Covariate-dependent Reporting Bias:
Methods and Application to the LGBQ Earnings Gap

Cameron Deal[†]

[†]Harvard University

Table of Contents

Table of Contents

LGBQ Economic and Mental Health Differences

Lesbian, Gay, Bisexual, and Queer (LGBQ) individuals appear to have distinct economic and health outcomes:

- 5-10% of the US population (Carpenter et al 2021)
- **10% lower earnings** than comparable heterosexuals (Drydakis 2019)
- **83% higher mental distress** scores (NSDUH 2015-2019)
- **3x higher suicidal ideation** (NSDUH 2015-2019)
- Discrimination and stigma are two prominent drivers

LGBQ Economic and Mental Health Differences

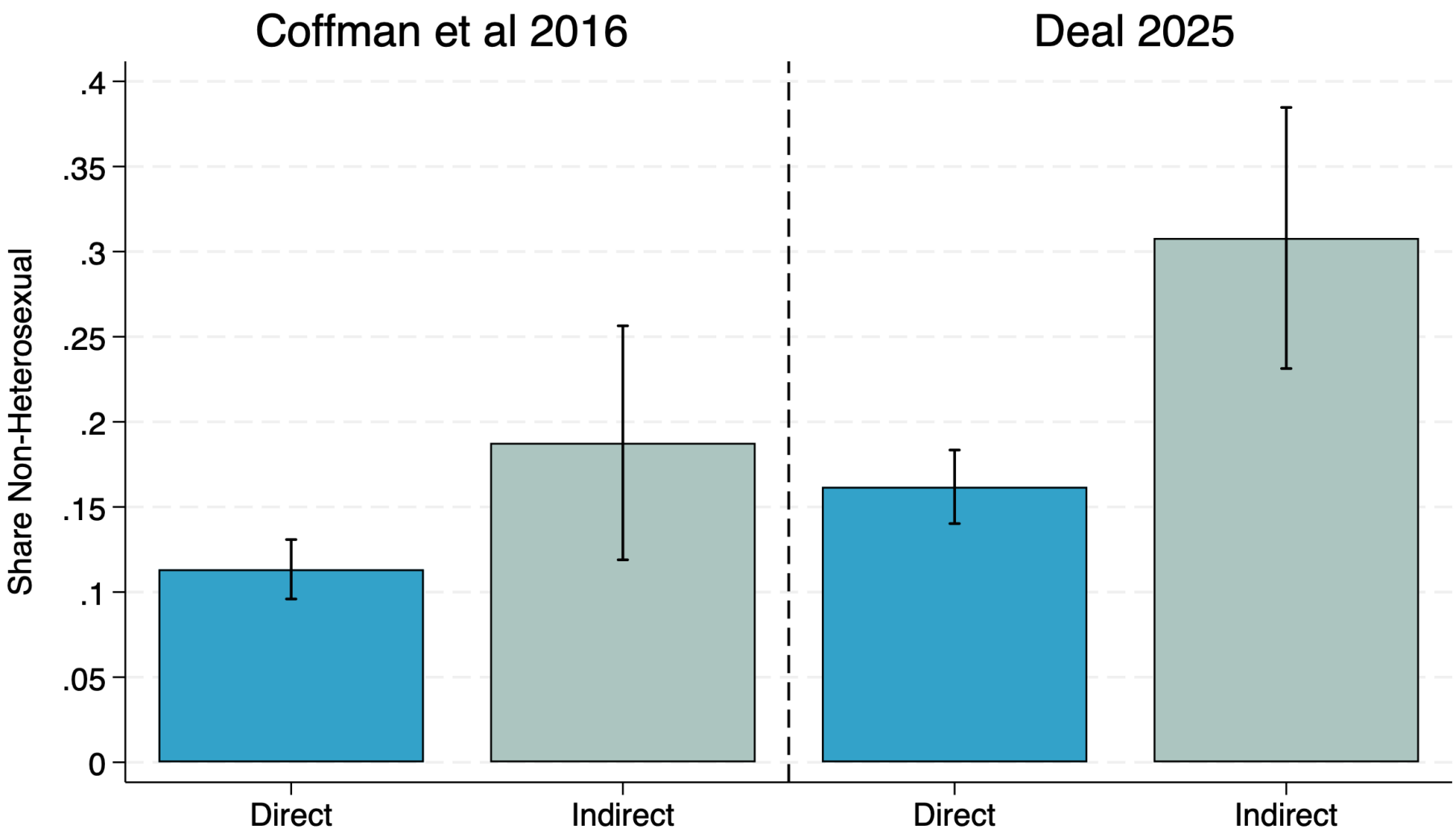
Lesbian, Gay, Bisexual, and Queer (LGBQ) individuals appear to have distinct economic and health outcomes:

- 5-10% of the US population (Carpenter et al 2021)
- **10% lower earnings** than comparable heterosexuals (Drydakis 2019)
- **83% higher mental distress** scores (NSDUH 2015-2019)
- **3x higher suicidal ideation** (NSDUH 2015-2019)
- Discrimination and stigma are two prominent drivers

Fundamental limitation: sexual identity is **by definition self-reported.**

- High stigma against LGBQ individuals: 33% of Americans say gay/lesbian relations morally wrong. (Gallup 2025)
- Substantial evidence of underreporting.

Underreporting of Minority Sexual Identity



This Project

Want to go beyond the **level** of misreporting and instead ask: **what characteristics** are associated with misreporting?

- Helps understand group disparities and test hypotheses about the decision to disclose.
- Depending on misreporters' earnings, LGBQ gap could range from -\$47k to +\$57k.

This paper develops methods to address description under misreporting and applies them to LGBQ earnings/mental health gaps.

- Use tools from IV literature on characterizing compliers to describe indirect reporters.
- Combine with direct report estimates to characterize misreporters.
- Field survey of 2501 respondents on Prolific.

Preview of Results

1. Non-heterosexual identity underreported: 16 $\rightarrow$ 31%
2. Misreporters have **lower mental distress** and **higher earnings** than their direct report counterparts.
3. Including indirect reporters **eliminates mental health gaps** ($+0.48 \rightarrow +0.02$ SDs) and **flips earnings penalty** ($-\$19$k $\rightarrow +\$30$k) to premium.
4. Suggestive evidence that demographics and costs of disclosing play a role in generating these patterns.

Table of Contents

List Experiment/Item Count Technique

Randomize individuals into 2 groups:

- Direct report ($Z_i = 0$) see list of 4 control items
- Veiled report ($Z_i = 1$) see list of same items + sensitive item

Direct Report	Veiled Report
• I remember where I was the day of the *Challenger* space shuttle disaster. • I spent a lot of time playing video games as a kid. • I would vote to legalize marijuana if there was a ballot question in my state. • I have voted for a political candidate who is pro-life. Please fill in the bubble that corresponds to the total number of statements above that apply to you. 0 1 2 3 4	• I remember where I was the day of the *Challenger* space shuttle disaster. • I spent a lot of time playing video games as a kid. • I would vote to legalize marijuana if there was a ballot question in my state. • I have voted for a political candidate who is pro-life. • I consider myself to be heterosexual. Please fill in the bubble that corresponds to the total number of statements above that apply to you. 0 1 2 3 4 5

Estimate mean difference in $R_i = \#\text{items}$: $E[R_i|Z_i = 1] - E[R_i|Z_i = 0]$ to get prevalence of sensitive item S_i.

Does This Work? Why?

Randomized Response Validation:

- Does better than direct reports in predicting county-level vote shares for anti-abortion referendum (Rosenfeld et al 2016)
- Increases reporting of harassment by reducing fear of retaliation (Boudreau et al 2024)
- Closes gender gaps in reported lifetime sexual partners (Krumpal et al 2018)
- No differences for non-sensitive placebo items (Coffman et al 2017)
- Does not work when participants confused (Chuang et al 2016)

Does This Work? Why?

Randomized Response Validation:

- Does better than direct reports in predicting county-level vote shares for anti-abortion referendum (Rosenfeld et al 2016)
- Increases reporting of harassment by reducing fear of retaliation (Boudreau et al 2024)
- Closes gender gaps in reported lifetime sexual partners (Krumpal et al 2018)
- No differences for non-sensitive placebo items (Coffman et al 2017)
- Does not work when participants confused (Chuang et al 2016)

Why Indirectly Report:

- Privacy concerns/plausible deniability
- Self-image
- In this setting, likely most salient for those not "out," though unable to validate

Method Overview

1. **Reframe as IV:** Individuals who add one to their response total when they see long vs. short list are *compliers*—and the **full population of indirect reporters**.
2. **Characterize Indirect Reporters:** Use standard tools to estimate covariates of compliers to describe indirect reporters.
3. **Back Out Misreporters:** Using indirect reporter estimates, direct reporter estimates and relative shares, solve for misreporter means.

▸ First Stage Bar Graph

Key Assumption and Intuition

- Define **potential response totals** $R_i(z)$ for $z \in \{0, 1\}$.
- **No Design Effects (NDE)**: adding the sensitive item (S_i) does not change answers to control items.

$$\Rightarrow R_i(1) - R_i(0) = S_i \in \{0, 1\}$$

- Thus, we can reframe list experiment as IV: the people whose $R_i(1) > R_i(0)$ (respond to the veiled report) are **compliers**.
- This is also the full population with $S_i = 1$ (indirect reporters).
  $\Rightarrow S_i = 1 \Leftrightarrow R_i(1) > R_i(0)$

IV View and Characterizing Indirect Reporters

- Use tools to **characterize compliers.**
- Let X_i be covariate of interest and $g(\cdot)$ be measurable function.
- Ex: Wald ratio identifies complier (i.e., $S_i = 1$) characteristics (Angrist & Imbens 1994, Abadie 2003):

$$\frac{\mathrm{E}[g(X_i)R_i \mid Z_i = 1] - \mathrm{E}[g(X_i)R_i \mid Z_i = 0]}{\mathrm{E}[R_i \mid Z_i = 1] - \mathrm{E}[R_i \mid Z_i = 0]} = \mathrm{E}[g(X_i) \mid S_i = 1]\,.$$

- Works for any measurable g: means, indicators for CDF/quantiles, transformations.

▸ Identification Proof

Notation and Types of Reporters

- Let $S_i = R_i(1) - R_i(0) \in \{0, 1\}$ be latent binary outcome of interest (LGBQ status). We see direct report $D_i \in \{0, 1\}$ in sample with $Z_i = 0$ (short list).

We can decompose the sample into three groups:

1. **Direct Reporters**: Individuals that report the trait regardless of the method of elicitation—$S_i = D_i = 1$.
2. **Nonreporters**: Individuals that do not report the trait regardless of the method of elicitation—$S_i = D_i = 0$.
3. **Misreporters**: These individuals only report anonymously—$S_i \neq D_i$.

Previous method gives us means for Indirect = Direct + Misreporters.

Characterizing Misreporters

We may want to go beyond the full population and consider the misreporter mean specifically.
Consider the following identity (which assumes underreporting):

$$E[X_i|S_i=1] = \underbrace{\Pr(D_i=1|S_i=1)E[X_i|S_i=1,D_i=1]}_{\text{Direct Reporters}}$$

$$+\underbrace{\Pr(D_i=0|S_i=1)E[X_i|S_i=1,D_i=0]}_{\text{Misreporters}}$$

- We already have the LHS from the previous slide.
- $\Pr(D_i=1|S_i=1) = \frac{E[D_i]}{E[S_i]}$ (observed)
- $\Pr(D_i=0|S_i=1) = 1 - \frac{E[D_i]}{E[S_i]}$ (observed)
- $E[X_i|S_i=1,D_i=1] = E[X_i|D_i=1]$ (observed)

Assumptions

1. **Randomization:** Z_i independent of $R_i(z)$ and X_i.
2. **NDE on controls:** long list doesn't change control-item answers $\Rightarrow R_i(1) - R_i(0) = S_i$.
3. **First stage:** $p_S = \Pr(S_i = 1) > 0$ (nonzero denominator).
4. **SUTVA:** $R_i = R_i(Z_i)$ and X_i is pre-treatment.
5. **Reporting monotonicity:** either $S_i \geq D_i$ or $D_i \geq S_i$ $\forall i$.

More Details/Extensions:

▸ Proof of Concept ▸ Inverse Procedure ▸ Improving Precision ▸ Prediction ▸ Causal Inference

Placebo Validation

Can validate by estimating covariate means for items that are not sensitive.

Coffman et al (2017) run list experiments with placebo items, e.g.,:

- Are you completing this survey from a laptop computer?
- Are you wearing a long-sleeved shirt right now?
- Are you wearing a wristwatch right now?

They find **no differences** in direct and indirect reporting for these items.

I use complier methods on their data to characterize the indirect reporters of these items on political and religious scales—should expect no difference vs. direct reporters.

Good Targeting for Placebo Items

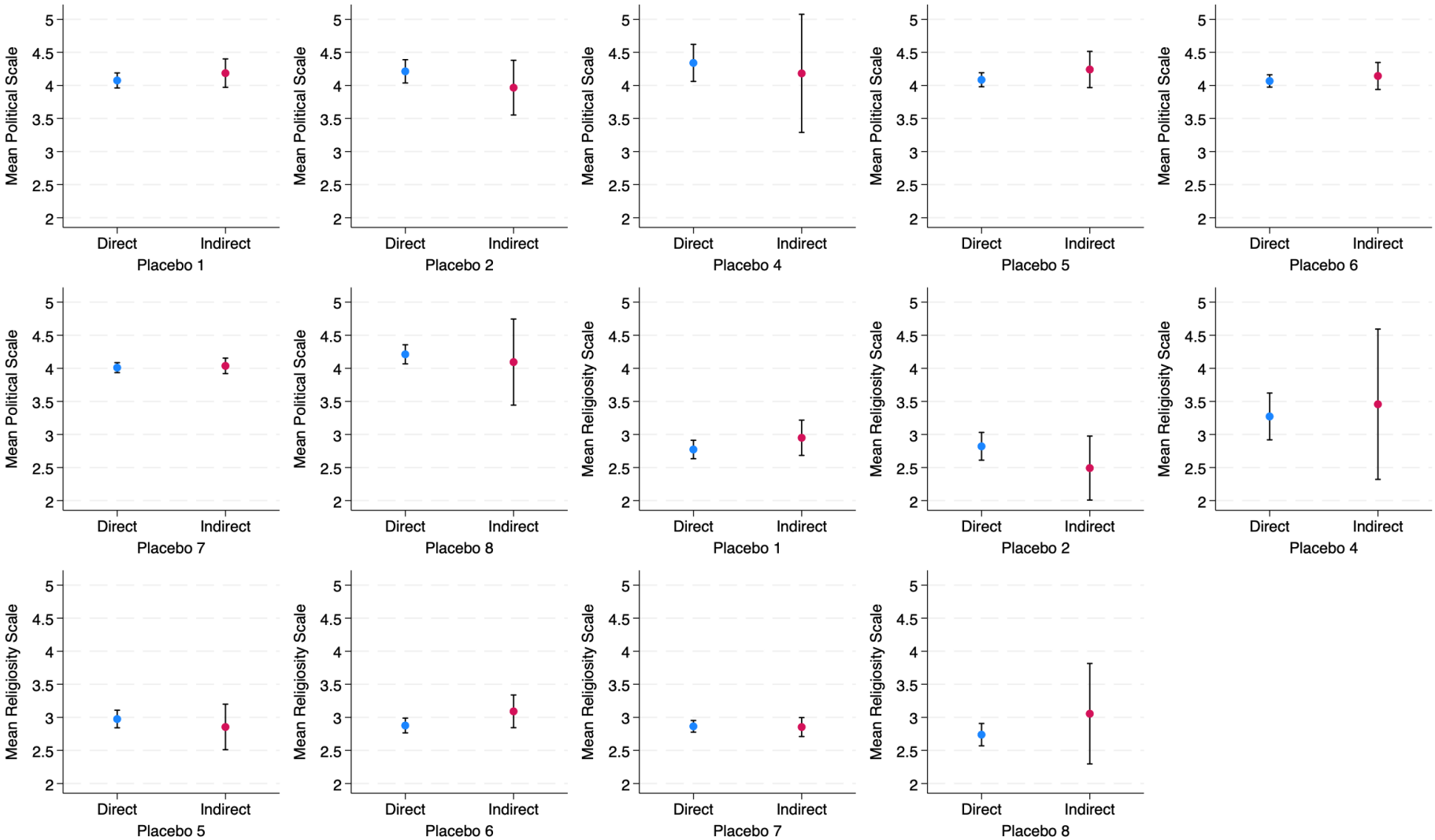


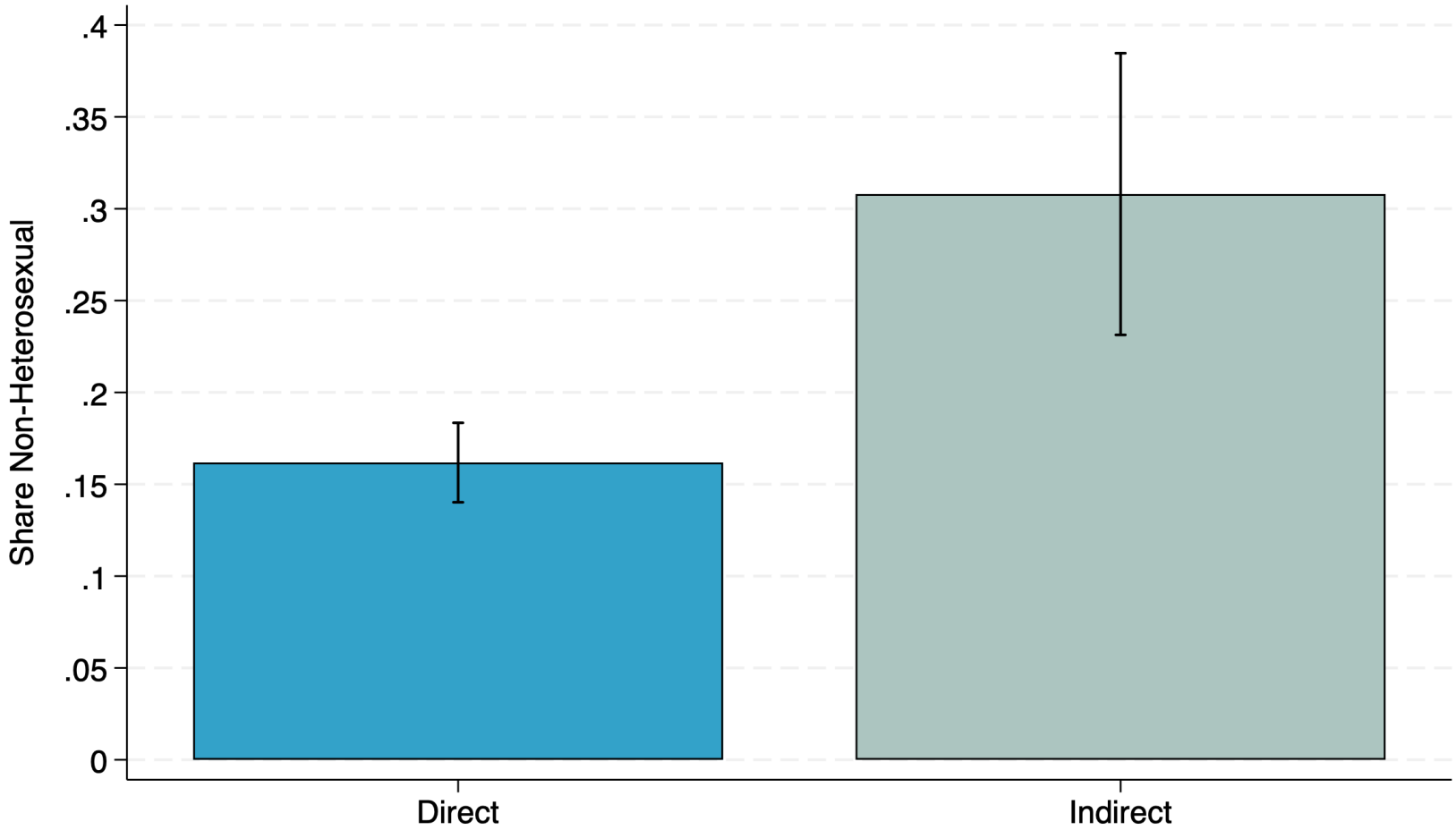
Table of Contents

Sample

Field survey on Prolific

- 2501 respondents
- Representative on sex, age, political affiliation ▸ ACS Comparison
- Needed > 100 previous surveys, $> 95\%$ approval rate, passed 2 of 3 attention checks ▸ Sample Validity Checks
- Collect standard demographics, self-reported income, PHQ-8 (Depression) and GAD-7 (Anxiety) scores, current and birth zip code
- Instructed on how list experiment works with examples. ▸ Examples
- 1,253 respondents saw the veiled report treatment (long lists), 1,248 saw direct report (short lists). ▸ Randomization Worked

First Stage on LGBQ Identification



First Stage Results

▸ Other Measures of Sexuality

Validity Concerns

1. **No Design Effects:** Adding the sensitive item doesn't change the answers to control items.
   - **Middle Option or Random Choice:** If people were simply picking randomly or the middle option, mechanical increase in their response total.
2. **No Liars:** People answer truthfully for the sensitive item.
   - **Ceiling/Floor Effects:** If a large fraction of the sample is answering $R_i = 5$ or $R_i = 0$, then they may be shading their answers due to the lack of privacy, which narrows the pool we can characterize.
3. **Reporting Monotonicity:** All direct reporters would also indirectly report.

Validity Concerns

1. **No Design Effects:** Adding the sensitive item doesn't change the answers to control items.
   - **Middle Option or Random Choice:** If people were simply picking randomly or the middle option, mechanical increase in their response total.
2. **No Liars:** People answer truthfully for the sensitive item.
   - **Ceiling/Floor Effects:** If a large fraction of the sample is answering $R_i = 5$ or $R_i = 0$, then they may be shading their answers due to the lack of privacy, which narrows the pool we can characterize.
3. **Reporting Monotonicity:** All direct reporters would also indirectly report.

I conduct a series of tests to detect design and ceiling/floor effects and find no evidence of either for my sexuality measure. Similarly, find evidence that aligns with reporting monotonicity.

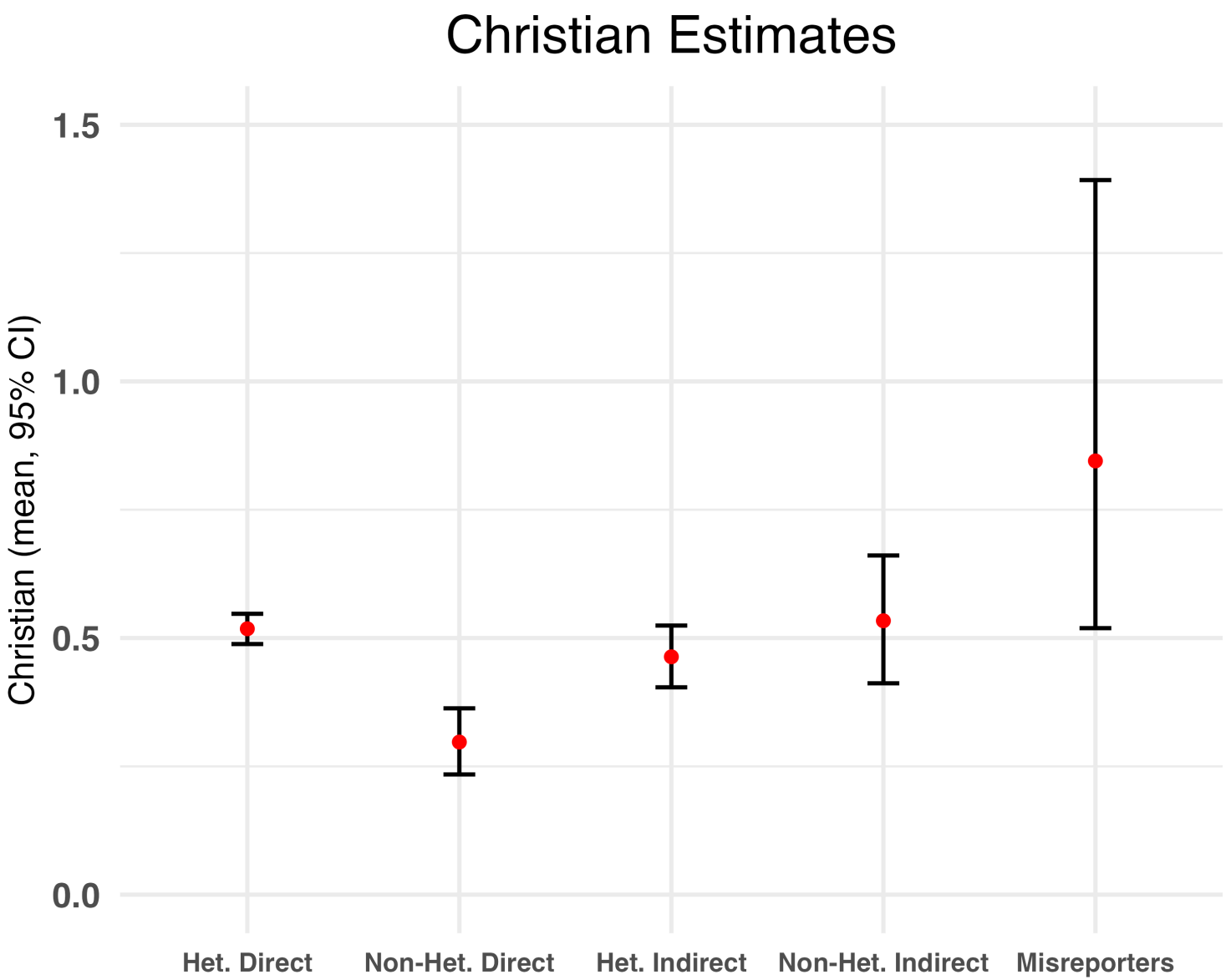
▸ Validity Tests

Table of Contents

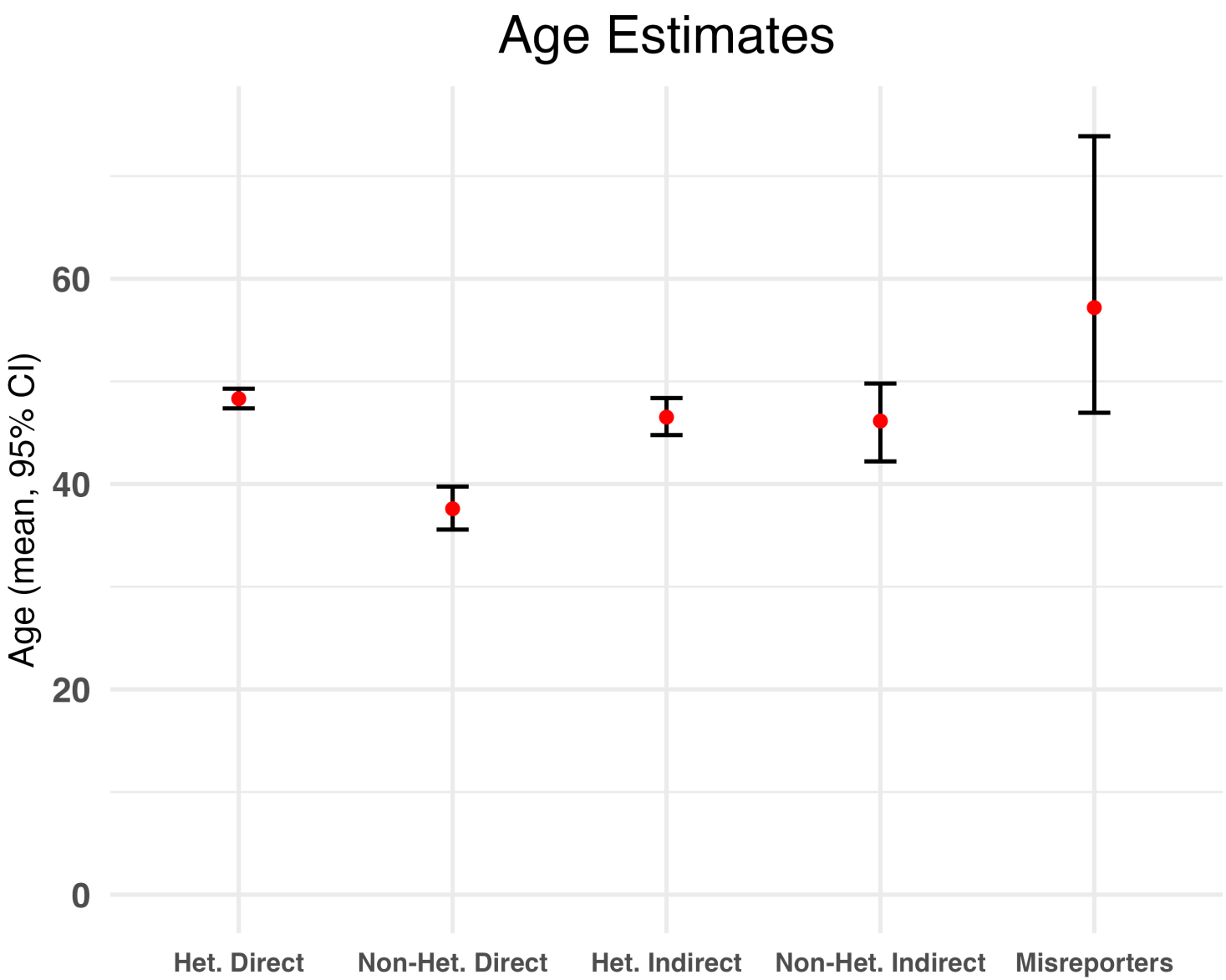
Results

- Estimate:
  1. **Direct means** for heterosexual and non-heterosexual ($E[X_i|\tilde{Y}_i = 1], E[X_i|\tilde{Y}_i = 0]$)
  2. **Indirect means** for heterosexual and non-heterosexual (Wald ratios from earlier slides)
  3. **Correction**: How the indirect method results compare to direct method (and previous literature).
  4. **Misreporter mean** using weighted average identity
- Inference:
  1. Bayesian Bootstrap with 10,000 replications
  2. Control for covariates to improve precision, similar results w/o controls

Christian Share



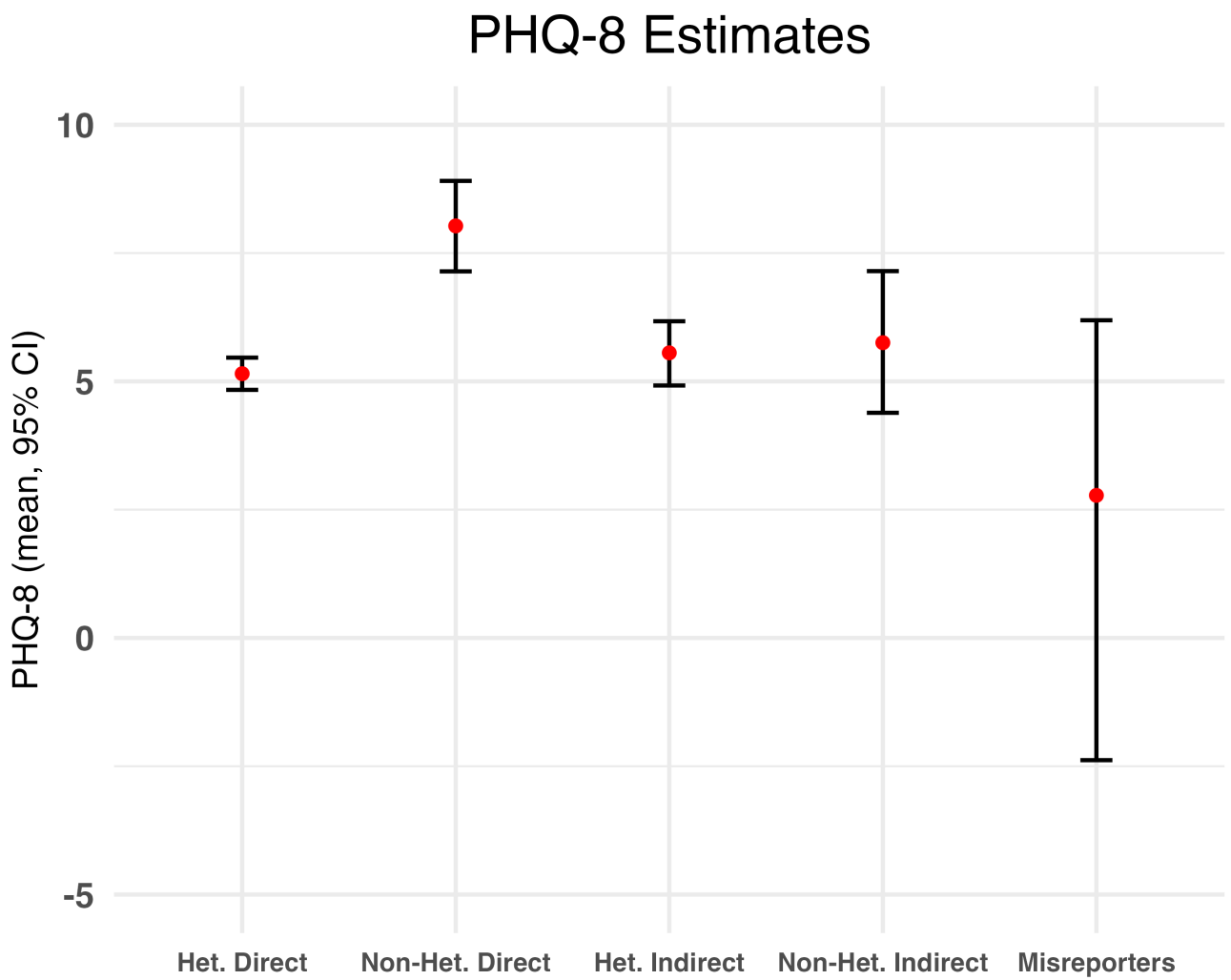
Age


Age Estimates
Age (mean, 95% CI)
60
40
20
0
Het. Direct
Non-Het. Direct
Het. Indirect
Non-Het. Indirect
Misreporters

Understanding LGBQ Mental Health and Earnings

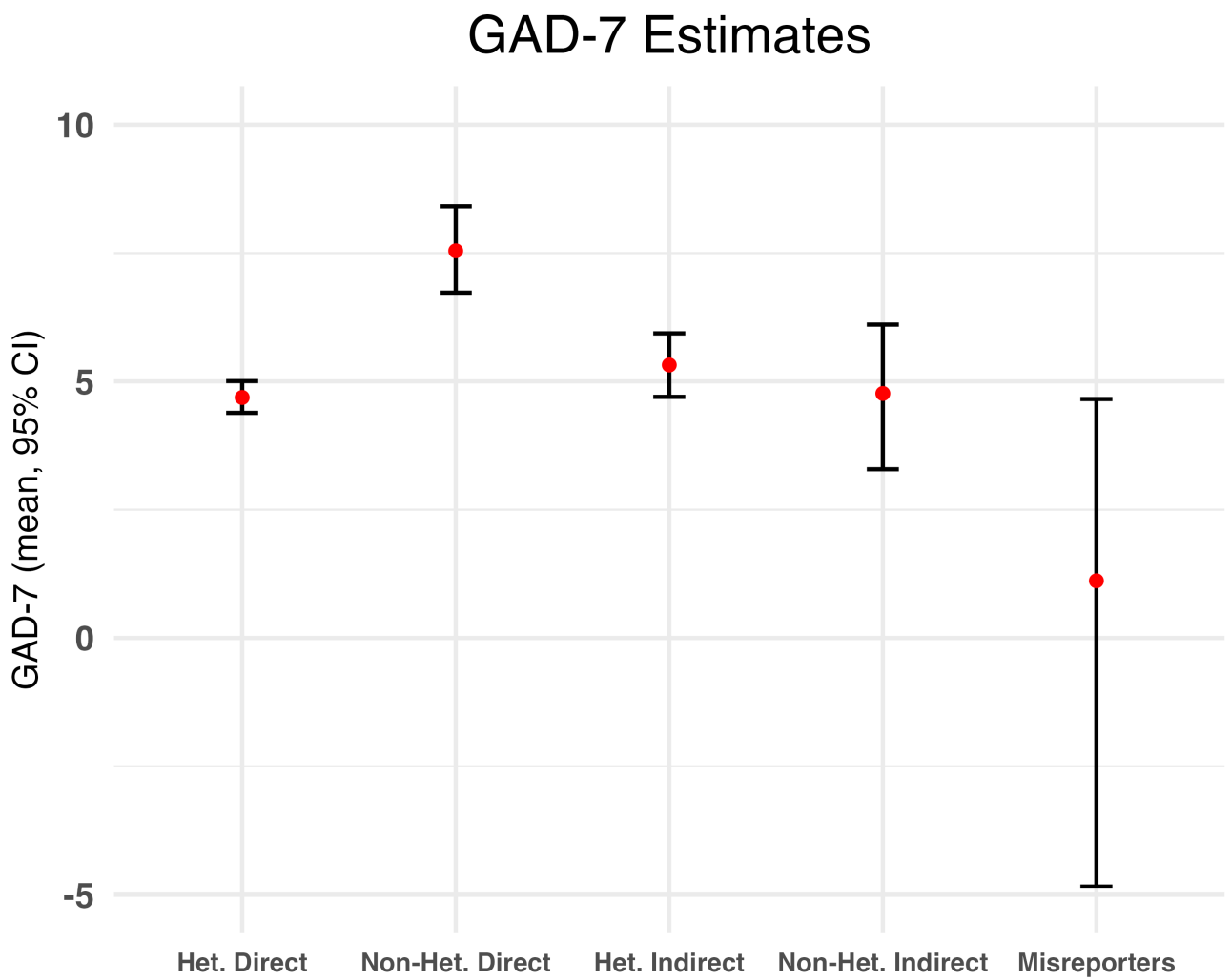
- Examine three main covariates:
  1. PHQ-8 Score: Standard depression screening tool, ranges from 0-24. Higher is worse.
  2. GAD-7 Score: Standard anxiety screening tool, ranges from 0-21. Higher is worse.
  3. Self-reported income: Reported past year individual income.

PHQ-8 Differences



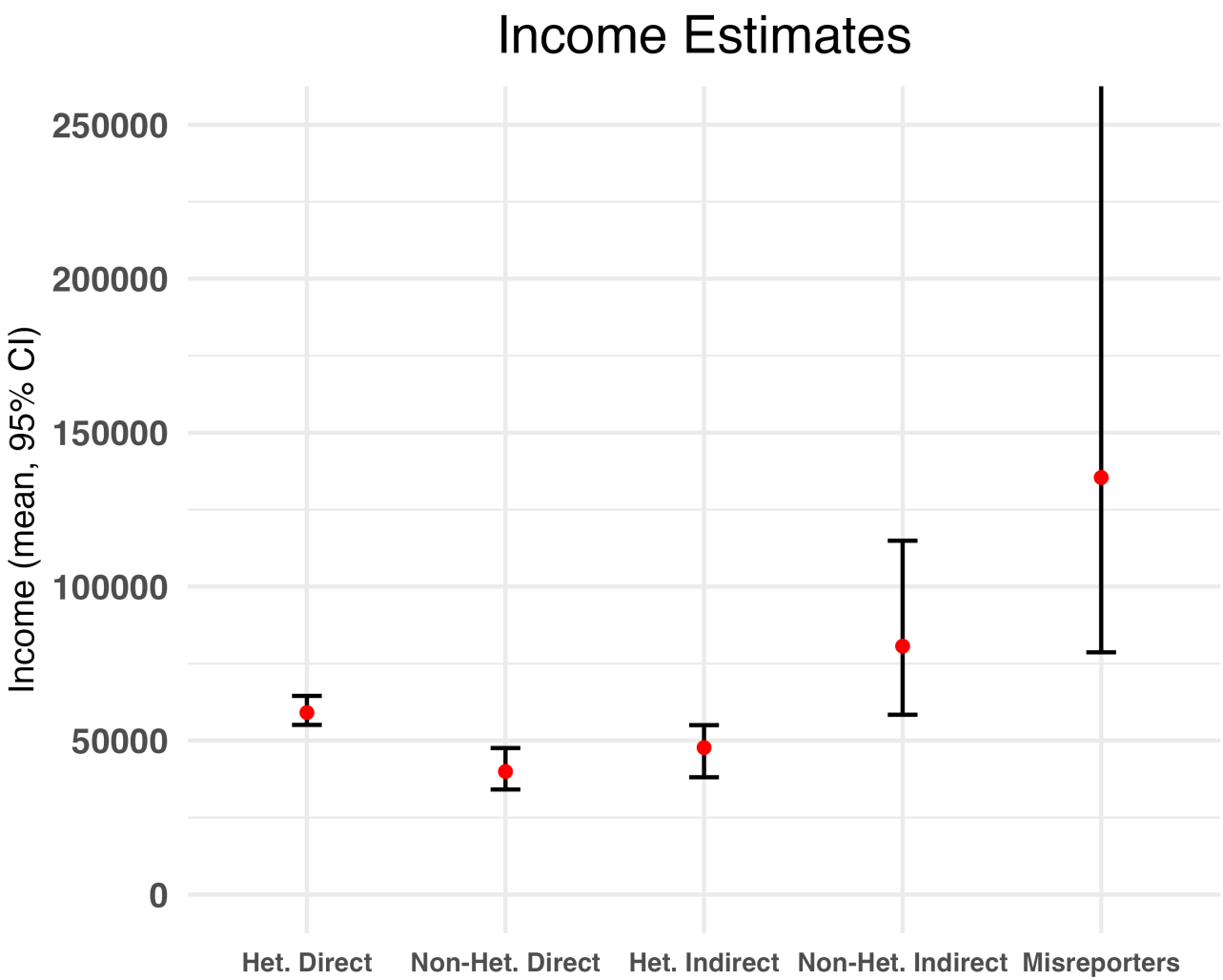
▸ First Stage Heterogeneity ▸ Differences by Gender

GAD-7 Differences



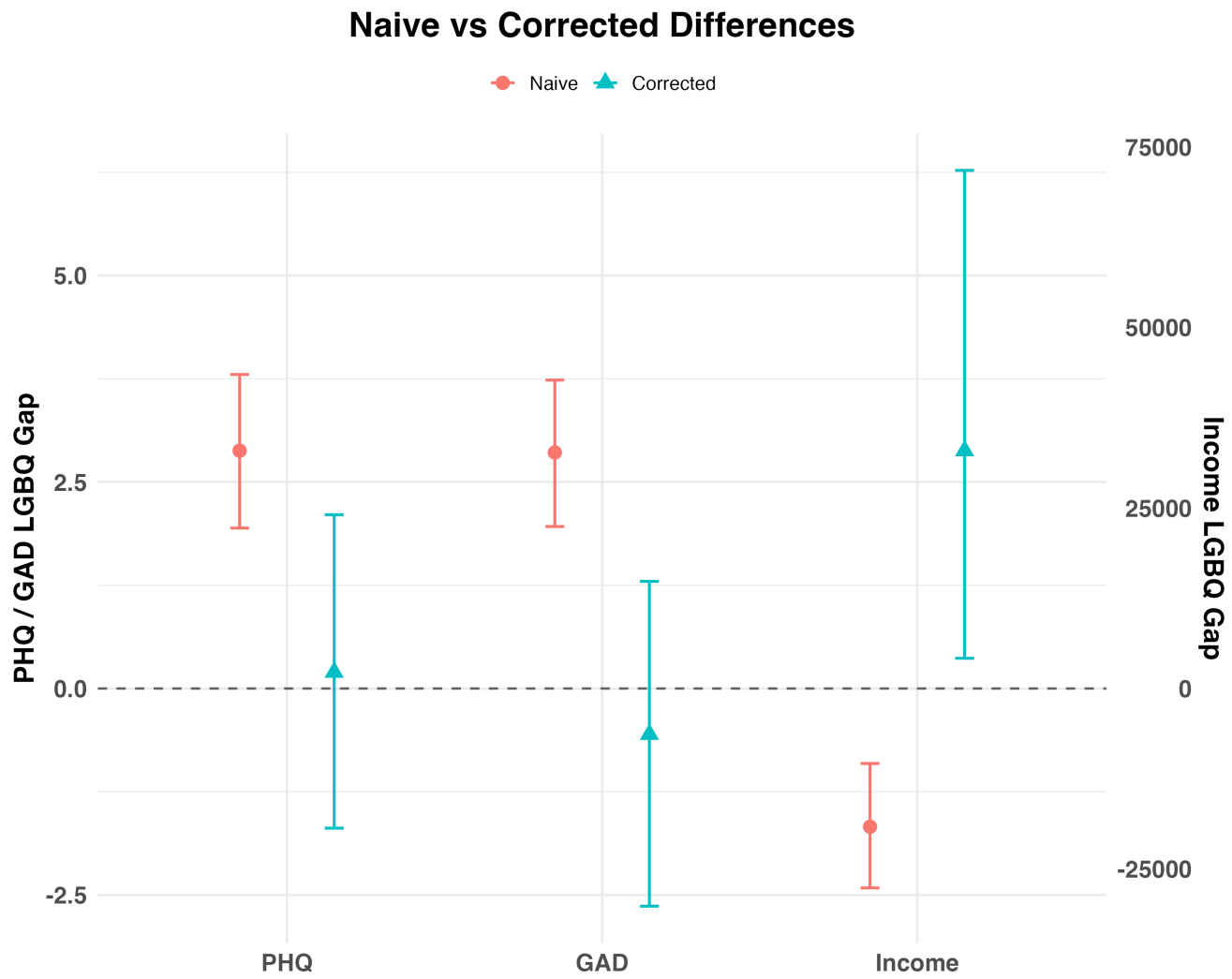
▸ First Stage Heterogeneity

Income Differences



▸ First Stage Heterogeneity ▸ Differences by Gender

Do the Corrections Matter?



▸ Benchmarking Against Prior Literature

Potential Explanations for Misreporters

1. **Composition Effects:** These respondents are different on demographic characteristics (age, sex, race) that correlate with lower mental distress, higher income.
    - Test by first residualizing on demographics (age, sex, race, religion).

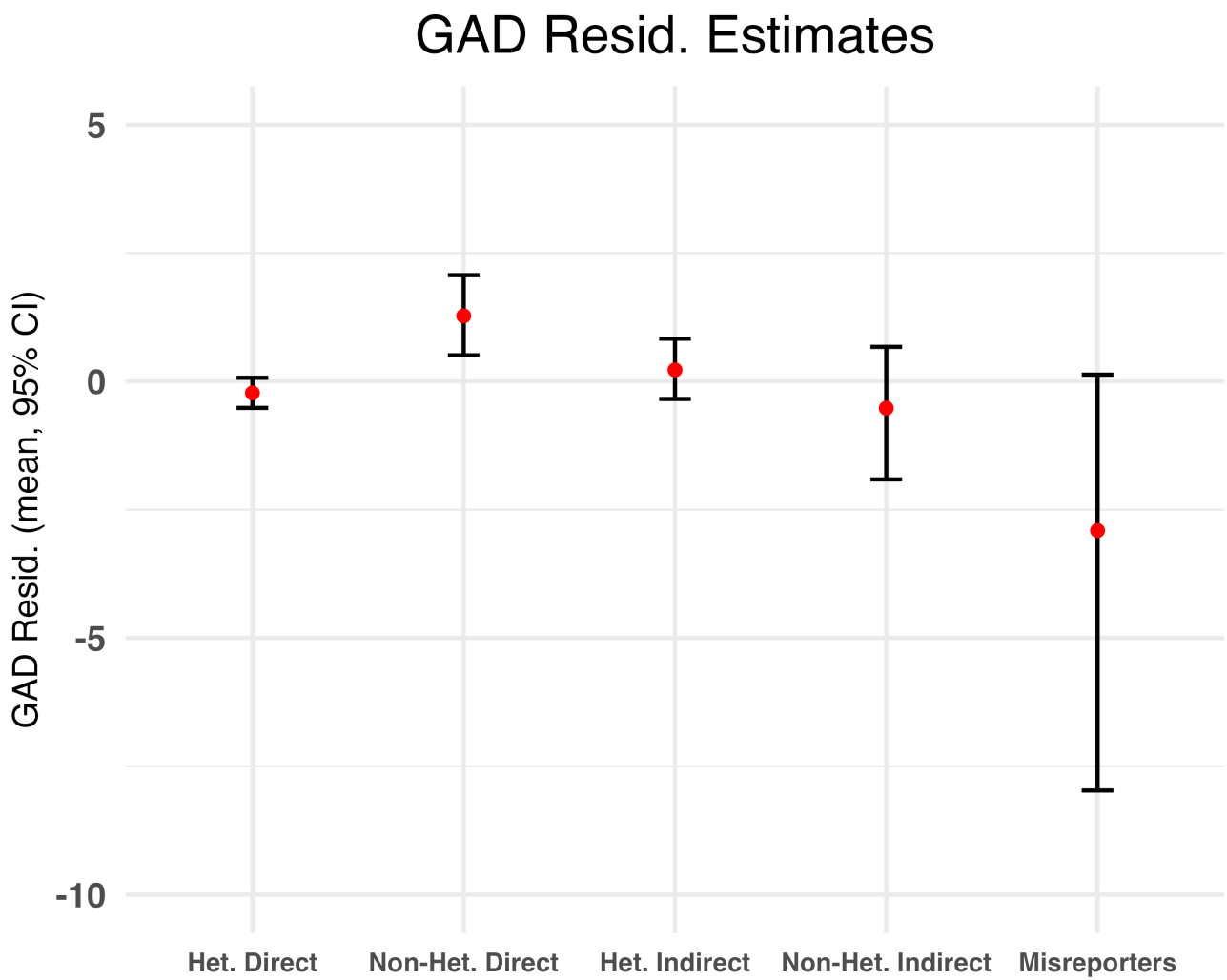
Potential Explanations for Misreporters

1. **Composition Effects:** These respondents are different on demographic characteristics (age, sex, race) that correlate with lower mental distress, higher income.
   - Test by first residualizing on demographics (age, sex, race, religion).
2. **Underreporting Across Domains:** These respondents underreport both minority sexual identity and mental distress
   - Find no evidence of underreporting of mental distress (separately).
   - Test with anonymous elicitation mental health measure.

Potential Explanations for Misreporters

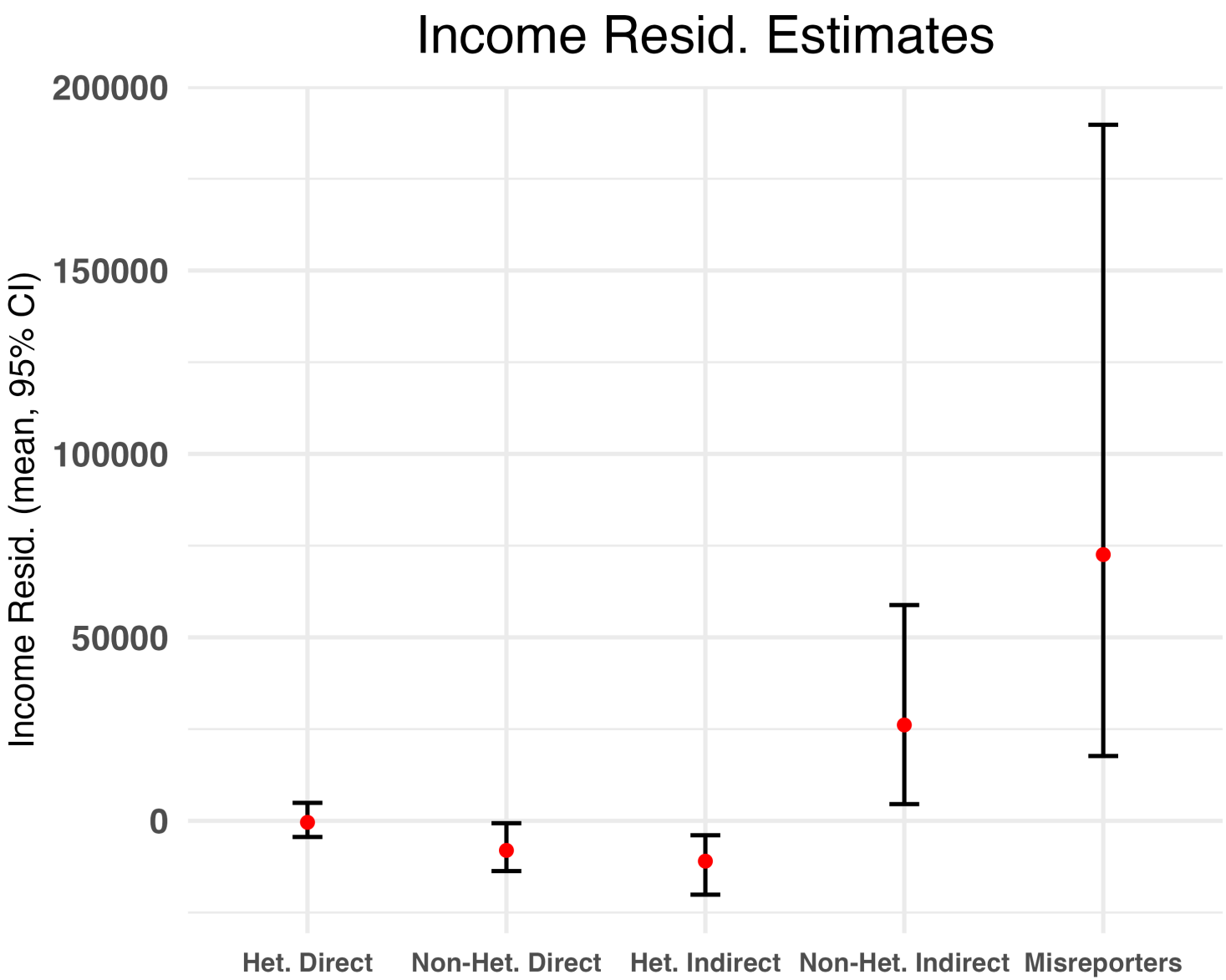
1. **Composition Effects:** These respondents are different on demographic characteristics (age, sex, race) that correlate with lower mental distress, higher income.
   - Test by first residualizing on demographics (age, sex, race, religion).
2. **Underreporting Across Domains:** These respondents underreport both minority sexual identity and mental distress
   - Find no evidence of underreporting of mental distress (separately).
   - Test with anonymous elicitation mental health measure.
3. **Costs of Reporting exceed Costs of Concealment:** Perhaps these people are rationally underreporting because they perceive costs to coming out that exceed costs to concealing identity.
   - Examine by characterizing LGBQ+ support/political attitudes in local ZIP code.
   - Future: Elicit discrimination expectations/perceptions

Residualized GAD Differences

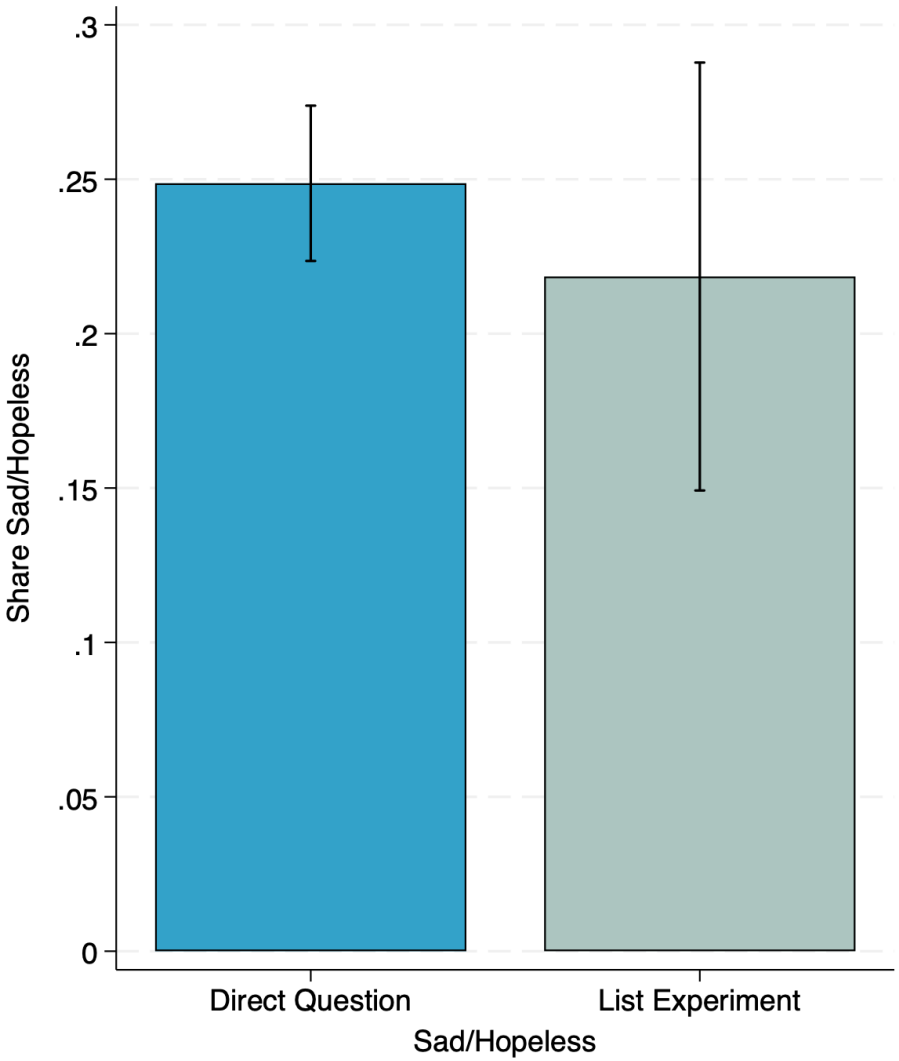


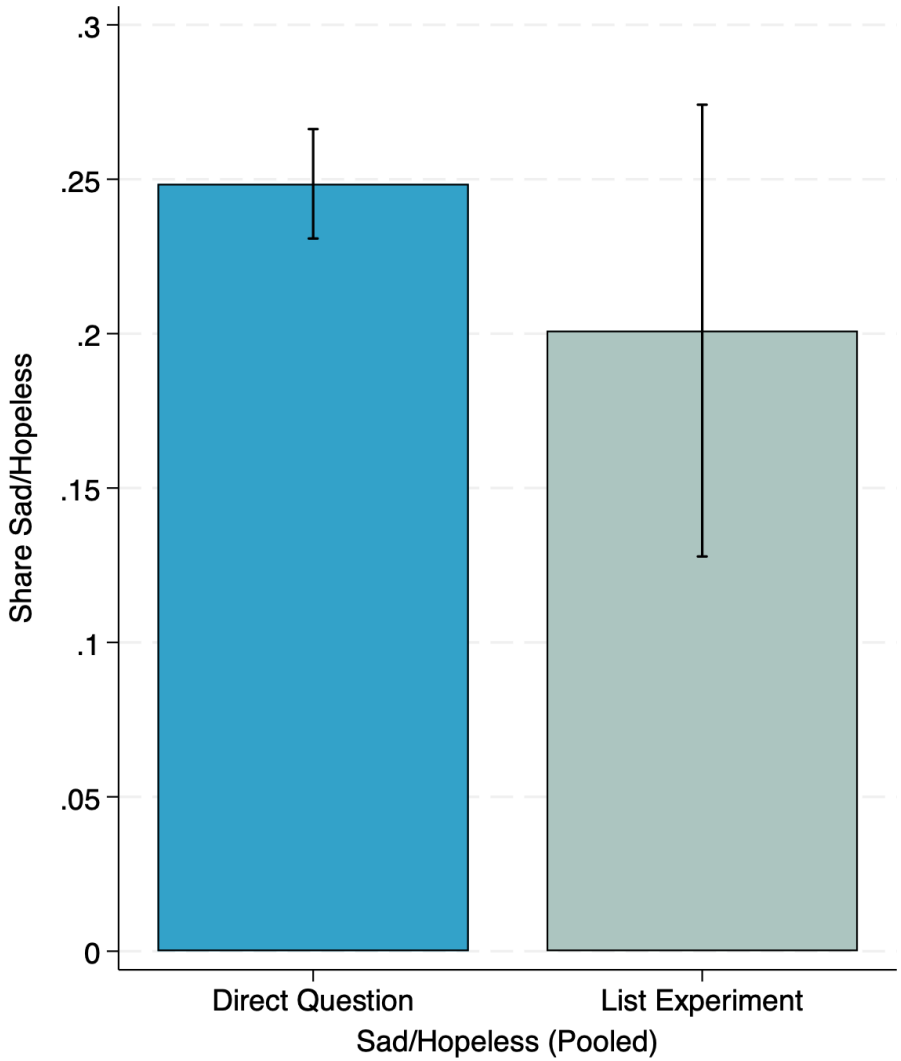
▸ PHQ Residualized Gaps

Residualized Income Differences

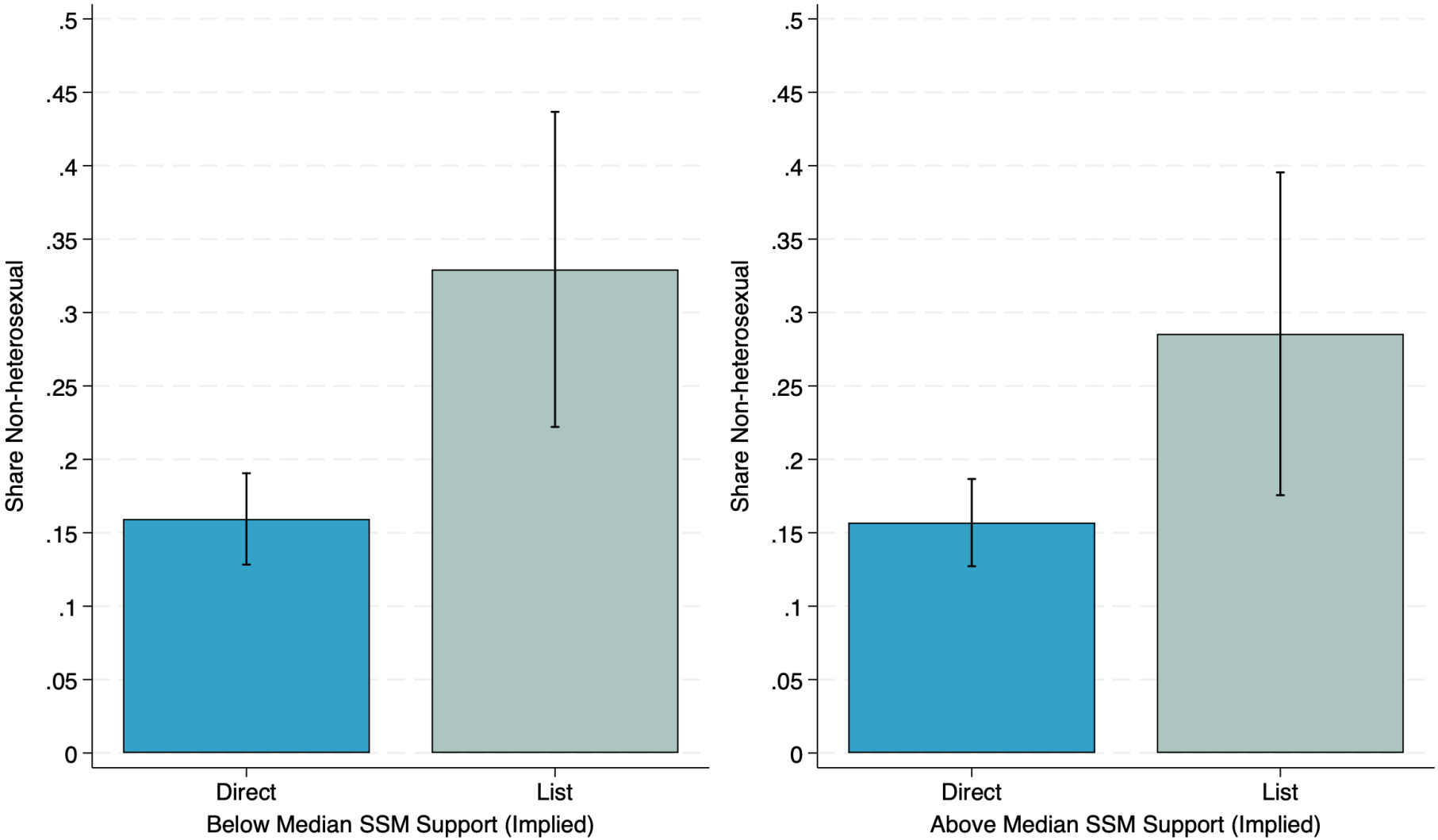


No Mental Health Underreporting



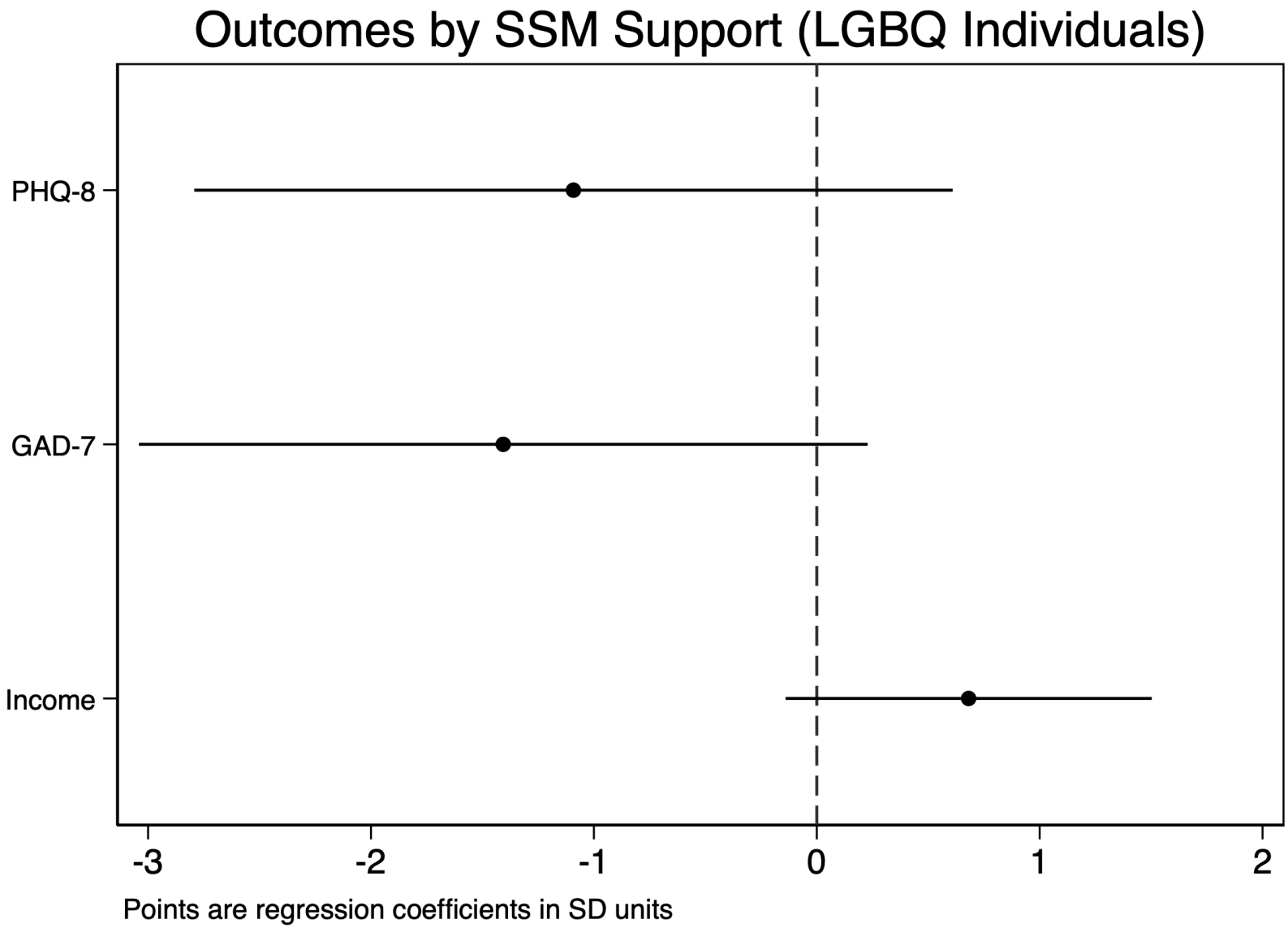


Underreporting Slightly Larger in Low Support Areas



▸ Mean Support by Reporting Status

Outcome Gaps by SSM Support



Evidence on Interpretation

- **Demographics reduces differences**: Direct vs. Misreporter gap shrinks:
  - 5.2 $\Rightarrow$ 2.8 PHQ Score
  - 6.3 $\Rightarrow$ 4.1 GAD Score
  - Still significant difference in residualized GAD score, income.

Evidence on Interpretation

- **Demographics reduces differences**: Direct vs. Misreporter gap shrinks:
  - 5.2 $\Rightarrow$ 2.8 PHQ Score
  - 6.3 $\Rightarrow$ 4.1 GAD Score
  - Still significant difference in residualized GAD score, income.
- **No evidence for general underreporting**: No differences in MH reporting for direct vs. indirect.

Evidence on Interpretation

- **Demographics reduces differences**: Direct vs. Misreporter gap shrinks:
  - 5.2 $\Rightarrow$ 2.8 PHQ Score
  - 6.3 $\Rightarrow$ 4.1 GAD Score
  - Still significant difference in residualized GAD score, income.
- **No evidence for general underreporting**: No differences in MH reporting for direct vs. indirect.
- **Suggestive evidence for costs of revealing>costs of concealment**
  - Misreporters might live in less supportive areas.
  - Direct reporters might have worse MH and income in less supportive areas.
  - Next steps: elicit discrimination perceptions to test precisely, use panel data to test dynamics of differences.

Conclusion

- Develop toolkit to characterize indirect reporters and misreporters.
- Apply to re-estimate LGBQ earnings and mental health gaps using Prolific survey.
- Findings:
  1. Non-heterosexual share doubles when elicited indirectly
  2. Indirect reporters have same mental health and higher income relative to heterosexuals—in contrast to previous work finding large mental health and earnings penalties.
  3. Misreporters have lower mental distress and higher income than direct reporters.
- Suggestive evidence on role of demographics and costs to reporting; more work to come.

Objects and Assumptions

Objects:

- X_i: Covariate of interest (e.g., earnings, mental health), $g(X_i)$ is arbitrary function
- Z_i: List assignment (=1 if long list, =0 if short list)
- $D_i \in \{0, 1\}$: Direct report
- $S_i \in \{0, 1\}$: Indirect report (sensitive item)
- $C_{ij}(z)$: Item-level potential response for control item j
- $R_i(z) = \sum_{j=1}^{J} C_{ij}(z) + zS_i$: Potential response totals for $z \in \{0, 1\}$

Assumptions:

1. **Independence:** $Z_i \perp\!\!\!\perp (R_i(0), R_i(1), X_i)$
2. **No Design Effects:** $C_{ij}(1) = C_{ij}(0)$
3. **First Stage:** $\Pr[R_i(1) > R_i(0)] \neq 0$
4. **Integrability:** $E[|g(X_i)|] < \infty$
5. **SUTVA:** $R_i = R_i(Z_i)$

No Design Effects ⇒ Monotonicity and Binary Treatment

$$R_i(1) - R_i(0) = \sum_{j=1}^{J} C_{ij}(1) + S_i - \sum_{j=1}^{J} C_{ij}(0)$$

$$= \sum_{j=1}^{J} [C_{ij}(1) - C_{ij}(0)] + S_i$$

$$= S_i \quad \text{(No Design Effects)}$$

$$\Rightarrow R_i(1) - R_i(0) \geq 0; \quad R_i(1) - R_i(0) \in \{0, 1\}$$

- Binary treatment is key for identification of covariate means across single threshold.
- Angrist & Imbens (1995) show that with more levels of ordered/multi-valued treatment, this targets a k-weighted average across complier types.

Identification of Covariate Mean for Compliers

Goal:

$$\frac{E[g(X_i)R_i|Z_i=1]-E[g(X_i)R_i|Z_i=0]}{E[R_i|Z_i=1]-E[R_i|Z_i=0]} = E[g(X_i)|R_i(1)>R_i(0)]$$

Numerator:

$$\begin{aligned}
&E[g(X_i)R_i|Z_i=1]-E[g(X_i)R_i|Z_i=0] \\
&= E[g(X_i)R_i(1)|Z_i=1]-E[g(X_i)R_i(0)|Z_i=0] \\
&= E[g(X_i)R_i(1)]-E[g(X_i)R_i(0)] \quad \text{(Independence)} \\
&= E[g(X_i)\cdot[R_i(1)-R_i(0)]] \quad \text{(Lin. of E[$\cdot$])} \\
&= E[g(X_i)\cdot[R_i(1)-R_i(0)]|R_i(1)>R_i(0)]\cdot \Pr[R_i(1)>R_i(0)] \\
&+E[g(X_i)\cdot[R_i(1)-R_i(0)]|R_i(1)=R_i(0)]\cdot \Pr[R_i(1)=R_i(0)] \quad \text{(LTP)} \\
&= E[g(X_i)|R_i(1)-R_i(0)=1]\cdot \Pr[R_i(1)-R_i(0)=1]
\end{aligned}$$

Continuation

Denominator:

$$E[R_i|Z_i = 1] - E[R_i|Z_i = 0] = E[R_i(1)|Z_i = 1] - E[R_i(0)|Z_i = 0]$$
$$= E[R_i(1) - R_i(0)] \quad \text{(Independence \& Lin. of E[·])}$$
$$= \Pr[R_i(1) - R_i(0) = 1] = \Pr[S_i = 1] = E[S_i]$$

Thus:

$$\frac{E[g(X_i)R_i|Z_i = 1] - E[g(X_i)R_i|Z_i = 0]}{E[R_i|Z_i = 1] - E[R_i|Z_i = 0]}$$
$$= \frac{E[g(X_i)|R_i(1) > R_i(0)] \Pr[R_i(1) - R_i(0) = 1]}{\Pr[R_i(1) - R_i(0) = 1]}$$
$$= E[g(X_i)|R_i(1) > R_i(0)]$$
$$= E[g(X_i)|R_i(1) - R_i(0) = 1]$$
$$= E[g(X_i)|S_i = 1]$$

Giving us the expectation of the function of covariates among indirect reporters.

Identification of Misreporter Mean

Given $E[g(X_i)|S_i = 1]$, we might be interested in further characterizing $E[g(X_i)|D_i \neq S_i]$, where the direct report D_i does not match the indirect report S_i.

We need a further monotonicity assumption:

1. **Reporting Monotonicity**: Either $D_i \geq S_i \; \forall i$ (overreporting) or $S_i \geq D_i \; \forall i$ (underreporting)

If $S_i \geq D_i$, then $D_i = 1 \Rightarrow S_i = 1$ and by law of total expectation:

$$E[g(X_i)|S_i = 1] \Pr[S_i = 1]$$

$$= \underbrace{E[g(X_i)|D_i = 1] \Pr[D_i = 1]}_{\text{Direct Reporters}} + \underbrace{E[g(X_i)|S_i \neq D_i] \Pr[S_i \neq D_i]}_{\text{Misreporters}}$$

Continued

Hence, with $p_S = \Pr[S_i = 1]$, $p_D = \Pr[D_i = 1]$, $\mu_S = E[g(X_i) \mid S_i = 1]$ and $\mu_D = E[g(X_i) \mid D_i = 1]$,

$$E[g(X_i) \mid S_i \neq D_i] = \frac{\mu_S p_S - \mu_D p_D}{p_S - p_D}$$

(If $D_i \geq S_i$ instead, replace numerator and denominator by $\mu_D p_D - \mu_S p_S$ and $p_D - p_S$.)

- μ_S identified via Wald ratio method
- $\mu_D = E[g(X_i) \mid D_i = 1] = E[g(X_i) \mid D_i = 1, Z_i = 0]$ (support), calculated from short list portion of sample
- $p_S = E[S_i] = E[R_i(1) - R_i(0)]$ identified by LE
- $p_D = E[D_i] = E[D_i | Z_i = 0]$ (support), calculated from short list portion of sample.

◂ Back to Method

Proof of Concept using Coffman et al 2017

- I characterize religion for individuals who do/do not report same-sex sexual experiences.
- Among respondents reporting same-sex sexual experiences, only 23% identify as Christian—compared to 36% in the full sample.
- Is this a true difference or social desirability bias?
- Using my method, the estimated Christian share among compliers (direct + misreporters) is 33.8%, much closer to the full sample:

$$\frac{E[C_i R_i | Z_i = 1] - E[C_i R_i | Z_i = 0]}{E[R_i | Z_i = 1] - E[R_i | Z_i = 0]} = 0.338$$

Proof of Concept for Misreporters

- We can also characterize religion for the misreporters specifically (assuming underreporting).
- First, we need the relative shares of misreporters ($S_i > D_i$) and direct reporters ($S_i = D_i = 1$):

$$E[D_i] = 0.172, \quad E[S_i] - E[D_i] = 0.288 - 0.172 = 0.116$$

- $\Rightarrow \Pr(D_i = 1|S_i = 1) = 59.7\%, \Pr(D_i = 0|S_i = 1) = 40.3\%$
- Compliers include both direct reporters (59.7%) and misreporters (40.3%).

$$\Rightarrow 0.338 = 0.597 \cdot 0.234 + 0.403 \cdot C$$

Solving for C, we find that 49.1% of misreporters are Christian. ▸ Return

Inverse Procedure

My approach estimates $E[C_i|S_i = 1]$, the mean Christian share when sensitive item S_i (same-sex sexual experiences) is 1.

Equivalently, because C_i is binary, by Bayes rule:

$$E[C_i|S_i = 1] = \Pr(C_i = 1|S_i = 1) = \frac{\Pr(C_i = 1, S_i = 1)}{\Pr(S_i = 1)}$$

$$= \frac{\Pr(S_i = 1|C_i = 1)\Pr(C_i = 1)}{\Pr(S_i = 1)}$$

- Thus, if we estimate the first stage among Christians ($\Pr(S_i = 1|C_i = 1)$), multiply by the Christian share, and divide by the overall first stage, we can recover the same quantity.
- Works nicely for binary characteristics, but need additional assumptions for continuous variables.

▸ Return

Improving Precision via Controls

In practice, will estimate complier mean using regressions of the form (defining $Y_i \equiv X_i R_i$):

$$Y_i = \beta Z_i + \varepsilon_i$$

$$R_i = \delta Z_i + \varepsilon_i$$

Use ratio β/δ to estimate complier mean, but is very noisy.

Solution: Use regressions of form (where W_i is other covariates):

$$Y_i = \tilde{\beta} Z_i + \gamma X_i + \alpha W_i + \varepsilon_i$$

$$R_i = \tilde{\delta} Z_i + \phi X_i + \kappa W_i + \varepsilon_i$$

Because $Z_i \perp\!\!\!\perp X_i, W_i$ by construction, can get large precision gains while reducing residual variance in Y_i.

OVB formula tells us $\beta = \tilde{\beta} + \gamma \cdot \text{Coef}(X_i, Z_i) + \alpha \cdot \text{Coef}(W_i, Z_i)$

$\Rightarrow \beta = \tilde{\beta}$; $\quad (\text{Coef}(X_i, Z_i) = 0, \text{Coef}(W_i, Z_i) = 0$ by independence) ▸ Return

Prediction with Misreporting

- **Goal:** Predict binary outcome $Y_i \in \{0, 1\}$ from covariates X_i, but training data contains misreporting.
- Common with socially sensitive topics: substance use, cheating, AI use, voting, etc.
- Misreporting skews estimates of $\hat{f}(X_i) = \Pr(Y_i = 1|X_i)$, leading to:
  - Biased predictions.
  - Unequal bias across groups if misreporting varies by covariates.
- List experiments (LE) can estimate overall prevalence of $Y_i = 1$, but cannot identify who misreports.
- I propose an alternative: correct the covariate distribution using LE-derived estimates.

Example: Linear Regression

- Goal: Predict a binary outcome $Y_i \in \{0, 1\}$ from covariates $X_i \in \mathbb{R}^K$
- True model: $E^*[Y_i \mid X_i] = X_i^\top \beta$
- But we observe misreported outcome $\tilde{Y}_i$, so OLS fits:

$$\tilde{\beta} = \hat{\Sigma}_{XX}^{-1} \cdot E[X_i \tilde{Y}_i]$$

- This leads to biased predictions if:
  - Respondents under-report or over-report Y_i
  - Misreporting varies with covariates X_i
- Need a way to recover $E[X_i Y_i]$ from observed data.

Bias-Corrected Linear Prediction

- For binary Y_i, population moment:

$$E[X_i Y_i] = E[X_i \mid Y_i = 1] \cdot \Pr(Y_i = 1)$$

- A list experiment delivers both components:
  1. Prevalence: $\hat{\pi} = \Pr(Y_i = 1) = E[R_i \mid Z_i = 1] - E[R_i \mid Z_i = 0]$
  2. Conditional means:

$$\widehat{E[X_i \mid Y_i = 1]} = \frac{E[X_i R_i \mid Z_i = 1] - E[X_i R_i \mid Z_i = 0]}{\hat{\pi}}$$

- Then, compute bias-corrected coefficients:

$$\hat{\beta} = \hat{\Sigma}_{XX}^{-1} \left(\widehat{E[X_i \mid Y_i = 1]} \cdot \hat{\pi} \right) \quad \text{with} \quad \hat{\Sigma}_{XX} = \frac{1}{n} \sum_{i=1}^{n} X_i X_i^\top$$

Logistic/Probit Regression

For logistic and probit regression, the moment condition(s) can be written as:

$$E[(Y_i - P(X_i))X_i] = 0$$

Where $P(\cdot)$ is $\Phi(X_i\beta)$ for probit regression and $\frac{1}{1+\exp(-X_i\beta)}$ for logistic regression. Thus, the moment condition $E[Y_iX_i]$ can be corrected, and we can obtain unbiased predictions in these settings.

- Extendable to polynomial terms, interactions, series regressions.
- Can we extend to LASSO, Classification Tree, Random Forest?

▸ Return

Digression into Causal Inference

Say we're interested in measuring the effect of an treatment on a misreported outcome (e.g., mental health intervention on suicidal thoughts).

- Could cross-randomize treatment ($D_i = 1$) with list experiment, calculate $E[R_i|Z_i = 1] - E[R_i|Z_i = 0]$ separately for both arms. May not be available (separate data from experiment).
- Alternatively, use auxiliary list experiment data. Assume:
  1. Treatment D_i doesn't affect level of misreporting.
  2. Misreporting is fully characterized by covariates X_i
- Then, correct for misreporting and evaluate treatment. Use predictions $\hat{Y}_i$ from corrected predictive algorithm and regress on treatment indicator D_i.

▸ Return

Sample Representativeness

Return

	Sample	Benchmark	Difference
High school or less	0.139	0.370	-0.231
Some college	0.348	0.250	0.098
College graduate	0.511	0.370	0.141
White	0.816	0.753	0.064
Black	0.116	0.137	-0.020
American Indian / Alaska Native	0.023	0.013	0.010
Asian	0.069	0.064	0.005
NH / Pacific Islander	0.002	0.003	-0.001
Other / multiracial	0.025	0.024	0.001
Hispanic (any race)	0.093	0.195	-0.102
Employed (16+)	0.666	0.603	0.063
Unemployed (16+, LF)	0.082	0.036	0.046
Individual income, mean	57466	59430	-1964
Individual income, median	45000	40480	4520

Sample Validity

- **ReCaptcha:** 90.4% of sample got perfect score, 99.6% above 0.7 threshold.
  - Results unchanged when restricting to perfect scores.
- **Difficult Attention Check**: Respondents were told early in survey they would be asked their favorite food at the end and to respond "Blueberries" no matter what.
  - 82.5% of sample passed.
  - Results unchanged when restricting to this sample.
- **Attrition:** 11% of sample that began survey attrited—89% completion rate.

◂ Return

Instructions of List Experiment

Instructions

In the following pages, you will be presented with lists of statements that may or may not be true for you. The statements will be about yourself and your views on social issues. We would like to know how many of the statements within each list are true for you.

In these lists, we are **not** asking which specific statements are true for you, we are only asking **how many** of them are true for you.

On the following page, we will give you an example. Please click next when you are ready.

Instructions of List Experiment

We want to be sure that you understand how this works. Here is an example:

- I own an orange t-shirt.
- My household has at least two pets.
- I regularly recycle.

Please select the option that corresponds to the total number of statements above that apply to you.

- 0
- 1
- 2
- 3

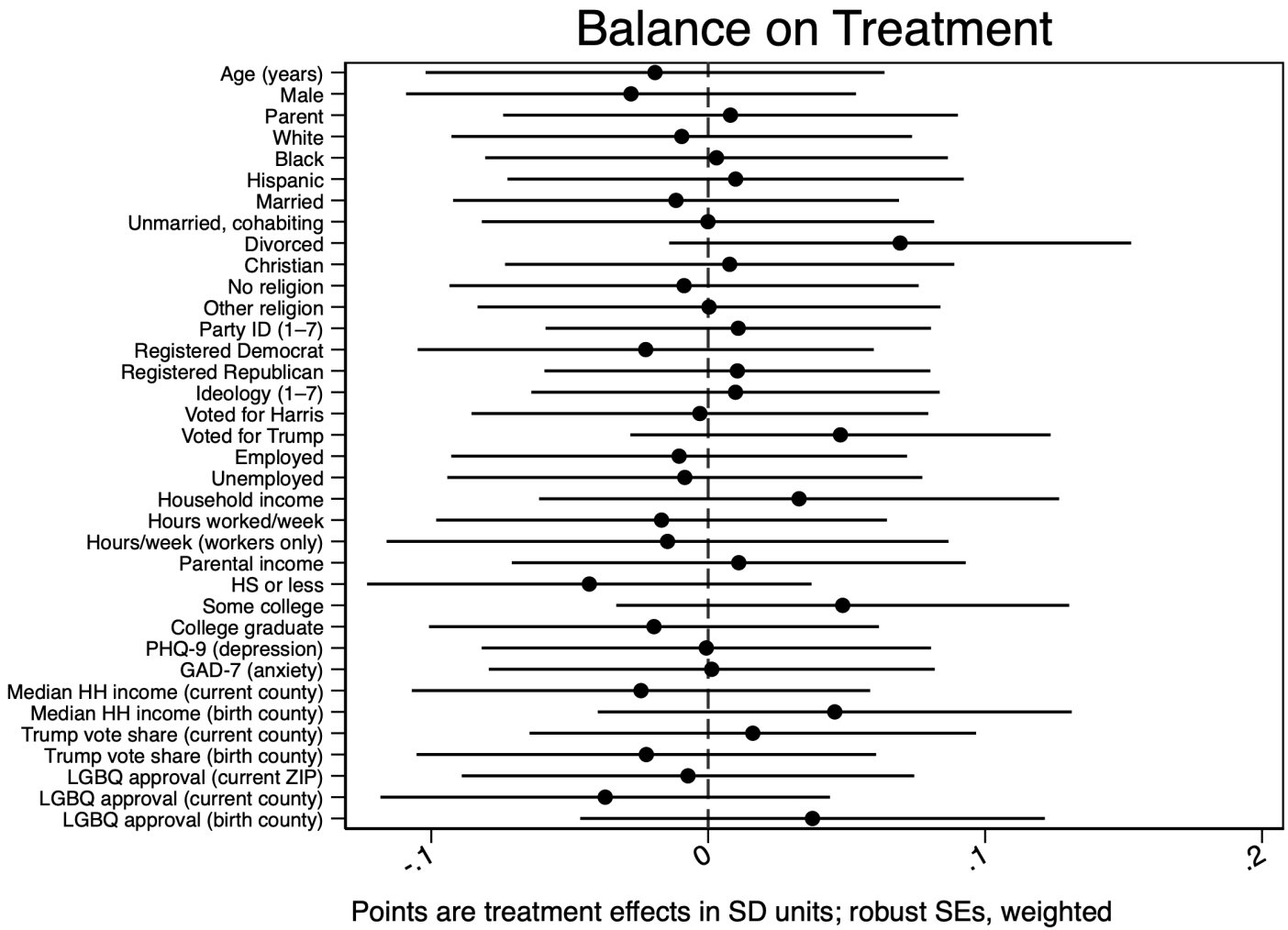
Instructions of List Experiment

Suppose that you do own an orange t-shirt and your household has at least two pets. But you do not recycle. In that case, two of the above statements are true for you. Hence, you would indicate this by entering 2 in the answer box.

Please Note: We are not asking which specific statements in these lists are true for you. We are only asking how many of them are true for you.

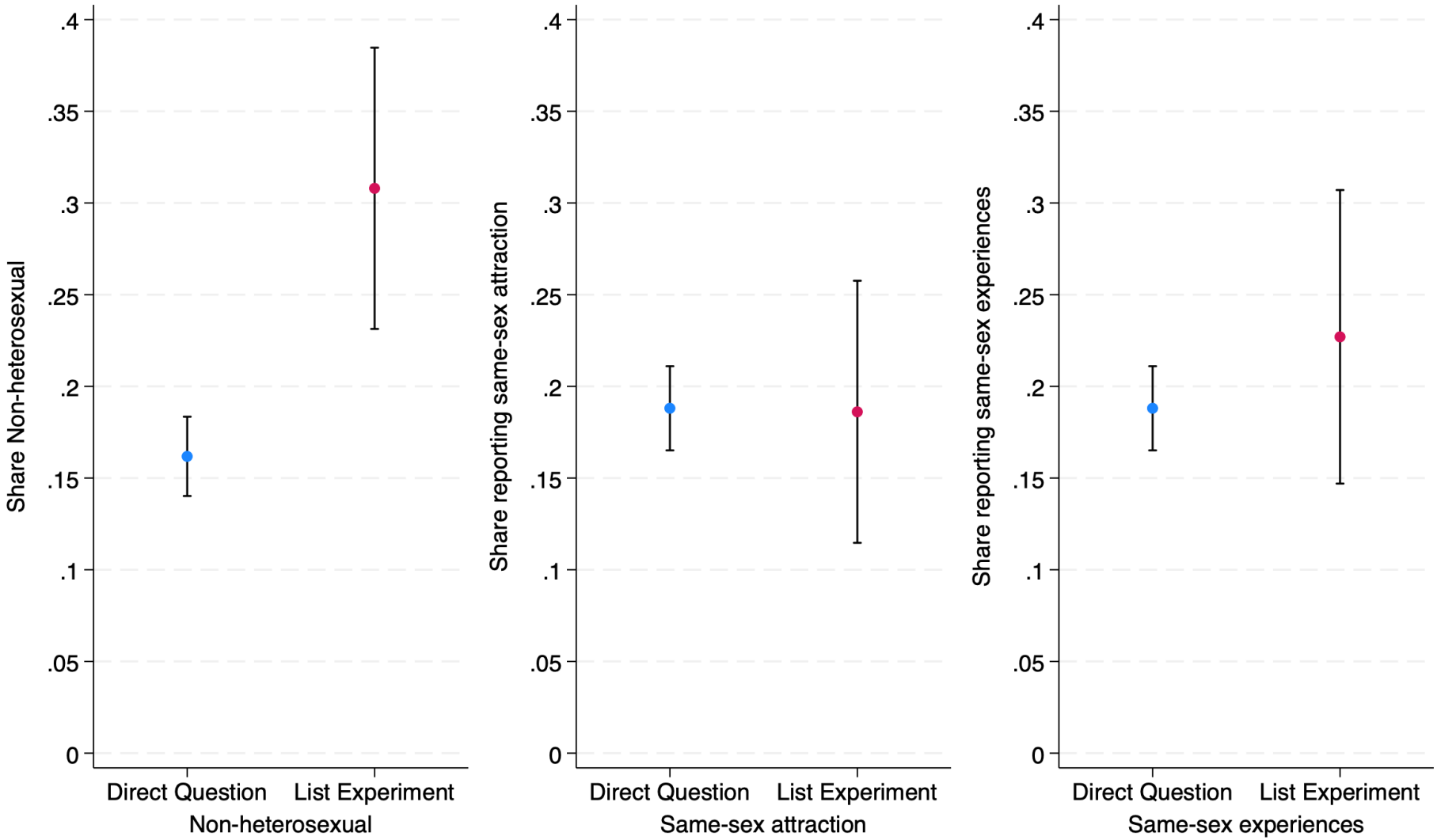
◂ Return

Randomization Worked

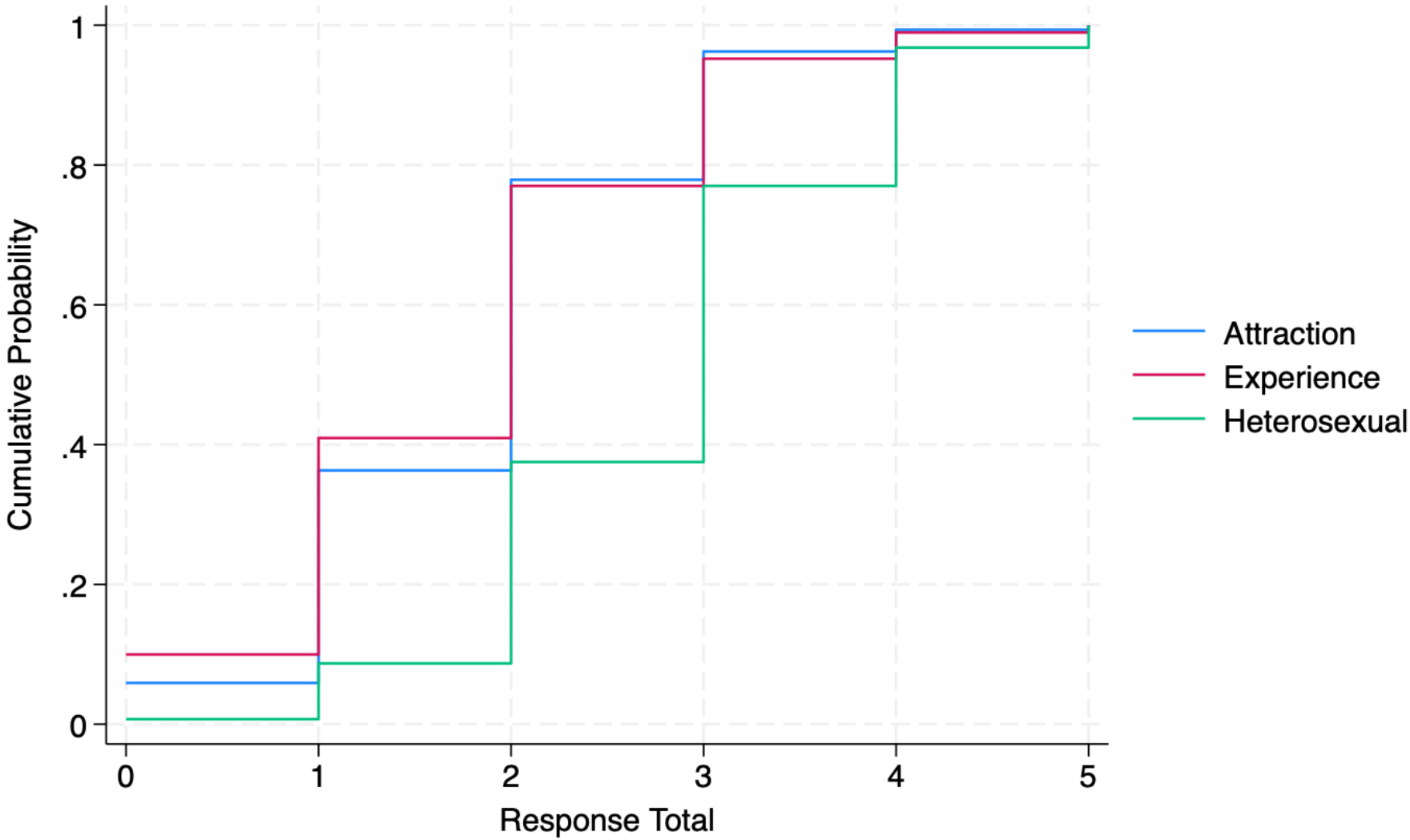


◂ Return

First Stage by Sexuality Measure



CDF of Response Totals by Sexuality Measure



Back to First Stage

Validity Tests

1. Blair and Imai (2012): Tests for Design Effects by examining whether response distributions consistent with the long list increasing response total by greater than 0 and less than 1.
2. Chuang et al (2021): Tests jointly for Design + Ceiling/Floor Effects by examining similarity of LE estimate using two (or more) different control item lists.
3. Ceiling/Floor Bunching: Assess Ceiling/Floor effects by examining level of bunching at bottom and top response level (in treatment arm).
4. Heterogeneity Across Sensitive Items: If respondents choosing randomly or not following instructions, we would not expect heterogeneity in the LE estimate

Blair and Imai (2012)

Blair and Imai are formally testing the joint null hypothesis that:

$$F(r|Z_i = 0) - F(r|Z_i = 1) \geq 0 \, \forall r$$

$$F(r|Z_i = 1) - F(r-1|Z_i = 1) \geq 0 \, \forall r$$

Where r is response total, Z_i is treatment assignment, and $F(\cdot)$ is the CDF of R_i.

- The first inequality tests whether adding the sensitive item can **only increase** the total number of "yes" responses (monotonicity of treatment).
- The second inequality ensures that the treatment can increase the response total by **at most one** (no multiple-response effects).

Chuang et al (2021)

Chuang et al (2021) test whether using different control items affects the list experiment estimate. If we consider:

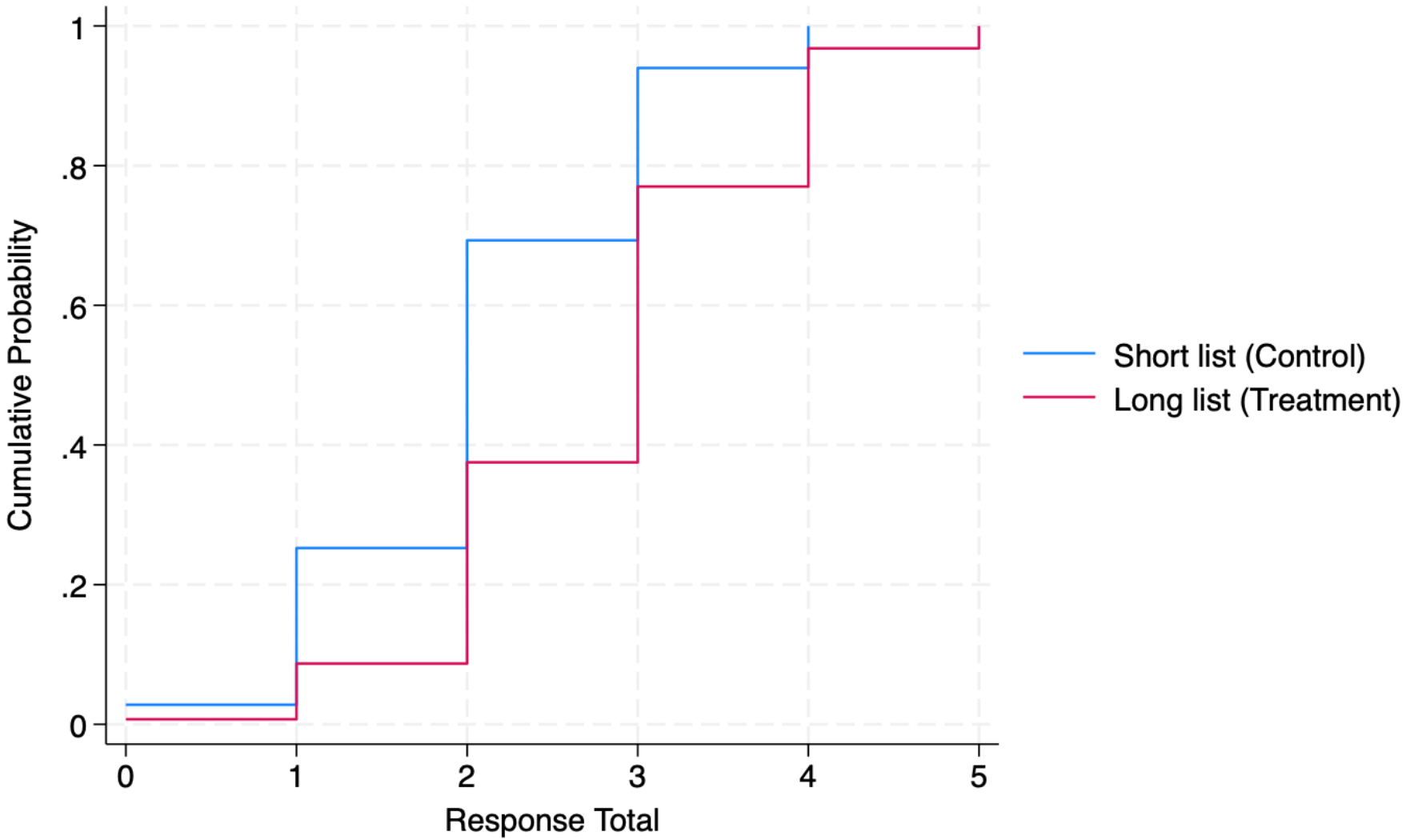
- $R_i^{C_1}$: response totals for control list C_1 and sensitive item S_i
- $R_i^{C_2}$: response totals for control list C_2 and sensitive item S_i
- Z_i treatment indicator—can be same or different for each control list.

Then, we test the null hypothesis that:

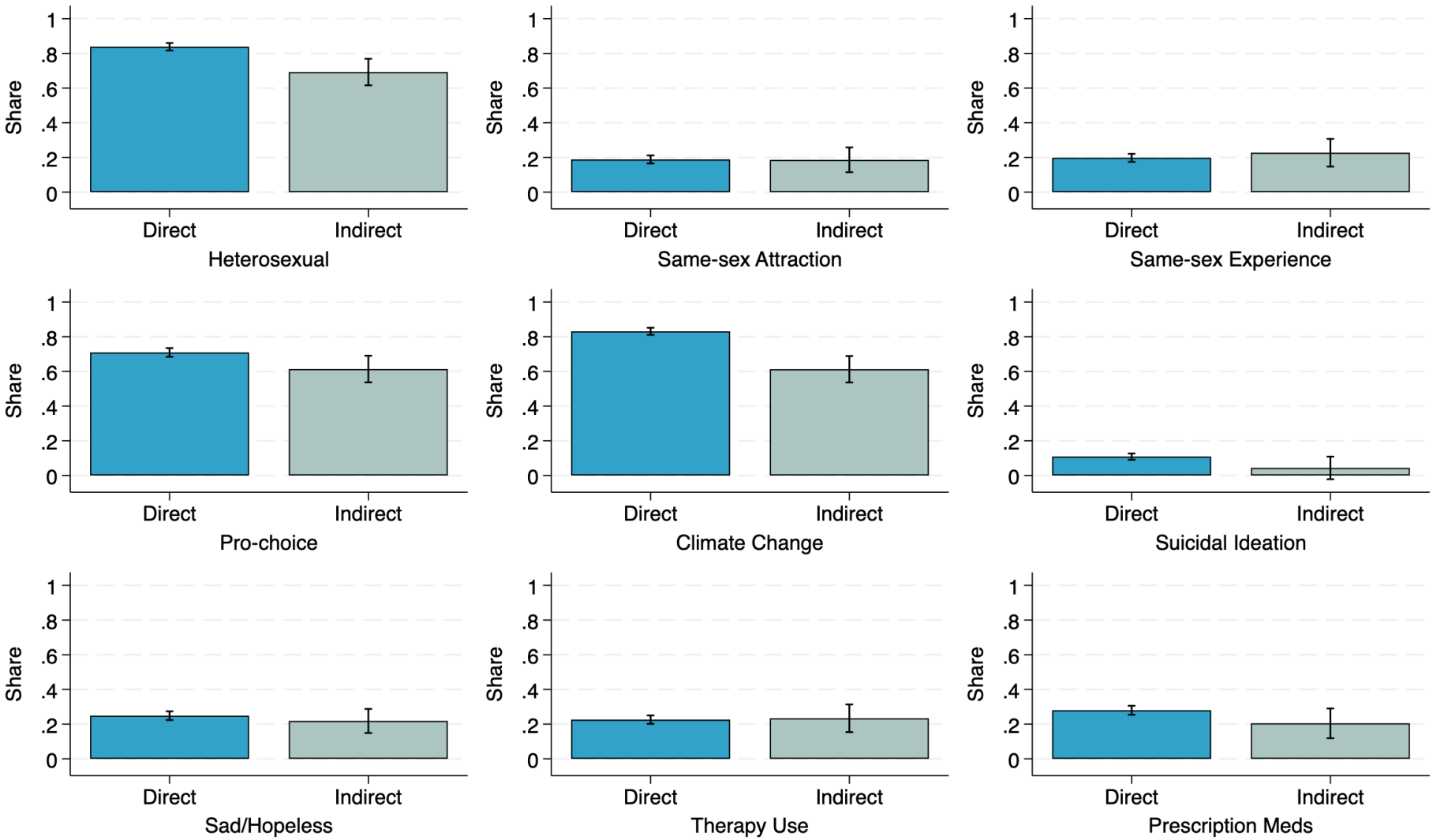
$$E[R_i^{C_1}|Z_i = 1] - E[R_i^{C_1}|Z_i = 0] = E[R_i^{C_2}|Z_i = 1] - E[R_i^{C_2}|Z_i = 1]$$

If this test rejects, we can interpret that as evidence for some combination of design effects and ceiling/floor effects on at least one of the lists.

CDF for Non-Heterosexual Question



Prevalence Across Items



Full Sample Results

1. I conduct the Blair and Imai (BI) test on my heterosexual list experiment and find no evidence of design effects ($p = 1$)
2. I conduct the Chuang test using two sets of control items for suicidal ideation and sadness/hopelessness. I fail to reject the estimates are the same ($p = 0.531$ for suicide, $p = 0.675$ for sadness/hopelessness).
3. Visual inspection of the heterosexual LE response totals suggests minimal role for ceiling/floor effects.
4. Vastly different estimates across sensitive items that are correlated with direct estimates—would need sophisticated bias response pattern.

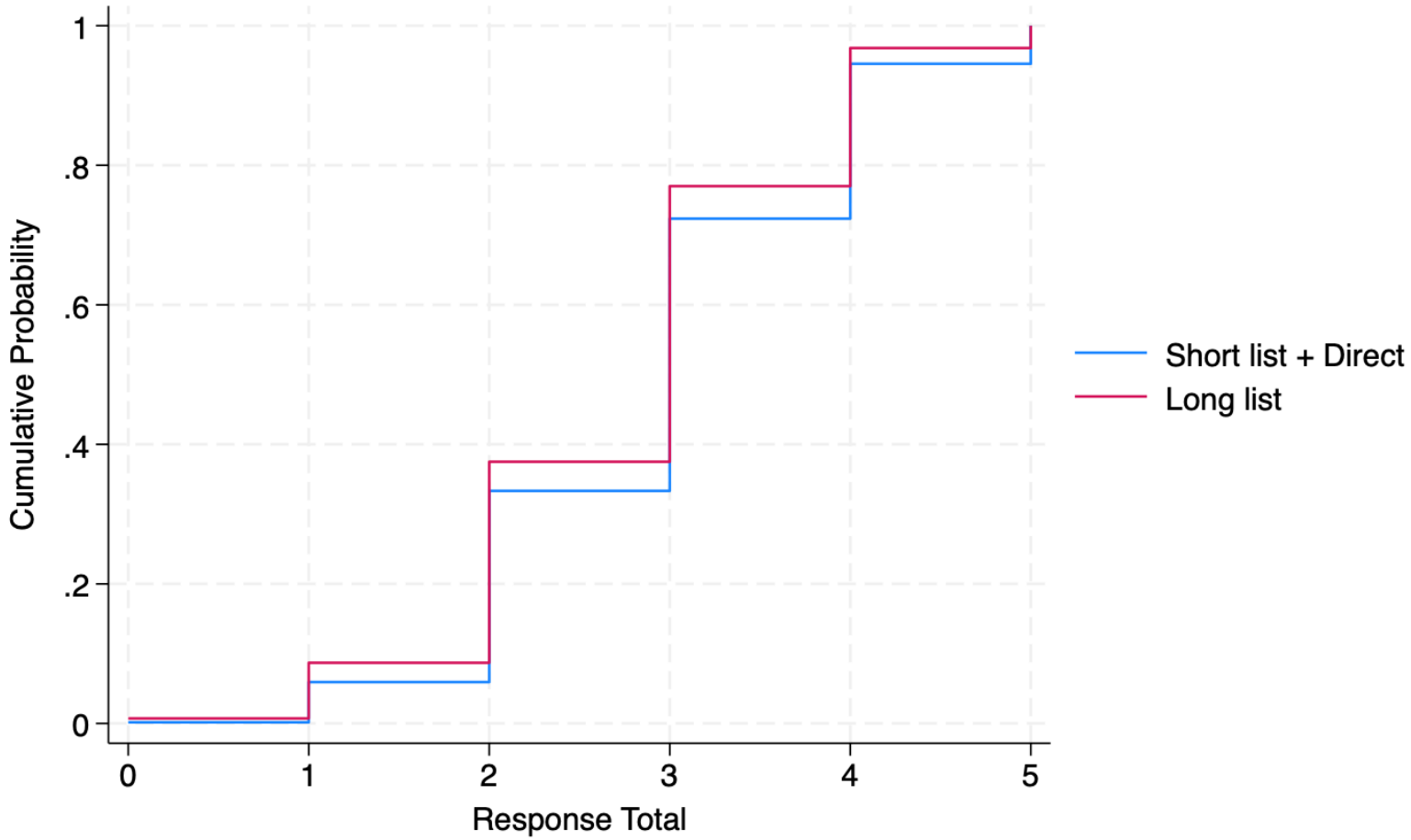
Testing across Covariate Cells

My methods require that NDE holds not just overall, but **within each covariate cell** for the Wald ratio to target an unbiased estimate of covariate means among indirect reporters.

- I divide my three main covariates of interest (PHQ, GAD, and income) into above- and below-median and conduct the BI and Chuang tests in each cell.
- Neither test rejects in any covariate cell.

Reporting Monotonicity

Intuitively, long list total should be weakly greater than short list total at every point in the distribution.



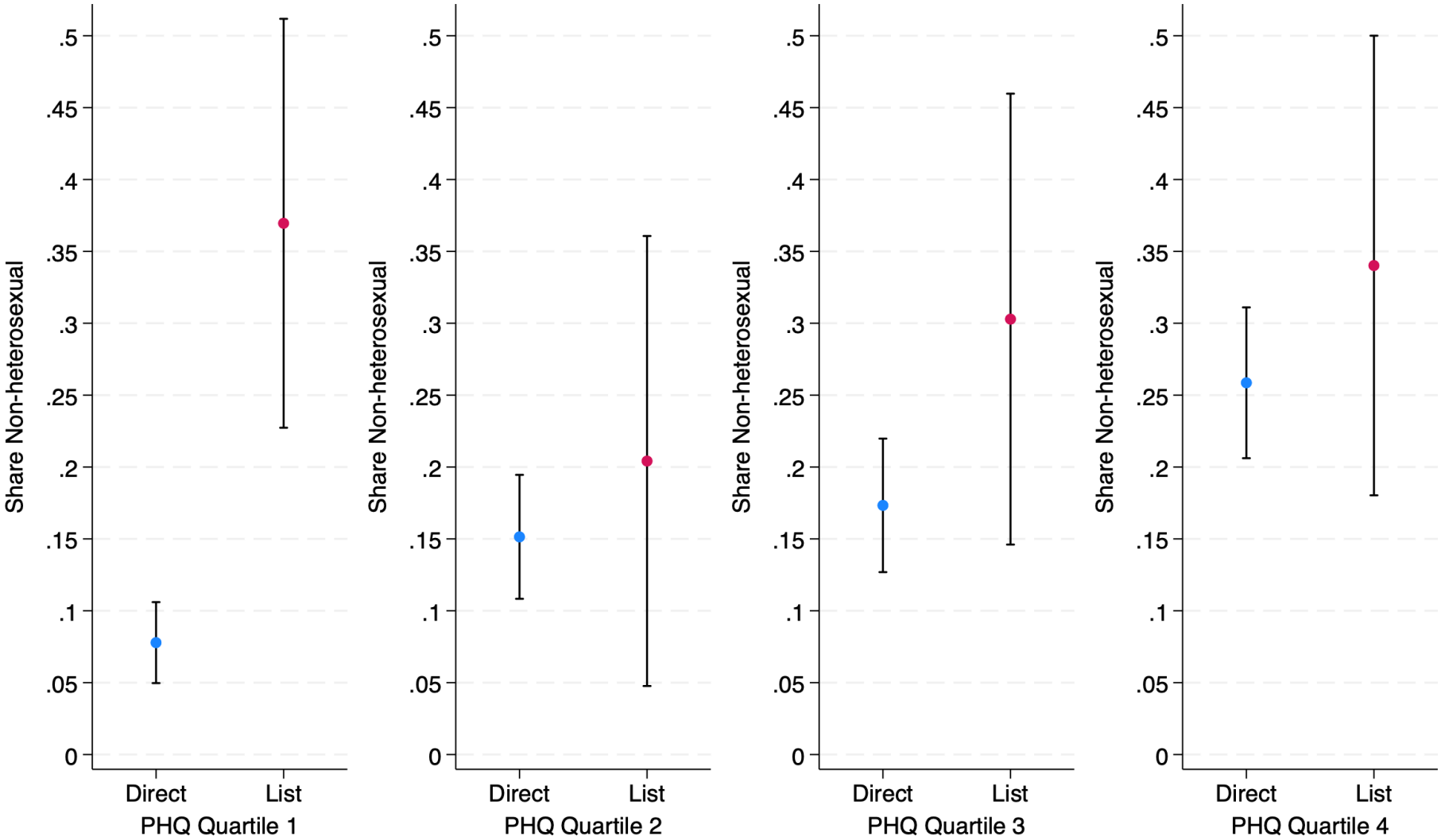
◂ Return

Benchmarking Gaps Against Literature

- **Mental Health:** I find direct non-heterosexual respondents to have +0.48 SDs PHQ-8 score and +0.49 SDs GAD-7 score.
  - NSDUH: +0.77 SDs Kessler-6 distress score
  - NHIS: +0.70 SDs GAD-7 score, +0.75 SDs PHQ-8 score
- **Income:** I find unadjusted income gap of $19,000 or 32%. Controlling for demographics + education reduces to $6500 or 11%.
  - Recent meta-analyses suggest 10% income gap relative to comparable heterosexuals. (Klawitter 2015, Drydakis 2019)

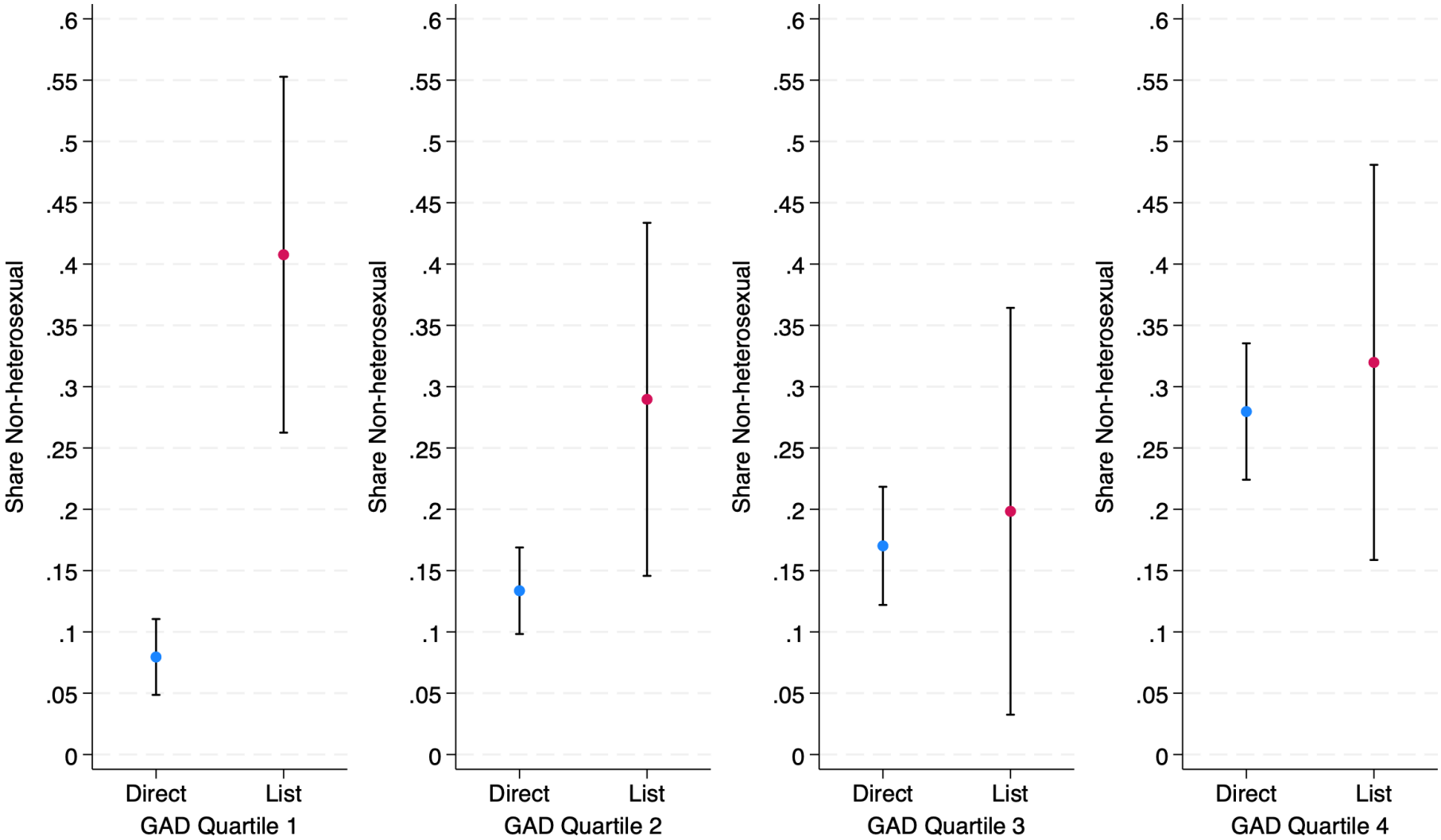
◂ Return

First Stage by PHQ Score



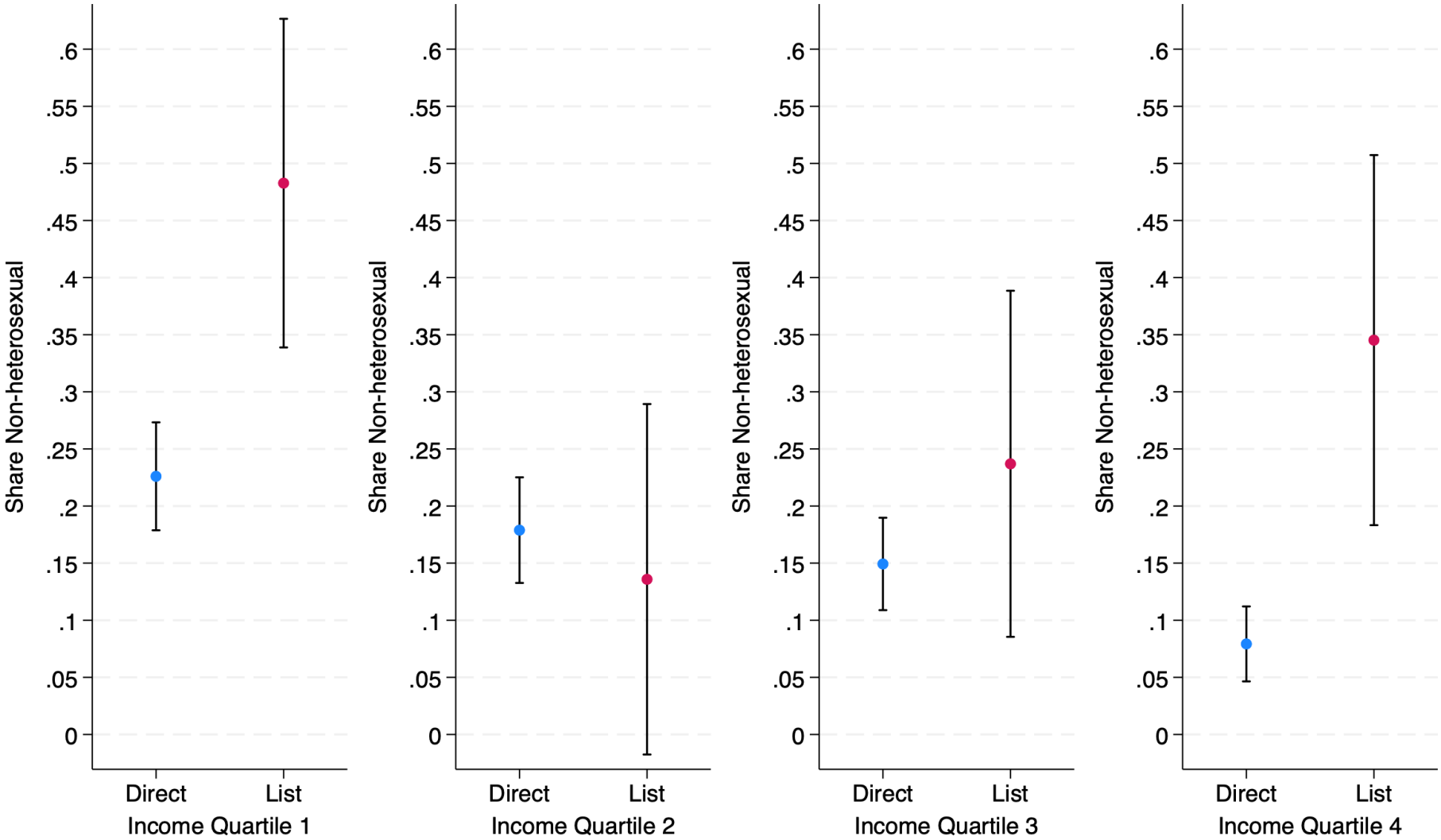
◂ Return

First Stage by GAD Score



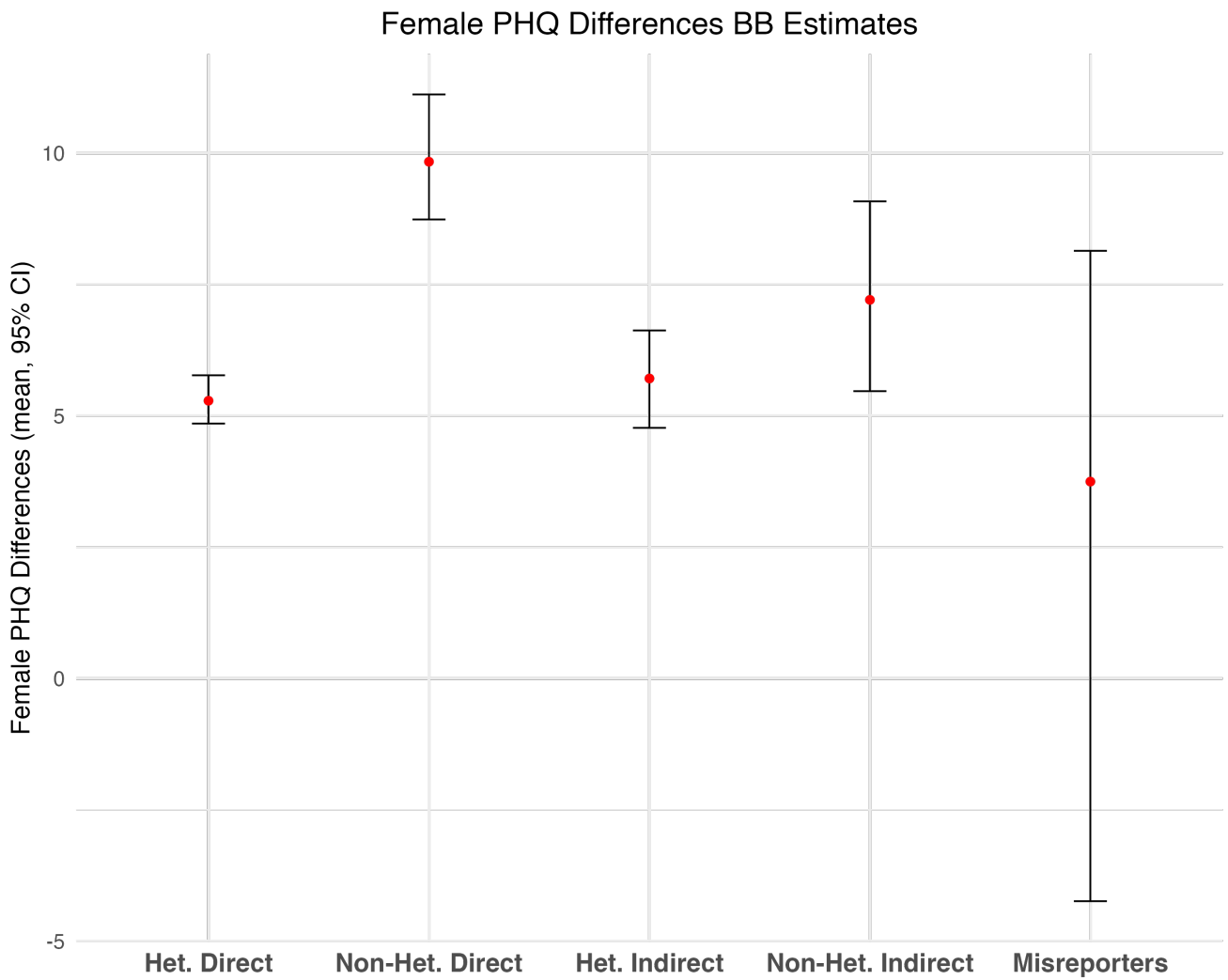
Return

First Stage by Income



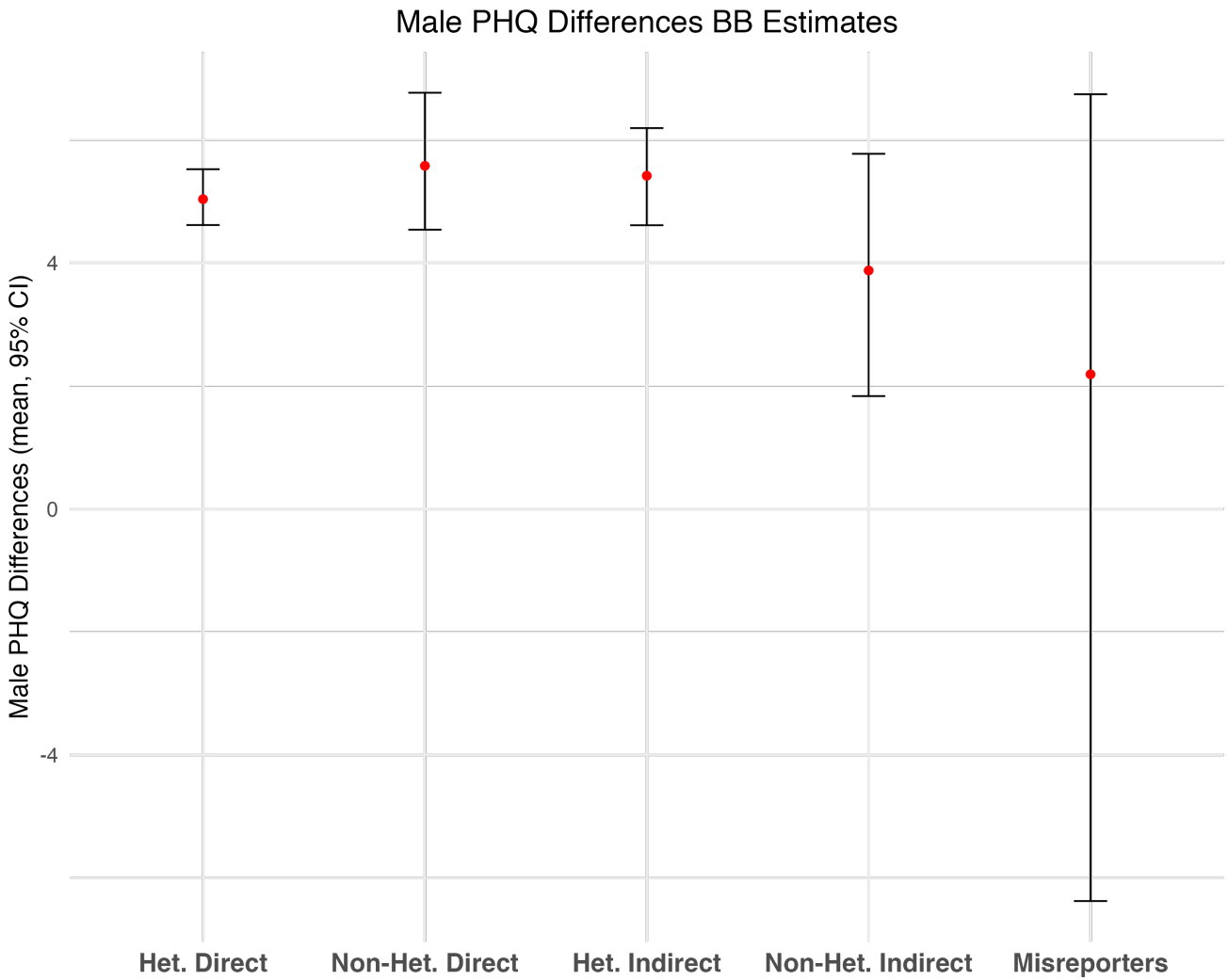
Return

PHQ Score Differences (Female)



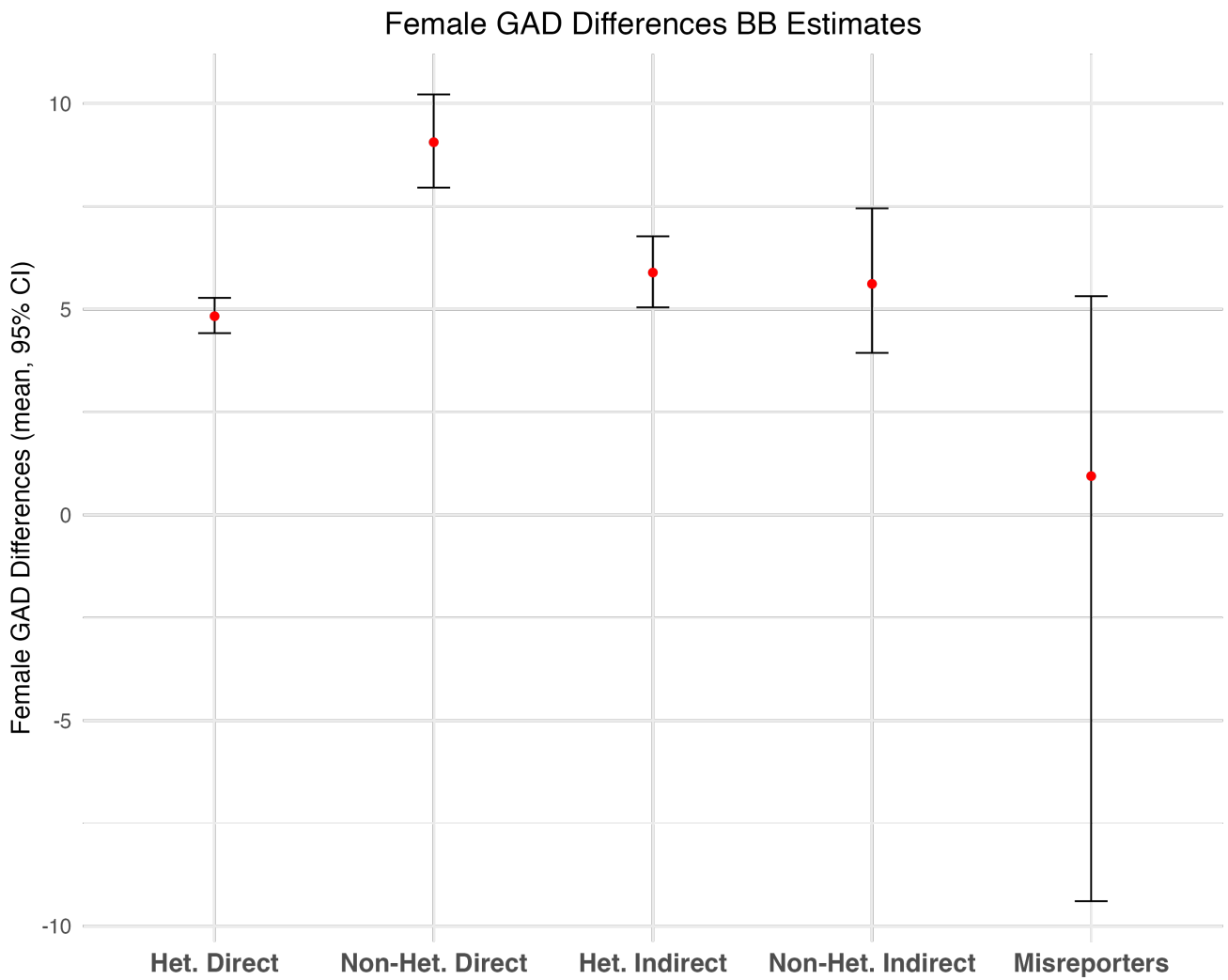
Return

PHQ Score Differences (Male)



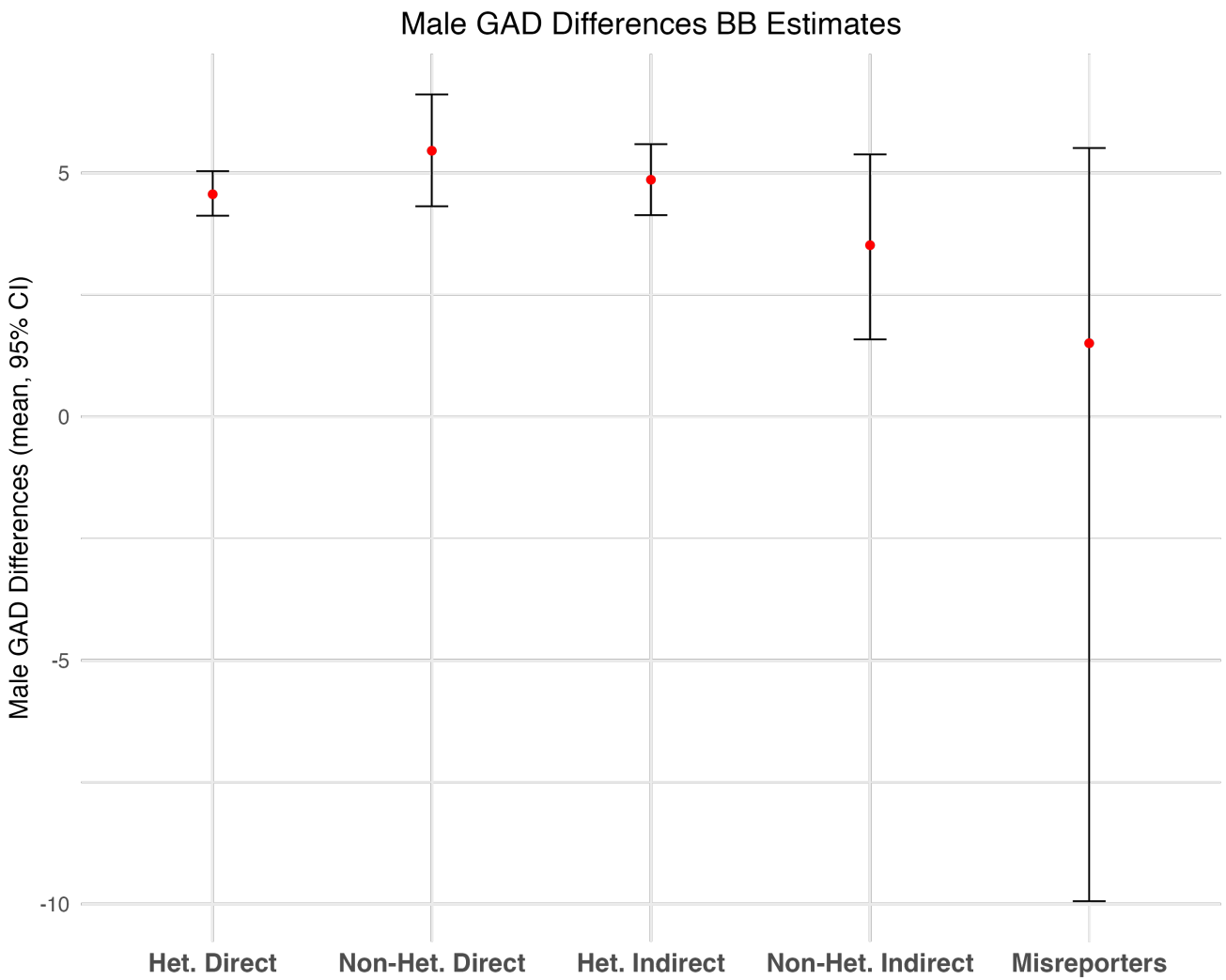
Return

GAD Score Differences (Female)



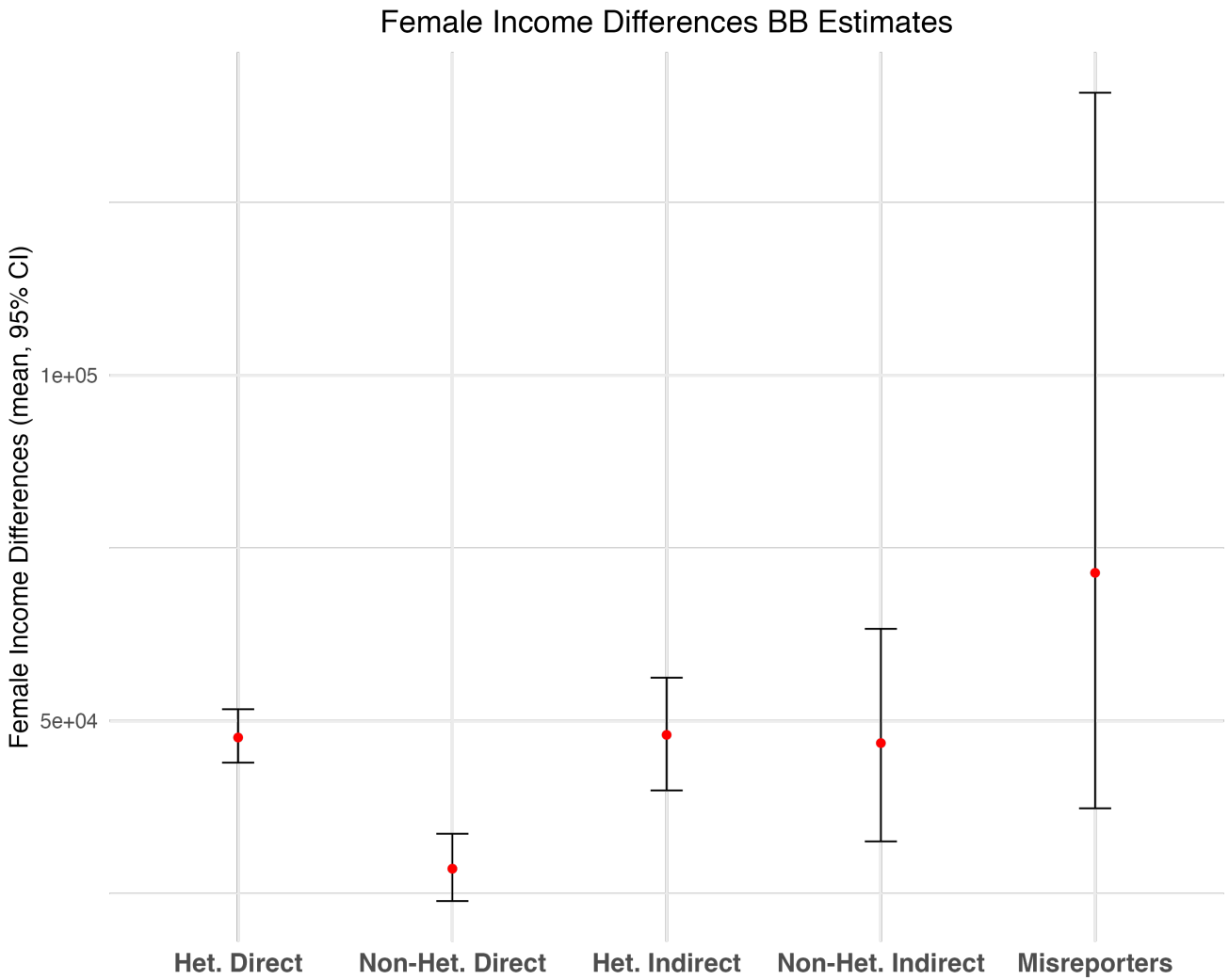
Return

GAD Score Differences (Male)



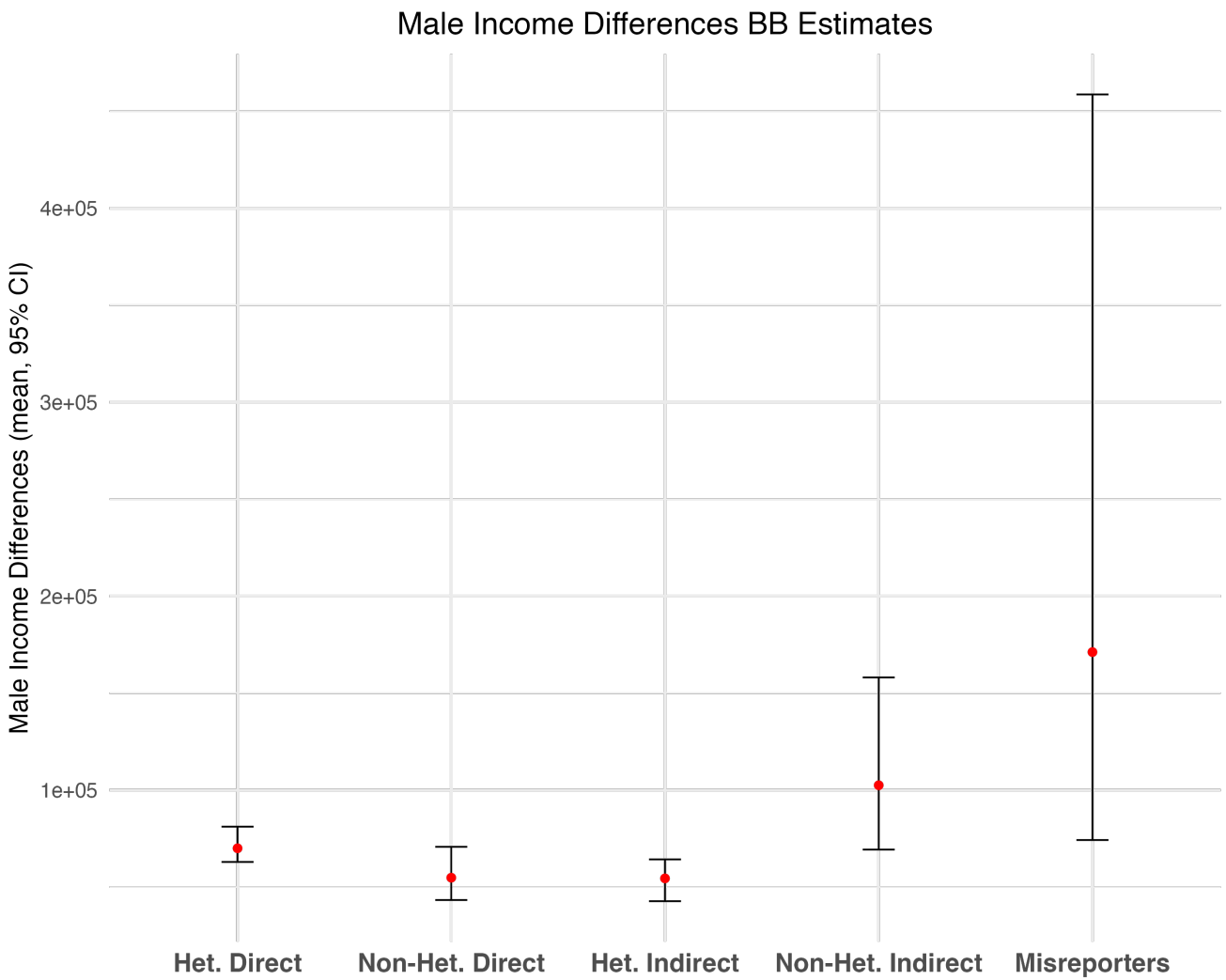
◂ Return

Income Differences (Female)



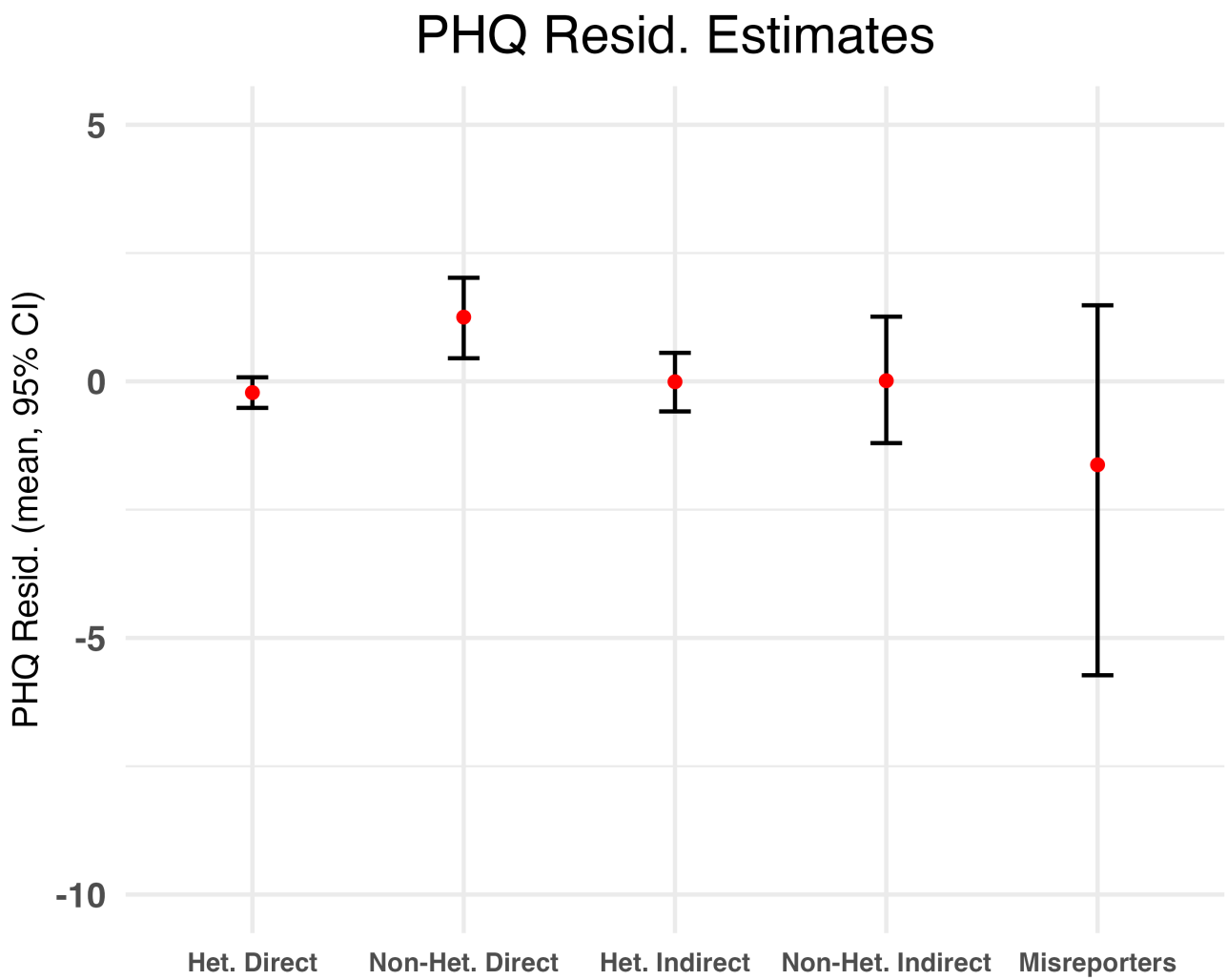
Return

Income Differences (Male)



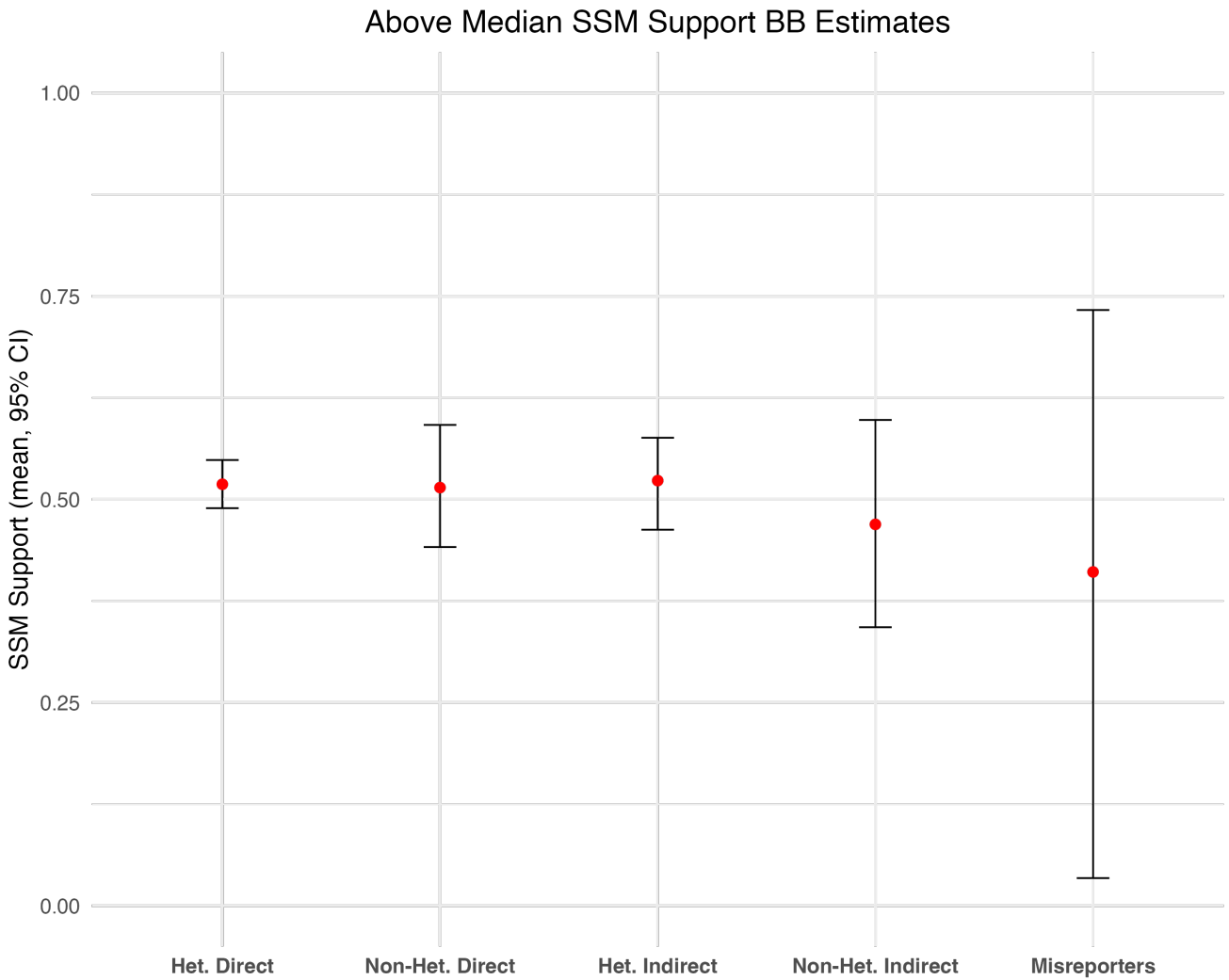
◂ Return

Residualized PHQ Differences



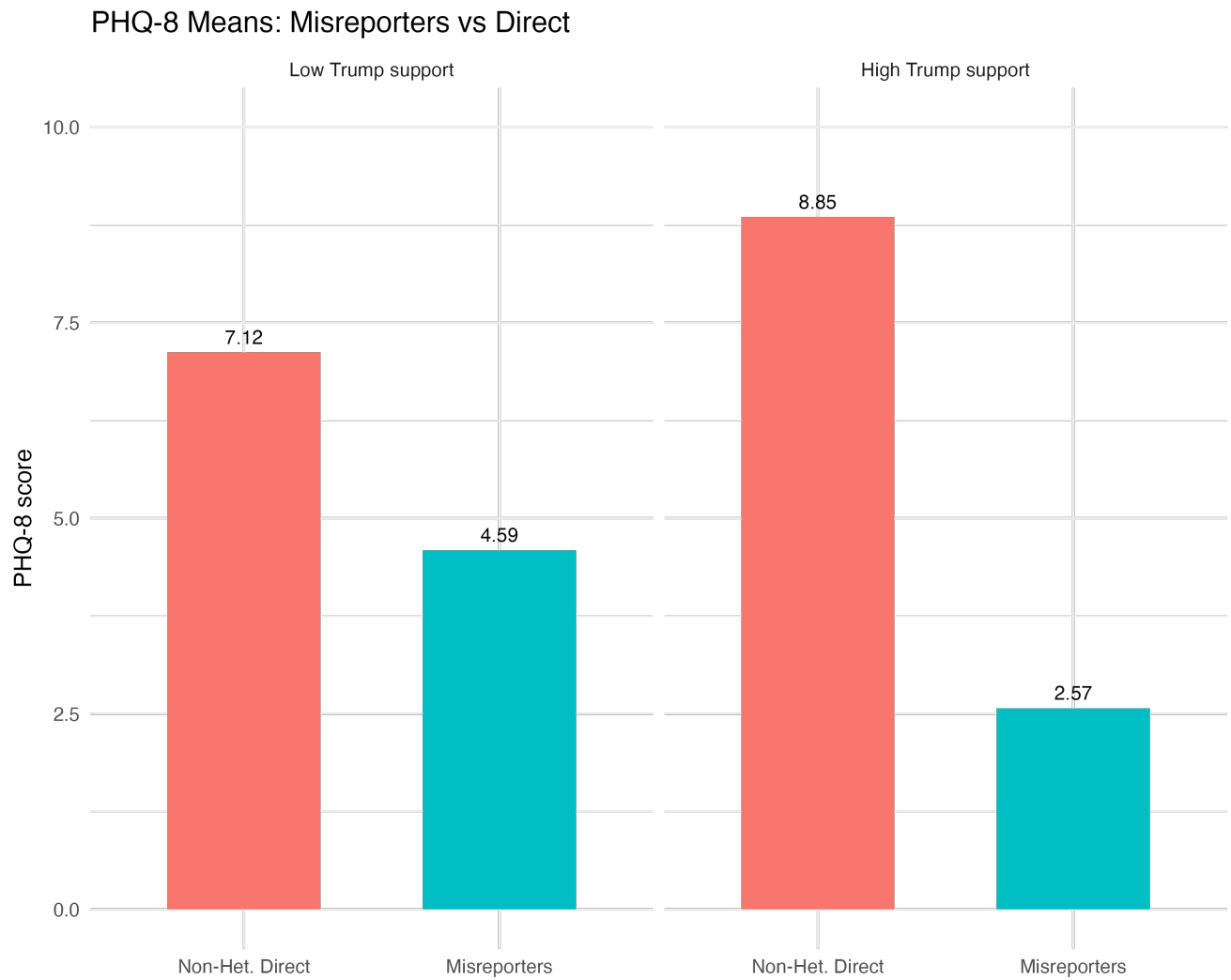
Return

LGBQ Support in Zip Code



Return

PHQ Gaps by Trump Support



◂ Return

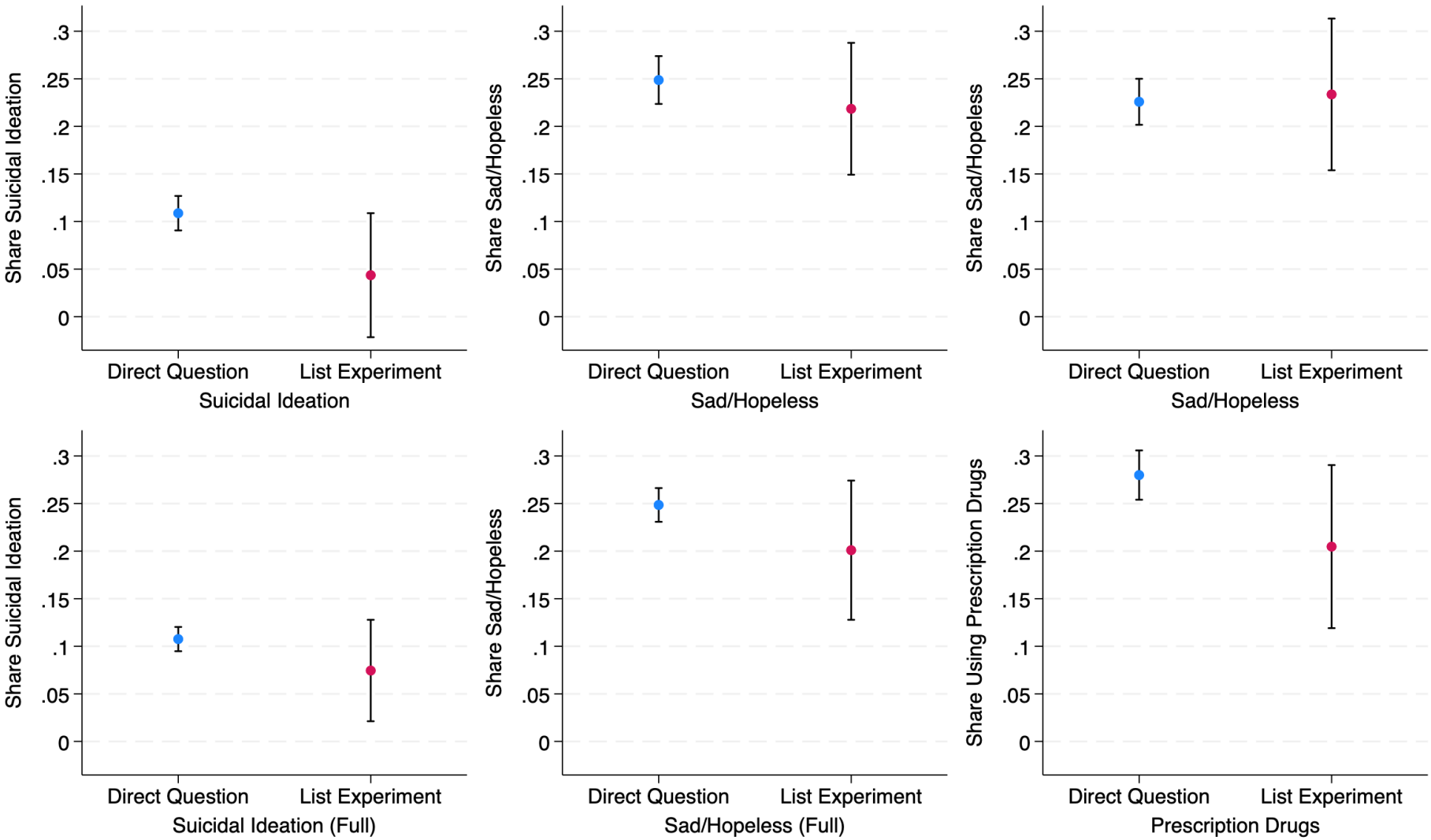
Mental Health Reporting

Examine four mental health and care utilization measures:

1. **Suicidal Ideation**
2. **Sadness/Hopelessness**
3. **Therapy Usage**
4. **Prescription Medication Usage**

Measure suicidal ideation and sadness/hopeless twice with a cross-randomization, examine both initial and pooled estimates.

Mental Health First Stages



Next Steps for Mental Health Reporting

- Interestingly, no evidence of under-reporting. If anything, some slight evidence of over-reporting on suicide and prescription drugs.
- Might still be differences between the direct reporters and the indirect reporters—is this interesting if there's not a difference in level of reporting?
- Was thinking of covariates like age, race, sex, and mental health scores.

Polarization First Stages