated outcome data under a mixture distribution with three Gaussian components, corresponding to the three principal strata of always-takers, compliers, and never-takers, with potentially different locations. Let $\nu = 0.15\nu_\mathrm{a} + 0.70\nu_\mathrm{c} + 0.15\nu_\mathrm{n}$ (so that the compliance rate is 70%), where $\nu_\mathrm{c} = N(0,1)$, $\nu_\mathrm{a} = N(-s,1)$, and $\nu_\mathrm{n} = N(s,1)$ with $s = 0, 1, 2,$ and 3. Fig. 4 provides a graphical illustration of the mixture distributions with different separation distances of the three components. Empirical power of the three tests were computed under the location-shift model (16) as a function of θ. The results are plotted in Fig. 5. At $s = 0$, the three com-

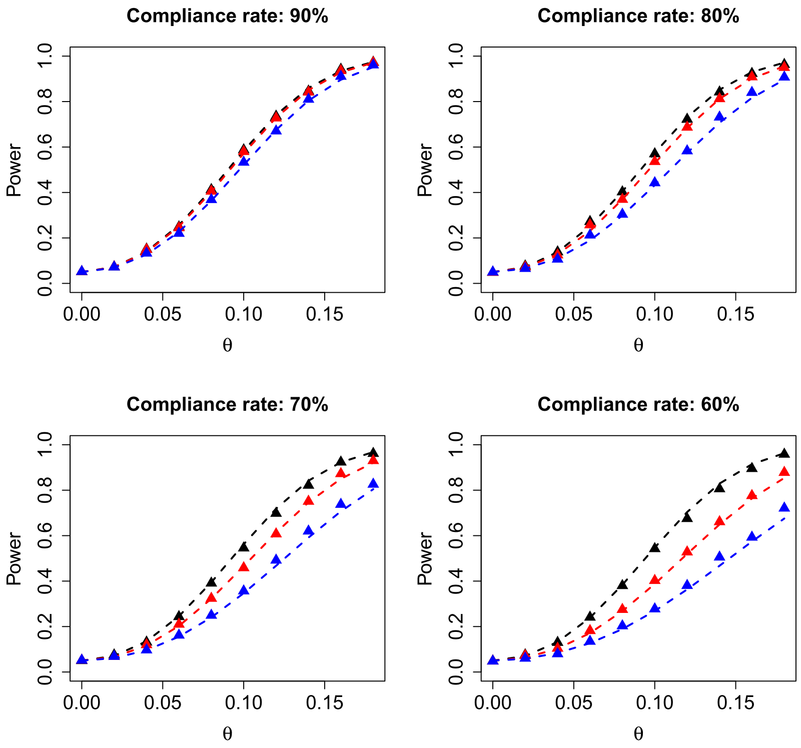

FIG 3. *Empirical and theoretical power for WMW-IV, WMW-ITT, and the t-test under the Gaussian symmetric non-compliance model with sample size* $n = 2000$. *Dashed lines: theoretical power function; triangular points: empirical power based on 2,000 replicates. Black: WMW-IV; Red: WMW-ITT; Blue: t-test.*

ponent distributions are identical, leading to random non-compliance. In this case, as established in Theorem 3.1(a) and Section 3.4, WMW-IV and WMW-ITT are asymptotically equivalent, with relative efficiency against the t-test the same as that under perfect compliance. In this case with the Gaussian distribution, the relative efficiency is $3/\pi$ (see, e.g., Table 14.2 of van der Vaart and Wellner (1996)) with the t-test slightly more powerful than the other two. This minor advantage of the t-test is reflected, though barely distinguishable, in the empirical power curves in first panel of Fig. 5. However, as s increases (i.e., non-compliance becoming more informative), WMW-IV again becomes more powerful than WMW-IV and the t-test.

5. A real example

The Job Training Partnership Act (JTPA) was enacted in 1982 to fund federal training programs to prepare youths and economically disadvantaged adults in the U.S. for (re-)entry into the workforce. The effectiveness of the training

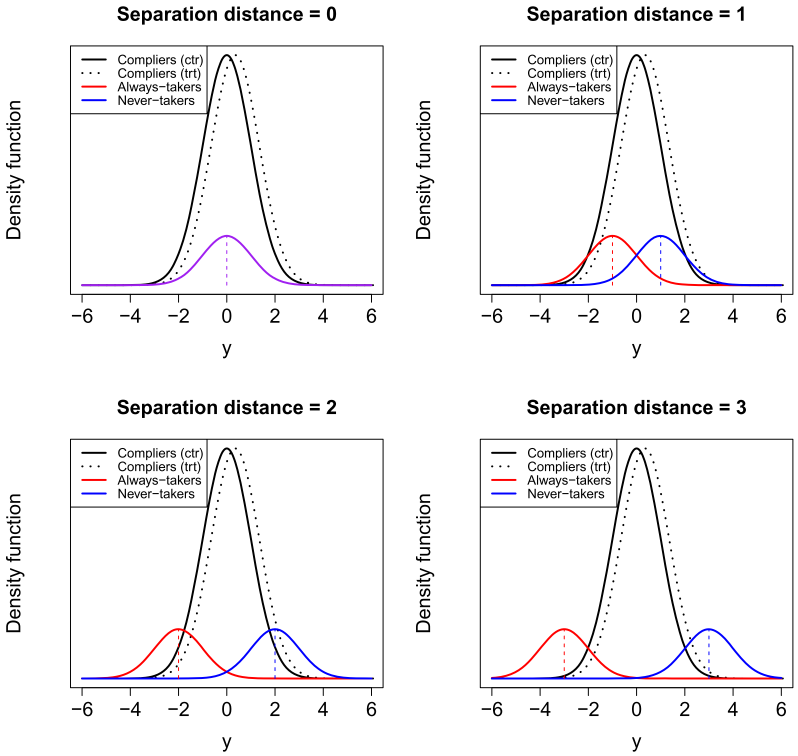

FIG 4. *Graphical illustration of mixture distributions of the form* $\nu = 0.15\nu_{\mathrm{a}} + 0.70\nu_{\mathrm{c}} + 0.15\nu_{\mathrm{n}}$*, where* $\nu_{\mathrm{c}} = N(0,1)$*,* $\nu_{\mathrm{a}} = N(-s,1)$*, and* $\nu_{\mathrm{n}} = N(s,1)$ *with* $s = 0, 1, 2,$ *and 3.*

programs was subsequently evaluated in the National JTPA Study, a large-scale randomized controlled trial involving more than twenty thousand participants. To illustrate the proposed methodology, we analyze a study cohort consisting of 11,204 of adult participants. This dataset was previously analyzed by Abadie, Angrist and Imbens (2002) using quantile regression models.

Among the participants, 7,484 were randomized to the treatment group ($Z = 1$) and were offered job training. However, 2,683 (35.8%) of them chose not to engage. The remaining 3,720 participants were randomized to the control group ($Z = 0$) and were excluded from job training for a period of 18 months. However, 54 (1.5%) of them managed to get training elsewhere. Thus, the estimated compliance rate is $\widehat{p}_{\mathrm{c}} = 1 - 35.8\% - 1.5\% = 62.7\%$. The outcome Y is the sum of earnings in 30 months. The average earning in the randomized treatment group is \$16,200 and that in the control group is \$15,040.

We first use the methods described in Sections 2.1 and 3 to estimate the local MWTE and to test the treatment effect in various subgroups defined by gender, race, ethnicity, and age. The results are summarized in Table 3. For ease of interpretation, we focus on the local MW odds (i.e., $\tau_{\mathrm{c}}/(1 - \tau_{\mathrm{c}})$), estimated

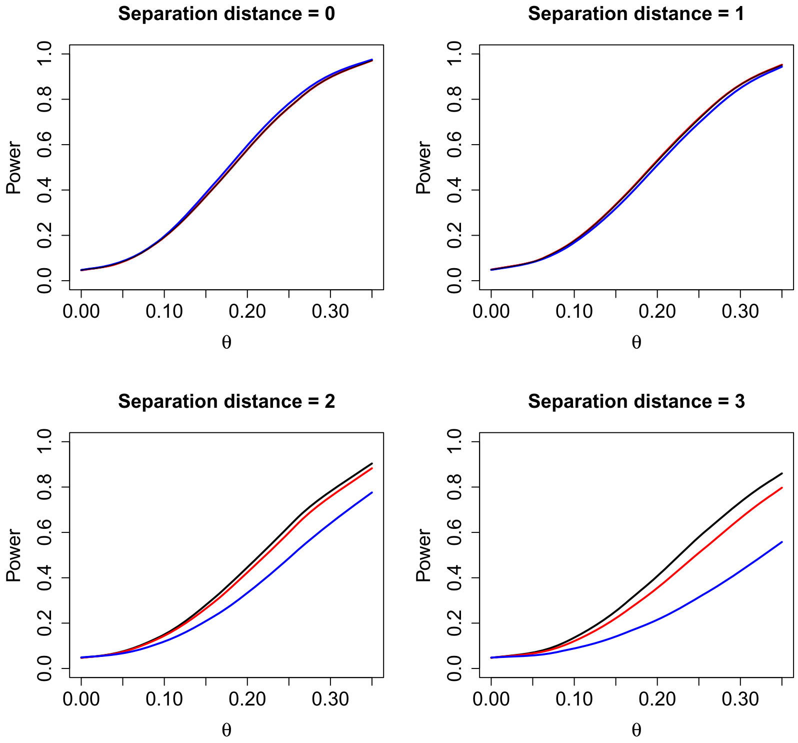

FIG 5. *Empirical power (based on 2,000 replicates) for WMW-IV, WMW-ITT, and the t-test under outcome distributions illustrated in Fig. 4.*

based on $\widehat{\tau}_{\mathrm{c}}$. For comparison, we also compute the ITT MW odds. It can be seen that the ITT effects are attenuated versions of the corresponding local treatment effects. Overall, compliers going through job training are $1.15 - 1 = 15\%$ more likely to earn more than those without training ($p < 0.001$). The local effect is also significant (at the 0.05 level) in most of the subgroups, except for males and Hispanics (the latter likely due to small sample size). In particular, there is a notable difference in treatment effect between females and males. Among complying females, the trained are 22% more likely to have a higher income than the untrained; whereas among complying males, the trained are only 10% more likely to have a higher income than the untrained. Similar differential effects are observed between the younger and older age groups.

In keeping with the theoretical results about their relative efficiency by Theorem 3.1, the p-values produced by WMW-IV are consistently more significant than those by WMW-ITT across the subgroups. The t-test on the original scale of the outcome ($) yields comparable results to WMW-IV. However, a log-transformation of the outcome, i.e., $\log(1 + Y)$, leads to substantial power loss. This shows that the performance of the t-test is rather sensitive to the outcome scale.

TABLE 3
Estimation and testing of treatment effect in different subgroups of the JTPA study.

Subgroup	$N(p_c)$	MW Odds ITT	MW Odds Local (95% CI)	p-value (10^{-2}) IV	p-value (10^{-2}) ITT	p-value (10^{-2}) t	p-value (10^{-2}) t (log)
Overall	11,204 (63%)	1.11	1.15 (1.07, 1.24)	<0.1	0.1	<0.1	2.3
Female	6,102 (64%)	1.16	1.22 (1.11, 1.35)	<0.1	0.1	0.1	0.7
Male	5,102 (61%)	1.07	1.10 (0.98, 1.23)	7.0	16.6	5.0	67.8
AA	2,909 (59%)	1.11	1.17 (1.00, 1.36)	2.8	9.6	2.5	50.5
Non-AA	8,295 (64%)	1.11	1.15 (1.06, 1.25)	<0.1	0.4	0.4	2.4
Hisp	1,225 (69%)	1.13	1.15 (0.94, 1.41)	11.4	28.4	30.9	51.8
Non-Hisp	9,979 (62%)	1.11	1.15 (1.07, 1.25)	<0.1	0.2	0.1	2.9
<30 yr	4,926 (63%)	1.13	1.20 (1.08, 1.34)	<0.1	0.3	0.7	2.1
≥30 yr	6,278 (62%)	1.09	1.12 (1.01, 1.23)	1.4	9.3	2.3	32.0

AA: African American; Hisp: Hispanic.
IV: WMW-IV; ITT: WMW-ITT; t: t-test on income; t (log): t-test on log-transformed income.

Next, we use the methods proposed in Section 2 for sensitivity analyses of the local MWTE under violated assumptions and of the unidentifiable global MWTE. In particular, Theorem 2.2 and Corollary 2.1 are used to bound the local MW treatment odds under violations of Exclusion Restriction (ER) and Monotonicity, respectively. (In this context, ER means that the offering of job training affects future earnings only through the subject's decision whether to take it; Monotonicity means that no subject would refrain from training when assigned to it and take it when assigned otherwise.) The results are summarized in Table S1 of the Supplementary Materials. The bounding intervals for both the local treatment effect under violated ER and the global treatment effect are rather wide. This is because both types of bounds are functions of the overall non-compliance rate, which in this case is substantial (37.3%). In contrast, the bounding intervals for the local treatment effect under violated Monotonicity are fairly tight due to a small proportion of always-takers (1.5%; cf. $\lambda_{\max} = \rho_{\mathrm{a}} \wedge \rho_{\mathrm{n}}$ in Corollary 2.1). These bounds suggest that our conclusions about the beneficial effects of job training on the compliers are largely robust to possible contamination by defiers.

Finally, we use the graphical procedure suggested in Proposition 2.1 to assess the plausibility of the ER assumption. The estimated cumulative distribution functions for the compliers and their associated bounds are plotted in Fig. 6. Since the estimated functions are well within the bounds implied by the ER assumption, we conclude that no evidence yet exists to refute the absence of randomization direct effects.

6. Discussions and extensions

Under standard IV assumptions for randomization, we have developed estimation and inference procedures, along with sensitivity analysis techniques, for the local MWTE to address non-compliance. Moreover, the IV-based WMW test is shown asymptotically superior to standard WMW-ITT test and t-test in some

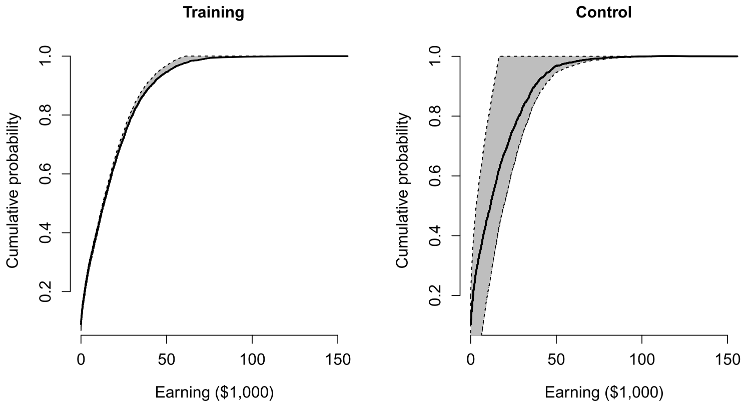

FIG 6. *The estimated outcome cumulative distributions for the compliers and their bounding regions implied by the Exclusion Restriction assumption according to Proposition 2.1. (The lower bound in the left panel is visually indistinguishable from the estimated curve due to the low proportion of always-takers (1.5%).)*

scenarios, especially under high rates of informative non-compliance. Our work extends the classical IV approach of Angrist, Imbens and Rubin (1996) from the ATE to MWTE. It provides empirical researchers with an alternative tool for causal inference in the presence of non-ignorable non-compliance.

When the treatment is available only to those assigned to it, non-compliance can only be one-sided as those assigned to the control have no means to cross over. In that case, it can be shown that our MWTE on the compliers is also the MWTE on the treated. On other occasions, there may be multiple treatment groups with graduated doses, such as under the“randomized encouragement design” (West et al., 2008), where the Monotonicity assumption would mean that a subject receiving “encouragement” (i.e., some incentive to take the treatment, say, job training) would at least take as much treatment as he/she otherwise would without encouragement. To extend our approach, we can divide the treatments into two groups based on a cutoff dose, calculate the two-group local MWTE estimate $\widehat{\tau}_{\mathrm{c}}$ (the multi-treatment Monotonicity carries over to the dichotomized version), and, at the end, perform a Kruskal–Wallis-type joint test for all possible cutoffs using the robust (co)variances of the $\widehat{\tau}_{\mathrm{c}}$ discussed in Section 2.1. Although this overall test is expected to be valid under the sharp null, its causal interpretation is suspect because the compliers are defined differently for different cutoffs. A consistent definition would be challenged by identifiability issues inherent in non-compliance between multiple treatments. Some details are offered in the Supplementary Materials.

The “mixture-data” approach of Section 2.2 underlies a number of existing sensitivity bounds for the ATE and quantile treatment effect (QTE) (see, e.g., Imai, 2008; Blanco, Flores and Flores-Lagunes, 2013; Flores and Flores-Lagunes, 2013; Huber and Mellace, 2015; Blanco et al., 2020; Mao, 2022). However, our study appears to be the first to apply this idea (Manski, 2003, Ch. 4) to Mann–

Whitney-type estimand. As a byproduct, we have exploited the ER-implied stochastic-order constraints to check the validity of the assumption (see Proposition 2.1 and Fig. 6). Compared with the weaker inequalities on the mean (Huber and Mellace, 2015), the stochastic order is more holistic and may be better at detecting violations.

To make the sensitivity analysis more precise, it helps to quantify the uncertainty in the empirical bounds. While we have derived the confidence intervals for the bounds in (4) (see Supplementary Materials), doing so for those in Theorem 2.1 and Corollary 2.2 would be more difficult, as they are "intersection bounds" involving the infimum or supremum of functions or scalers (Imbens and Manski, 2004; Stoye, 2009; Chernozhukov, Lee and Rosen, 2013; Kaido, Molinari and Stoye, 2019). For those of which we do have the means to calculate the variances, we can use the bound estimates to test the treatment effect. Unlike the tests studied in Section 3, however, the power of tests on the bounds will depend not only on the effect size and compliance rate, but also the degree to which the identifying assumptions are violated (e.g., the magnitude of randomization effect or the proportion of defiers), which determines how far the true bounds are from the effect size.

An intriguing question remains as to whether WMW-IV is universally more powerful than WMW-ITT under location-shift alternatives. Yet a universal lower bound for the ARE between the two tests may not be straightforward. On the other hand, our approach can be easily applied to other nonparametric tests based on, e.g., the (local) QTE (Doksum, 1974; Firpo, 2007; Rosenbaum, 2013; Bickel and Doksum, 2015). To do so, we only need to replace the MW functional $\mathcal{M}(\eta_1, \eta_0) = \int \eta_0(y)\eta_1(\mathrm{d}y)$ with, e.g., the quantile difference functional $\mathcal{Q}(\eta_1, \eta_0) = \eta_1^{-1}(\pi) - \eta_0^{-1}(\pi)$ $(0 < \pi < 1)$, and use its functional derivative (along the lines of Section S1.9 of Supplementary Materials) to derive Pittman efficiency results similar to Lemma 3.1. Some initial results are described in our recent work (Mao, 2022).

We have considered estimation and testing of the local MWTE in a completely nonparametric setting. In practice, pre-treatment variables such as participant demographics may be utilized to improve the efficiency and robustness of inference (see, e.g., Abadie, Angrist and Imbens, 2002; Tchetgen Tchetgen et al., 2015; Tchetgen Tchetgen and Wirth, 2017; Wang and Tchetgen Tchetgen, 2018). Let $\boldsymbol{X}_i$ and $\boldsymbol{X}_j$ denote the covariates associated with (independent) subjects i and j, respectively. Following the probability index model (PIM; a standard regression model for the MWTE) (Thas et al., 2012), we could specify the following (structural) model for the local MWTE on the compliers:

$$\mathbb{P}\big\{Y_j(a) \geq Y_j\big(a'\big) \mid \mathcal{C}_i = \mathcal{C}_j = 1, \boldsymbol{X}_i, \boldsymbol{X}_j\big\} = g\big\{\gamma\big(a - a'\big) + \boldsymbol{\beta}^{\mathrm{T}}\big(\boldsymbol{X}_i - \boldsymbol{X}_j\big)\big\}, \quad (17)$$

where γ and $\boldsymbol{\beta}$ are regression coefficients for the treatment and covariates, respectively, and $g(\cdot)$ is a suitable link function, e.g., $g(x) = \exp(x)/\{1 + \exp(x)\}$. Unlike the standard PIM, the outcomes (and conditioning variables) on the left hand side of (17) are not fully observed. As a result, we may need weights such as employed in the regression of local quantile treatment effect (Abadie, Angrist and Imbens, 2002) to fit the latent model with observed data. The use of such weights in a pairwise regression setting has not been explored in the literature. This will be our future work.

Acknowledgments

I thank the associate editor and two anonymous referees for helpful comments.

Funding

This research was supported by National Science Foundation grant DMS2015526 and National Institutes of Health grant R01HL149875.

Supplementary Material

Program Codes
(doi: 10.1214/23-EJS2209SUPPA; .zip).

Supplementary Materials to "Wilcoxon-Mann-Whitney Statistics in Randomized Trials With Non-Compliance"
(doi: 10.1214/23-EJS2209SUPPB; .pdf).

References

Abadie, A., Angrist, J. and Imbens, G. (2002). Instrumental variables estimates of the effect of subsidized training on the quantiles of trainee earnings. *Econometrica* **70** 91–117. MR1926256

Acion, L., Peterson, J. J., Temple, S. and Arndt, S. (2006). Probabilistic index: an intuitive non-parametric approach to measuring the size of treatment effects. *Statistics in Medicine* **25** 591–602. MR2222116

Angrist, J. D., Imbens, G. W. and Rubin, D. B. (1996). Identification of Causal Effects Using Instrumental Variables. *Journal of the American Statistical Association* **91** 444–455.

Bickel, P. J. and Doksum, K. A. (2015). *Mathematical Statistics: Basic Ideas and Selected Topics, Volumes I–II Package*. Chapman and Hall/CRC. MR3287337

Blanco, G., Flores, C. A. and Flores-Lagunes, A. (2013). Bounds on average and quantile treatment effects of Job Corps training on wages. *Journal of Human Resources* **48** 659–701.

Blanco, G., Chen, X., Flores, C. A. and Flores-Lagunes, A. (2020). Bounds on average and quantile treatment effects on duration outcomes under censoring, selection, and noncompliance. *Journal of Business & Economic Statistics* **38** 901–920. MR4154896

Brumback, L. C., Pepe, M. S. and Alonzo, T. A. (2006). Using the ROC curve for gauging treatment effect in clinical trials. *Statistics in Medicine* **25** 575–590. MR2222115

Burgess, S., Small, D. S. and Thompson, S. G. (2017). A review of instrumental variable estimators for Mendelian randomization. *Statistical Methods in Medical Research* **26** 2333–2355. MR3712236

Chernozhukov, V., Lee, S. and Rosen, A. M. (2013). Intersection bounds: Estimation and inference. *Econometrica* **81** 667–737. MR3043345

Cliff, N. (1993). Dominance statistics: Ordinal analyses to answer ordinal questions. *Psychological Bulletin* **114** 494.

Doksum, K. (1974). Empirical probability plots and statistical inference for nonlinear models in the two-sample case. *The Annals of Statistics* **2** 267–277. MR0356350

Fay, M. P., Brittain, E. H., Shih, J. H., Follmann, D. A. and Gabriel, E. E. (2018). Causal estimands and confidence intervals associated with Wilcoxon-Mann-Whitney tests in randomized experiments. *Statistics in Medicine* **37** 2923–2937. MR3848163

Fieller, E. C. (1954). Some problems in interval estimation. *Journal of the Royal Statistical Society: Series B (Methodological)* **16** 175–185. MR0093076

Firpo, S. (2007). Efficient semiparametric estimation of quantile treatment effects. *Econometrica* **75** 259–276. MR2284743

Flores, C. A. and Flores-Lagunes, A. (2013). Partial identification of local average treatment effects with an invalid instrument. *Journal of Business & Economic Statistics* **31** 534–545. MR3173699

Frangakis, C. E. and Rubin, D. B. (2002). Principal stratification in causal inference. *Biometrics* **58** 21–29. MR1891039

Grissom, R. J. (1994). Probability of the superior outcome of one treatment over another. *Journal of Applied Psychology* **79** 314.

Halperin, M., Gilbert, P. R. and Lachin, J. M. (1987). Distribution-free confidence intervals for $\Pr(X_1 < X_2)$. *Biometrics* 71–80. MR0882776

Hauck, W. W., Hyslop, T. and Anderson, S. (2000). Generalized treatment effects for clinical trials. *Statistics in Medicine* **19** 887–899.

Horowitz, J. L. and Manski, C. F. (2000). Nonparametric analysis of randomized experiments with missing covariate and outcome data. *Journal of the American Statistical Association* **95** 77–84. MR1803142

Huber, M. and Mellace, G. (2015). Testing instrument validity for LATE identification based on inequality moment constraints. *Review of Economics and Statistics* **97** 398–411.

Imai, K. (2008). Sharp bounds on the causal effects in randomized experiments with "truncation-by-death". *Statistics & Probability Letters* **78** 144–149. MR2382067

Imbens, G. W. and Manski, C. F. (2004). Confidence intervals for partially identified parameters. *Econometrica* **72** 1845–1857. MR2095534

Imbens, G. W. and Rubin, D. B. (1997). Estimating outcome distributions for compliers in instrumental variables models. *The Review of Economic Studies* **64** 555–574. MR1485828

Kaido, H., Molinari, F. and Stoye, J. (2019). Confidence intervals for projections of partially identified parameters. *Econometrica* **87** 1397–1432. MR3994276

Kieser, M., Friede, T. and Gondan, M. (2013). Assessment of statistical significance and clinical relevance. *Statistics in Medicine* **32** 1707–1719. MR3060636

Kraemer, H. C. and Kupfer, D. J. (2006). Size of treatment effects and their importance to clinical research and practice. *Biological Psychiatry* **59** 990–996.

Mann, H. B. and Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics* 50–60. MR0022058

Manski, C. F. (2003). *Partial Identification of Probability Distributions*. Springer Science & Business Media. MR2151380

Mao, L. (2018). On causal estimation using U-statistics. *Biometrika* **105** 215–220. MR3768875

Mao, L. (2022). Nonparametric inference of complier quantile treatment effects in randomized trials with imperfect compliance. *Biostatistics & Epidemiology* **6** 249–265.

McGraw, K. O. and Wong, S. (1992). A common language effect size statistic. *Psychological Bulletin* **111** 361.

Nelson, C. R. and Startz, R. (1990). The distribution of the instrumental variables estimator and its t-ratio when the instrument is a poor one. *Journal of Business* S125–S140.

Newcombe, R. G. (2006a). Confidence intervals for an effect size measure based on the Mann–Whitney statistic. Part 2: Asymptotic methods and evaluation. *Statistics in Medicine* **25** 559–573. MR2222114

Newcombe, R. G. (2006b). Confidence intervals for an effect size measure based on the Mann–Whitney statistic. Part 1: general issues and tail-area-based methods. *Statistics in Medicine* **25** 543–557. MR2222113

Robins, J. M. and Greenland, S. (1996). Identification of causal effects using instrumental variables: comment. *Journal of the American Statistical Association* **91** 456–458.

Rosenbaum, P. (2013). *Observational Studies*. New York: Springer. MR1353914

Rubin, D. B. (1978). Bayesian inference for causal effects: the role of randomization. *The Annals of Statistics* 34–58. MR0472152

Stoye, J. (2009). More on confidence intervals for partially identified parameters. *Econometrica* **77** 1299–1315. MR2547075

Tchetgen Tchetgen, E. J. and Wirth, K. E. (2017). A general instrumental variable framework for regression analysis with outcome missing not at random. *Biometrics* **73** 1123–1131. MR3744526

Tchetgen Tchetgen, E. J., Walter, S., Vansteelandt, S., Martinussen, T. and Glymour, M. (2015). Instrumental variable estimation in a survival context. *Epidemiology* **26** 402.

Thas, O., Neve, J. D., Clement, L. and Ottoy, J.-P. (2012). Probabilistic index models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **74** 623–671. MR2965954

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge: Cambridge University Press. MR1652247

van der Vaart, A. W. and Wellner, J. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Science & Business Media. MR1385671

Wang, L. and Tchetgen Tchetgen, E. J. (2018). Bounded, efficient and multiply robust estimation of average treatment effects using instrumental variables. *Journal of the Royal Statistical Society. Series B, Statistical Methodology* **80** 531. MR3798877

West, S. G., Duan, N., Pequegnat, W., Gaist, P., Des Jarlais, D. C., Holtgrave, D., Szapocznik, J., Fishbein, M., Rapkin, B., Clatts, M. et al. (2008). Alternatives to the randomized controlled trial. *American Journal of Public Health* **98** 1359–1366.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin* **1** 80–83. MR0025133

Wu, P., Han, Y., Chen, T. and Tu, X. (2014). Causal inference for Mann–Whitney–Wilcoxon rank sum and other nonparametric statistics. *Statistics in Medicine* **33** 1261–1271. MR3238234

Zhang, Z., Liu, C., Ma, S. and Zhang, M. (2019). Estimating Mann–Whitney-type causal effects for right-censored survival outcomes. *Journal of Causal Inference* `https://doi.org/10.1515/jci-2018-0010`. MR4350062

Zhao, Q., Wang, J., Hemani, G., Bowden, J., Small, D. S. et al. (2020). Statistical inference in two-sample summary-data Mendelian randomization using robust adjusted profile score. *Annals of Statistics* **48** 1742–1769. MR4124342

Zhou, W. (2008). Statistical inference for $P(x < y)$. *Statistics in Medicine* **27** 257–279. MR2412707