Causal Mediation Analysis Assignment 2

Bin Yu

May 2, 2025

Question 1

Define Average Total Effect (ATE)

$$\text{ATE} \;=\; E\big[Y(1) - Y(0)\big] \;=\; E\Big[Y\big(1, M(1)\big) - Y\big(0, M(0)\big)\Big]$$

In the JOBSII study, this estimand represents the average causal effect of assigning all participants to the job-training workshop versus assigning all participants to receive only the booklet. Thus, it is the difference in the probability of working at least 20 hours per week when assigning all participants to the job-training workshop versus assigning all participants to receive only the booklet.

Decomposition of ATE

We start with:

$$\text{ATE} = \; E\big[Y\big(1, M(1)\big) \;-\; Y\big(0, M(0)\big)\big]$$

now add and subtract the "cross-world" term $Y(1, M(0))$ inside the expectation, which means that we separate the effect mediated by M and not mediated by M

$$= \; E\Big[\big\{Y(1, M(1)) - Y(1, M(0))\big\} \;+\; \big\{Y(1, M(0)) - Y(0, M(0))\big\}\Big]$$

use linearity of expectation, $E[A + B] = E[A] + E[B]$

$$= \; E\big[Y(1, M(1)) - Y(1, M(0))\big] \;+\; E\big[Y(1, M(0)) - Y(0, M(0))\big]$$

recognize the two terms as the natural indirect and direct effects

$$= \; \underbrace{E[\,Y(1, M(1)) - Y(1, M(0))\,]}_{\text{NIE}} \;+\; \underbrace{E[\,Y(1, M(0)) - Y(0, M(0))\,]}_{\text{NDE}}$$

Thus we have: Let $M(d)$ denote the mediator value under treatment $D = d$, and $Y(d, m)$ the outcome if we set $D = d$ and mediator $M = m$. The total effect admits the decomposition

$$\underbrace{E\big[Y\big(1, M(1)\big) - Y\big(0, M(0)\big)\big]}_{\text{ATE}} \;=\; \underbrace{E\big[Y\big(1, M(0)\big) - Y\big(0, M(0)\big)\big]}_{\text{NDE}} \;+\; \underbrace{E\big[Y\big(1, M(1)\big) - Y\big(1, M(0)\big)\big]}_{\text{NIE}}.$$

- NDE $= E[Y(1, M(0)) - Y(0, M(0))]$
- NIE $= E[Y(1, M(1)) - Y(1, M(0))]$

Interpretations

ATE

The overall increase in the probability of working at least 20 hours/week when all participants are assigned to the job-training workshop ($D = 1$) rather than to receive only the booklet ($D = 0$).

NDE

$$\text{NDE} \;=\; E\big[Y\big(1,\, M(0)\big) \;-\; Y\big(0,\, M(0)\big)\big].$$

- Fix each participant's self-efficacy at the level they would have under only receiving booklet, $M(0)$, which is the self-efficacy level they would have naturally for them under only receiving booklet. .
- Compare the difference of probability of employment if they instead attend the workshop ($D = 1$) versus receive the booklet ($D = 0$) after fixing each participant's self-efficacy at the level they would have under only receiving booklet, $M(0)$, which is the self-efficacy level they would have naturally for them under only receiving booklet.
- This "deactivates" the $D \to M \to Y$ chain, capturing all other pathways by which attending the workshop may affect reemployment. It can be interpreted as: How much would the workshop change reemployment probability if it did not change self-efficacy at all?

NIE

$$\text{NIE} \;=\; E\big[Y\big(1,\, M(1)\big) \;-\; Y\big(1,\, M(0)\big)\big].$$

- Hold the treatment fixed at participating the "workshop" ($D = 1$) for every participant.
- It captures the difference of probability in employment when pariticipants' self-efficacy is (counterfactually) set to the level it would naturally attain under participating the workshop, $M(1)$, versus the level they would naturally attain under getting the booklet, $M(0)$.
- This isolates the effect transmitted solely along the pathway

$$D \;\to\; M \;\to\; Y \quad \text{(treatment} \to \text{self-efficacy} \to \text{reemployment)}.$$

- In the JOBSII context, NIE answers: "If everyone attends the workshop, how much of the gain in reemployment probability is due to the boost in self-efficacy that the workshop produces?"

Identification requires cross-world counterfactuals such as $Y\big(1, M(0)\big)$, which are never jointly observed. Estimation thus relies on untestable no-confounding assumptions.

Fundamental Problem

In standard causal inference we face the "fundamental problem" that for each individual we can never observe both potential outcomes under two different treatments:

$$Y(1) \quad \text{and} \quad Y(0).$$

Whichever treatment an individual actually receives, the counterfactual outcome under the other treatment remains unobserved. Hence we can never directly observe an individual-level causal effect

$$Y(1) - Y(0).$$

Mediation analysis introduces an additional layer of counterfactuals—so-called "cross-world" outcomes of the form

$$Y\big(d,\, M(d^*)\big),$$

meaning "the outcome if we set treatment to d but (counterfactually) fix the mediator at the value it would have taken under treatment d^*."

By design:

- We never observe $Y(d)$ for individuals not assigned to d.
- We never observe $Y(d^*)$ for individuals not assigned to d^*.
- And crucially, we can never observe the cross-world quantity $Y\big(d,\, M(d^*)\big)$ for any one person.

Thus mediation analyses confront *two* fundamental missing-data problems simultaneously:

$$\underbrace{Y(1) \text{ vs. } Y(0)}_{\text{treatment counterfactual}} \qquad \text{and} \qquad \underbrace{Y\big(1,\, M(0)\big) \text{ vs. } Y\big(0,\, M(1)\big)}_{\text{cross-world counterfactual}}.$$

Because these cross-world potential outcomes can never be observed in reality, identification of natural direct and indirect effects depends critically on some assumptions (no unmeasured confounding, cross-world independence, etc.).

Question 2

Assumptions for Nonparametric Identification

Let $D \in \{0, 1\}$ be treatment, M the mediator, Y the outcome, and C baseline covariates. Write $Y(d, m)$ for the potential outcome under $D = d$ and mediator fixed at m, and $M(d)$ for the potential mediator under d. The average total effect and its decomposition are nonparametrically identified under the following conditions:

NE.1: No unmeasured D-Y confounding

$$Y(d, m) \;\perp\; D \;\mid\; C.$$

Formally, the joint potential outcomes $Y(d, m)$ are independent of treatment assignment D once we condition on measured background characteristics (education, income, age, race, financial strain, etc.) C. Substantively, there must be no unobserved confounders that confound the exposure-outcome relationship, which is the relationship between assignment to the workshop (D) and the employment outcome Y after controlling for covariates C in this example

NE.2: No unmeasured M-Y confounding

$$Y(d, m) \;\perp\; M \;\mid\; D, C$$

Formally, for each d, m, the mediator M must be independent of potential outcome $Y(d, m)$ conditional on treatment D and baseline covariates C.

Substantively, this assumption requires that there must not be any unobserved factors that confound the mediator-outcome relationship after control for exposure and baseline covariates. In our example, this means that once we account for workshop assignment and all measured background characteristics (education, income, age, race, financial strain, etc.), there are no unobserved factors that simultaneously affect both a participant's job-search

self-efficacy and their subsequent employment. Because M was not randomized, this assumption is not ensured by design and relies on having measured all relevant confounders in C.

NE.3: No unmeasured D–M confounding

$$M(d) \perp D \mid C.$$

Formally, exposure D must be statistically independent of the potential values of the mediator $M(d)$, conditional on the baseline confounders C.

Substantively, this assumption requires that there must not be any unobserved factors that confound the exposure-mediator relationship. In our example, this means that once we control for measured background characteristics (education, income, age, race, financial strain, etc.), there are no hidden factors that influence both assignment to the workshop and a participant's job-search self-efficacy.

NE.4: No exposure-induced confounding

$$Y(d, m) \perp M(d^*) \mid C.$$

Formally, the potential outcome under treatment d and mediator value m is independent of the mediator value that would have been observed under a possibly different treatment d^*, once we condition on baseline covariates C.

This "cross-world" independence assumption rules out any variable that is both (a) influenced by the exposure D and (b) itself influences the mediator–outcome relationship. In other words, there must be no exposure-induced confounders, whether measured or unmeasured, that lie on paths

$$D \rightarrow \underbrace{L}_{\text{induced confounder}} \rightarrow M \quad \text{or} \quad D \rightarrow L \rightarrow Y,$$

other than through the mediator of interest. If such an L existed, then $M(d^*)$ and $Y(d, m)$ would become statistically dependent even after adjusting for C, violating NE.4.

In our example, suppose attending the workshop ($D = 1$) changes a participant's social network in ways beyond measured covariates C, and that altered network both affects self-efficacy and directly affects reemployment. Such an exposure-induced confounder L would invalidate NE.4.

NE.5: Positivity

$$\Pr\big(D = d,\, M = m \mid C = c\big) > 0 \quad \text{if} \quad \Pr(C = c) > 0$$

Formally, for each treatment value d and mediator value m, and for any covariate value c such that $\Pr(C = c) > 0$, there must be a positive probability.

This "positivity" (or "overlap") assumption ensures that, within every subpopulation defined by the baseline covariates $C = c$, there is a nonzero chance of observing each treatment level $D = d$ and each mediator value $M = m$. that is there must be at least some chance that individuals experience all possible levels of the exposure and mediator within every subpopulation defined by the confounders

In our example, it means that, for each combination of covariates c, some individuals must sometimes be assigned to the workshop ($D = 1$) and some to the booklet ($D = 0$). Also, within each c and each d, the mediator M must take all values in its relevant range with positive probability.

NE.6: Consistency

$$Y \;=\; Y(D) \;=\; Y\big(D,\, M(D)\big) \;=\; Y\big(D,\, M\big).$$

This assumption means that:

- $Y = Y(D)$: An individual's observed outcome equals their potential outcome under the treatment level they actually received.

- $Y(D) = Y\big(D,\, M(D)\big)$: The potential outcome under the observed treatment D equals the "nested" potential outcome when we set the mediator to the value it would naturally take under that same D.
- $Y\big(D,\, M(D)\big) = Y\big(D,\, M\big)$: The nested potential outcome (with mediator fixed at its natural value) equals the joint potential outcome under the observed treatment and the mediator value actually experienced.
- By the same logic, we also have the mediator consistency $M = M(D)$: the observed mediator equals its potential value under the observed treatment.

Therefore, all versions of "what would have happened" align with what was actually observed once we plug in the real treatment and mediator. There is no ambiguity or measurement error in mapping observed data to the corresponding potential-outcome expressions.

In our example, consistency holds if each participant's recorded job-search self-efficacy and employment status truly reflect the hypothetical values under their assigned workshop or booklet condition, with no misclassification or interference.

Which assumptions hold in experimental design

Below we review each NE assumption in turn, state whether randomization or design secures it, and discuss potential threats.

NE.1: No unmeasured DY confounding ($Y(d,m) \perp D \mid C$). Holds by design. Treatment D (workshop vs. booklet) was randomly assigned, so—regardless of any baseline covariates C—there are no hidden common causes of assignment and outcome. Even if some C were omitted from our analysis, randomization ensures D is independent of all pre–treatment factors.

NE.3: No unmeasured DM confounding ($M(d) \perp D \mid C$). Holds by design. For the same reason as NE.1, randomization guarantees that D is independent of the potential mediator values $M(0)$ and $M(1)$. There can be no pre–treatment factor that simultaneously predicts assignment and self-efficacy.

NE.5: Positivity ($\Pr(D = d, M = m \mid C = c) > 0$ if $\Pr(C = c) > 0$). Mostly holds by design. Randomization ensures $\Pr(D = 0 \mid C) > 0$ and $\Pr(D = 1 \mid C) > 0$, that is, every baseline subgroup has both workshop and booklet participants. For the mediator, it is plausible if the self-efficacy index has sufficient variation in each subgroup. One should examine histograms of M by D and by key C-strata to confirm that no score levels are "missing." If, for example, older participants never report the highest confidence level under control, then $(D = 0, M = 5, C = \text{old})$ would have zero probability, and NIE/NDE would not be identified for that subgroup. As long as the self-efficacy index M varies continuously (no perfect floor/ceiling) within each treatment–covariate cell, mediator positivity holds. **Threats:** extreme clustering of self-efficacy (e.g. everyone in one subgroup scores at the maximum or minimum), which would leave some (d, m, c) combinations unobserved.

NE.6: Consistency ($Y = Y(D, M(D))$ and $M = M(D)$). Likely holds by design. If each participant's observed self-efficacy and employment truly reflect what would happen under their own assigned condition—and there is no interference between units—then consistency holds. Howeverfirst, if self-efficacy M or employment Y are mis-recorded or measured in a biased way, then $M \neq M(D)$ or $Y \neq Y(D, M(D))$. second, if workshop participants share materials with control participants, one person's D may influence another's M or Y. last, if there are multiple "versions" of the workshop (different instructors, curricula), then D is not a single well-defined intervention. In these situations, the assumptions could be violated. **Threats:** measurement error in M or Y, or spillover of workshop effects across participants.

NE.2: No unmeasured MY confounding ($Y(d,m) \perp M \mid D, C$). Not guaranteed by design. Because M (self-efficacy) was not randomized, there may exist unmeasured factors—motivation, baseline mental health, unobserved social support—that influence both M and Y even after controlling for C.

NE.4: No exposure-induced confounding ($Y(d,m) \perp M(d^*) \mid C$). Not guaranteed by design. This "cross-world" assumption is not guaranteed by design since there might be some variables L that is itself affected by D and

then confounds $M \to Y$. For example, if workshop attendance changes participants' social networks in unmeasured ways that both boost self-efficacy and directly affect reemployment, NE.4 fails. Such exposure-induced confounders cannot be ruled out by randomization and require substantive justification.

Summary. Randomization secures NE.1, NE.3 and, probably ensure NE.5 and NE.6, but identification of mediation further requires NE.2 (no mediator–outcome confounding) and NE.4 (no exposure-induced confounders), neither of which is ensured by the experimental design.

Question 3

DAG for JOBSII under the "only C–confounding" supposition

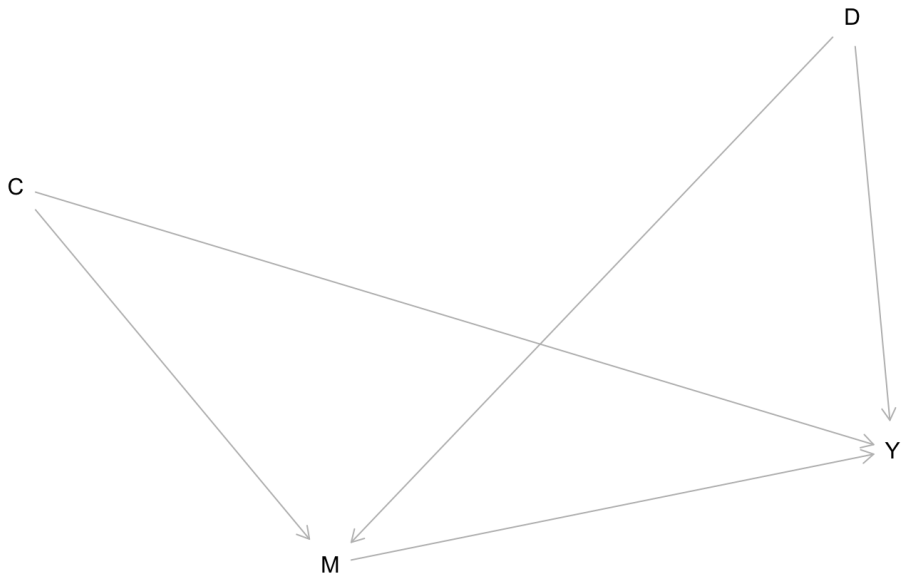

Figure 1: DAG

Identification

The key to identification is that all back-door (confounding) paths between each pair of variables are blocked once we condition on the baseline covariates C, and since that treatment D itself is randomized. The six "mediation-identification" assumptions hold:

1. *No unmeasured D–Y confounding:*
$$Y(d, m) \perp D \mid C.$$
Because D was randomized, D is independent of any potential outcome once we condition on nothing (and a fortiori once we condition on C).

2. *No unmeasured D–M confounding:*
$$M(d) \perp D \mid C.$$
Randomization of D likewise ensures no unmeasured common cause of D and the mediator M, after adjusting for C.

3. *No unmeasured M–Y confounding (given D):*

$$Y(d,m) \;\perp\; M \;\big|\; D,\, C.$$

By hypothesis all common causes of self-efficacy M and employment Y lie in C, so once we condition on C (and on D) there is no remaining confounding of the $M \to Y$ relation.

4. *No exposure–induced confounding:*

$$Y(d,m) \;\perp\; M(d^*) \;\big|\; C.$$

There is no variable that (i) is affected by D and then (ii) itself confounds $M \to Y$. In our DAG all arrows out of D go straight into M or Y, and no child of D is a parent of both M and Y.

5. *Positivity:*

$$\Pr(D = d,\, M = m \mid C = c) > 0 \quad \text{whenever } \Pr(C = c) > 0.$$

Random assignment ensures $\Pr(D = 0 \mid C)$ and $\Pr(D = 1 \mid C)$ are both bounded away from zero; and the mediator M has support over a continuous range within each D, C cell.

6. *Consistency:*

$$Y = Y\big(D,\, M(D)\big), \quad M = M(D).$$

Each observed outcome and mediator value coincides with the corresponding potential outcome under the realized treatment, and there is no interference between units.

Because all six conditions hold, we can identify the ATE, NDE and NIE using observed data, Let $d = 1$ (workshop) and $d^* = 0$ (booklet). Then under the standard mediation assumptions:

ATE

$$\text{ATE} = E[Y(1, M(1)) - Y(0, M(0))] = \sum_c \Big(E[Y \mid D = 1, C = c] - E[Y \mid D = 0, C = c] \Big) \Pr(C = c).$$

NDE

$$\begin{aligned}
\text{NDE}(1,0) &= E\big[Y\big(1,\, M(0)\big) - Y\big(0,\, M(0)\big)\big] \\
&= \sum_c \sum_m \Big[E\big[Y \mid C = c,\, D = 1,\, M = m\big] - E\big[Y \mid C = c,\, D = 0,\, M = m\big] \Big] P\big(M = m \mid C = c,\, D = 0\big)\, P(C = c) \\
&= E_C\Big[E_{M \mid C,\, D=0}\big(E[Y \mid C,\, D = 1,\, M] - E[Y \mid C,\, D = 0,\, M] \big) \Big],
\end{aligned}$$

NIE

$$\begin{aligned}
\text{NIE}(1,0) &= E\big[Y\big(1,\, M(1)\big) - Y\big(1,\, M(0)\big)\big] \\
&= \sum_c \sum_m E\big[Y \mid C = c,\, D = 1,\, M = m\big] \Big[P\big(M = m \mid C = c,\, D = 1\big) - P\big(M = m \mid C = c,\, D = 0\big) \Big] P(C = c) \\
&= E_C\Big[E_{M \mid C,\, D=1}\big(E[Y \mid C,\, D = 1,\, M]\big) \;-\; E_{M \mid C,\, D=0}\big(E[Y \mid C,\, D = 1,\, M]\big) \Big].
\end{aligned}$$

Using these formulas, ATE, NDE and NIE can be estimated using observed data:

Let C be the vector of baseline covariates, $D \in \{0, 1\}$ treatment, M is mediator, and Y is outcome. Write

$n_{c,d} = \#\{i : C_i = c,\, D_i = d\}$, $\quad n_{c,d,m} = \#\{i : C_i = c,\, D_i = d,\, M_i = m\}$, $\quad n_c = \#\{i : C_i = c\}$, $\quad n =$ total sample size.

Define

$$\bar{Y}_{c,d} = \frac{1}{n_{c,d}} \sum_{i : C_i = c, D_i = d} Y_i, \qquad \bar{Y}_{c,d,m} = \frac{1}{n_{c,d,m}} \sum_{i : C_i = c, D_i = d, M_i = m} Y_i,$$

and

$$\hat{\pi}_c = \frac{n_c}{n}, \qquad \hat{\pi}_{m|c,d} = \frac{n_{c,d,m}}{n_{c,d}}.$$

Estimator for ATE:

$$\widehat{\mathrm{ATE}}^{np} = \sum_c \left[\bar{Y}_{c,1} - \bar{Y}_{c,0}\right] \hat{\pi}_c = \sum_c \left(\bar{Y}_{c,1} - \bar{Y}_{c,0}\right) \hat{\pi}_c.$$

Estimator for NDE:

$$\widehat{\mathrm{NDE}}^{np}(1,0) = \sum_c \sum_m \left[\bar{Y}_{c,1,m} - \bar{Y}_{c,0,m}\right] \hat{\pi}_{m|c,0}\, \hat{\pi}_c.$$

Estimator for NIE:

$$\widehat{\mathrm{NIE}}^{np}(1,0) = \sum_c \sum_m \bar{Y}_{c,1,m} \left[\hat{\pi}_{m|c,1} - \hat{\pi}_{m|c,0}\right] \hat{\pi}_c.$$

Question 4

Randomization

We stratify by treatment assignment and conduct two-sample t-tests (for continuous covariates) and chi–square tests (for categorical covariates). Results are shown below:

	Stratified by treat			
	control	exp	p	test SMD
n	301	601		
econ_hard (mean (SD))	3.03 (1.01)	3.02 (0.97)	0.928	0.006
sex (mean (SD))	0.57 (0.50)	0.52 (0.50)	0.094	0.118
age (mean (SD))	37.38 (10.75)	37.69 (10.31)	0.672	0.030
nonwhite = non.white1 (%)	52 (17.3)	100 (16.6)	0.883	0.017
educ (%)			0.838	0.085
lt-hs	16 (5.3)	34 (5.7)		
highsc	93 (30.9)	180 (30.0)		
somcol	109 (36.2)	212 (35.3)		
bach	43 (14.3)	103 (17.1)		
gradwk	40 (13.3)	72 (12.0)		
income (%)			0.688	0.105
lt15k	56 (18.6)	108 (18.0)		
15t24k	66 (21.9)	140 (23.3)		
25t39k	70 (23.3)	149 (24.8)		
40t49k	44 (14.6)	68 (11.3)		
50k+	65 (21.6)	136 (22.6)		

Thus, randomization successfully balanced all six baseline covariates (all $p > 0.05$).

Total Effect on Reemployment

We fit the linear probability model

$$\text{work1}_i = \alpha + \tau\ \text{treat}_i + \varepsilon_i$$

by OLS. The output is:

```
Call:
lm(formula = work1 ~ treat, data = jobs_clean)

Residuals:
    Min      1Q  Median      3Q     Max
-0.3461 -0.3461 -0.2890  0.6539  0.7110

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.28904    0.02703  10.695   <2e-16 ***
treat        0.05705    0.03311   1.723   0.0852 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4689 on 900 degrees of freedom
Multiple R-squared:  0.003288, Adjusted R-squared:  0.002181
F-statistic: 2.969 on 1 and 900 DF,  p-value: 0.0852
```

Thus, the estimated average treatment effect (ATE) can be written in the potential-outcomes notation as

$$\widehat{E}\big[Y(1) - Y(0)\big] \;=\; 0.0571.$$

In other words, had everyone in the study been assigned to the workshop ($D = 1$) rather than the booklet ($D = 0$), their average probability of working at least 20 hrs/week would have been about 5.7 percentage points higher.

The two-sided p-value for testing $H_0 \colon E[Y(1) - Y(0)] = 0$ is 0.081, so we deem this ATE marginally significant at the 10% level (but not at 5%).

Total Effect on Job-Search Self-Efficacy

We fit the model

$$\text{job\_seek}_i = \beta + \gamma\, \text{treat}_i + \eta_i$$

by OLS. The output is:

```
Call:
lm(formula = job_seek ~ treat, data = jobs_clean)

Residuals:
     Min       1Q   Median       3Q      Max
-3.06733 -0.40067  0.09933  0.59933  1.00166

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.99834    0.04195  95.318   <2e-16 ***
treat        0.06899    0.05139   1.343     0.18
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7278 on 900 degrees of freedom
Multiple R-squared:  0.001999, Adjusted R-squared:  0.0008899
F-statistic: 1.802 on 1 and 900 DF,  p-value: 0.1798
```

We estimate the average causal effect of the workshop on self-efficacy in the potential-outcomes and the outcome is:

$$\widehat{E}\big[M(1) - M(0)\big] = 0.06899$$

In other words, had everyone been assigned to participating the workshop rather than only receiving the booklet, their job-search self-efficacy score would on average be 0.069 points higher on the 5-point scale—an effect that is not statistically significant ($p > 0.10$).

Question 5

Let D = treatment (workshop vs. control), M = job-search self-efficacy, Y = employment indicator. We report point estimates of

$$\begin{aligned} \text{ATE} &= E[Y(1, M(1)) - Y(0, M(0))], \\ \text{NDE} &= E[Y(1, M(0)) - Y(0, M(0))], \\ \text{NIE} &= E[Y(1, M(1)) - Y(1, M(0))]. \end{aligned}$$

Specification	ATE	NDE	NIE
1. Additive, no controls	0.057	0.055	0.002
2. Additive, with controls C	0.057	0.054	0.003
3. Additive + $D \times M$ interaction with controls C	0.057	0.056	0.002

Table 1: Point estimates of total, natural direct, and natural indirect effects

Interpretation

ATE ≈ 0.057

This is the average total effect of being assigned to the job-training workshop versus receiving only the booklet. Numerically, it means

$$E[\, Y(1, M(1)) - Y(0, M(0))\,] \approx 0.057,$$

Thus, if everyone in our sample had attended the workshop instead of just getting the booklet, the probability of working at least 20 hours/week would increase by about 5.7 percentage points on average.

NDE (0.054–0.056)

The natural direct effect holds each individual's self-efficacy at the level it would naturally have taken under control ($M(0)$) and captures how employment would change purely from changing the workshop assignment.

Formally,

$$E\big[Y(1,\, M(0)) - Y(0,\, M(0))\big] \approx 0.054\text{–}0.056.$$

Thus, if we hold every participant's job search self-efficacy at the level of they will naturally take under receiving booklet. That is, even if the workshop did not alter anyone's self-efficacy in job searching, merely participating (versus receiving the booklet) would raise reemployment probability by about 5.4–5.6 percentage points.

This direct effect captures all mechanisms of the workshop other than changes in self-efficacy—such as learning concrete search strategies, tapping into instructor networks, or gaining motivational support.

NIE (0.002–0.003) Holding treatment fixed at $D = 1$ (everyone attends the workshop), we compare the difference in the probability of employment when each individual's self-efficacy is set to its natural workshop level $M(1)$ versus its control-group level $M(0)$, in other words, the natural indirect effect holds every participant's treatment at $D = 1$ (everyone attends the workshop) and compare two scenarios for the mediator:

- $M(1)$: each person's self-efficacy when they actually attend the workshop,
- $M(0)$: the level of self-efficacy they would have had under the control.

it quantifies how much of the workshop's total effect on reemployment is mediated by its effect on job-search self-efficacy. It is defined as

$$\text{NIE} \;=\; E\big[Y\big(1, M(1)\big) \;-\; Y\big(1, M(0)\big)\big] \;\approx\; 0.002\text{–}0.003.$$

This implies that the change of participants' job search self-efficacy followed by participating the workshop increases the probability of their reemployment by about 0.2–0.3 percentage points. Thus the $D \to M \to Y$ pathway via self-efficacy is negligible.

Summary

Thus, adding baseline covariates (C) or allowing the effect of self-efficacy on employment to differ by treatment (the $D \times M$ interaction) changes ATE, NDE, NIE by at most 0.002. This stability shows that the workshop appears to boost reemployment primarily through mechanisms other than boosting self-efficacy.

Question 6

We use the simulation approach (2000 Monte Carlo draws) to estimate ATE, NDE, NIE from two models: 1. **Mediator model (linear)**

$$M_i = \alpha_M + \delta\, D_i + \theta^\top C_i + \varepsilon_{M,i},$$

where M_i is job-search self-efficacy, D_i treatment, and C_i baseline covariates.

2. **Outcome model (logit)**

$$\Pr(Y_i = 1 \mid D_i, M_i, C_i) = \text{logit}^{-1}\big(\alpha_Y + \tau\, D_i + \beta\, M_i + \phi^\top C_i\big),$$

where Y_i indicates employment.

We then draw 2000 parameter vectors from the joint asymptotic normal distribution of each model's coefficients, compute counterfactuals $Y(d, m)$ and $M(d)$, and average contrasts to obtain:

Using 2000 Monte Carlo draws from the linear mediator model and logistic outcome model, we obtain:

Specification	ATE	NDE	NIE
No $D\times M$ interaction	0.05829	0.05509	0.00320
With $D\times M$ interaction	0.05778	0.05531	0.00247

Table 2: Point estimates of ATE, NDE, NIE (2000 simulation draws).

Interpretation

$ATE \approx 0.058$

Participating the workshop raises the probability of reemployment by about 5.8 percentage points if everyone attended versus everyone received the booklet.

$NDE \approx 0.055$

Holding each individual's self-efficacy at the level they would have under receiving the booklet ($M(0)$), simply attending the workshop raises reemployment by about 5.5 percentage points.

$$E\big[Y(1, M(0)) - Y(0, M(0))\big] \approx 0.055.$$

This direct effect captures all mechanisms of the workshop other than changes in self-efficacy—such as learning concrete search strategies, tapping into instructor networks, or gaining motivational support.

$NIE \approx 0.0025$–0.0032

Holding the treatment at $D = 1$ (everyone in the workshop) and comparing each individual's self-efficacy under workshop versus the level they would have under the booklet, we isolate the mediated effect:

$$E\big[Y(1, M(1)) - Y(1, M(0))\big] \;\approx\; 0.0025\text{–}0.0032.$$

This means that the increase in job-search self-efficacy generated by the workshop contributes only about 0.25–0.32 percentage points to the probability of reemployment.

Interaction vs. no interaction.

Allowing treatment to modify the effect of self-efficacy (the $D \times M$ term) slightly lowers ATE (5.78 pp vs. 5.83 pp) and NIE (0.25 pp vs. 0.32 pp), while NDE remains near 5.5 pp. This minimal change confirms that non-linearity or effect heterogeneity does not alter the substantive conclusion: the workshop's benefit on reemployment flows almost entirely through direct mechanisms other than self-efficacy.

Comparison to Question 5 (Linear/Additive Models)

In Question 5 we obtained by additive linear mediation:

$$\text{ATE} \approx 0.057, \quad \text{NDE} \approx 0.054\text{–}0.056, \quad \text{NIE} \approx 0.002\text{–}0.003.$$

Our simulation-based estimates under a logistic outcome model are:

$$\begin{aligned} \text{ATE} &\approx 0.058, \\ \text{NDE} &\approx 0.055, \\ \text{NIE} &\approx 0.0025\text{–}0.0032. \end{aligned}$$

Thus the two approaches yield nearly identical point estimates. This close agreement demonstrates that the simple linear/additive mediation analysis was robust to (a) the binary nature of the outcome and (b) allowing for a non-linear (logistic) link and interaction. In all cases, the indirect effect mediated by self-efficacy remains negligible.

Question 7

Using inverse-probability weighting (IPW) with stabilized weights censored at the 1st–99th percentiles, we estimate ATE, NDE, NIE via two logistic models for $P(D \mid C)$. The point estimates are:

Estimator	ATE	NDE	NIE
IPW (stabilized, censored)	0.055	0.054	0.002

Interpretation and Comparison

ATE = 0.055. If everyone attended the workshop rather than receiving only the booklet, the probability of reemployment would increase by about 5.5 percentage points.

NDE = 0.054

Holding every participant's job-search self-efficacy at the level they will have at its control-group level $M(0)$, participating the workshop still raises the probability of working at least 20 hrs/week by about 5.4 percentage points.

This direct effect captures all mechanisms of the workshop other than changes in self-efficacy—such as learning concrete search strategies, tapping into instructor networks, or gaining motivational support.

NIE = 0.002.

Holding the treatment at $D = 1$ (everyone in the workshop) and comparing each individual's self-efficacy under workshop versus the level they would have under the booklet, we isolate the mediated effect:

$$E\big[Y(1, M(1)) - Y(1, M(0))\big] \approx 0.002.$$

This means that the increase in job-search self-efficacy generated by the workshop contributes only about 0.2 percentage points to the probability of reemployment.

Comparison to previous methods:

Linear/additive models (Q5) gave ATE = 0.057, NDE = 0.054–0.056, NIE =0.002–0.003.

Simulation/logit models (Q6) gave ATE= 0.058, NDE= 0.055, NIE =0.0025–0.0032.

The IPW estimates (0.055, 0.054, 0.002) are nearly identical to both the linear-additive and simulation-based estimates, confirming robustness of the finding that the workshop's benefit on reemployment is almost entirely "direct" (not mediated by self-efficacy).

Question 8

We first compared two versions of the simulation-based estimator from Question6:

- *No interaction model:*
$$\Pr(Y = 1 \mid D, M, C) = \alpha + \tau D + \beta M + \phi^\top C,$$
- *Interaction model:*
$$\Pr(Y = 1 \mid D, M, C) = \alpha + \tau D + \beta M + \theta\, D{:}M + \phi^\top C.$$

A likelihood-ratio test between these nested models yields:

```
Model 1 (no interaction):  LogLik = {546.85,  df = 14
Model 2 (with interaction):LogLik = {546.03,  df = 15
  df = 1,  ² = 1.636,  p = 0.2008
```

Since p=0.2098, the extra $D \times M$ term is not statistically signficant.

Therefore we select the simpler no-interaction specification for final inference.

Advantages of the Simulation Estimator (Q6) over Alternatives

- **Versus linear/additive mediation (Q5):** By using a logistic link for the binary employment outcome, the simulation estimator is more applicable given the $[0, 1]$ range and better matches the true distribution of Y. It also accommodates non-Gaussian residuals and naturally produces valid confidence intervals via Monte Carlo sampling. The simulation approach therefore yields more reliable, efficient estimates and more accurate uncertainty quantification.
- **Versus IPW mediation (Q7):** IPW relies on estimated weights $w_i = P(D)/P(D \mid C_i)$, which can become extreme and inflate variance when some $P(D \mid C)$ are small. The simulation approach instead models both the mediator and outcome directly, avoiding unstable weights and yielding lower-variance estimates in finite samples. It also facilitates straightforward inclusion of covariates and diagnostics of model fit.

Thus, we select the simulation estimator from Question 6 without the $D \times M$ interaction.

Using this estimator with 2000 nonparametric bootstrap replications, we obtain:

```
point.est ll.90ci ul.90ci pval
TE( treat = 1 , treat *= 0 ) 0.057731707 0.0058434313 0.111453492 0.064
NDE( treat = 1 , treat *= 0 ) 0.054900776 0.0029851718 0.108066075 0.075
NIE( treat = 1 , treat *= 0 ) 0.002830931 -0.0007615022 0.009098392 0.269
```

In table form:

Effect	Estimate	90%CI Lower	90%CI Upper	p-value
ATE(1,0)	0.058	0.0058	0.1115	0.064
NDE(1,0)	0.055	0.0030	0.1080	0.075
NIE(1,0)	0.003	–0.0008	0.0091	0.269

Table 3: Point estimates, 90% percentile bootstrap confidence intervals, and two-sided p-values

Interpretation

ATE = 0.058 (90%CI: [0.0058, 0.1115], $p = 0.064$**).**

Participating in the workshop rather than receiving the booklet increases the probability of working at least 20 hrs/week by about 5.8 percentage points; the 90% interval excludes 0, indicating marginal significance at the 10% level.

NDE =0.055 (90% CI: [0.0030, 0.1080], $p = 0.075$**).**

Holding job-search self-efficacy at its control-group level, the workshop still raises reemployment by about 5.5 percentage points, confirming that the direct pathways (other than self-efficacy) drive most of the effect.

NIE = 0.003 (90% CI: [–0.0008, 0.0091], $p = 0.269$**).** The mediated effect through self-efficacy is near zero and its 90% interval overlaps zero, indicating no reliable indirect effect.

Question 9

We wish to test

$$H_0 : \text{NIE}(1,0) = 0 \quad \text{vs.} \quad H_a : \text{NIE}(1,0) \neq 0$$

at significance level $\alpha = 0.1$, using our bootstrap distribution for the natural indirect effect.

Percentile-bootstrap p-value

From 2000 bootstrap replicates of $\widehat{\text{NIE}}$, we obtained the 90% percentile interval

$$\big[\,-0.0008,\ 0.0091\,\big].$$

By inversion, an approximate two-sided p-value is

$$p \;=\; 2\,\min\{\Pr(\widehat{\text{NIE}} < 0),\ \Pr(\widehat{\text{NIE}} > 0)\} \approx 0.269.$$

Decision and interpretation

Since $p = 0.269 > 0.1$, we fail to reject the null hypothesis H_0: $\mathrm{NIE}(1, 0) = 0$ at the $\alpha = 0.1$ level. Equivalently, when we hold everyone's treatment fixed at $D = 1$ (all attend the workshop) and compare the counterfactual outcomes under their workshop-induced self-efficacy $M(1)$ versus the self-efficacy they would have under the booklet $M(0)$, the average difference in employment is indistinguishable from zero. That is, the pathway

$$D \;\rightarrow\; M \;\rightarrow\; Y$$

via job-search self-efficacy does not contribute detectably to the workshop's effect on reemployment.

Question 10

Summary of findings

According to the results in Q5-9, we conclude that job-search self-efficacy does not meaningfully mediate the workshop's effect on reemployment.

For the question that whether it transmit, at least in part, the effect of treatment on subsequent employment? Our answer is no, since all three estimators (additive linear, simulation/logistic, IPW) yield a natural indirect effect $\widehat{\mathrm{NIE}} \approx 0.002$–$0.003$ that is statistically indistinguishable from zero (bootstrap $p \approx 0.25$, 90% CI contains 0).

Potential threats to validity

1. **Unmeasured mediator–outcome confounders.** There may be baseline traits (e.g. intrinsic motivation, mental health, social support) not recorded in C that influence both self-efficacy and reemployment, biasing the NIE estimate.
2. **Exposure-induced confounding.** The workshop could alter participants' social networks or stress levels, which then affect both M and Y. Such post-treatment confounders violate the cross-world assumption.
3. **Measurement error in M.** Self-efficacy is self-reported on a 5-point scale; misreporting or scale-use heterogeneity can attenuate the mediator effect.
4. **Model misspecification.** Linear models for M and logistic (or linear probability) models for Y may fail to capture nonlinearity, interactions, or heteroskedasticity, leading to biased effect estimates.
5. **Positivity / overlap violations.** If some covariate strata rarely exhibit certain mediator values under one treatment arm, the integrals in the NIE expression rely on extrapolation, increasing bias and variance.
6. **Interference (spillovers).** Participants may share workshop lessons with each other, violating the no-interference assumption and contaminating direct and indirect effect estimates.