Methodology Note: Detecting Heterogeneous Treatment Effects

Random Forest vs OLS Interactions vs GenericML BLP

Helena Montoya Calero · May 2026

The question. Does the effect of SDM training on scientific thinking (ΔSI) vary across founder types? Formally: is the Conditional Average Treatment Effect $\tau(z) = E[Y_i(1) - Y_i(0) \mid Z_i = z]$ a function of observable baseline characteristics Z? Three methods exist to answer this. They differ in what they assume, what they can find, and what mistakes they make.

Method 1 — Random Forest (variable importance only).

What it does. Train a random forest on $Y \sim Z$ separately within each arm. Compare which variables the forest uses most to predict Y in the treated arm vs. the control arm. High importance in treated = potential moderator.

Why it is useful. Non-parametric, captures non-linearities, handles many variables simultaneously. Good for exploration and shortlisting.

Why it is not *sufficient for HTE.*

- Variable importance measures *predictive* relevance, not *causal* relevance. A variable can rank high because it correlates with Y, not because it moderates the treatment effect.
- Gives no p-values, no confidence intervals, no test of significance.
- Cannot distinguish: "IoC predicts SI" from "IoC moderates the treatment effect on SI." In our data: IoC is the top RF predictor of ΔSI in the treated arm, but the causal HTE test (GenericML BLP) gives $\hat{\beta}_2 = -0.008$, $p = 0.66$ — zero effect. RF was seeing a correlation, not heterogeneity.

Our use. Scripts 01–02: RF used to shortlist the 4 most promising moderators (IoC, FEI, idea breadth, startup phase) before running causal tests. Not used for inference.

Method 2 — OLS with interactions.

What it does. For each moderator W, estimate:

$$Y_i = \alpha + \beta_1 T_i + \beta_2 (T_i \times W_i) + \gamma W_i + X_i'\delta + \varepsilon_i$$

$\hat{\beta}_2$ is the interaction coefficient: does the treatment effect increase/decrease with W? HC1-robust standard errors. Test $H_0 : \beta_2 = 0$.

Why it is useful. Transparent, interpretable, easy to communicate. Gives a direct estimate of "one more unit of W changes the treatment effect by $\hat{\beta}_2$." Sufficient when the researcher has strong priors about which moderator matters and the relationship is approximately linear.

Limitations for our setting.

- **Overfitting / multiple testing.** Testing 12 moderators × 5 periods × 8 outcomes = 480 cells at $\alpha = 0.05$ yields ∼24 expected false positives. Without a correction for the search, any "significant" result is unreliable.
- **Misspecified B(Z).** The interaction assumes the heterogeneity in $\tau(z)$ is linear in W. If the true effect is non-linear (e.g., only founders above a threshold respond), the linear interaction will have low power.
- **No separation of estimation and inference.** Using the same data to find the pattern and test it inflates type I error.

Our use. Scripts 13–16 (mediation, translation HTE, control dynamics, attrition): used after the GenericML null is established, for targeted follow-up questions with strong priors. Not the primary HTE test.

Method 3 — GenericML: Best Linear Projection (BLP).

Reference. Chernozhukov, Demirer, Duflo, Fernández-Val (*Econometrica* 93(4), 2025).

The core idea. Separate the problem into two steps: (1) **estimate** the heterogeneous effect using ML (without worrying about inference), (2) **test** whether that estimated heterogeneity is real using a causal regression (without overfitting from step 1).

Step 1 — Estimate B(Z). Split the sample in half. On the first half, train a machine learner to predict $Y \sim Z$ in the treated arm minus the control arm. The learner produces $B(Z_i)$ — a score that ranks individuals by predicted treatment effect magnitude. The ML can be anything: lasso, random forest, XGBoost, neural net. Because it is used only for ranking (not inference), overfitting does not bias the test.

Step 2 — BLP regression. On the *second* half (held-out), estimate:

$$Y_i = \underbrace{\alpha_1 B(Z_i)}_{\text{baseline}} + \underbrace{\beta_1(D_i - p_i)}_{\text{ATE proxy}} + \underbrace{\beta_2(D_i - p_i)(B(Z_i) - \bar{B})}_{\text{HTE}} + \varepsilon_i$$

- $\hat{\beta}_1$: average treatment effect (ATE proxy). Should match OLS ATE.
- $\hat{\beta}_2$: heterogeneity coefficient. If $\hat{\beta}_2 > 0$ and significant: founders with high $B(Z)$ benefit more. If $\hat{\beta}_2 = 0$: treatment effect is flat.

To test a specific moderator W, replace $(B(Z_i) - \bar{B})$ with $(W_i - \bar{W})$.

Sample splitting repeated 100–250 times. Each split gives one estimate of $(\hat{\beta}_1, \hat{\beta}_2)$. Final estimate = median across splits. p-values are valid because each split uses held-out data for inference.

Why it dominates for our setting.

	RF importance	**OLS interaction**	**GenericML BLP**
Causal inference	No	Yes	Yes
Valid p-values	No	Yes (single test)	Yes (repeated splits)
Multiple testing	No protection	No protection	Controlled via B(Z)
Non-linear effects	Captures	Misses	Captured in B(Z)
Overfitting	High	Medium	Eliminated by splitting
Power vs many W	N/A	Low	High
Interpretability	Low	High	Medium

What we did: three specifications with increasing B(Z) quality.

- **D5 — Manual lasso BLP** (script 08, 480 cells, 250 splits): B(Z) estimated by lasso on ∼25 baseline controls + site dummies. Conservative baseline. Result: all null in P1–P2.
- **D6 — 5-learner manual ensemble** (script 09, 96 cells, 250 splits): Best B(Z) selected per split from lasso, elastic net, ridge, RF, SVM by CV-MSE. Better ranking of heterogeneity if it exists. Result: all null.
- **D7 — Official GenericML package** (script 10, 120 cells, 100 splits): 6-learner ensemble: lasso, elastic net, ridge, random forest, SVM, XGBoost. Automatic best-learner selection. Most powerful B(Z) estimator. Result: 2 borderline hits (`fei_resid` $p = 0.036$, `months_working_1` $p = 0.096$ on log sales P2).
- **Robustness check** (script 11): Re-ran D7 hits with lasso-only B(Z) (D5 spec). `fei_resid`: $p = 0.751$. `months_working_1`: $p = 0.566$. Both disappear. **Conclusion**: hits are artifacts of better B(Z) finding noise, not real heterogeneity. With 120 cells at $\alpha = 0.05$, ∼6 false positives are expected; finding 2 at $p < 0.10$ is consistent with the null.

Verdict. The null is robust to learner specification. GenericML with a poor B(Z) finds nothing; GenericML with the best available B(Z) finds 2 borderline signals that are machine-learning overfitting, not causal heterogeneity. SDM training moves all founder types uniformly on ΔSI.

Reference: Chernozhukov V., Demirer M., Duflo E., Fernández-Val I. (2025). "Generic Machine Learning Inference on Heterogeneous Treatment Effects in Randomized Experiments." *Econometrica* 93(4): 1289–1334.

Implementation: `GenericML` R package (Klaassen, Kueck, Calonico et al.). Scripts 08–11 in `02_analysis/scripts/01_initial_analyses/`.