Research Article

Soojin Park* and Esra Kürüm

A Two-Stage Joint Modeling Method for Causal Mediation Analysis in the Presence of Treatment Noncompliance

https://doi.org/10.1515/jci-2019-0019
Received Jul 06, 2019; accepted Apr 05, 2020

Abstract: Estimating the effect of a randomized treatment and the effect that is transmitted through a mediator is often complicated by treatment noncompliance. In literature, an instrumental variable (IV)-based method has been developed to study causal mediation effects in the presence of treatment noncompliance. Existing studies based on the IV-based method focus on identifying the mediated portion of the intention-to-treat effect, which relies on several identification assumptions. However, little attention has been given to assessing the sensitivity of the identification assumptions or mitigating the impact of violating these assumptions. This study proposes a two-stage joint modeling method for conducting causal mediation analysis in the presence of treatment noncompliance, in which modeling assumptions can be employed to decrease the sensitivity to violation of some identification assumptions. The use of a joint modeling method is also conducive to conducting sensitivity analyses to the violation of identification assumptions. We demonstrate our approach using the Jobs II data, in which the effect of job training on job seekers' mental health is examined.

Keywords: Treatment noncompliance, Two-stage method, Sensitivity analysis, Compliers-average causal mediation effect

2020 Mathematics Subject Classification: 62D20

1 Introduction

In randomized experiments, the interest is often not only in the effect of a randomized treatment but also in the effect transmitted through a mediator. This is because investigating mediating mechanisms provides a complete explanation of the effect of the treatment. Recently, there have been many studies on causal mediation analysis, which focuses on how to identify and estimate the average effect of a treatment transmitted through a mediator (See, e.g., [1–4]). One complication that arises when conducting this type of analysis is non-ignorability of a mediator because mediators are seldom randomized, even in randomized experiments.

Another complication in randomized experiments is that some participants do not adhere to the assigned treatment. In this article, we refer this non-adherence to the assigned treatment to *treatment noncompliance*. In the presence of treatment noncompliance, the treatment receipt status is no longer random even when the treatment is assigned randomly because participants self-select to adhere to the treatment or not. One analytical option to address this issue is to focus on the effect of the assigned treatment, namely, intention-to-treat (ITT) effects. Under a randomized treatment and the stable unit treatment value assumption (SUTVA),[1] the ITT effect is identified as the difference in the average outcome value between those who are assigned to

***Corresponding Author: Soojin Park:** University of California, Riverside, E-mail: soojinp@ucr.edu
Esra Kürüm: University of California, Riverside, E-mail: esra.kurum@ucr.edu

1 This assumption consists of two sub-assumptions: 1) the treatment is clearly defined (no variation in the treatment) and 2) the treatment assignment of an individual do not affect the potential outcomes (or mediators) of others.

Open Access. © 2020 S. Park and E. Kürüm, published by De Gruyter. CC BY This work is licensed under the Creative Commons Attribution 4.0 License

the treatment and those who are not. ITT analysis avoids the problem of treatment noncompliance because inference relies only on the randomization of the treatment [5].

Further challenge arises when identifying the mediated portion of this ITT effect. Simply employing the mediation formula [6] with the assigned treatment (as if the assigned treatment is actual receipt of the treatment) does not provide a valid result [7] because this approach violates an important assumption of causal mediation analysis: *no treatment-induced mediator and outcome confounding* [1, 3, 6]. In the presence of treatment noncompliance, the actual treatment receipt status impacts both mediator and outcome and is influenced by the assigned treatment (i.e., treatment-induced mediator and outcome confounding). One way of circumventing this issue is to identify this mediated portion of the ITT effects on the basis of the average causal mediation effect (ACME) among compliers. Among compliers, the assigned treatment always coincides with the treatment received and thus, the ACME can be estimated without this issue of treatment induced mediator and outcome confounding. Yamamoto [7] proposed this way of identifying the mediated portion of the ITT effect using the instrumental variables (IV) approach.

While the IV approach successfully addresses the issue of identifying the mediated ITT effect, a concern remains. Estimating the mediated ITT effect on the basis of the ACME among compliers requires multiple identification assumptions. Due to these multiple identification assumptions, validating results in the IV approach is often challenging. Previous research by Yamamoto [7] left assessing the validity of results to the violation of identification assumptions to future study. Another study by Park and Kürüm [8] assessed the validity of results by assuming a worst case scenario but failed to assess the sensitivity of the results systematically to all possible scenarios. Therefore, it is necessary to develop an approach that can mitigate the impact of violations of identification assumptions and/or be more conducive to conducting sensitivity analyses.

In this article, we propose a two-stage joint modeling method to estimate the mediated ITT effect because of its potential benefit of employing modeling assumptions such as distributional assumptions [9] and additional covariates [10, 11] that can mitigate the impact of the violation of some identification assumptions. Another benefit of this method is that it provides a relatively convenient setting to conduct sensitivity analyses to the violation of identification assumptions, compared to the IV-based method. These benefits are demonstrated using the JOBS II data, in which the effect of job training on job-seekers' mental health is examined.

The rest of the article is organized as follows. We introduce our motivating example in Section 2. In Section 3, we present the identification result of the mediated and unmediated ITT effects. In Section 4, we propose a two-stage joint modeling estimation method, which is followed by a simulation study that examines the role of modeling assumptions when identification assumptions are violated (Section 5). In Section 6, we propose sensitivity analyses based on the proposed joint modeling method. In Section 7, we show how sensitivity analyses to the violation of identification assumptions can be conducted in the context of our example. We conclude with a discussion.

2 JOBS II Intervention Project

This study is motivated by the JOBS Search Intervention Study (JOBS II) [12]. Job loss can lead to harmful effects on a worker's mental, physical, and social health [13–15]. The JOBS II study was designed as a randomized trial to examine the effects of a job training intervention on unemployed individuals' mental health. The goal of this intervention was to prevent the negative effects of job loss by equipping job seekers with efficient job search strategies. The randomized treatment group was assigned to five half-day job searching seminars. Both treatment and control groups received a booklet describing job searching skills. In the JOBS II study, the job training intervention seminars were only available to subjects in the treatment group; subjects in the control group had no way of participating in the seminars. In line with many previous studies [16–18], we define the treatment receipt status as attending at least one out of five job-searching seminars. Forty-eight percent of those who were assigned to the treatment did not attend any job searching seminars.

Project recruitment consisted of a short screening questionnaire (T0) to determine eligibility, resulting in 1,801 participants. The pre-treatment survey was mailed (T1), and follow-up surveys were mailed two months

(T2), six months (T3), and two years (T4) after the week of job training seminars. Data collected in this study included demographic variables such as age, gender, race, and marital status, as well as measures of depression, self-esteem, job-search efficacy, internal control orientation, and reemployment status. Descriptive statistics for the variables used in our analysis are presented in Table 1.

Previous analysis of JOBS II data showed that the job-training intervention produced beneficial effects, including increased reemployment rates and improved mental health [8, 19, 20]. More specifically, Price et al. [21] showed that the intervention had beneficial effects on those who were identified as being at high risk for experiencing mental health setbacks such as episodes of depression. They also identified sense of mastery as a mediator for the relationship between the intervention and depression. Our analysis will differ from these analyses in that we will investigate the association between job-training seminars and depression in the presence of the mediator, sense of mastery, and by addressing treatment noncompliance using a two-stage joint method, which provides a convenient setting for systematic analyses of sensitivity to the violation of identification assumptions. The outcome variable, depression, was measured using responses to an 11-item list based on the Hopkins Symptom Checklist [22]. The mediator variable, sense of mastery, was computed as the mean score of job-search efficacy, self-esteem, and internal control orientation.

Table 1: Descriptive Statistics for JOBSII data

Variables	Mean	Variance	Min.[1]	Max. [2]
Depression (post)[3]	1.755	0.442	1	4.9
Sense of Mastery	2.591	1.211	1	4
Sex (X_1)	0.531[4]	-	0	1
Motivation (X_2)	5.228	0.732	1	6.5
Nonwhite (X_3)	0.762[5]	-	0	1
Marital (X_4)	1.142	1.265	0	4
Education (X_5)	1.845	1.171	0	4
Assertiveness (X_6)	3.453	0.838	1	5
Age (X_7)	36.12	18.8	16.5	76.9
Depression (pre)[6] (X_8)	1.863	0.331	1	3.5
Economic Hardship (X_9)	3.092	0.979	1	5

1. Minimum value, 2. maximum value, 3. depression levels measured after (T3) the training, 4. represent the ratio of males in our data,
5. represent the ratio of nonwhite subjects in our data, and 6. depression levels measured before the training.

3 Identification

In order to precisely define the effects of interest, consider an experimental setting that mimics the JOBS II project, where some subjects did not comply with the assigned treatment. Let Z_i represent the *assigned treatment*, where $Z_i = 0$ if individual i is assigned to the control condition and $Z_i = 1$ otherwise; let T_i represent the *actual treatment received*, where $T_i = 0$ if individual i did not receive the treatment and $T_i = 1$ if individual i attended at least one job training seminar; M_i and Y_i represent the mediator and outcome, respectively; and X is a vector of multiple observed pre-treatment covariates. The supports of the distributions of X_i, M_i, and Y_i are represented as $\mathcal{X}$, $\mathcal{M}$ and $\mathcal{Y}$, respectively. Under the SUTVA, $T_i(z)$ represents the treatment receipt status if individual i was assigned to $Z_i = z$; $M_i(z)$ represents the potential mediator of M under $Z_i = z$; $Y_i(z, m)$ represents the potential outcome Y under $Z_i = z$, and $M_i = m$ for individual i for $z \in \{0, 1\}$ and $m \in \mathcal{M}$. P_i is an indicator for compliance type that includes compliers ($P_i = c$) and never takers ($P_i = n$).

Throughout the paper, we assume the randomization of the treatment assignment.

Assumption 1: Randomization. Treatment assignment is random.

Effects of Interest. Our primary effects of interest are the mediated and unmediated portion of the ITT effect. These are the average effect of offering the treatment on the outcome transmitted through (mediated ITT) or not through (unmediated ITT) a mediator. Since the decomposition is based on the average effect of offering the treatment, we include both those who did and did not comply with the assigned treatment in the analysis. In other words, ITT analysis tests the effectiveness of a randomized intervention regardless whether the subjects actually received the treatment or not. Therefore, the mediated and unmediated ITT effects are of interest for those who want to evaluate the overall effect of an intervention and investigate underlying mechanisms of the effect in a usual setting, in which not every subject complied with the treatment. Throughout this paper, we focus on the mediated and unmediated portion of the ITT effects that include both compliers and non compliers.

Following Yamamoto [7], the mediated and unmediated portion of the ITT effect will be identified and estimated on the basis of the ACME and average natural direct effect among compliers, respectively. Therefore, we first define the complier average causal mediation effect (CACME) and complier average natural direct effect (CANDE), as

$$\begin{aligned} \delta_c(z) &\equiv E[Y_i(z, M_i(1)) - Y_i(z, M_i(0))|P_i = c] \text{ and} \\ \zeta_c(z) &\equiv E[Y_i(1, M_i(z)) - Y_i(0, M_i(z))|P_i = c], \end{aligned} \tag{1}$$

where $z \in \{0, 1\}$. In our example, $\delta_c(1)$ indicates among the compliers to what degree the level of depressive symptoms has changed in response to the change in the sense of mastery (from the value that would have resulted under the training to the value that would have resulted under the control) under the job training condition. Likewise, $\zeta_c(1)$ indicates among compliers the average change in the level of depressive symptoms in response to the change in treatment status (that is, from being assigned to job training vs no training), while holding the mediator at the value under the job training condition.

In order to obtain the CACME and CANDE, distributions of mediator and outcome need to be modeled. We use the likelihood to model the distribution of Y, M, and T, given X and Z. For $t \in \{0, 1\}$ and $z \in \{0, 1\}$, let S_{tz}^{TZ} denote a set of observations with $T = t$ and $Z = z$. Under assumption 1, the likelihood is

$$\begin{aligned} L(\alpha, \beta, \lambda|data) &= \prod_{t,z} \prod_{i \in S_{tz}^{TZ}} f\{Y(z_i, M(z_i)) = y_i, M(z_i) = m_i, T(z_i) = t_i|Z = z_i, X = x_i; \beta_{tz}, \alpha_{tz}, \lambda\} \\ &= \prod_{t,z} \prod_{i \in S_{tz}^{TZ}} f\{Y(z_i, m_i) = y_i|M(z_i) = m_i, T(z_i) = t_i, Z = z_i, X = x_i; \beta_{tz}\} \\ &\quad \times f\{M(z_i) = m_i|T = t_i, Z = z_i, X = x_i; \alpha_{tz}\} f\{T(z_i) = t_i|X = x_i; \lambda\}, \end{aligned} \tag{2}$$

where $f(\cdot|\cdot)$ is a conditional probability density function of a random variable of M and Y; α_{tz} and β_{tz} are the vectors of coefficients in the mediator and outcome models, respectively, when $T = t$ and $Z = z$; and λ is the vector of coefficients for treatment receipt status.

From this likelihood, however, it is not possible to model the distributions within the subpopulation of compliers because compliance type is unknown. According to Angrist et al. [23], an individual compliance type can be expressed as the difference in the actual treatment receipt status that would have been observed under the treatment and control conditions. For example, compliers are those who adhere to their assigned treatment (that is, $T_i(1) - T_i(0) = 1$). Always takers are those who receive the treatment regardless of assignment, and never takers are those who do not receive the treatment regardless of assignment (that is, $T_i(1) - T_i(0) = 0$). Defiers are those who do not comply with the treatment protocol and do the opposite of what they are assigned to (that is, $T_i(1) - T_i(0) = -1$). The compliance type for each individual is unknown because subjects are assigned to either the treatment or control condition but not to both (that is, $T_i(1)$ or $T_i(0)$). Therefore, we need to invoke more assumptions to identify the distributions of mediator and outcome by compliance type, which are strong monotonicity and exclusion restriction for never takers.

Assumption 2: Strong Monotonicity [23]. This assumption states that there are no defiers or always takers. Formally, $T_i(0) = 0$ for all i.

In a study where program protocol prohibits subjects in the control group from having access to the intervention, $T_i(0) = 0$ for all i. This implies that we can rule out the possibility of defiers and always takers. After excluding defiers, those who are assigned to the training but did not attend ($T_i(1) = 0$) are uniquely identified as never takers. After excluding always takers, those who are assigned to the training and attended ($T_i(1) = 1$) are uniquely identified as compliers. However, the compliance type for those who are assigned to the control group is still not identified. Therefore, we make the exclusion restriction assumption for never takers.

Assumption 3: Exclusion restriction (ER) for never takers. This assumption was discussed by Little and Yau [16] in the absence of a mediator, and we extend it to a mediation setting. This assumption states that the never-taker distribution in terms of the mediator (or the outcome) is the same under either assignment, given covariates. In formal expression,

$$\begin{aligned} &f(M(z)|P = n, X = x; \alpha_{nz}) = f(M(z')|P = n, X = x; \alpha_{nz'}), \text{ and} \\ &f(Y(z, m)|M = m, P = n, X = x; \beta_{nz}) = f(Y(z', m)|M = m, P = n, X = x; \beta_{nz'}), \end{aligned} \tag{3}$$

for $z \in \{0, 1\}$, $z' = 1 - z$, $m \in \mathcal{M}$, and $x \in \mathcal{X}$, where α_{pz} and β_{pz} are the vector of coefficients in the mediator and outcome models, respectively, when $P = p$ and $Z = z$.

This assumption implies that the direct and indirect effects are allowed only for compliers (but not for never takers), given baseline covariates. This assumption enables us to identify the complier distributions of the mediator and the outcome by fixing the parameters for never-taker distributions at the same value under either assignment, given covariates.

The plausibility of this assumption is often questionable due to psychological effects unless a double-blind design was used to prevent these effects. For example, this assumption would be violated if those who are assigned to but did not receive the job training (i.e., never takers) regretted their failure to take advantage of the intervention and improved job-searching skills by reading a book. Therefore, we develop a sensitivity analysis to assess the effect of violating this assumption for studies in which this assumption might be violated or not plausible, and we demonstrate this sensitivity analysis approach in the JOBS II example.

Under assumptions 1-3, the likelihood can be rewritten as

$$\begin{aligned} &L(\beta, \alpha, \lambda|data) \\ &= \prod_{i \in S_{11}^{TZ}} f\{Y(z_i, m_i) = y_i|M(z_i) = m_i, P = c, X = x_i; \beta_{c1}\} f\{M(z_i) = m_i|P = c, X = x_i; \alpha_{c1}\} \pi_c(x_i; \lambda) \\ &\times \prod_{i \in S_{01}^{TZ}} f\{Y(z_i, m_i) = y_i|M(z_i) = m_i, P = n, X = x_i; \beta_{n1}) f\{M(z_i) = m_i|P = n, X = x_i; \alpha_{n1}\} \pi_n(x_i; \lambda) \\ &\times \prod_{i \in S_{00}^{TZ}} [f\{Y(z_i, m_i) = y_i|M(z) = m_i, P = c, X = x_i; \beta_{c0}) f\{M(z) = m_i|P = c, X = x_i; \alpha_{c0}\} \pi_c(x_i; \lambda) \\ &\quad + f\{Y(z_i, m_i) = y_i|M(z_i) = m_i, P = n, X = x_i; \beta_{n1}\} f\{M(z_i) = m_i|P = n, X = x_i; \alpha_{n1}\} \pi_n(x_i; \lambda)], \end{aligned} \tag{4}$$

where π_p is the probability of $P_i = p$, given covariates. We offer four remarks regarding this likelihood. First, the compliance type for those who are assigned to the treatment is uniquely identified under strong monotonicity. Second, even with strong monotonicity, the compliance type for those who are assigned to the control condition is not uniquely identified. Therefore, the likelihood is expressed as the mixture between complier and never taker distributions, as shown in the last two lines of equation (4). Third, under the exclusion restriction for never takers, parameters for never-taker distributions are fixed to α_{n1} and β_{n1} under either assignment. Fourth, parameters by compliance type for the mediators and outcome models can be consistently estimated from this likelihood although the estimates may not be necessarily given a causal interpretation. Since Assumption 4 (LSI) is not assumed, the estimates are obtained given the correlation between the errors in the mediator and outcome models generated from the data.

Based on the parameters among compliers obtained from the likelihood, we can write the following linear structural equation models (LSEM) with varying coefficients as

$$\begin{aligned} Y_i(z) =& \gamma_{c,i} + \gamma_{cz,i} z + \gamma_{x,i} X_i + e_{c1,i} \\ M_i(z) =& \alpha_{c,i} + \alpha_{cz,i} z + \alpha_{x,i} X_i + e_{c2,i} \\ Y_i(z, m) =& \beta_{c,i} + \beta_{cz,i} z + \beta_{cm,i} m + \beta_{czm,i} zm + \beta_{x,i} X_i + e_{c3,i}, \end{aligned} \tag{5}$$

for $z \in \{0, 1\}$ and $m \in \mathcal{M}$, where $e_{cj,i} \sim N(0, \sigma_{cj})$, in which $j \in \{1, 2, 3\}$. We define $\gamma_c \equiv E(\gamma_{c,i})$, $\gamma_{cz} \equiv E(\gamma_{cz,i})$, $\gamma_x \equiv E(\gamma_{x,i})$, $\alpha_c \equiv E(\alpha_{c,i})$, $\alpha_{cz} \equiv E(\alpha_{cz,i})$, $\alpha_x \equiv E(\alpha_{x,i})$, $\beta_c \equiv E(\beta_{c,i})$, $\beta_{cz} \equiv E(\beta_{cz,i})$, $\beta_{cm} \equiv E(\beta_{cm,i})$, $\beta_{czm} \equiv E(\beta_{czm,i})$, and $\beta_x \equiv E(\beta_{x,i})$ where these terms are the mean parameters of corresponding varying coefficients.

Under assumption 1, we can causally identify the complier average effect of treatment on the mediator (i.e., α_{cz}) and on the outcome (i.e., γ_{cz}). However, the complier average effect of mediator on the outcome (i.e., β_{cm} and β_{czm}) is not causally identified due to possible confounding in the mediator and outcome relationship among compliers. Therefore, we need to additionally invoke the local sequential ignorability assumption.

Assumption 4: Local sequential ignorability (LSI) [7]. This assumption asserts ignorability of the mediator with respect to the potential outcome among compliers, given treatment and pretreatment covariates. This assumption implies that 1) among compliers, there is no pre-treatment confounding between M and Y, given baseline covariates and 2) among compliers, there is no treatment-induced confounding in the M and Y relationship, given baseline covariates. In formal expression,

$$Y_i(z', m) \perp M_i(z) | Z_i = z, P_i = c, X_i = x,$$

for $z \in \{0, 1\}$, $z' = 1 - z$, and $m \in \mathcal{M}$.

Instead of requiring no unmeasured confounding in the $M - Y$ relationship for every participant as in standard causal mediation literature, the local sequential ignorability assumption requires the unconfoundness between the mediator and outcome to be met only for compliers. Although LSI is required for a smaller subset of participants, this assumption is still challenging to meet in practice. Therefore, it is essential to examine the sensitivity of results against this assumption.

Under assumptions 1-4 and given the LSEM, we can identify the CACME and CANDE as $\delta_c(z) = \alpha_{cz} \times (\beta_{cm} + \beta_{czm} z)$ and $\zeta_c(z) = \beta_{cz} + \beta_{czm}(\alpha_c + \alpha_{cz} z)$, respectively[2]. Under assumptions 2 and 3, the mediated and unmediated ITT effects are estimated by multiplying the proportion of compliers to the CACME and CANDE estimate respectively, as $\delta(z) = \delta_c(z) \times \pi_c$ and $\zeta(z) = \zeta_c(z) \times \pi_c$. The proof is provided in Appendix A.

4 Estimation

In this section, we propose a two-stage estimation method based on a joint modeling approach, in which distributional assumptions or additional covariates can be used to reduce the impact of violating some identification assumptions. The proposed estimation method consists of two stages. In the first stage, using joint modeling, we estimate the densities of $f(y|m, x, p; \beta_{pz})$ and $f(m|x, p; \alpha_{pz})$, which depend on parameters β_{pz} and α_{pz}, respectively; and the probability of compliers $\pi_c(x, \lambda)$, which depend on parameters λ. In the second stage, the CACME and CANDE are estimated based on the identification results presented in the previous section. Subsequently, the mediated and unmediated ITT effects are estimated on the basis of the CACME and CANDE estimates, respectively.

First Stage. In the first stage, we use joint modeling, which has been used for estimating the complier-average causal effect (CACE) [16, 23]. We generalize that work by formulating and fitting a model to investigate

2 We assume that X is mean-centered for convenience.

CACME and CANDE. The estimation procedure of this joint modeling approach is based on the expectation-maximization (EM) algorithm, in which the unobserved compliance type for each subject in the control group is treated as missing data. The E-step computes the expected values of sufficient statistics, given data and current estimates, and the M-step maximizes the likelihood shown in equation (4), given the updated sufficient statistics obtained from the E-step. These steps iterate until the estimates of the parameters become stabilized (See [11, 16, 24, 25] for further details on this procedure).

Using the EM algorithm, we can obtain the probability of compliers. We assume that the distribution of P_i given covariates is assumed to have a Bernoulli distribution with a probability of compliance $\pi_c(x_i;\lambda)$, where

$$\pi_c(x_i;\lambda) = \frac{exp(x_i\lambda)}{1+exp(x_i\lambda)}, \text{and} \quad \pi_n(x_i;\lambda) = 1-\pi_c(x_i;\lambda), \quad \text{for } x \in \mathcal{X}, \tag{6}$$

where λ is a vector of logistic regression coefficients. Compared to the previous IV-based method [7, 8], the proposed method provides additional information about the probability of compliers. This information will be used to create a pseudo-population of compliers in order to conduct a sensitivity analysis to violation of LSI.

The conditional probability density functions of random variables M and Y are obtained using the following parametric models. Given that we have two compliance types (compliers and never takers), the mediator and outcome models can be expressed as a mixture distribution between these two compliance types as

$$\begin{aligned} M_i =& N_i\alpha_n + C_i\alpha_c + N_i\alpha_{nz}Z_i + C_i\alpha_{cz}Z_i + \alpha_x X_i + N_i e_{n2,i} + C_i e_{c2,i}, \text{and} \\ Y_i =& N_i\beta_n + C_i\beta_c + N_i\beta_{nz}Z_i + C_i\beta_{cz}Z_i + N_i\beta_{nm}M_i + C_i\beta_{cm}M_i + \\ & N_i\beta_{nzm}Z_iM_i + C_i\beta_{czm}Z_iM_i + \beta_x X_i + N_i e_{n3,i} + C_i e_{c3,i}, \end{aligned} \tag{7}$$

where C_i and N_i are indicators for compliers and never takers, respectively; α_p, and α_{pz} are the mean parameters of the mediator model coefficients; and β_p, β_{pz}, β_{pm}, and β_{pzm} are the mean parameters of the outcome model coefficients when $p \in \{c, n\}$. The error terms for the mediator and outcome models are $e_{p2,i}$ and $e_{p3,i}$ for p $\in \{c, n\}$, respectively. These error terms follow a bivariate normal distribution with a mean of zero and covariance of $\sum_p = \begin{pmatrix} \sigma^2_{p2} & \rho_p\sigma_{p2}\sigma_{p3} \\ \rho_p\sigma_{p3}\sigma_{p3} & \sigma^2_{p3} \end{pmatrix}$, where ρ_p is the correlation between $e_{p2,i}$ and $e_{p3,i}$; and σ_{p2} and σ_{p3} are standard deviations of the two error terms.

To impose ER, we fixed the effect of treatment on the mediator and the outcome among never takers to zero (that is, $\alpha_{nz} = \beta_{nz} = \beta_{nmz} = 0$) thus not allowing a treatment effect among never takers. To impose LSI, we fixed the the covariance among compliers between errors obtained from mediator and outcome models to be zero as $\sum_c = \begin{pmatrix} \sigma^2_{2c} & 0 \\ 0 & \sigma^2_{3c} \end{pmatrix}$.

Second Stage. Based on parameter estimates obtained from the first stage, the CACME and CANDE can be estimated as $\hat{\delta}_c(z) = \hat{\alpha}_{cz} \times (\hat{\beta}_{cm} + \hat{\beta}_{czm}z)$ and $\hat{\zeta}_c(z) = \hat{\beta}_{cz} + \hat{\beta}_{czm}(\hat{\alpha}_c + \hat{\alpha}_{cz}z)$, respectively. The mediated and unmediated ITT effects are estimated by multiplying the proportion of compliers to the CACME and CANDE estimates respectively, as $\hat{\delta}(z) = \hat{\delta}_c(z)\times\hat{\pi}_c$ and $\hat{\zeta}(z) = \hat{\zeta}_c(z)\times\hat{\pi}_c$. Two-stage estimation is known to be inefficient in terms of standard errors [24], so we employed a bootstrap procedure to obtain correct standard errors for mediated and unmediated ITT effects.

5 Simulation Study

The purpose of this simulation study is to 1) assess the performance of the proposed joint modeling method and 2) examine statistical power in the method. In addition, we examine the sensitivity of the estimates to violations of identification assumptions and we explore changes in this sensitivity when the normality assumption is met or when a strong predictor of compliance exists. In the context of CACE, the impact of violation of ER can be mitigated by using additional covariates [11]. However, the reliance on modeling assumptions in case of violating the ER assumption is not well known in a mediation setting. This will be addressed in our simulation study. For simplicity, we focus on the decomposition of $\tau = \delta(1) + \zeta(0)$ in this simulation study.

Data Generation. Our simulation results are based on 1000 replications with the sample sizes of 200, 400, and 600. The assigned treatment Z is a binary variable that takes the value of 1 or 0. The two values of Z are randomly assigned for each observation with the proportion of 0.5. In line with the JOBS II data, we assume that there are two compliance types: compliers and never takers. The compliance type for each observation is determined by a pretreatment covariate following the logistic regression shown in equation (6), in which the pretreatment covariate (X) is generated to follow a standard normal distribution. The true ratio of compliers and never takers is 50 : 50. The mediator (M) and outcome (Y) are generated for each compliance type following the regression shown in equation (7). For simplicity, the average complier treatment effect on the mediator is set to $\alpha_{cz} = 1$, and the average complier mediator effect and its interaction with the treatment on the outcome are set to $\beta_{cm} = \beta_{czm} = 1$, respectively. Thus, the true values of the mediated and unmediated ITT effects are assumed to be $\delta(1) = \zeta(0) = 1$. The true residual variance is 1 for compliers and never takers (i.e., $\sigma^2_{p2} = \sigma^2_{p3} = 1$, where $p \in \{c, n\}$).

One of the important conditions that we vary is the strength of the predictor (X) of compliance. In order to reflect the strong, medium, and small impact of the predictor, we vary the true values of $\lambda_n = \{2.3, 1.2, \text{and } 0.7\}$, which are equivalent to the odds ratios of 0.1, 0.3, and 0.5. This setting is in line with Jo and Stuart [25] and Stuart and Jo [26], which investigated the impact of predictors of compliance on estimating treatment effects conditional on compliance types.

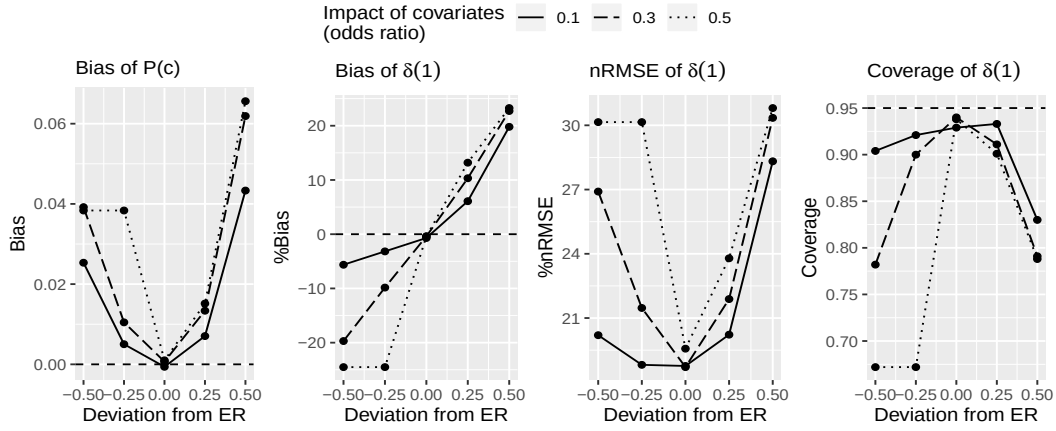
In addition, we generated three types of data in which 1) both mediator and outcome follow a normal distribution, 2) the outcome follows a normal distribution but the mediator does not, and 3) the mediator follows a normal distribution but the outcome does not. For the case in which both mediator and outcome follow a normal distribution, we generated errors for the mediator and outcome from the standard normal distribution. When either the mediator or the outcome violated the normality assumption, we generated two normal distributions that follow $N(-1, 1)$ and $N(3, 1)$ separately and combined them, which generates a bimodal distribution.

In order to create a situation in which the ER is violated, the effect of the treatment on the mediator and outcome among never takers is varied to $\alpha_{nz} = \beta_{nz} = \beta_{nzm} = \{-0.5, 0.25, 0, 0.25, 0.5\}$. Since the residual variance is 1, these deviations of the ER can be considered as standard deviation (SD) units. We chose these ranges of values because the treatment effect on the mediator and the outcome for compliers is set to 1. We set the maximum values of α_{nz} and β_{nz} to half the size of the corresponding complier effect (i.e., α_{cz} and β_{cz}) because never takers did not actually receive the treatment. In the analytical model in which we estimate mediated and unmediated ITT effects using the generated model, we assumed the ER and LSI. The rest of the parameters are specified as follows: $\alpha_x = \beta_x = \beta_{nm} = 1$.

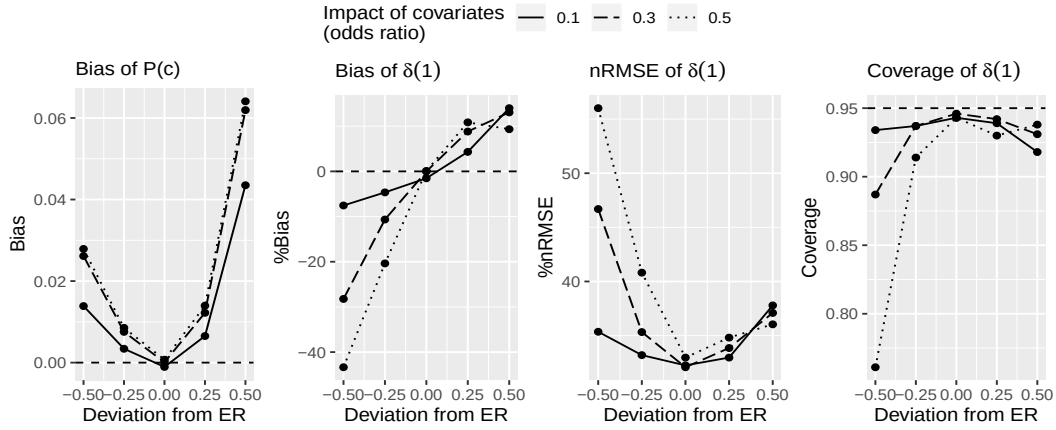
To assess the performance of the proposed method in various settings, we first examine the bias of the probability of compliers. This is crucial because this information will be used for sensitivity analysis in the later section. Then, we examine the percent bias (%bias), the percent normalized root mean square errors (%nRMSE), and coverage rate for the mediated and unmediated ITT effects to summarize our simulation results. The %bias measures the difference between the average of estimates and the true value relative to the true value. The %nRMSE measures the square root of the average of squared difference between the estimate and the true value relative to the true value. The coverage rate is defined as the proportion of replications where the true value is covered by the 95% confidence interval out of 1000 replications. To examine the statistical power in the method, we calculate the power under different sample sizes and distributions of the mediator and the outcome. The power is defined as the proportion of replications where the effect estimate is significantly different from zero ($\alpha = 0.05$) out of 1000 replications.

Simulation Results. The simulation results are summarized in Figures 1a-1c. The top plots present the bias of P(c) as well as %bias, %nRMSE, and 95% confidence interval coverage rates of $\delta(1)$ with a normally distributed mediator and outcome. The middle and bottom plots present the same quantities with a non-normally distributed mediator and a non-normally distributed outcome, respectively.

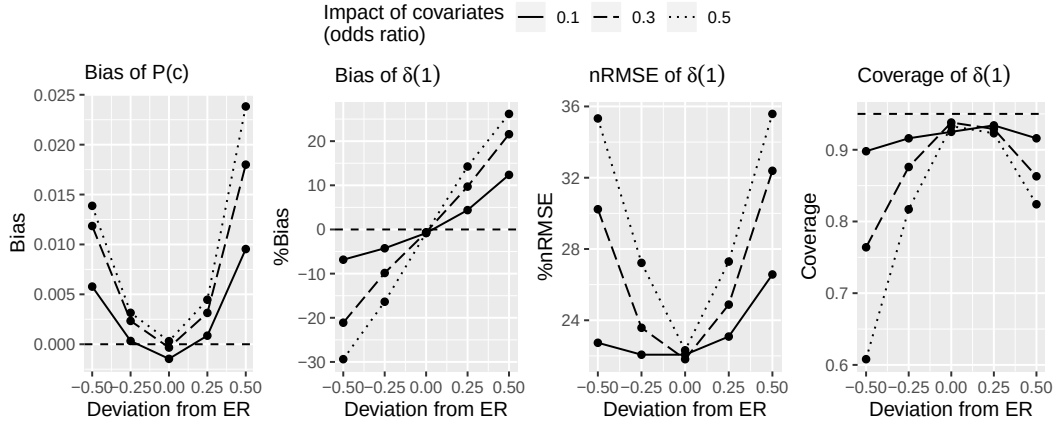
The estimates of $P(c)$ under the deviation of zero from ER are unbiased regardless of whether or not normality holds. The estimates of $P(c)$ tends to be biased when the data deviate from ER although the bias is relatively small. Even when the ER is violated by the 0.5 S.D, the bias is less than 0.07 with the small impact of covariates (OR of 0.5).



(a) Normal mediator and Normal outcome



(b) Nonnormal mediator and normal outcome



(c) Normal mediator and nonnormal outcome

Figure 1: Sensitivity of the estimates ($P(c)$ and $\delta(1)$) when the ER is violated

Note. 1) True effect: $P(c) = 0.5$ *and* $\delta(1) = 1$*, sample size: 600. 2) The results for* $\zeta(0)$ *are similar to the ones for* $\delta(1)$*. Given the similarity, we present the results for* $\zeta(0)$ *in the e-Appendix.*

Not surprisingly, the estimates of $\delta(1)$ under the deviation of zero from ER are unbiased, and the 95% coverage rate reaches the nominal level even when normality is not met. Although the bias is almost zero regardless of whether or not normality holds, the nRMSE tends to be large if the normality does not hold for either the mediator or outcome distribution. When normality holds, the nRMSE is less than 19% with a strong predictor of compliance. With the same setting, the nRMSEs are 32% and 22%, respectively, when normality is violated for the mediator and outcome. This indicates that standard errors tend to be large if normality is violated for the mediator or outcome distribution when all identification assumptions are met.

As expected, the effect estimates of $\delta(1)$ become biased when the data deviate from ER regardless of whether or not normality holds. If normality does not hold, the nRMSE becomes larger. When normality is met for both the mediator and outcome and the ER is violated by the 0.25 S.D, the bias is less than 10% and the nRMSE is 21% with the medium impact of covariate (OR of 0.3) (Figure 1a). With the same setting but when the normality is violated for the mediator, the bias is same as 10% but the nRMSE is 35% (Figure 1b). The nRMSE is also larger (24%) when normality is violated for the outcome (Figure 1c).

Also, the bias is smaller in cases with a stronger predictor of compliance. In cases with a covariate with a strong effect size (OR of 0.1), the biases are about half what they are with a covariate with a medium effect size. In the same setting (with a covariate with a medium effect size), the bias is less than 5% when normality is met (Figure 1a), and the bias is same when normality is not met for the mediator and outcome (Figure 1b and Figure 1c).

In summary, when normality is met and a strong predictor of compliance exists, the bias due to the relatively smaller deviation from ER (one fourth of the complier average effect) may be negligible given that the bias is less than 5% of the true value. However, when normality is violated for either the mediator or outcome, the nRMSE becomes larger, which will result in large standard errors.

The statistical power for the mediated ITT effect ($\delta(1)$) under different sample sizes and distributions of the mediator and the oucome is shown in Figure 2. The figure illustrates that statistical power to detect the mediated ITT effect is greatly influenced by whether or not normality holds (Figure 2a). For example, if normality holds, statistical power is greater than 0.8 regardless of whether strong or small impact of covariates were used. If normality in the mediator does not hold, statistical power ranges from 0.4 (sample size of 200) to 0.9 (sample size of 600) (Figure 2b). Statistical power does not appear to be different if normality in the outcome does not hold. In summary, statistical power to detect the mediated ITT effect reaches a desirable level if normality holds even with a small sample size (N=200).

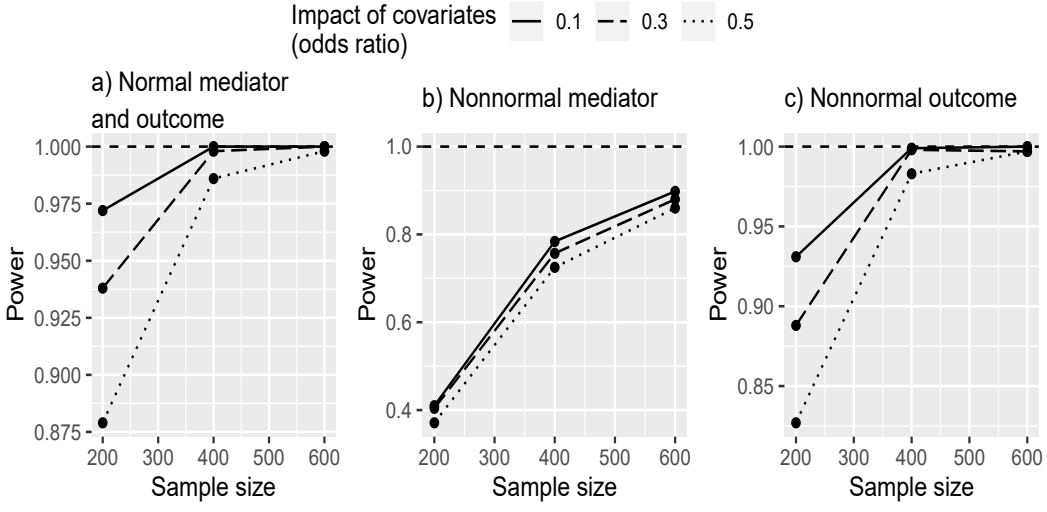


Figure 2: Statistical power of $\delta(1)$

Note. 1) True effect: $\delta(1) = 1$. *2) The results for* $\zeta(0)$ *are similar to the ones for* $\delta(1)$. *Given the similarity, we present the results for* $\zeta(0)$ *in the e-Appendix.*

6 Joint Modeling-based Sensitivity Analysis

In this section, we propose sensitivity analyses that can assess the validity of results to a possible violation of ER for never takers and LSI. We focus on sensitivity analyses with respect to these two assumptions because the identification of the mediated and unmediated ITT effects crucially rely on them. The proposed sensitivity analyses can be employed when investigating a mediating mechanism with any randomized experiments that suffer from treatment noncompliance, in which access to the treatment is prohibited for those who are assigned to the control condition.

Sensitivity analysis for ER for never takers. The ER assumption for never takers requires that there is no effect of the assigned treatment on the mediator (or on the outcome) and, hence, the treatment effect is zero for never takers. As shown in our simulation study, the impact of violation of ER is smaller if there is a strong predictor of compliance and the normality assumption is met. However, the validity of results may still be questioned if these modeling assumptions do not hold and/or the degree to which ER is violated could be severe.

Although many sensitivity analyses have been developed for ER, very few sensitivity analyses are available for a mediation setting. For example, an alternative sensitivity analysis technique has been developed by Park and Kürüm [8] on the basis of the IV-based method. This technique involves specifying a ratio of the predicted outcome (mediator) value given Z=1 to the predicted outcome (mediator) value given Z=0 among never takers relative to a corresponding ratio among compliers. This approach is similar to our proposed sensitivity analysis technique. However, an IV-based sensitivity analysis technique does not have any means to decrease the impact of violating ER and thus provides a relatively large range of estimates for the change in the sensitivity parameters. In contrast, our proposed sensitivity analysis technique provides a smaller range of results for the change in the sensitivity parameters when normality is met or additional covariates exist.

If ER is violated, we can no longer assume that the distributions of mediator and outcome among never takers are the same under either assignment. Therefore, our sensitivity parameters are based on expected difference in the mediator and outcome distributions among never takers between those who are assigned to the treatment and control conditions. Specifically, let ϵ_m be the expected difference in the mediator value among never takers between those who are assigned to the treatment and control conditions, given covariates. Let $\epsilon_{y1} + \epsilon_{y2}m$ be the expected difference in the outcome value among never takers between those who are assigned to the treatment and control conditions, given covariates for every $m \in \mathcal{M}$. Formally,

$$\begin{aligned} \epsilon_m &= E[M(1) - M(0)|P = n, X = x] \text{ and} \\ \epsilon_{y1} + \epsilon_{y2}m &= E[Y(1, m) - Y(0, m)|P = n, X = x], \text{ for all m} \in \mathcal{M} \text{ and x} \in \mathcal{X}. \end{aligned}$$

Suppose that ER is violated but other assumptions are met. Then, given particular values of ϵ_m, ϵ_{y1}, and ϵ_{y2}, the mediated and unmediated ITT effects are identified, respectively, as

$$\begin{aligned} \delta(z) &= \tilde{\pi}_c\{\tilde{\alpha}_{cz} \times (\tilde{\beta}_{cm} + \tilde{\beta}_{czm}z)\} + \tilde{\pi}_n\{\epsilon_m \times (\tilde{\beta}_{nm} + \epsilon_{y2}z)\} \text{and} \\ \zeta(z) &= \tilde{\pi}_c\{\tilde{\beta}_{cz} + \tilde{\beta}_{czm}(\tilde{\alpha}_c + \tilde{\alpha}_{cz}z)\} + \tilde{\pi}_n\{\epsilon_{y1} + \epsilon_{y2}(\tilde{\alpha}_n + \epsilon_m z)\}, \end{aligned} \tag{8}$$

where $\tilde{\pi}_c, \tilde{\pi}_n, \tilde{\alpha}_n, \tilde{\alpha}_{cz}, \tilde{\beta}_{cz}, \tilde{\beta}_{cm}, \tilde{\beta}_{nm}$, and $\tilde{\beta}_{czm}$ are obtained from the maximized complete-data likelihood given particular values of ϵ_m, ϵ_{y1}, and ϵ_{y2}. The proof of this result is provided in Appendix B.

Sensitivity analysis for LSI. The first part of assumption 2 states that among compliers, there is no unmeasured confounding in the mediator and outcome relationship given baseline covariates. In many cases, the more covariates we observe, the more plausible the assumption is. However, we may not be able to measure all the covariates to remove confounding between the mediator and outcome among compliers. Many studies have addressed this issue of unmeasured mediator and outcome confounding when perfect compliance was assumed (e.g., [1, 2, 29, 30]). However, very few studies have addressed this issue when perfect compliance was not assumed. The previous study based on the IV-based method [8] examined the sensitivity of the results to the violation of LSI by assuming the worst case scenario. In this study, we provide a systematic sensitivity analysis technique that can be used for all possible scenarios of unobserved confounding between the mediator and the outcome.

Imai et al. [1] identified the ACME given a value of the correlation between two error terms obtained from the mediator and outcome models when perfect compliance to the treatment was assumed. However, we cannot apply this approach in the presence of treatment noncompliance because the previously developed IV-based method does not provide any information on individual compliance status. Unlike the IV-based method, the joint modeling method provides the probability of an individual being a complier and this information can be used to assess the sensitivity to a possible violation of LSI.

Development of the sensitivity analysis for LSI relies on using an individual's probability of being a complier as a weight to create a pseudo-population of compliers. The term "pseudo-population" is often used in the field of survey sampling that mimics the original population by replicating sample units based on the probability of being sampled. Here, we define pseudo-population as the original population of compliers, which is partially observed. For the treatment group, those who attended the job training will be assigned a weight of 1, and those who did not attend the training will be assigned a weight of 0 because the probability of being a complier is measured without any error under strong monotonicity. For the control group, we cannot uniquely identify compliance types for each individuals because they are not observed; yet, we can create a weighted sample based on the probability of compliers. Each individual will be assigned a weight of $\pi_c(x)/\pi_c$, where $\pi_c(x)$ is the probability of being a complier given pretreatment covariates from equation (6) and π_c is the proportion of compliers. By giving a weight of $\pi_c(x)$, those who have a high chance of being a complier will be given more weight and those who have low chance of being a complier will be given less weight. By dividing the weight by the proportion of compliers (π_c), we can recover the total sample size of the control group. For example, an individual in the control group with the probability of compliers of $\pi_c(x) = 0.8$ will be replicated $\frac{0.8}{0.5} = 1.6$ times (when $\pi_c = 0.5$), delivering 1.6 clones for the pseudo-population. The same logic was used in Ding and Lu [27].

Based on this pseudo-population of compliers, the sensitivity of the results will be examined across the varying values of the correlation between the errors obtained from the mediator and the outcome models as in Imai et al. [1].

Suppose that LSI is violated, but the other assumptions are met. Let the correlation between the error terms from the mediator and outcome models fitted among the pseudo-population of compliers be denoted as ρ_c. Then, given a value of ρ_c, the mediated and unmediated ITT effects are identified as

$$\begin{aligned}\delta(z) =& \pi_c \alpha_{cz} \left\{ \frac{\sigma_{c1}}{\sigma_{c2}} \left(\tilde{\rho}_{cz} - \rho_c \sqrt{\frac{1-\tilde{\rho}_{cz}^2}{1-\rho_c^2}} \right) \right\} \text{ and} \\ \zeta(z') =& ITT - \delta(z),\end{aligned}$$

where $z' = 1 - z$ for $z \in \{0, 1\}$. The term $\tilde{\rho}_{cz}$ is the correlation between the error terms $\epsilon_{c1,i}$ and $\epsilon_{c2,i}$ (from equations (5)) when $Z_i = z$; and σ_{c1} and σ_{c2} are standard deviations of the error terms, respectively, which are fixed to be constant across the values of Z_i. The proof of this result is given in Appendix C.

7 Application to Jobs II Study

Our question of interest is whether the effect of the JOBS II intervention on reducing job-seekers' depression is transmitted through increased sense of mastery. To answer this question, we estimate the mediated and unmediated portion of the ITT effect via sense of mastery using the proposed joint modeling method. We then show how the sensitivity of the estimated mediated and unmediated ITT effects to the violation of ER and LSI can be investigated using the results from the previous section.

Results. Table 2 shows the estimates of the mediated and unmediated ITT effects given assumptions 1-4. The difference in the outcome value between treatment and control subjects of -0.07 estimates the ITT estimand. The mediated portion of the ITT effect for treated and controlled conditions are negatively significant as -0.03 and -0.04, which occupy the 43.1% and 61.1% of the ITT effect, respectively. In contrast, the unmediated ITT effects for the treated and controlled conditions are not significant. This implies that the mediating

mechanism through which the job training impacts job-seekers' depression includes enhanced sense of mastery under assumptions 1-4.

Table 2: Estimates of the mediated and unmediated ITT effects

	Compliers effects				***ITT effects***		
Parameter	Est.	S.E.	P-Value	Parameter	Est.	S.E	P-Value
$\delta_c(1)$	−0.057	0.020	0.005	$\delta(1)$	−0.031	0.011	0.005
$\delta_c(0)$	−0.081	0.032	0.012	$\delta(0)$	−0.044	0.017	0.012
$\zeta_c(1)$	−0.053	0.054	0.324	$\zeta(1)$	−0.029	0.029	0.322
$\zeta_c(0)$	−0.077	0.066	0.244	$\zeta(0)$	−0.041	0.035	0.242
CACE	−0.134	0.069	0.052	ITT	−0.072	0.037	0.050

Note. Est.=estimates; S.E.=standard errors; CACE=compliers average causal effect; ITT= intention-to-treat effect

However, for a valid causal interpretation of the estimates, it is crucial to examine the sensitivity of the estimates to a violation of the identification assumptions. We require randomization, strong monotonicity, ER for never takers, and LSI. Randomization is satisfied because job training is assigned randomly. Strong monotonicity is also guaranteed to be met because program protocol prohibits subjects in the control group to have access to the job search seminar. However, ER for never takers is controversial. ER might be violated due to psychological effects. For example, some participants who were assigned to the job training but failed to attend (never takers) may feel more depressed, which violates ER. Another controversial assumption is LSI because there could be unobserved confounding between sense of mastery and depression given the treatment level and pretreatment covariates. Therefore, we conduct sensitivity analyses for ER and LSI.

Sensitivity analysis for ER. In our study, we assume that this psychological effect is unlikely to be large because never takers did not actually attend the training. Hence, we limit the violation of ER to be at most half the size of the complier average effect. The sensitivity parameters of ϵ_m, ϵ_{y1}, and ϵ_{y2} were given a value of one fourth (0.25) or half (0.5) the size of the corresponding complier average effect.

Table 3 shows the adjusted estimates of the mediated ITT effect by varying values of ϵ_m, ϵ_{y1}, and ϵ_{y2}. It appears that the mediated ITT effect for those who are assigned to the treatment ($\delta(1)$) is robust to the violation of ER with respect to both mediator and outcome. For example, the mediated ITT effect for those who are assigned to the treatment is still negative and significant when the treatment effect among never takers is half the size of the corresponding complier average effect for either mediator and outcome model ($\epsilon_m = 0.5$, or $\epsilon_{y1} = \epsilon_{y2} = 0.5$). In contrast, the mediated ITT effect for those who are assigned to the control ($\delta(0)$) is relatively vulnerable to the violation of ER with respect to both mediator and outcome. The mediated ITT effect for those who are assigned to the control is still negative but loses its significance when the treatment effect among never takers is the one fourth of the size of the complier average effect for either mediator and outcome model ($\epsilon_m = 0.25$ or $\epsilon_{y1} = \epsilon_{y2} = 0.25$).

Sensitivity analysis for LSI. We next examine whether our conclusion about the mediated ITT effect changes if there are unmeasured pre-treatment covariates between the mediators and outcome among compliers while assuming other assumptions are satisfied.

Figures 3a and 3b show the sensitivity of the mediated ITT effect estimates under treatment and control conditions, respectively, to the violation of the LSI while assuming other assumptions are satisfied. These figures show how the change in ρ_c affects the mediated ITT effect estimates. The sensitivity parameter ρ_c represents the correlation among compliers between the errors obtained from the mediator and outcome model, and a non-zero value of ρ_c indicates the existence of unmeasured confounding among compliers in the mediator and outcome relationship. The bold line in the middle represents the changed mediated ITT effect estimates depending on the value of ρ_c, and the solid lines represent the lower and upper values of 95% confidence intervals.

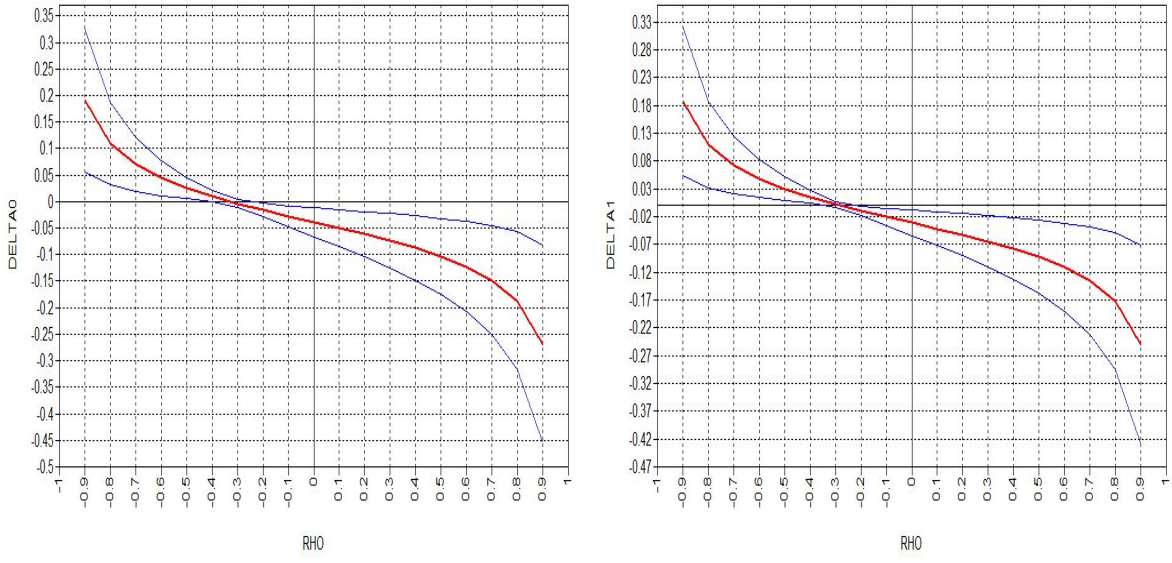
Table 3: Sensitivity of estimates with the deviation from the ER

	ϵ_Y=0× c.e.		ϵ_Y=0.25× c.e.		ϵ_Y=0.5× c.e.	
ϵ_M	$\delta(1)$	$\delta(0)$	$\delta(1)$	$\delta(0)$	$\delta(1)$	$\delta(0)$
0× c.e.	-0.031**	-0.044*	-0.026**	-0.033*	-0.025*	-0.026
	(0.011)	(0.017)	(0.009)	(0.016)	(0.010)	(0.015)
0.25× c.e.	-0.032**	-0.031*	-0.031**	-0.026	-0.030**	-0.018
	(0.009)	(0.014)	(0.010)	(0.016)	(0.010)	(0.013)
0.5 × c.e.	-0.037**	-0.025	-0.037**	-0.022	-0.036**	-0.014
	(0.010)	(0.014)	(0.011)	(0.018)	(0.011)	(0.013)

Note. 1) Standard errors are in parentheses. 2) c.e.=corresponding complier-average effect. 3) **: p<0.01, and *: p<0.05

As shown in Figure 3a, the mediated ITT effect estimate for those who are assigned to the control will be close to zero if ρ_c is -0.4. However, the 95% confidence interval of the effect estimate will cover zero with a smaller value of ρ_c, which is -0.3. This value of ρ_c is equivalent to the amount of confounding that explains the variances of mediator and outcome, for example, by 25% and 36%, respectively[3]. This amount of confounding can be considered very large given that the strongest covariate (i.e., pre-measured depression) in the existing model explains the variances of mediator and outcome by 5.8% and 16.8%, respectively.

As shown in Figure 3b, the mediated ITT effect estimate for those who are assigned to the treatment will be zero if ρ_c is -0.3. However, the 95% confidence interval of the effect estimate will cover zero if ρ_c is -0.2, which is equivalent to the amount of confounding that explains the variances of mediator and outcome, for example, by 16% and 25%, respectively[4]. This amount of confounding can still be considered very large.



(a) Mediated ITT effect under the control condition **(b)** Mediated ITT effect under the treatment condition

Figure 3: Sensitivity of the mediated or unmediated portions of ITT effects to the violation of LSI

3 This is because $-0.3 = -0.5 \times 0.6$, and the corresponding $R^2 = (-0.5)^2 = 0.25$, and $R^2 = 0.6^2 = 0.36$, respectively.
4 This is because $-0.2 = -0.5 \times 0.4$, and the corresponding $R^2 = (-0.5)^2 = 0.25$, and $R^2 = 0.4^2 = 0.16$, respectively.

In summary, the significant mediation effect for those who are treated is robust to a potential violation of ER and it is robust to a potential violation of LSI while other assumptions are assumed to be satisfied. However, the mediation effect for those who are controlled may lose its significance if the effect of never takers are as large as one fourth of the corresponding complier-average effect; however, it is robust to a potential violation of the LSI when other assumptions are met. For these sensitivity analyses, we used Mplus [28]. Annotated Mplus code can be found in the online appendix.

8 Summary and Conclusions

In this article, we proposed a two-stage joint modeling method that combines a mediation analysis with a mixture analysis to conduct causal mediation analysis in the presence of treatment noncompliance. On the basis of the mediation analysis, the mediator and outcome models can be specified and estimated. On the basis of the mixture analysis, the compliance-specific parameters can be specified and estimated, considering the mixed distributions of compliers and never takers.

One useful feature of the joint modeling method is that it is conducive to conducting sensitivity analyses to the violation of identification assumptions. In this study, we offer a systematic sensitivity analysis that addresses the two identification assumptions (the ER and LSI), which was not available in the previous instrumental variables approach. Sensitivity analysis is an important component in any causal inference framework because many identification assumptions are not verifiable with empirical data. The proposed sensitivity analysis can be easily used by applied researchers to test their results against violation of these identification assumptions.

Another useful feature of the joint modeling method is that we can invoke modeling assumptions such as normality or the existence of strong predictors of compliance that can decrease the sensitivity of violating some identification assumptions such as the ER. In the context of CACE, including a strong predictor of compliance can decrease the bias due to violation of the ER and increase precision of the estimates. We demonstrate in our simulation study that these benefits also apply when estimating the mediated ITT effect. Normality also plays a role in estimating compliance type more precisely, and the simulation study suggests that estimating compliance type is more affected by the outcome distribution than the mediator distribution.

However, these benefits come with a cost. From the simulation study, we observe a large variance in the estimates even when all identification assumptions are met if normality is violated. If normality is violated, advantages of the proposed joint modeling method disappear. In this case, one should consider using a propensity score method, suggested by Jo and Stuart [25], Ding and Lu [27], which relies only on pre-treatment covariates to identify unobserved compliance types and, thus, reduces reliance on particular parametric assumption such as normality.

In this article, we introduced a two-stage joint modeling method to estimate the mediated and unmediated portion of the ITT effect and demonstrated the benefits of employing this method through simulation and case studies. A next logical step for future research is to compare relative performance between the proposed joint modeling method and the previous approach using the IV method [7, 8]. Unlike the joint modeling method, the IV method does not require modeling assumptions and hence, the identification of the mediated ITT effect relies only on identification assumptions. Comparing relative performance when modeling assumptions are met or not met would be an interesting subject for future study.

References

[1] K. Imai, L. Keele, and D Tingley. A general approach to causal mediation analysis. *Psychological Methods*, 15:309–334, 2010.

[2] K. Imai and T Yamamoto. Identification and sensitivity analysis for multiple causal mechanisms: Revisiting evidence from framing esperiments. *Political Analysis*, 21:141–171, 2013.

[3] T VanderWeele. *Explanation in causal inference: methods for mediation and interaction*. Oxford University Press, 2015.
[4] Johan Steen, Tom Loeys, Beatrijs Moerkerke, and Stijn Vansteelandt. Flexible mediation analysis with multiple mediators. *American journal of epidemiology*, 186(2):184–193, 2017.
[5] Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.
[6] J Pearl. The causal mediation formula—a guide to the assessment of pathways and mechanisms. *Prevention Science*, 13(4): 426–436, 2012.
[7] T Yamamoto. Identification and estimation of causal mediation effects with treatment noncompliance. Unpublished manuscript, 2014.
[8] Soojin Park and Esra Kürüm. Causal mediation analysis with multiple mediators in the presence of treatment noncompliance. *Statistics in medicine*, 37(11):1810–1829, 2018.
[9] JL Zhang, DB Rubin, and F Mealli. Likelihood-based analysis of causal effects via principal stratification: new approach to evaluating job-training programs. *Journal of the American Statistical Association*, 104:166–176, 2009.
[10] Peng Ding, Zhi Geng, Wei Yan, and Xiao-Hua Zhou. Identifiability and estimation of causal effects by principal stratification with outcomes truncated by death. *Journal of the American Statistical Association*, 106(496):1578–1591, 2011.
[11] Booil Jo, Tihomir Asparouhov, Bengt O Muthén, Nicholas S Ialongo, and C Hendricks Brown. Cluster randomized trials with treatment noncompliance. *Psychological methods*, 13(1):1–18, 2008.
[12] Amiram D Vinokur, Michelle Van Ryn, Edward M Gramlich, and Richard H Price. Long-term follow-up and benefit-cost analysis of the jobs program: a preventive intervention for the unemployed. *Journal of Applied Psychology*, 76(2):213–219, 1991.
[13] R. Catalano and C. D Dooley. Economic predictors of depressed mood and stressful life events in a metropolitan community. *Journal of Health and Social Behavior*, 18:292–307, 1977.
[14] Sidney Cobb and Stanislav V Kasl. *Termination; the consequences of job loss*, volume 77. NIOSH, 1977.
[15] R Catalano. The health effects of economic insecurity. *American Journal of Public Health*, 81(9):1148–1152, 1991.
[16] R. J. Little and L. H. Y Yau. Statistical techniques for analyzing data from prevention trials: Treatment of no-shows using rubin's causal model. *Psychological Methods*, 3(2):147–159, 1998.
[17] C. E. Frangakis and D. B Rubin. Principal stratification in causal inference. *Biometrics*, 58(1):21–29, 2002.
[18] L. H. Y Yau and R. J Little. Inference for the complier-average causal effect from longitudinal data subject to noncompliance and missing data, with application to a job training assessment for the unemployed. *Journal of the American Statistical Association*, 96(456):1232–1244, 2001.
[19] R. D. Caplan, A. D. Vinokur, R. H. Price, and M Van Ryn. Job seeking, reemployment, and mental health: a randomized field experiment in coping with job loss. *Journal of applied psychology*, 74(5):759–769, 1989.
[20] Michael E Sobel and Bengt Muthén. Compliance mixture modelling with a zero-effect complier class and missing data. *Biometrics*, 68(4):1037–1045, 2012.
[21] R. H. Price, M. Van Ryn, and A. D Vinokur. Impact of a preventive job search intervention on the likelihood of depression among the unemployed. *Journal of Health and Social Behavior*, 33:158–167, 1992.
[22] L. R. Derogatis, R. S. Lipman, K. Rickels, E. H. Uhlenhuth, and L Covi. The hopkins symptom checklist (hscl): A self-report symptom inventory. *Systems Research and Behavioral Science*, 19(1):1–15, 1974.
[23] J. D Angrist, G. W. Imbens, and D. B Rubin. Identification of causal effects using instrumental variables. *Journal of the American statistical Association*, 91(434):444–455, 1996.
[24] Edward Bein, Jonah Deutsch, Guanglei Hong, Kristin E Porter, Xu Qin, and Cheng Yang. Two-step estimation in ratio-of-mediator-probability weighted causal mediation analysis. *Statistics in medicine*, 37(8):1304–1324, 2018.
[25] Booil Jo and Elizabeth A Stuart. On the use of propensity scores in principal causal effect estimation. *Statistics in medicine*, 28(23):2857–2875, 2009.
[26] Elizabeth A Stuart and Booil Jo. Assessing the sensitivity of methods for estimating principal causal effects. *Statistical methods in medical research*, 24(6):657–674, 2015.
[27] Peng Ding and Jiannan Lu. Principal stratification analysis using principal scores. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(3):757–777, 2017.
[28] Linda K Muthén, Bengt O Muthén, et al. Mplus (version 5.1). *Los Angeles, CA: Muthén & Muthén*, 2008.
[29] Guanglei Hong, Xu Qin, and Fan Yang. Weighting-based sensitivity analysis in causal mediation studies. *Journal of Educational and Behavioral Statistics*, 43(1):32–56, 2018.
[30] T. J VanderWeele. Bias formulas for sensitivity analysis for direct and indirect effects. *Epidemiology (Cambridge, Mass.)*, 21(4):540–551, 2010.

9 Appendix A: Identification of $\delta(z)$ and $\zeta(z)$

The mediated and unmediated portion of ITT effects are identified on the basis of CACME and CANDE, respectively. Therefore, we first identify the CACME and CANDE. From equations (5), note that parameters in the second line of equations (5) are identified under randomization because $e_{c1}(z) \perp Z|X = x$ holds. The parameters in the third line of equations (5) are identified under randomization and LSI because $e_{c3}(z, m) \perp Z|X = x$ and $e_{c3}(z, m) \perp M|Z = z', X = x, P = c$. Given these parameters, the CACME is identified as

$$\begin{aligned}
\delta_c(z) =& E[Y_i(z, M_i(1)) - Y_i(z, M_i(0))|P_i = c] \\
=& E[\beta_{c,i} + \beta_{cz,i}z + \beta_{cm,i}\{M_i(1)|P_i = c\} + \beta_{czm,i}z\{M_i(1)|P_i = c\} \\
& - \beta_{c,i} - \beta_{cz,i}z - \beta_{cm,i}\{M_i(0)|P_i = c\} - \beta_{czm,i}z\{M_i(0)|P_i = c\}], \\
=& E[(\beta_{cm,i} + \beta_{czm,i}z)\{M_i(1) - M_i(0)|P_i = c\}] \\
=& E[(\beta_{cm,i} + \beta_{czm,i}z)\alpha_{cz,i}] \\
=& (\beta_{cm} + \beta_{czm}z)\alpha_{cz}
\end{aligned} \tag{A-1}$$

The first equality is from the definition of CACME. The second equality holds after incorporating the outcome model (i.e., the third line of equations (5)). The fourth equality holds after incorporating the mediator model (i.e., second line of equations (5)). The fifth equality holds due to LSI (assumption 2). Specifically, given compliers, $Y_i(z, m) - Y_i(z, m') = \beta_{cm,i} + \beta_{czm,i}z$ is independent from $M_i(z)$ for any $z \in \{0, 1\}$ as in LSI.

Likewise, the CANDE is identified as

$$\begin{aligned}
\zeta_c(z) =& E[Y_i(1, M_i(z)) - Y_i(0, M_i(z))|P_i = c] \\
=& E[\beta_{c,i} + \beta_{cz,i} + \beta_{cm,i}\{M_i(z)|P_i = c\} + \beta_{czm,i}\{M_i(z)|P_i = c\} - \beta_{c,i} - \beta_{cm,i}\{M_i(z)|P_i = c\}] \\
=& E[\beta_{cz,i} + \beta_{czm,i}\{M_i(z)|P_i = c\}] \\
=& E[\beta_{cz,i} + \beta_{czm,i}(\alpha_{c,i} + \alpha_{cz,i}z)] \\
=& \beta_{cz} + \beta_{czm}(\alpha_c + \alpha_{cz}z).
\end{aligned} \tag{A-2}$$

The first equality is from the definition of CANDE. The second equality holds after incorporating the outcome model (i.e., third line of equations (5)). The fourth equality holds after incorporating the mediator model (i.e., second line of equations (5)). The fifth equality holds due to LSI (assumption 2). Specifically, given compliers, $Y_i(1, m) - Y_i(0, m) = \beta_{czm,i}z$ is independent from $M_i(z)$ for any $z \in \{0, 1\}$ as in LSI.

Next, we identify the the mediated and unmediated ITT effects on the basis of CACME and CANDE as

$$\begin{aligned}
\delta(z) =& \delta_c(z)\pi_c + \delta_n(z)\pi_n = \delta_c(z)\pi_c, \text{ and} \\
\zeta(z) =& \zeta_c(z)\pi_c + \zeta_n(z)\pi_n = \zeta_c(z)\pi_c,
\end{aligned}$$

where δ_n and ζ_n are ACME and average natural direct effects among never takers, respectively. The first equality holds because of strong monotonicity. The second equality holds because of ER for never takers. This completes the proof.

10 Appendix B: Sensitivity analysis for ER

Our sensitivity parameters depend on the deviation from the ER as

$$\begin{aligned}\epsilon_m =& E[M(1) - M(0)|P = n, X = x] \text{ and}\\ \epsilon_{y1} + \epsilon_{y2} m =& E[Y(1, m) - Y(0, m)|P = n, X = x] \text{ for all m} \in \mathcal{M}, \text{and x} \in \mathcal{X}.\end{aligned}$$

This implies that among never takers, the parameter for the treatment on the mediator is fixed to $\alpha_{nz} = \epsilon_m$, and the parameters for the treatment on the outcome among never takers are fixed to $\beta_{nz} = \epsilon_{y1}$ and $\beta_{nzm} = \epsilon_{y2}$. Given particular values of ϵ_m, ϵ_{y1}, and ϵ_{y2}, we can rewrite linear structural models as

$$\begin{aligned}M_i(z) =& N_i\tilde{\alpha}_{n,i} + C_i\tilde{\alpha}_{c,i} + N_i\epsilon_m z + C_i\tilde{\alpha}_{cz,i}z + \tilde{\alpha}_{x,i}X_i + N_i e_{n2,i} + C_i e_{c2,i}, \text{ and}\\ Y_i(z, m) =& N_i\tilde{\beta}_{n,i} + C_i\tilde{\beta}_{c,i} + N_i\epsilon_{y1} z + C_i\tilde{\beta}_{cz,i}z + N_i\tilde{\beta}_{nm,i}m + C_i\tilde{\beta}_{cm,i}m +\\ & N_i\epsilon_{y2} zm + C_i\tilde{\beta}_{czm,i}zm + \tilde{\beta}_{x,i}X_i + N_i e_{n3,i} + C_i e_{c3,i},\end{aligned} \tag{A-3}$$

where $\tilde{\alpha}_{p,i}, \tilde{\beta}_{p,i}, \tilde{\beta}_{pm,i}$, and $\tilde{\beta}_{x,i}$ for $p \in \{c, n\}$ are obtained from the maximized complete-data likelihood given particular values of $\epsilon_m, \epsilon_{y1}$, and ϵ_{y2}. We define $\tilde{\alpha}_p \equiv E(\tilde{\alpha}_{p,i})$, $\tilde{\alpha}_{pz} \equiv E(\tilde{\alpha}_{pz,i})$, $\tilde{\alpha}_x \equiv E(\tilde{\alpha}_{x,i})$, $\tilde{\beta}_p \equiv E(\tilde{\beta}_{p,i})$, $\tilde{\beta}_{pz} \equiv E(\tilde{\beta}_{pz,i})$, $\tilde{\beta}_{pzm} \equiv E(\tilde{\beta}_{pzm,i})$, and $\tilde{\beta}_x \equiv E(\tilde{\beta}_{x,i})$ for $p \in \{c, n\}$.

Based on equations (A-3), the ACME among never takers given particular values of ϵ_m, ϵ_{y1}, and ϵ_{y2} ($\delta_n^\epsilon(z)$) is identified as

$$\begin{aligned}\delta_n^\epsilon(z) =& E[Y_i(z, M_i(1)) - Y_i(z, M_i(0))|P_i = n]\\ =& E[\tilde{\beta}_{n,i} + \epsilon_{y1} z + \tilde{\beta}_{nm,i}\{M_i(1)|P_i = n\} + \epsilon_{y2} z\{M_i(1)|P_i = n\}\\ & - \tilde{\beta}_{n,i} - \epsilon_{y1} z - \tilde{\beta}_{nm,i}\{M_i(0)|P_i = n\} - \epsilon_{y2} z\{M_i(0)|P_i = n\}],\\ =& E[(\tilde{\beta}_{nm,i} + \epsilon_{y2} z)\{M_i(1) - M_i(0)|P_i = c\}]\\ =& E[(\tilde{\beta}_{nm,i} + \epsilon_{y2} z)\epsilon_m]\\ =& (\tilde{\beta}_{nm} + \epsilon_{y2} z)\epsilon_m.\end{aligned} \tag{A-4}$$

The first equality is due to the definition of ACME among never takers. The second equality holds after incorporating the second line of equations (A-3). The fourth equality holds after incorporating the first line of equations (A-3). The fifth equality holds because ϵ_m is constant. In the same way, the ANDE among never takers ($\zeta_n^\epsilon(z)$) is identified as $\epsilon_{y1} + \epsilon_{y2}(\tilde{\alpha}_n + \epsilon_m z)$.

Given equations (A-3), we can also obtain CACME and CANDE. Under LSI, $\delta_c^\epsilon(z) = \tilde{\alpha}_{cz} \times (\tilde{\beta}_{cm} + \tilde{\beta}_{czm} z)$, as in equations (A-1), and $\zeta_c^\epsilon(z) = \tilde{\beta}_{cz} + \tilde{\beta}_{czm}(\tilde{\alpha}_c + \tilde{\alpha}_{cz} z)$, as in equations (A-2).

Based on $\delta_c^\epsilon(z)$, $\delta_n^\epsilon(z)$, $\zeta_c^\epsilon(z)$, and $\zeta_n^\epsilon(z)$, the mediated and unmediated ITT effects are identified, respectively, as

$$\begin{aligned}\delta(z) = \tilde{\pi}_c\delta_c^\epsilon(z) + \tilde{\pi}_n\delta_n^\epsilon(z) =& \tilde{\pi}_c\{\tilde{\alpha}_{cz} \times (\tilde{\beta}_{cm} + \tilde{\beta}_{czm} z)\} + \tilde{\pi}_n\{\epsilon_m \times (\tilde{\beta}_{nm} + \epsilon_{y2} z)\}, \text{ and}\\ \zeta(z) = \tilde{\pi}_c\zeta_c^\epsilon(z) + \tilde{\pi}_n\zeta_n^\epsilon(z) =& \tilde{\pi}_c\{\tilde{\beta}_{cz} + \tilde{\beta}_{czm}(\tilde{\alpha}_c + \tilde{\alpha}_{cz} z)\} + \tilde{\pi}_n\{\epsilon_{y1} + \epsilon_{y2}(\tilde{\alpha}_n + \epsilon_m z))\},\end{aligned} \tag{A-5}$$

where $\tilde{\pi}_c, \tilde{\pi}_n, \tilde{\alpha}_n, \tilde{\alpha}_{cz}, \tilde{\beta}_{cz}, \tilde{\beta}_{cm}, \tilde{\beta}_{nm}$, and $\tilde{\beta}_{czm}$ are obtained from the maximized complete-data likelihood given particular values of $\epsilon_m, \epsilon_{y1}$, and ϵ_{y2}. The first equality is due to strong monotonicity. The second equality is due to incorporating results for $\delta_p^\epsilon(z)$ and $\zeta_p^\epsilon(z)$ for $p \in \{c, n\}$. This completes the proof.

11 Appendix C: Sensitivity analysis for LSI

For this proof, we follow Imai et al. [1]'s work. We assumed homogeneous effects as in Imai et al. [1] but expand their work by conditioning on pseudo-population of compliers. We omit pre-treatment confounding in equation (5) for simplicity, but the result remains the same even with covariates. Under randomization (assumption 1), we can consistently estimate γ_c, γ_{cz}, α_c, and α_{cz} as well as a variance measure for each error term σ_{c1} and σ_{c2}, and the correlation between the errors $\tilde{\rho}_{c1} = cor(e_{c1,i}, e_{c2,i}|Z_i = 1)$ and $\tilde{\rho}_{c0} = cor(e_{c1,i}, e_{c2,i}|Z_i = 0)$. We assume that σ_{c1}, σ_{c2} are constant between $Z = 1$ and $Z = 0$.

Using equations (5), $Y_i(0, M_i(0))$ among pseudo-population of compliers can be expressed as

$$\begin{aligned} Y_i(0, M_i(0)) =& \beta_c + \beta_{cm} M_i(0) + e_{c3,i} \\ =& \beta_c + \beta_{cm}(\alpha_c + e_{c2,i}) + e_{c3,i} \\ =& \beta_c + \beta_{cm}\alpha_c + e_{c3,i} + \beta_{cm} e_{c2,i}. \end{aligned} \tag{A-6}$$

By comparing this result with $Y_i(0) = \gamma_c + e_{c1,i}$ (using the first line of equations (5)), we know that $e_{c1i} = e_{c3i} + \beta_{cm} e_{c2i}$ under $Z = 0$. Let ρ_c be the correlation among the pseudo-population of compliers between the error terms obtained from the mediator and outcome models (the second and third lines of equations (5)). Given a value of ρ_c, we have $\tilde{\rho}_{c0}\sigma_{c1}\sigma_{c2} = \rho_c\sigma_{c3}\sigma_{c2} + \beta_{cm}\sigma_2^2$ and $\sigma_{c1}^2 = \sigma_{c3}^2 + \beta_{cm}^2\sigma_{c2}^2 + 2\beta_{cm}\rho_c\sigma_{c3}\sigma_{c2}$. Now assume that $\rho_c \neq 0$, which indicates the violation of LSI. Then, solving these equations with respect to the value of β_{cm}, we have

$$\beta_{cm} = \left\{ \frac{\sigma_{c1}}{\sigma_{c2}} \left(\tilde{\rho}_{c0} - \rho_c \sqrt{\frac{1 - \tilde{\rho}_{c0}^2}{1 - \rho_c^2}} \right) \right\}. \tag{A-7}$$

Likewise, $Y(1, M(1))$ among the pseudo-population of compliers can be expressed as

$$\begin{aligned} Y(1, M(1)) =& \beta_c + \beta_{cz} + \beta_{cm} M(1) + \beta_{czm} M(1) + e_{c3,i} \\ =& \beta_c + \beta_{cz} + (\beta_{cm} + \beta_{czm}) M(1) + e_{c3,i} \\ =& \beta_c + \beta_{cz} + (\beta_{cm} + \beta_{czm})(\alpha_c + \alpha_{cz} + e_{c2,i}) + e_{c3,i} \\ =& \beta_c + \beta_{cz} + (\beta_{cm} + \beta_{czm})(\alpha_c + \alpha_{cz}) + e_{c3,i} + (\beta_{cm} + \beta_{czm}) e_{c2,i}. \end{aligned} \tag{A-8}$$

When comparing this result with $Y(1) = \gamma_c + \gamma_{cz} + e_{c1,i}$ (using the first line of equations (5)), we know that $e_{c1,i} = e_{c3,i} + (\beta_{cm} + \beta_{czm}) e_{c2,i}$ under $Z = 1$. Given a value of ρ_c, we have $\tilde{\rho}_{c1}\sigma_{c1}\sigma_{c2} = \rho_c\sigma_{c3}\sigma_{c2} + (\beta_{cm} + \beta_{czm})\sigma_{c2}^2$ and $\sigma_{c1}^2 = \sigma_{c3}^2 + (\beta_{cm} + \beta_{czm})^2\sigma_{c2}^2 + 2(\beta_{cm} + \beta_{czm})\rho_c\sigma_{c3}\sigma_{c2}$. Then solving these equations with respect to the value of $\beta_{cm} + \beta_{czm}$, we have

$$\beta_{cm} + \beta_{czm} = \left\{ \frac{\sigma_{c1}}{\sigma_{c2}} \left(\tilde{\rho}_{c1} - \rho_c \sqrt{\frac{1 - \tilde{\rho}_{c1}^2}{1 - \rho_c^2}} \right) \right\}. \tag{A-9}$$

Therefore, given a particular value of ρ_c, the CACME and CANDE are identified as

$$\begin{aligned} \delta_c(z) =& \alpha_{cz} \left\{ \frac{\sigma_{c1}}{\sigma_{c2}} \left(\tilde{\rho}_{cz} - \rho_c \sqrt{\frac{1 - \tilde{\rho}_{cz}^2}{1 - \rho_c^2}} \right) \right\}, \quad \text{and} \\ \zeta_c(z') =& CACE - \delta_c(z), \end{aligned} \tag{A-10}$$

where $z' = 1 - z$ for $z \in \{0, 1\}$. Under assumptions 2 and 3, the mediated and unmediated ITT effects are estimated by multiplying the proportion of compliers to the CACME and CANDE estimate respectively, as $\delta(z) = \delta_c(z) \times \pi_c = \pi_c\alpha_{cz} \left\{ \frac{\sigma_{c1}}{\sigma_{c2}} \left(\tilde{\rho}_{cz} - \rho_c \sqrt{\frac{1 - \tilde{\rho}_{cz}^2}{1 - \rho_c^2}} \right) \right\}$ and $\zeta(z') = \zeta_c(z) \times \pi_c = ITT - \delta(z)$. This completes the proof.