EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

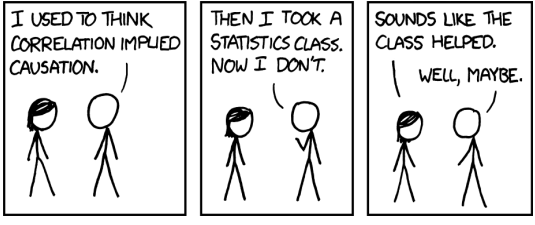
METHODS CENTER

G-Estimation for Latent Mediation: Accounting for Unobserved Outcome-Mediator Confounding via Instrumental Variables

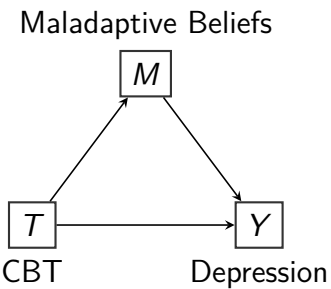
Sofia Morelli, Roberto Faleh, & Holger Brandt

05.06.2026

Psicostat Meeting



The Setting: What is Mediation Analysis?



- Investigate the pathways linking $T \rightarrow M \rightarrow Y$ to understand **how or why** a treatment works.
- Identify the **underlying mechanism** of the treatment

Example: maladaptive beliefs mediating the effect of a Cognitive Behavioral Therapy (CBT) intervention on depression (e.g., Driessen & Hollon, 2010)

Effect Definitions in the Potential Outcome Framework

Define the potential outcome for an individual i at treatment $T = t$ and mediator $M = m$:

$$Y_i^{tm} = \theta_t t + \theta_m m + f_{yx}(\mathbf{x}_i) + \epsilon_{y_i}^{tm}$$

where $f_{mx}(\cdot)$ is a function of the measured covariates $\mathbf{x}_i$. The effects are defined as causal contrasts that estimate the controlled effects:

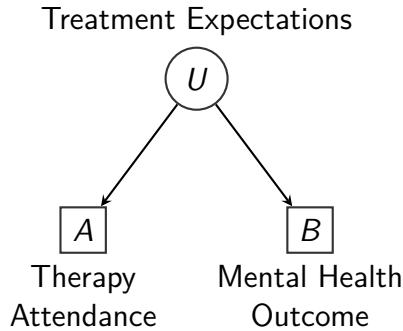
$$\text{CDE} = \text{E}\left[Y_i^{1m} - Y_i^{0m}\right] = \theta_t$$

$$\text{CME} = \text{E}\left[Y_i^{tm^2} - Y_i^{tm^1}\right] = \theta_m(m^2 - m^1)$$

The Core Issue: Why worry about Unobserved Confounders?

- U causally affects both variables of interest, A and B
- If U is not measured or adjusted for, it may induce a spurious association between A and B

→ Biased estimation of the causal effect

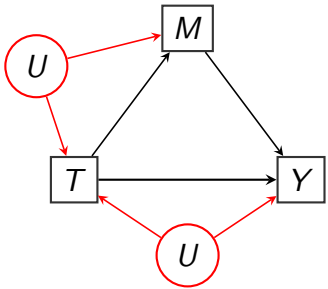


Potential Confounders that might be Omitted

Example: maladaptive beliefs mediating the effect of CBT on depression

- **Likely to be measured:** Baseline depression, age, gender, length of depression
- **Hard to Retrieve:** Social support, thought patterns
- **Expensive:** Genetic predispositions, stress levels (cortisol), brain activity
- **Likely Forgotten:** Daily sunshine hours

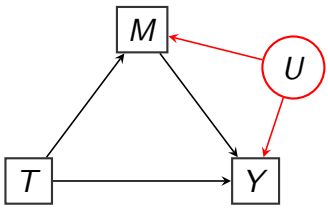
Leveraging Treatment Randomization



- **Randomizing** the treatment T removes confounding on both the treatment–outcome and treatment–mediator paths.
- Standard approach in practice via **randomized controlled trials** (RCTs)

→ **Direct effect** estimation is **simplified** since treatment assignment is independent of baseline covariates.

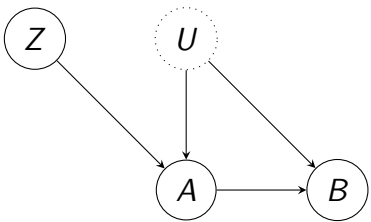
Remaining Unobserved Confounders between Mediator and Outcome



- Randomizing the mediator M to remove mediator-outcome confounding is usually not part of the experimental design.
- Regression-based mediation analysis after Baron and Kenny (1986) ($>$ 135,000 citations) assumes no unobserved confounders (**sequential ignorability**).

What if we could use the Treatment Randomization again?

- This is the foundation of the approach I will present.
- Specifically, we will use treatment-covariate interactions that influence the mediator but have no direct effect on the outcome as an Instrument*



*An Instrument/Instrumental Variable (IV) Z that influences A but does not directly affect B, except through A, can be used to adjust for unmeasured confounding (U).

Identifying Assumptions: Independence

Z must be independent of any unmeasured confounders U that affect B.

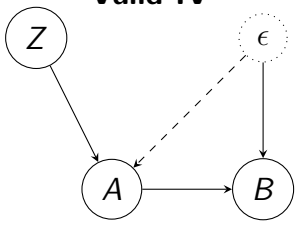
$$Z \perp U$$

This ensures that Z does not inherit bias from confounding, so that the variation in A induced by Z is as if random.

IV Identifying Assumptions: Exclusion Restriction

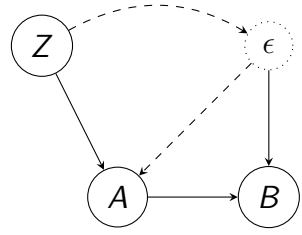
An instrumental variable Z affects the outcome Y *only through* A:

Valid IV



→ IV estimates are consistent for the causal effect of A on B.

Invalid IV



→ part of the effect of Z on B bypasses. $A \rightarrow$ biased estimates

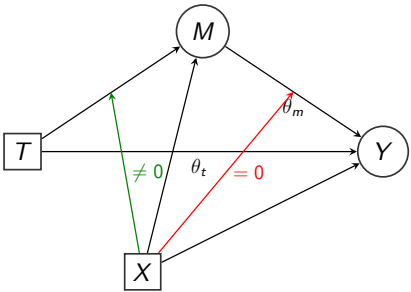
IV Identifying Assumptions: Relevance

Z must be correlated with the treatment A.

$$\text{Cov}(Z, A) \neq 0$$

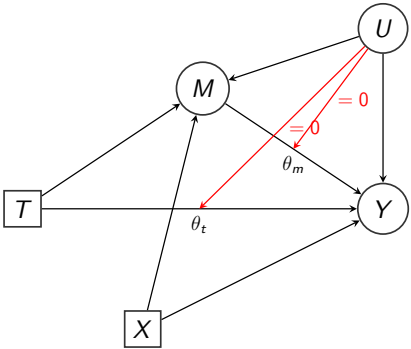
Without this, the IV provides no information about A and estimates are weak or undefined.

Treatment-Covariate Interaction as an Instrument in Mediation



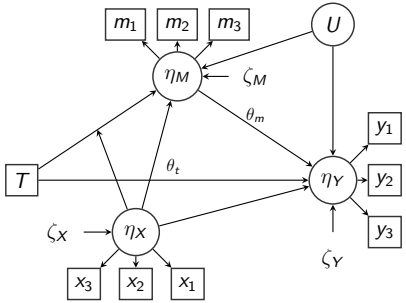
- **Independence** is automatically fulfilled through Randomization
- **Relevance** requires that the treatment-covariate interaction(s) influence(s) the mediator
- **Exclusion Restriction** requires that the treatment-covariate interaction(s) do(es) not influence the outcome

From Sequential Ignorability to No-Unobserved-Effect-Modification



- Unmeasured confounders (U) may affect levels of Y and M
- The unmeasured confounder may not modify the causal effects (θ_t, θ_m)
- No interactions between confounder U and the effects of the treatment T or mediator M on the outcome Y

Latent Variable Approach: Rank Preserving Structural Equation Model



- Ten Have et al. (2007): original rank preserving model (RPM) for mediation
- Zheng and Zhou (2015): extension to a more general form allowing for multivalue treatment, multivariate mediation and interactions with covariates
- **Morelli, Faleh, & Brandt (under review): latent variable version RAPSEM**

Model Specification

Structural Model

Outcome: $\boldsymbol{\eta_y = \mathbf{X}_y\theta + \zeta_y}$

rows of $\mathbf{X}_y$: $\mathbf{x}_{y,i} = \begin{pmatrix} t_i & \eta_{m_i} & f_{yx}(\boldsymbol{\eta}_{x_i}) \end{pmatrix}$

parameters: $\boldsymbol{\theta} = \begin{pmatrix} \theta_r & \theta_m & \boldsymbol{\theta}_x \end{pmatrix}^\top$

Mediator: $\boldsymbol{\eta_m = \mathbf{X}_m\gamma + \zeta_m}$

rows of $\mathbf{X}_m$:
$\mathbf{x}_{m,i} = \begin{pmatrix} t_i & f_{mx}(\boldsymbol{\eta}_{x_i}) & t_i \cdot f_{mrx}(\boldsymbol{\eta}_{x_i}) \end{pmatrix}$

parameters: $\boldsymbol{\gamma} = \begin{pmatrix} \gamma_r & \boldsymbol{\gamma}_x & \boldsymbol{\gamma}_{rx} \end{pmatrix}^\top$

Measurement Model

$\boldsymbol{\xi_i = \boldsymbol{\tau} + \boldsymbol{\Lambda}\boldsymbol{\eta}_i + \boldsymbol{\epsilon}_i}$

observed indicators: $\boldsymbol{\xi}_i = \begin{pmatrix} \mathbf{m}_i & \mathbf{x}_i & \mathbf{y}_i \end{pmatrix}^\top$

factor scores: $\boldsymbol{\eta}_i = \begin{pmatrix} \eta_{m_i} & \boldsymbol{\eta}_{x_i} & \eta_{y_i} \end{pmatrix}^\top$

intercepts: $\boldsymbol{\tau} = \begin{pmatrix} \boldsymbol{\tau}_{\text{free}} & \mathbf{0}_{\mathbf{k}\times\mathbf{1}} \end{pmatrix}^\top$

factor loadings: $\boldsymbol{\Lambda} = \begin{pmatrix} \boldsymbol{\Lambda}_{\text{free}} & \mathbf{I}_k \end{pmatrix}^\top$

measurement error: $\boldsymbol{\epsilon}_i$

Estimation Strategy

RAPSEM uses a two-stage, limited-information estimation approach.

- **Stage 1:** Estimate latent factor scores and propagate estimation error. (Wall & Amemiya, 2000)
- **Stage 2a:** Fit structural equations via *G*-estimation. (Ten Have et al., 2007; Zheng & Zhou, 2015)
- **Stage 2b:** Adjust structural estimates for latent variable interactions and higher-order terms. (Wall & Amemiya, 2000)

Study 1: Robustness to violations of no-unmeasured-confounder and no-unobserved-effect-modifier assumptions

$$\eta_{y_i} = 0.224 \cdot t_i + 0 \cdot \eta_{m_i} + 0.240 \cdot (\eta_{x1_i} + \eta_{x2_i}) + \delta_u\, u_i + \delta_{um}\, u_i \eta_{m_i} + \zeta_{y_i},$$
$$\eta_{m_i} = 0.316 \cdot t_i + 0.277 \cdot (\eta_{x1_i} + \cdot \eta_{x2_i}) + 0.232 \cdot (t_i \eta_{x1_i} + \cdot t_i \eta_{wx_i}) + \delta_u\, u_i + \zeta_{m_i}$$

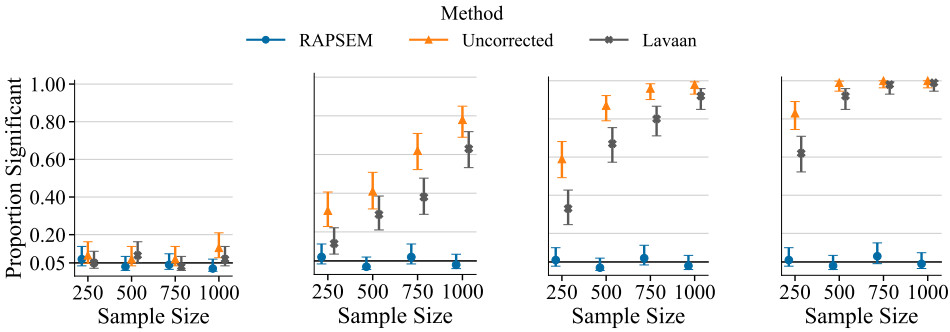
Confounding effect size δ_u	0.100	0.245	0.316	0.374*
Modfication effect size δ_{um}	0.100	0.245	0.316	0.374*
Sample size N	250	500	750	1000

*Cohen's d values of $0.2, 0.5, 0.65$, and 0.8

Method Compared

- **RAPSEM**: Our proposed framework utilizing latent *G*-estimation implemented as R package `rapsem`
- **Uncorrected**: A two-stage latent regression approach using the standard predictor matrix in place of the weight matrix. This represents factor score regression without *G*-estimation corrections.
- **Lavaan**: A standard `lavaan` implementation utilizing a two-group model to capture treatment-covariate interactions, with factor loadings and intercepts constrained to equality across groups.

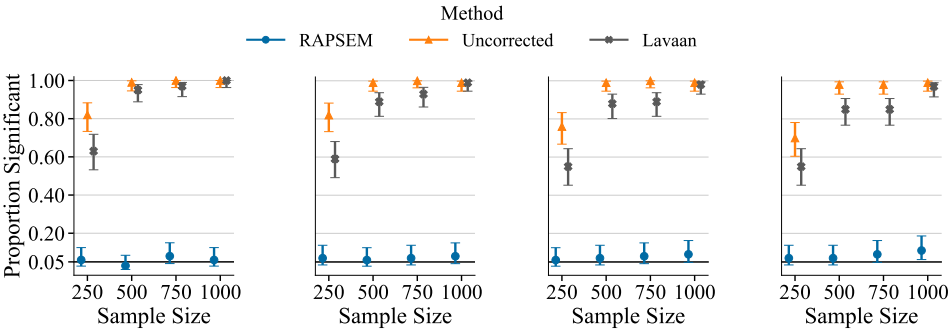
Unobserved Confounding



(a) $\delta_u = 0.100$ **(b)** $\delta_u = 0.245$ **(c)** $\delta_u = 0.316$ **(d)** $\delta_u = 0.374$

Figure 1: False positive rates of $\check{\theta}_m$ with Wilson score 95% confidence intervals under varying confounding levels and sample size.

Effect-Modification by Unobserved Confounders



(a) $\delta_{um} = 0.100$ **(b)** $\delta_{um} = 0.245$ **(c)** $\delta_{um} = 0.316$ **(d)** $\delta_{um} = 0.374$

Figure 2: False positive rates of $\breve{\theta}_m$ with Wilson score 95% confidence intervals under varying confounding levels and sample size.

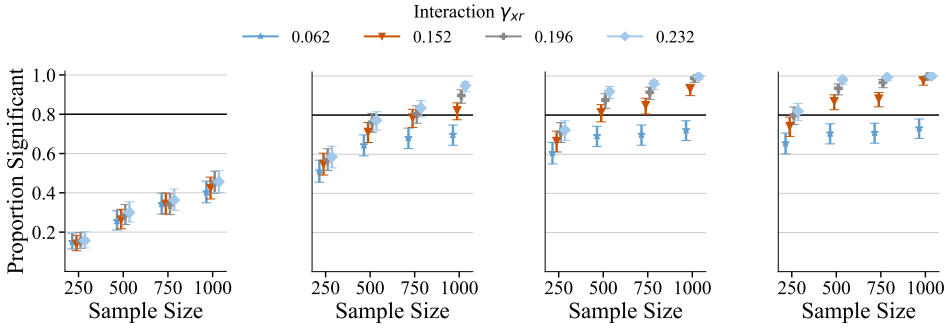
Study 2: Power of RAPSEM under varying CME effect sizes and treatment-covariate interaction effect sizes

$$\eta_{y_i} = 0.224\, t_i + \theta_m \cdot \eta_{m_i} + 0.240(\eta_{x1_i} + \eta_{x2_i}) + \zeta_{y_i},$$
$$\eta_{m_i} = 0.316 \cdot t_i + 0.277(\eta_{x1_i} + \eta_{x2_i}) + \gamma_{xr}\,(t_i\eta_{x1_i} + t_i\eta_{x2_i}) + \zeta_{m_i}.$$

CME effect size θ_m	0.100	0.243	0.316	0.371*
Interaction effect size γ_{xr}	0.062	0.152	0.196	0.232*
Sample size N	250	500	750	1000

*Cohen's d values of $0.2, 0.5, 0.65$, and 0.8

Power to Detect Mediator Effect



(a) $\theta_m = 0.100$ **(b)** $\theta_m = 0.243$ **(c)** $\theta_m = 0.361$ **(d)** $\theta_m = 0.374$

Figure 3: Power to detect CME under varying CME effect sizes, interaction strengths, and effect sizes.

Conclusion

- **Standard SEM for mediation:**
  - Type I error rates can reach 100% for spurious mediator effects
  - Problem worsens with increasing sample size
- **Alternative: RAPSEM**
  - Robust to unobserved confounding, at the cost of lower power
  - Power highly depends on treatment-covariate interaction

Thank you for your attention!

Our Preprint *RAPSEM: Identifying Latent Mediators Without Sequential Ignorability via a Rank-Preserving Structural Equation Model* by Sofia Morelli, Roberto Faleh & Holger Brandt:

`https://arxiv.org/abs/2509.23935` `https://github.com/PsychometricsMZ/RAPSEM`

Baron, R., & Kenny, D. (1986, 01). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, *51*, 1173-1182. doi: 10.1037//0022-3514.51.6.1173

Driessen, E., & Hollon, S. D. (2010). Cognitive behavioral therapy for mood disorders: efficacy, moderators and mediators. *Psychiatric Clinics*, *33*(3), 537–555. doi: 10.1016/j.psc.2010.04.005

Russ, H., Sibley, L., Flegr, S., Kuhn, J., Hoogerheide, V., Scheiter, K., & Lachner, A. (2026). Combining non-interactive teaching and drawing fosters conceptual knowledge but not monitoring accuracy from guided inquiry in science learning. *Journal of Educational Psychology*, *118*(3), 345–367. doi: 10.1037/edu0000971

Ten Have, T. R., Joffe, M. M., Lynch, K. G., Brown, G. K., Maisto, S. A., & Beck, A. T. (2007). Causal mediation analyses with rank preserving models. *Biometrics*, *63*, 926–934.

Wall, M. M., & Amemiya, Y. (2000). Estimation for polynomial structural equation models. *Journal of the Statistical American Association*, *95*, 929–940. doi: 10.1080/01621459.2000.10474283

Zheng, C., & Zhou, X.-H. (2015). Causal mediation analysis in the multilevel intervention and multicomponent mediator case. *Journal of the Royal Statistical Society (Series B)*, *77*, 581–615. doi: 10.1111/rssb.12082

Empirical Example: Russ et al. (2026)

- Data collected in 30 seventh- and eighth-grade classes from 11 German secondary schools, with a resulting sample size of $N = 590$ students
- Binary treatment: students explained the content to a fictitious, non-present peer via voice messages vs control ($N = 150$) - students restudied the lecture content on their own
- Outcome: conceptual knowledge, measured with 15 binary items, aggregated into three parcels
- Mediator: task interest, measured with two items on a 4-point Likert scale

Empirical Example: Treatment-Covariate Interaction

Testing individual and joint treatment-covariate interaction effects on the mediator: reported are the interaction coefficients: interaction coefficient γ_{xr}, partial F-statistic, and p-value for the comparison between models with and without interaction terms.

Covariate	**Effect Size**	F-**Statistic**	p-**value**
interest in physics	-0.231	6.6	0.011
conscientiousness in physics	-0.220	5.4	0.020
perceived teacher support	-0.290	6.2	0.013
All Combined	—	5.1	0.002

Empirical Example: Models for Comparison

- **Full Model M1**: Included all the aforementioned variables.
- **Reduced Model M2**: Omitted the confounder perceived cognitive activation, artificially treating it as an unobserved confounder. This reflects a scenario with unobserved confounding where a critical dimension of the learning environment might be overlooked during study design.

Empirical Example: Method Comparison

Comparison of mediation effects ($\breve{\theta}_m$) across estimation methods. Values represent point estimates and 95% confidence intervals ($\breve{\theta} \pm 1.96\,\breve{\sigma}_{\text{pooled}}$) aggregated across imputations.

Model	Estimation Method		
	G-Estimator	Uncorrected	Lavaan
M1 (Full)	-0.08 $[-0.81, 0.65]$	0.08 $[-0.04, 0.19]$	-0.16 $[-0.47, 0.15]$
M2 (Reduced)	-0.08 $[-0.78, 0.61]$	0.10 $[0.00, 0.21]$	-0.10 $[-0.31, 0.12]$

Empirical Example: Extension with Synthetic Instrument

$\breve{\theta}_m$ estimates under artificially increased instrument strength: The parameter λ represents the artificial boost applied to the interaction between treatment and task interest within the mediator equation, F_{interact} denotes the F-statistic for all interaction terms.

λ	F_{interact}	$\hat{\theta}_m$ [95 % CI]
-0.2	11.3	-0.05 $[-0.48, 0.37]$
-0.4	21.0	-0.04 $[-0.36, 0.28]$
-0.6	34.2	-0.02 $[-0.27, 0.23]$

Structural Nested Model

Observed outcomes Y_i can be linked to counterfactuals Y_i^{am} by sequentially removing causal effects and past history:

$$
\begin{aligned}
Y_i^{t0} &= Y_i - \theta_m m_i \\
Y_i^{00} &= Y_i^{t0} - \theta_t t_i \\
Y_{i,\text{adj}}^{00} &= Y_i^{00} - t_{yx}(\mathbf{x}_i)
\end{aligned}
$$

Remaining variation in the outcome after blipping down and adjusting for covariates:

$$
\boxed{Y_{i,\text{adj}}^{00} = Y_i - \theta_M m_i - \theta_T t_i - t_{yx}(\mathbf{x}_i)}
$$

No Essential Heterogeneity Assumption (NEH)

Under SUTVA, consistency and correct model specification, we get

$$Y^{00}_{i,\text{adj}}(\boldsymbol{\theta}^*) = \mathrm{E}[\varepsilon^{tm}_{y_i} \mid t_i, m_i, \mathbf{x}_i],$$

where $\boldsymbol{\theta}^*$ contains the true parameters.
We assume this conditional mean of the error term to be the same for all potential outcomes under any levels of the treatment and the mediator

$$\boxed{\mathrm{E}[\varepsilon^{tm}_{y_i}(\mathbf{u}_i, \mathbf{x}_i) \mid t_i, m_i, \mathbf{x}_i] = F(t, m, \mathbf{x}_i)} \quad \forall t, m$$

G-Estimation

Estimation is based on orthogonality conditions between the blipped-down, adjusted residual outcome and weights W:

$$\text{Cov}(W_t, Y^{00}_{i,\text{adj}}) = 0, \quad \text{Cov}(W_m, Y^{00}_{i,\text{adj}}) = 0$$

Solving the corresponding estimating equations

$$\boxed{\sum_{i=1}^{n} W_t \cdot Y^{00}_{i,\text{adj}} = 0}, \quad \boxed{\sum_{i=1}^{n} W_m \cdot Y^{00}_{i,\text{adj}} = 0}$$

yields the causal effects θ_t and θ_m.

Randomization-based Instrument Functions

Randomized treatment assignment is conditionally independent of $Y^{00}_{i,\text{adj}}$, satisfying the exclusion restriction. This allows identification via **instrument functions**, defined as the projections of the treatment and mediator variables onto the instrument space:

- Treatment instrument function:

$$W_t = T - \mathrm{E}[T \mid X]$$

- Mediator instrument function:

$$W_m = W_t \cdot \Big(\mathrm{E}[M \mid T = 1, X] - \mathrm{E}[M \mid T = 0, X] \Big)$$

Stage 1: Estimation of Latent Factor Scores (Wall & Amemiya, 2000)

- **Factor score estimator:** $\tilde{\boldsymbol{\eta}}_i = \left(-\mathbf{H} \quad \mathbf{I}_k + \mathbf{H}\boldsymbol{\Lambda}_{\text{free}}\right)\left[\boldsymbol{\xi}_i - \left(\boldsymbol{\tau}_{\text{free}} \quad \mathbf{0}\right)^\top\right]$ with

  $\mathbf{H} = \left(\mathbf{0}_{k\times(p-k)} \quad \mathbf{I}_k\right)\boldsymbol{\Psi}\left(\mathbf{I}_{(p-k)} \quad -\boldsymbol{\Lambda}_{\text{free}}^T\right)^\top\left[\left(\mathbf{I}_{(p-k)} \quad -\boldsymbol{\Lambda}_{\text{free}}\right)\boldsymbol{\Psi}\left(\mathbf{I}_{(p-k)} \quad -\boldsymbol{\Lambda}_{\text{free}}^T\right)^\top\right]^{-1}$

- **Asymptotics:** $\hat{\boldsymbol{\eta}}_i = \tilde{\boldsymbol{\eta}}_i(\hat{\boldsymbol{\Upsilon}})$ is consistent if $\hat{\boldsymbol{\Upsilon}} = (\hat{\boldsymbol{\Upsilon}}_1, \hat{\boldsymbol{\Upsilon}}_2)$ is consistent, where $\hat{\boldsymbol{\Upsilon}}_1 = \left(\hat{\boldsymbol{\tau}}'_{\text{free}}, (\operatorname{vec}\hat{\boldsymbol{\Lambda}}_{\text{free}})', (\operatorname{vec}\hat{\mathbf{H}})'\right)'$ and $\hat{\boldsymbol{\Upsilon}}_2$ contains the higher-order moments of $\mathbf{e}_i$ required for the error propagation in the second stage.

- **Estimation error propagation:** $\underbrace{\tilde{\boldsymbol{\eta}}_i}_{\text{factor estimator}} = \underbrace{\boldsymbol{\eta}_i}_{\text{true factors}} + \underbrace{\mathbf{e}_i}_{\text{estimation error}}$

  with $\mathbf{e}_i = \left(-\mathbf{H} \quad \mathbf{I}_k + \mathbf{H}\boldsymbol{\Lambda}_{\text{free}}\right)\boldsymbol{\epsilon}_i$ and $\boldsymbol{\Sigma}_{ee} = \left(-\mathbf{H} \quad \mathbf{I}_k + \mathbf{H}\boldsymbol{\Lambda}_{\text{free}}\right)\boldsymbol{\Psi}\left(\mathbf{0} \quad \mathbf{I}_k\right)^\top$

Stage 2a: Structural Equation Estimation via g-Estimation (Ten Have et al., 2007; Zheng & Zhou, 2015)

- **Orthogonality condition**: $\boxed{\mathbf{W}^\top \zeta_y = 0}$ where $\mathbf{w}_i = \begin{pmatrix} w_{r_i} & w_{m_i} & t_{yx}(\boldsymbol{\eta}_{x_i})^\top \end{pmatrix}$ requiring residuals ζ_y to be orthogonal to treatment and mediator weights.
- **Weights** corresponding to the instrument functions:
  - Treatment: $\mathbf{w}_r = \mathbf{r} - \mathrm{E}[\mathbf{r}]$
  - Mediator: $\mathbf{w}_r \cdot \mathbf{w}_m = (\mathrm{E}[M \mid \boldsymbol{\eta}_X, R = 1] - \mathrm{E}[M \mid \boldsymbol{\eta}_X, R = 0])$
- **Closed-form estimator**: $\boxed{\bar{\boldsymbol{\theta}} = (\mathbf{W}^\top \mathbf{X}_y)^{-1} \mathbf{W}^\top \boldsymbol{\eta}_y}$

Stage 2b: Correction for Latent Variable Interactions (Wall & Amemiya, 2000)

- **Problem:** Structural outcome $\eta_{y,i}$ may involve interactions or higher-order terms. These require correction using measurement error moments $\boldsymbol{\Upsilon}_2$ derived from $\hat{\boldsymbol{\Sigma}}_{ee}$.
- **2SMM estimator:** $\hat{\boldsymbol{\theta}} = \hat{\mathbf{M}}^{-1}\hat{\mathbf{m}}$ with moment functions $\hat{\mathbf{M}} = \frac{1}{N}\sum_{i=1}^{N} M(\hat{\boldsymbol{\eta}}_{\text{pred},i}, \hat{\boldsymbol{\Upsilon}}_2)$ and $\hat{\mathbf{m}} = \frac{1}{N}\sum_{i=1}^{N} m(\hat{\boldsymbol{\eta}}_i, \hat{\boldsymbol{\Upsilon}}_2)$ with $\begin{pmatrix} M(\tilde{\boldsymbol{\eta}}_{\text{pred},i}, \boldsymbol{\Upsilon}_2) & m(\tilde{\boldsymbol{\eta}}_i, \boldsymbol{\Upsilon}_2) \end{pmatrix} = \sum_{j=0}^{J}(-1)^j A_j(\tilde{\boldsymbol{\eta}}_i, \boldsymbol{\Upsilon}_2)$ where A_0 = uncorrected estimates, and A_j $(j > 0)$ = higher-order corrections from error moments.
- **Asymptotics:** Under standard IV and measurement error assumptions, $\hat{\boldsymbol{\theta}}$ is consistent and asymptotically normal.