Institute for Economic Studies, Keio University

Keio-IES Discussion Paper Series

Treatment Effect Identification under Selection on Potential Outcomes

星野崇宏、篠田和彦、大津泰介

2026 年 6 月 10 日
DP2026-012
https://ies.keio.ac.jp/publications/27877/

Institute for Economic Studies, Keio University
2-15-45 Mita, Minato-ku, Tokyo 108-8345, Japan
ies-office-group@keio.jp
10 June, 2026

Treatment Effect Identification under Selection on Potential Outcomes

星野崇宏、篠田和彦、大津泰介

IES Keio DP2026-012
2026 年 6 月 10 日
JEL Classification: C14, C21, C26, C36, J24
キーワード: Roy model, selection on gains, average treatment effect, auxiliary measurements, inverse problems, sieve GMM, orthogonal moments, weak identification

【要旨】

This paper develops an auxiliary-measurement approach to identifying average treatment effects in generalized Roy environments where treatment choice may depend directly on potential outcomes. Identification is formulated as a primal–dual inverse problem. A latent selection-odds representer anchors a causally correct element in an observed calibration set, which may be nonunique. An adjoint outcome representer certifies that the target mean is invariant over that set, so identification does not require point identification of the calibrating function itself. This separation yields a trichotomy between non-invariance, irregular identification, and regular orthogonal-moment representation, according to the position of the outcome signal in the adjoint range. The same geometry delivers an orthogonal estimating equation and an exact product-bias identity, supporting sieve GMM and cross-fitted estimation. Simulations illustrate regular, weak, and failed range regimes.
An application to retirement and cognition in the Health and Retirement Study shows that specifications restricted to observed adjustment and those allowing selection on gains yield materially different estimates, illustrating the framework's empirical content under maintained calibration and adjoint-representation assumptions.

星野崇宏
慶應義塾大学経済学部
hoshino@econ.keio.ac.jp

篠田和彦
名古屋大学経済学部
k-shinoda@nagoya-u.jp

大津泰介
ロンドスクールオブエコノミクス　経済学部
t.otsu@lse.ac.uk

謝辞：This work is supported by JSPS Kakenhi 25K2165.

Treatment Effect Identification under Selection on Potential Outcomes

A Generalized Roy Model with Auxiliary Measurements

Takahiro Hoshino* Kazuhiko Shinoda†
Keio University and RIKEN Nagoya University

Taisuke Otsu‡
London School of Economics and Political Science

June 10, 2026

Abstract

This paper develops an auxiliary-measurement approach to identifying average treatment effects in generalized Roy environments where treatment choice may depend directly on potential outcomes. Identification is formulated as a primal–dual inverse problem. A latent selection-odds representer anchors a causally correct element in an observed calibration set, which may be nonunique. An adjoint outcome representer certifies that the target mean is invariant over that set, so identification does not require point identification of the calibrating function itself. More generally, the paper characterizes the maximal class of arm-specific linear expectation functionals of the marginal potential-outcome distributions whose calibrated representation is invariant over the observed calibration set: invariance holds if and only if the selected-arm signal lies in the closure of the adjoint range, and under the latent anchor conditions the common value equals the causal target, with the average treatment effect as the leading special case. This separation yields a trichotomy between non-invariance, irregular identification, and regular orthogonal-moment representation, according to the position of the outcome signal in the adjoint range. The same geometry delivers an orthogonal estimating equation and an exact product-bias identity, supporting sieve GMM and cross-fitted estimation. Simulations illustrate regular, weak, and failed range regimes. An application to retirement and cognition in the Health and Retirement Study shows that specifications restricted to observed adjustment and those allowing selection on gains yield materially different estimates, illustrating the framework's empirical content under maintained calibration and adjoint-representation assumptions.

Keywords: Roy model; selection on gains; average treatment effect; auxiliary measurements; inverse problems; sieve GMM; orthogonal moments; weak identification.
JEL Classification: C14, C21, C26, C36, J24.

*Email: hoshino@econ.keio.ac.jp
†Email: k-shinoda@nagoya-u.jp
‡Email: t.otsu@lse.ac.uk

1 Introduction

When workers choose whether to migrate, firms whether to export, and students whether to major in STEM, they typically know more about their own potential gains than the econometrician does. Self-selection on these private gains is the source of much of the empirical content of microeconomic data—and the source of a recurring identification problem: who selects into treatment is informative about what they would have earned, but it also means that the observed difference between treated and untreated outcomes confounds gains with selection. Roy (1951) recognized this tension and used a deterministic-sorting model to extract its empirical content. The subsequent literature has searched for weaker structural assumptions that still pin down population-level treatment effects.

A large identification toolkit exists for selection problems, but each of its standard branches changes the question. Adjustment for observed covariates removes only the part of selection that observables explain and has no purchase on selection driven by latent gains. Instrument-based approaches—local average treatment effects and marginal treatment effects—exploit exogenous variation in the participation margin and typically identify margin-local effects, or population effects only under additional support or extrapolation conditions. Proximal causal inference uses proxies to address latent confounding, but under restrictions that rule out direct dependence of treatment choice on the potential outcomes themselves—precisely the dependence that defines selection on gains. The object studied here, the population average treatment effect under such direct dependence, falls outside each of these branches.

This paper studies the population average treatment effect in a generalized Roy environment where treatment choice depends nonseparably and nonlinearly on *both* potential outcomes, on multidimensional latent heterogeneity, and on observed environmental covariates. We do not impose deterministic sorting, monotonicity, instrument-induced participation margins, or scalar latent indices. Instead, identification flows from two pieces of auxiliary information: a latent-type measurement W that carries information about latent heterogeneity, and an outcome-representation variable V that supplies the adjoint-side variation the identifying conditions require (one transparent sufficient interpretation is that it shifts the conditional distribution of latent type). These ingredients are not free—we explain in detail what each must satisfy economically.

The central move of the paper is to recast identification as a *primal–dual* inverse problem in which the calibrating function is not the object of interest; the mean is. A latent selection-odds representer places a causally correct element in an observed calibration set, which need not be a singleton. The paper does not require identifying which observed calibration solution is the latent selection-odds representer. Instead, an adjoint outcome representer certifies that the target functional—the population mean of the potential outcome—is invariant over the entire nonunique calibration set. Identification of the mean therefore holds even when the calibrating functions themselves are not identified.

Four conditions organize the argument, and Table 1 previews their division of labor in the notation introduced below; a fuller assumption dictionary appears in Section 2.4. The latent anchor is deliberately split into a representability condition (R1-A) and a transfer condition (T): the first is a range condition on the latent measurement operator, the second is the exclusion restriction that converts the latent representer into an observed calibration solution.

Table 1: Identification logic at a glance.

Step	Object	Role	Failure mode
(R1-A)	Latent odds anchor: representability $T_{W,1}q_1^\ell = r_1^\ell$	Guarantees a latent selection-odds representer exists on the latent measurement operator	No causally correct calibrating function of (W, Y_t, C) exists
(T)	Conditional transfer / exclusion	Converts the latent representer into an element of the observed calibration set, $q^\ell \in \mathcal{Q}_1$	Observed calibration solutions lose causal meaning
(R1-B)	Observed balancing equation $A_1 q = r_1$	Guarantees that q^ℓ solves an estimable equation; the solution set $\mathcal{Q}_1$ may be nonunique	Empirical calibration solvability fails
(R2)	Adjoint representation equation $A_1^\star b = \ell_1$	Certifies that the target mean is invariant over all of $\mathcal{Q}_1$	Closure failure $\Rightarrow$ non-invariance; strict-range failure with closure $\Rightarrow$ only irregular

Contributions.

The paper makes three contributions. *First*, it derives a latent odds anchor for generalized Roy selection in which treatment may depend directly on potential outcomes; this anchor lies in the observed calibration set but is generally not point identified there. *Second*, it shows that the observed calibration solution need not be unique, and that an adjoint outcome representer is the certificate that makes the ATE invariant over the entire calibration set. Identification therefore separates into a latent causal anchor (R1-A), a transfer condition (T), an observed calibration implication (R1-B), and an adjoint range condition (R2) governing target invariance. *Third*, it connects this identification geometry to regularity: the paper gives a necessary-and-sufficient characterization of the maximal class of arm-specific linear expectation functionals of the marginal potential-outcome distributions whose calibrated representation is invariant over the observed calibration set—and which, under the latent anchor conditions (R1-A) and (T), coincide with the causal targets—through the closure and range of the observed adjoint operator (Theorem 3.3), with the ATE as the leading corollary; the position of the corresponding target signal relative to the closure of the adjoint range produces a three-case identification trichotomy (non-invariance, irregular identification, regular orthogonal-moment representation) and characterizes when a regular orthogonal score representation exists, functional by functional. The resulting framework provides a complementary route to identification of population-level treatment effects, distinct from but compatible with proximal causal inference and instrument-based MTE bounds. The appendix further shows that the latent anchor is observationally untestable without additional structure (Theorem D.9) and develops an exact latent-calibration bias identity with sensitivity bounds for departures from that anchor (Theorem D.1).

The baseline theory does not require an additive structural error representation. Let C denote the observed conditioning block used in the calibration equations; the formal results are stated for a generic C, whose content is application-specific. In the baseline two-environment illustration $C = (S, Q)$, where S and Q are treated- and untreated-side environments; the bridge specifications in the HRS application use exactly this Roy-side block $C_R = (S, Q)$, while additional baseline adjustment covariates X (in the HRS application, demographics and own schooling) enter only the observed-adjustment benchmark block $C_A = (X, S, Q)$ rather than the bridge. For one-sided means one may instead use a smaller target-specific block—for example $C_1 = S$ for μ_1 or $C_0 = Q$

for μ_0 when the two environments are asymmetric—but such reductions are separate identifying restrictions, not automatic consequences of observing S and Q. We take the conditional selection probability

$$D = \mathbf{1}\{\eta_D \leq \pi(Y_1, Y_0, U, C)\} \tag{1}$$

as primitive, where D is the treatment indicator, (Y_1, Y_0) are potential outcomes, U is latent heterogeneity, C is the conditioning covariate block, η_D is an exogenous scalar disturbance, and π is a conditional choice probability. This formulation nests additive threshold rules such as $D = \mathbf{1}\{f_D(Y_1, Y_0, U, S, Q) + \varepsilon_D > 0\}$, but it also covers multiplicative, nonlinear, and fully nonseparable choices.

The identifying information comes from equation-specific excluded covariates and an auxiliary measurement. The variable W is a latent-type measurement carrying information about latent heterogeneity or the potential-outcome type. The variable V supplies the adjoint-side variation needed for the adjoint range condition; in the most transparent interpretation it shifts the latent type U but has no direct path to treatment choice after conditioning on U, the relevant potential outcome, and C. The variables collected in C, typically S and Q, describe the treated- and untreated-side environments. In the full calibration formulation they may enter the selection margin directly. Thus $S \to D$ and $Q \to D$ are allowed; the crucial exclusions concern direct paths from W to D and from V to D or W.

Formally, for the treated potential outcome, let $X_1 = (W, Y_1, C)$ and $Z_1 = (V, C)$, where C is the conditioning block (in the baseline full-ATE formulation $C = (S, Q)$). Define the primal operator $A_1 g = \mathbb{E}[Dg(X_1) \mid Z_1]$ and its Hilbert adjoint $A_1^\star$. A latent selection-odds representer implies that the observed balancing equation $A_1 q = r_1$ has a solution, where $r_1(z) = \mathbb{E}[1 - D \mid Z_1 = z]$. The solution set $\mathcal{Q}_1$ may be nonunique. An adjoint outcome representer $b \in \mathcal{B}_1$ solves $A_1^\star b = \ell_1$, where $\ell_1(x) = \mathbb{E}[Y \mid X_1 = x, D = 1]$. Because $Y = Y_1$ on the treated arm and X_1 already contains Y_1, this signal is the treated outcome itself, so the adjoint equation $A_1^\star b = \ell_1$ asks whether functions of $Z_1 = (V, C)$ can reproduce Y_1 in conditional mean on $\{D = 1\}$; this is the substantive content of the range condition (R2). The main theorem states that, for every $q \in \mathcal{Q}_1$ and every $b \in \mathcal{B}_1$,

$$\mu_1 = \mathbb{E}[Y_1] = \mathbb{E}\big[D\{1 + q(X_1)\}\{Y - b(Z_1)\} + b(Z_1)\big].$$

The statement is not specific to the mean: for a generic linear functional $\theta_1(\varphi) = \mathbb{E}[\varphi(Y_1, C)]$ the same identity holds with Y replaced by $\varphi(Y, C)$ and b solving the adjoint equation for the signal $\ell_{1,\varphi} = \mathbb{E}[\varphi(Y, C) \mid X_1, D = 1]$, and invariance over the calibration set holds *if and only if* $\ell_{1,\varphi}$ lies in the closure of the adjoint range. The adjoint outcome representer is therefore a certificate that the target is invariant over the calibration set. Because q is evaluated only on the treated arm, where $Y_1 = Y$ is observed, the construction involves no counterfactual imputation.

This formulation separates identification from uniqueness of the calibrating functions. Point identification of q is unnecessary: what matters is the location of the outcome signal ℓ_1 relative to the adjoint range. The three exclusive and exhaustive cases

$$\ell_1 \notin \overline{\mathcal{R}(A_1^\star)}, \qquad \ell_1 \in \overline{\mathcal{R}(A_1^\star)} \setminus \mathcal{R}(A_1^\star), \qquad \ell_1 \in \mathcal{R}(A_1^\star)$$

correspond to non-invariance of the target over the calibration set, irregular identification, and regular orthogonal-moment representation. This trichotomy is a main result rather than a technical aside, and it organizes both the identification theory and the subsequent inference results.

The inferential contribution follows from this identification geometry. The signal

$$\Gamma_1(O; q, b) = D\{1 + q(X_1)\}\{Y - b(Z_1)\} + b(Z_1)$$

is not an ad hoc debiasing device. Its Neyman orthogonality follows from $A_1 q = r_1$ and $A_1^\star b = \ell_1$. The same structure gives the exact product-bias identity for plug-in estimators, $-\mathbb{E}[D(\hat{q} - q^\dagger)(\hat{b} - b^\dagger)]$, relative to a selected calibration representative (Lemma F.1). This identity motivates full-sample sieve generalized method of moments (Sieve GMM) and cross-fitted estimation. The resulting score is not imposed externally but follows directly from the primal-dual identification structure. The same framework also yields reduced calibration formulations for one-sided targets and practical implementations based on Sieve GMM and cross-fitted DML.

The empirical application revisits retirement and cognition in the Health and Retirement Study. It is presented as an illustration of implementation and identifying-assumption sensitivity: the estimated target changes as the maintained information structure moves from observed adjustment to latent-type measurements to selection on gains. It is not presented as a standalone resolution of the retirement–cognition question.

Relation to existing literatures

The framework relates to four literatures. Classical Roy and *generalized Roy* models obtain empirical content from deterministic comparative-advantage rules or monotone selection structures; recent work (Henry et al., 2024; Lee and Park, 2023; Mourifié et al., 2020) relaxes these restrictions but retains a single-index or latent-utility structure on the choice rule. A complementary negative benchmark is provided by Hoshino et al. (2026). They show that the deterministic Roy rule is a singular measurement condition for the joint latent law: when the known sorting rule $D = \mathbf{1}\{Y_1 > Y_0\}$ is locally relaxed to an unknown stochastic selection kernel, observationally equivalent latent joint laws open up at first order. The present paper studies a different positive question. Rather than trying to recover the full joint law from deterministic sorting, it identifies population mean functionals under stochastic, potential-outcome-dependent selection using latent-type and adjoint-side calibration restrictions. The present framework allows the treatment choice to be a nonlinear, nonseparable function of (Y_1, Y_0, U, C).

Instrumental-variable, local average treatment effect, and marginal treatment effect approaches (Imbens and Angrist, 1994; Heckman and Vytlacil, 2005; Carneiro et al., 2011; Mogstad et al., 2018; Brinch et al., 2017; Kline and Walters, 2019) use exogenous variation in participation margins and typically identify effects local to those margins; population targets are recovered only under additional support or extrapolation conditions. The present framework does not require an instrument that directly shifts treatment; identification operates through primal-side measurement information in W and adjoint-side variation in V to recover the population ATE rather than a margin-local average.

Proximal causal inference (Miao et al., 2018; Cui et al., 2024; Tchetgen Tchetgen et al., 2024) assumes that, conditional on latent confounders, treatment is independent of potential outcomes. Under selection on gains this conditional-independence restriction is violated, so a standard treatment-confounding bridge $q(W, C)$ that omits the potential-outcome coordinate does not in general identify the target studied here. The present selection-odds representer instead conditions on potential outcomes, and the adjoint outcome representer is not merely a nuisance correction but the object that makes the target invariant over a nonunique calibration set. The proximal

literature refers to related objects as bridge functions; in the present paper we call them selection-odds calibrators and adjoint outcome representers to emphasize their role in the generalized Roy inverse problem.

Finally, the general debiased-inference theory for inverse-problem functionals of Bennett et al. (2025) can be applied as the inferential layer. The contribution of the present paper is upstream: identification of the population ATE under generalized Roy selection on potential outcomes, organized as a primal–dual problem whose dual is a target-invariance certificate. Further detail and additional references are given in Appendix B.1.

The rest of the paper is organized as follows. Section 2 presents the generalized Roy model, the nonseparable formulation, the calibration-implication assumptions, and economic examples. Section 3 introduces latent selection-odds representers and the primal inverse problem. Section 3.2 gives the adjoint outcome representer, the main identification theorem, the trichotomy, and the calibrated orthogonal moment; extensions to distributional and quantile functionals are collected in the appendix. Section 4 records the observable moment conditions and Sieve GMM implementation. Section 5 reports simulations illustrating the regular, weak, and failed range regimes. Section 6 reports the empirical application to retirement and cognition in the Health and Retirement Study. Section 7 concludes. The appendices contain primitive sufficient conditions, singular-value arguments, efficient influence function details, DML theory, extensions, simulations, and additional empirical checks.

2 Setup

We observe $O = (D, Y, W, V, C)$ (with $C = (S, Q)$ in the baseline illustration), where $D \in \{0, 1\}$ is a binary treatment indicator and $Y = DY_1 + (1-D)Y_0$ is the observed outcome under potential outcomes (Y_1, Y_0). The target parameters are $\mu_d = \mathbb{E}[Y_d]$ for $d \in \{0, 1\}$ and the average treatment effect $\tau = \mu_1 - \mu_0$. The variable W is a primal-side measurement for latent heterogeneity or potential-outcome type, whereas V is an outcome-representation variable used in the adjoint range condition; (S, Q) denote conditioning variables for the treated and untreated environments. In some applications the conditioning block is reduced to a lower-dimensional sufficient index.

The identification results do not require additive separability. We consider the general structural representation

$$U = g_U(V, C, \eta_U), \quad Y_1 = g_1(U, C, \eta_1), \quad Y_0 = g_0(U, C, \eta_0), \quad W = g_W(U, C, \eta_W), \tag{2}$$

together with the treatment-selection rule

$$D = \mathbf{1}\{\eta_D \leq \pi(Y_1, Y_0, U, C)\}. \tag{3}$$

We use this threshold representation as a normalization of the conditional choice probability. After a probability integral transform, η_D may be taken to be uniformly distributed on $[0, 1]$ and independent of the threshold conditional on (Y_1, Y_0, U, C), so that $\pi(y_1, y_0, u, c) = \mathbb{P}(D = 1 \mid Y_1 = y_1, Y_0 = y_0, U = u, C = c)$; equivalently, any scalar disturbance with a strictly increasing conditional distribution can be transformed to this uniform normalization. Figure 1(a) summarizes the general full-calibration structure. Relative to standard proxy-control settings, treatment selection may directly depend on (Y_1, Y_0, U, C), allowing selection on gains. The maintained exclusions are the absence of direct effects $W \to D$, $V \to D$, $V \to W$, and $V \to Y_t$ conditional on (U, C).

The roles of W and V are deliberately asymmetric: W is a primal-side measurement that carries information about the latent type and may enter the calibrating function, whereas V supplies adjoint-side variation that enters only the conditioning block $Z_t = (V, C)$ (one transparent sufficient interpretation is that it shifts the conditional distribution of latent type); the two cannot be interchanged. Additional discussion of the structural DAG and path interpretations is deferred to Appendix B.2.

The analysis is stated at the level of calibration implications and range conditions. This separates the core identifying restrictions from stronger structural conditions that are sufficient but not necessary.

We first fix the notation used in the identifying conditions. For each $t \in \{0, 1\}$, let

$$X_t = (W, Y_t, C), \qquad Z_t = (V, C),$$

denote the primal and dual domains, and let

$$p_1(U, Y_1, C) = \mathbb{P}(D = 1 \mid U, Y_1, C), \qquad p_0(U, Y_0, C) = \mathbb{P}(D = 0 \mid U, Y_0, C)$$

denote the one-sided latent selection probabilities, with latent odds ratios $r_t^\ell = (1 - p_t)/p_t$. Let $T_{W,t} : \mathcal{H}_{X_t} \to \mathcal{H}_{U,Y_t,C}$ denote the latent measurement operator, $T_{W,t}q(u, y_t, c) = \mathbb{E}[q(W, y_t, c) \mid U = u, Y_t = y_t, C = c]$. The identifying restrictions separate into a latent *representability* condition and a distinct *transfer* (exclusion) condition.

Assumption 2.1 (Latent selection-odds representability (R1-A))**.** For each $t \in \{0, 1\}$ there exists a latent selection-odds representer $q_t^\ell \in L^2(W, Y_t, C)$ solving

$$\text{(R1-A)} \qquad \mathbb{E}\big[q_t^\ell(W, Y_t, C) \,\big|\, U, Y_t, C\big] = r_t^\ell(U, Y_t, C) \qquad \big(\text{equivalently } T_{W,t}\, q_t^\ell = r_t^\ell\big).$$

Assumption 2.2 (Conditional transfer to the observed calibration (T))**.** The latent selection-odds representers of Assumption 2.1 satisfy the conditional-transfer moments

$$\text{(T)} \qquad \begin{aligned} &\mathbb{E}\big[Dq_1^\ell(X_1) - (1 - D) \mid U, Y_1, C, V\big] = 0, \\ &\mathbb{E}\big[(1 - D)q_0^\ell(X_0) - D \mid U, Y_0, C, V\big] = 0. \end{aligned}$$

Condition (R1-A) is a range condition on the latent measurement operator: it asserts that a causally correct calibrating function of the observables (W, Y_t, C) exists. Condition (T) is the exclusion restriction that transfers this latent object to the observed data: it requires that, conditional on the latent state (U, Y_t, C) and the adjoint block V, treatment does not covary with the calibrating function beyond what the one-sided selection probability implies. The two conditions live on different objects and can fail separately; keeping them distinct is what allows the sensitivity analysis of Appendix D.2 and the non-testability construction of Appendix D.2 to be stated cleanly. Assumptions 2.1 and 2.2 together are the key identifying restrictions used in the analysis below. The transfer condition (T) can arise under weaker conditional-independence restrictions than those imposed in conventional proxy-control formulations. For the Y_t side, $t \in \{0, 1\}$, it is sufficient that

$$W \perp\!\!\!\perp (D, V) \mid U, Y_t, C, \qquad D \perp\!\!\!\perp V \mid U, Y_t, C. \tag{4}$$

Conditional-independence restrictions such as (4) are sufficient conditions that provide an eco-

nomic interpretation of the calibration structure, but they are not required for the identification results below. Additional discussion of conditional-independence interpretations and reduced formulations is deferred to Appendix B.2.

Proposition 2.1 (Sufficient conditional independences)**.** *If Assumption 2.1 holds and the conditional independences in* (4) *hold, then the transfer condition (T) of Assumption 2.2 holds.*

The proof is deferred to Appendix E.1. The transfer implication in Assumption 2.2 admits several economically distinct interpretations depending on the information set used to construct the calibrating function.

Classical proxy-based identification frameworks (Miao et al., 2018; Tchetgen Tchetgen et al., 2024; Cui et al., 2024) impose conditional treatment ignorability given latent heterogeneity. The present paper instead allows treatment selection to depend directly on potential outcomes, generating calibrating functions that depend on (W, Y_t, C) rather than only on (W, C). The resulting framework nests three identification regimes distinguished by the primal information set.

Branch A (selection on observables). The calibrating function depends only on an observed adjustment block C_A: $q^A(C_A)$. In the canonical theoretical comparison one may take $C_A = C$; in empirical implementations C_A may include additional observed adjustment covariates beyond (S, Q).

Branch U (latent heterogeneity without selection on gains). The calibrating function depends on the measurement and conditioning variables, $q^U(W, C)$.

Branch G (selection on gains). The calibrating function depends on the measurement, potential outcome, and conditioning variables, $q^G(W, Y_t, C)$. This is the principal new regime studied in the paper.

A first concern is that the Branch G calibration appears to depend on unobserved counterfactual outcomes. It does not. The Y_1-side calibration is evaluated only on the treated subsample, where $Y_1 = Y$, and the Y_0-side calibration is evaluated only on the untreated subsample, where $Y_0 = Y$. The counterfactual outcome is never imputed at the individual level; the identifying content enters through the latent-to-observed calibration implication and the adjoint range condition developed below.

In the canonical comparison with $C_A = C$, the three branches have nested primal information sets,

$$X_t^A = C \subseteq X_t^U = (W, C) \subseteq X_t^G = (W, Y_t, C).$$

More generally, each branch is defined by its own operator pair $(X_t^{\mathrm{b}}, Z_t^{\mathrm{b}})$, and the full primal–dual systems need not be literally nested. The corresponding dual systems are also branch-specific because the adjoint operators and target signals differ across branches. Empirically, comparing the branches provides a diagnostic decomposition across information sets rather than a formal specification test. Additional discussion of branch-specific dual systems, minimum-norm anchors, and evaluation details is deferred to Appendix B.2. Table 2 fixes the interpretation of the three branches as maintained identifying structures.

Table 2: Three maintained identifying structures.

Structure	Primal domain	Dual domain	Interpretation	What it is not
Branch A	C_A	C_A	Observed-adjustment benchmark	Not evidence of no latent selection
Branch U	(W, C)	(V, C)	Latent-type measurement adjustment	Not nested in Branch A as an inverse problem
Branch G	(W, Y_t, C)	(V, C)	Selection-on-gains calibration	Not a test that selection on gains is true

2.1 Hilbert-space formulation

For the Y_1 side define $X_1 = (W, Y_1, C)$ and $Z_1 = (V, C)$, where C is the conditioning block. Let $\langle a, b\rangle_{X_1} = \mathbb{E}[D\, a(X_1)b(X_1)]$ and $\langle c, d\rangle_{Z_1} = \mathbb{E}[c(Z_1)d(Z_1)]$, and define $A_1 g(z) = \mathbb{E}[Dg(X_1) \mid Z_1 = z]$. With the above inner products, the adjoint satisfies

$$(A_1^\star h)(x) = \mathbb{E}[h(Z_1) \mid X_1 = x, D = 1]. \tag{5}$$

Assumption 2.3 (Observed adjoint range)**.** Given the latent anchor and calibration implication of Assumption 2.2, the observed adjoint range condition holds: for the Y_1 side there exists $b_1 \in \mathcal{H}_{Z_1}$ such that

$$A_1^\star b_1 = \ell_1, \qquad \ell_1(x) := \mathbb{E}[Y \mid X_1 = x, D = 1].$$

The analogous condition holds for the Y_0 side.

Assumption 2.4 (Support and overlap)**.** There exists $\varepsilon > 0$ such that $\varepsilon \le p_1(U, Y_1, C) \le 1 - \varepsilon$ and $\varepsilon \le p_0(U, Y_0, C) \le 1 - \varepsilon$ almost surely on the relevant supports, where $p_1(U, Y_1, C) = P(D = 1 \mid U, Y_1, C)$ and $p_0(U, Y_0, C) = P(D = 0 \mid U, Y_0, C)$; the observed arm probabilities $P(D = t \mid Z_t)$ are likewise bounded away from zero and one; and all displayed calibrated orthogonal moments have finite second moments.

The latent representer condition induces the observed balancing equation

$$\text{(R1-B)} \qquad A_1 q_1^\ell = r_1, \qquad r_1(z) := \mathbb{E}[1 - D \mid Z_1 = z].$$

(R1-A) is the primitive latent representability condition involving the unobserved heterogeneity U, and (T) is the transfer condition that carries the latent representer to observed moments. Lemma 3.1 shows that (R1-A) together with (T) induce (R1-B), with $\mathcal{Q}_1 := \{q : A_1 q = r_1\}$ denoting the observed calibration set. The main identification result does not require uniqueness of the primal solution q; the role of the adjoint range in certifying invariance of the target over $\mathcal{Q}_1$ is developed in Section 3. Note that (R1-B) plus (R2) is not a causal model by itself: (R1-A) is the condition that makes one element of the observed calibration set causally meaningful.

Further discussion of equivalent dual-range formulations, latent-versus-observed range conditions, and reduced scalar-index formulations is collected in Appendix B.3, with observed dual-range primitives in Appendix D.

2.2 Target-specific conditioning blocks

Let C_t denote the conditioning block used for the target μ_t. In the baseline theoretical notation the full ATE formulation uses $C_1 = C_0 = C = (S, Q)$; the bridge specifications in the HRS application

Table 3: Typical calibration arguments under different targets

Target	X_t	Z_t
Full ATE, μ_1 side	(W, Y_1, C)	(V, C)
Full ATE, μ_0 side	(W, Y_0, C)	(V, C)
μ_1 only	(W, Y_1, S)	(V, S)
μ_0 only	(W, Y_0, Q)	(V, Q)
Index formulation	$(W, Y_t, \kappa_t(S, Q))$	$(V, \kappa_t(S, Q))$

of Section 6 use this same Roy-side block $C_R = (S, Q)$, with the additional demographic covariates X entering only the observed-adjustment benchmark $C_A = (X, S, Q)$. One-sided targets may admit reduced conditioning blocks: when the two environments are asymmetric—for instance under $(Y_1, S) \perp\!\!\!\perp (Y_0, Q) \mid U$—the μ_1 reduction may use $C_1 = S$ and the μ_0 reduction $C_0 = Q$, so that each side conditions only on its own environment. Such reductions require their own identifying conditions and are not automatic; see Appendix B.3.

Corollary 2.2 (One-sided target-specific conditioning)**.** *If the reduced calibration implication and range conditions hold for C_1, then μ_1 is identified using $X_1 = (W, Y_1, C_1)$ and $Z_1 = (V, C_1)$. The analogous statement holds for μ_0.*

Proof. Apply the one-sided identification theorem with C replaced by C_1 and the corresponding reduced calibration and range conditions imposed; the μ_0 statement is symmetric. □

Figure 1 illustrates the one-sided reductions. These reductions are not consequences of simply deleting nodes from the full causal graph; they require their own calibration implications and range conditions. Appendix G.3 gives finite-state examples in which scalar-index reductions succeed or fail.

2.3 Economic interpretation

The calibration formulation of Figure 1(a) is a primitive representation: identification rests on the calibration implication and range conditions, not on the absence of particular arrows. We use a firm export-entry example to illustrate the economic roles of the variables and the distinction from standard proximal causal inference.

Following the heterogeneous-firms trade literature initiated by Melitz (2003), let $D = 1$ denote export-market participation and $D = 0$ domestic-only operation. Let Y_1 and Y_0 denote firm profits under exporting and domestic-only operation, respectively. Export entry depends on expected gains from exporting relative to domestic operation and fixed entry costs. The latent heterogeneity U summarizes unobserved productivity, managerial quality, organizational capital, and related firm characteristics. This is a selection-on-gains environment: the treatment decision may depend directly on both Y_1 and Y_0, not merely on latent type.

The latent-type measurement W consists of pre-entry measures informative about latent firm type, such as lagged productivity or sales growth. The exclusion $W \not\to D$ requires that, conditional on the latent type, these variables do not directly affect the entry decision. The outcome-representation variable V shifts the latent type distribution without directly affecting treatment conditional on the latent type and conditioning variables. Examples include founding-cohort industrial conditions or parent-firm productivity. Unlike an MTE/LATE instrument, V is not

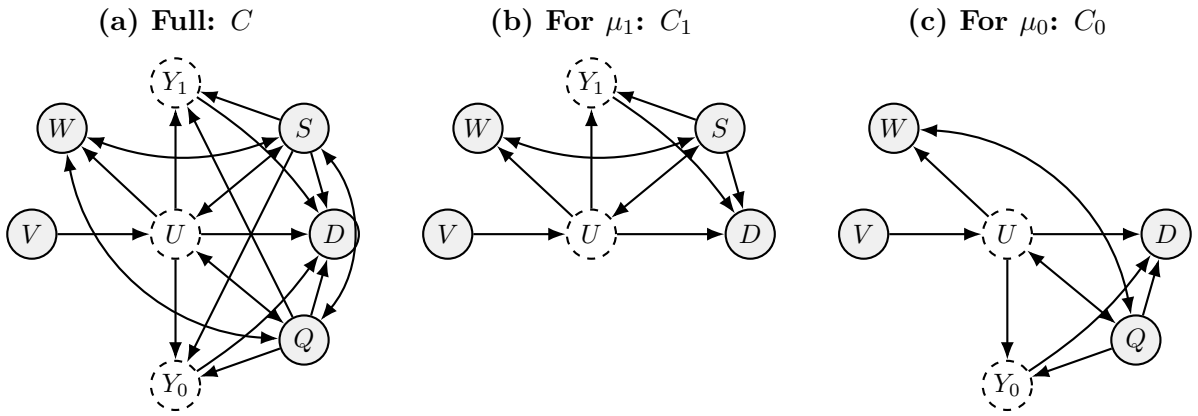


Figure 1: Target-specific reductions of the calibration formulation. Panel (a): the full formulation with conditioning block C for the average treatment effect $\tau = \mu_1 - \mu_0$; the latent type U affects the measurement, the potential outcomes, and treatment selection, and treatment selection may depend directly on the potential outcomes, allowing selection on gains. Bidirectional arrows denote unrestricted dependence paths. Panel (b): reduction for μ_1 using a target-specific block C_1, giving $X_1 = (W, Y_1, C_1)$ and $Z_1 = (V, C_1)$. Panel (c): symmetric reduction for μ_0 using C_0, giving $X_0 = (W, Y_0, C_0)$ and $Z_0 = (V, C_0)$. In the baseline two-environment illustration one may take $C = (S, Q)$, $C_1 = S$, and $C_0 = Q$, but these one-sided reductions require separate calibration and adjoint-range conditions and are not automatic. The dashed nodes Y_1 and Y_0 denote potential outcomes and are not jointly observed at the individual level.

required to shift treatment directly; rather, it provides the variation needed for the adjoint-side range condition. The variables S and Q represent environment variables affecting the treated and untreated potential outcomes, respectively. In the export example, S may include exchange rates or foreign demand conditions, while Q may include domestic demand or input-market conditions. Importantly, both S and Q may directly affect the treatment decision through expected gains from exporting. Thus $S \to D$ and $Q \to D$ are allowed. This differs fundamentally from standard proximal causal inference, where treatment assignment depends only on latent confounding and proxy variables are excluded from treatment.

The measurement side and the adjoint side are not interchangeable. The latent-type measurement W is used to recover the latent treatment odds structure, while V provides the adjoint-side variation needed for the adjoint range condition. If W contains insufficient information about latent heterogeneity, the latent selection-odds representer may fail to exist. If V has insufficient independent variation conditional on the conditioning variables, the adjoint representer may fail to exist or the parameter may become irregularly identified. Appendix G.3 formalizes this distinction.

For one-sided targets such as $\mu_1 = \mathbb{E}[Y_1]$, the untreated-side variables (Y_0, Q) may sometimes be marginalized out. In the export example, this may be plausible if domestic-side shocks affect export entry only through foregone domestic profit. Appendix B.2 gives a sufficient bilateral-independence condition under which such reduced formulations arise naturally. Additional economic interpretations and examples are provided in Appendix B.4.

2.4 Assumption dictionary

For navigability, Table 4 summarizes the four identifying conditions used throughout the paper and the regime where regular calibrated orthogonal moment inference applies. The conditions

(R1-A), (T), (R1-B), and (R2) live on *different* operators acting on *different* Hilbert spaces, and the paper consistently maintains this distinction.

Table 4: Identifying conditions and the regular regime.

Condition	Operator	Role	Observable?
(R1-A)	Latent measurement operator $T_W : \mathcal{H}_{W,Y_t,C} \to \mathcal{H}_{U,Y_t,C}$	Existence of a latent selection-odds representer q^ℓ solving $T_{W,t} q_t^\ell = r_t^\ell$ (Assumption 2.1); causal anchor in the latent state (U, Y_t, C)	No — latent structural condition; primitive sufficient conditions in Appendix B.3; observationally non-testable (Theorem D.9).
(T)	Conditional-transfer moment given (U, Y_t, C, V) (Assumption 2.2)	Exclusion restriction carrying q^ℓ into the observed calibration set, $q^\ell \in \mathcal{Q}_t$; deviations are measured by the latent calibration error $\eta_t(q)$ of Appendix D.2	No — latent exclusion; sufficient conditional independences in (4).
(R1-B)	Observed primal operator $A_1 : \mathcal{H}_{X_1} \to \mathcal{H}_{Z_1}$	Existence of an observed calibration solution $q \in \mathcal{Q}_1$; follows from (R1-A) and (T) by Lemma 3.1. Observed solvability alone is not a causal anchor.	Yes — A_1 is observable; diagnostics via finite-dim moment residuals.
(R2)	Observed adjoint $A_1^\star : \mathcal{H}_{Z_1} \to \mathcal{H}_{X_1}$	Existence of an adjoint outcome representer $b \in \mathcal{B}_1$; certifies μ_1 invariance over $\mathcal{Q}_1$.	Observed-side condition; holds under the primitive conditions of Appendix D.3, with finite-dimensional diagnostics (singular-value spectra, Picard coefficients) available as sanity checks.
R2(φ)	Target-specific adjoint range, $\ell_{t,\varphi} \in \mathcal{R}(A_t^\star)$	The displayed (R2) is the mean-specific case $\varphi(y, c) = y$; the closure membership $\ell_{t,\varphi} \in \overline{\mathcal{R}(A_t^\star)}$, equivalently $\varphi \in \mathfrak{V}_t^{\mathrm{cal}}$, governs invariance (Theorem 3.3)	Observed-side; residual, singular-value, and Picard diagnostics are available with ℓ_t replaced by $\ell_{t,\varphi}$, but exact population range membership is not nonparametrically testable.
Case R	$\ell_1 \in \mathcal{R}(A_1^\star)$	Finite-norm adjoint representer exists; a regular influence-function representation is available under finite-variance and local-path conditions, and root-N inference additionally requires the first-stage product-rate and regularization conditions of Appendix F.2.	Holds under the source conditions of Appendix D.4; diagnosable via Picard summability and Tikhonov path stability.

3 Primal–dual identification

Roadmap of the identification argument. The argument has five steps.

1. A latent selection-odds representer q^ℓ (R1-A) gives a causally correct IPW anchor (Step 1).

2. The transfer condition (T) transfers q^ℓ to the observed balancing equation $A_1 q = r_1$ (R1-B) (Step 2).
3. The observed calibration set $\mathcal{Q}_1$ is generally nonunique (Step 3).
4. An adjoint outcome representer $A_1^\star b = \ell_1$ certifies invariance of the target μ_t over $\mathcal{Q}_1$ (Step 4); more generally, invariance of a linear functional $\theta_1(\varphi) = \mathbb{E}[\varphi(Y_1, C)]$ holds if and only if its signal $\ell_{1,\varphi}$ lies in $\overline{\mathcal{R}(A_1^\star)}$ (Theorem 3.3).
5. The position of ℓ_1 (and of each $\ell_{1,\varphi}$) relative to $\mathcal{R}(A_1^\star)$ determines non-invariance, irregular identification, or regular orthogonal-moment representation (Step 5, Proposition 3.5).

The same primal–dual structure delivers an orthogonal score and the product-bias identity used in the inference theory.

3.1 Latent selection-odds representers and observed calibration

Under selection on gains, the latent treatment odds define a causal reweighting identity. The latent-type measurement and outcome-representation variables then convert this latent identity into the observed calibration inverse problem $A_1 q = r_1$. The possible nonuniqueness of the observed calibration motivates the target-invariance certificate developed in Section 3.2. Residualized and finite-dimensional implementations are numerical devices rather than additional population identification assumptions.

The substantive new content is the Branch G analysis, where treatment selection depends directly on potential outcomes, $D \leftarrow (Y_1, Y_0, U, C)$. In this branch the latent selection-odds representer depends on (U, Y_t, C) rather than on latent heterogeneity alone, which induces the primal–dual inverse-problem structure developed below. Branches A and U arise as simpler special cases obtained by restricting the calibration information set. Target-specific reductions for one-sided parameters are discussed in Section 2.2.

The latent odds ratio reweights treated outcomes to the population target. The resulting identity is then transferred to observed-data moments through the latent-type measurement W and the conditioning block $Z_1 = (V, C)$. Let

$$p_1(U, Y_1, C) = \mathbb{P}(D = 1 \mid U, Y_1, C), \qquad p_0(U, Y_0, C) = \mathbb{P}(D = 0 \mid U, Y_0, C).$$

Although the structural choice rule may depend on both potential outcomes, the Y_1-side probability $p_1(U, Y_1, C)$ denotes the one-sided conditional treatment probability obtained after integrating out the remaining potential outcome Y_0 under its conditional law given (U, Y_1, C). The Y_0-side probability $p_0(U, Y_0, C)$ is defined symmetrically.

Definition 3.1 (Latent selection-odds representers (Assumption 2.1 in operator form))**.** The latent selection-odds representers q_1^ℓ and q_0^ℓ satisfy

$$\mathbb{E}\big[q_1^\ell(W, Y_1, C) \mid U, Y_1, C\big] = \frac{1 - p_1(U, Y_1, C)}{p_1(U, Y_1, C)},$$
$$\mathbb{E}\big[q_0^\ell(W, Y_0, C) \mid U, Y_0, C\big] = \frac{1 - p_0(U, Y_0, C)}{p_0(U, Y_0, C)}.$$

In operator form, Definition 3.1 is exactly the representability condition (R1-A) of Assumption 2.1: $T_{W,t}\, q_t^\ell = r_t^\ell$ with $r_t^\ell = (1 - p_t)/p_t$. Definition 3.1 together with the transfer condition

(T) in Assumption 2.2 implies the normalization identity

$$\mathbb{E}\big[D\{1+q_1^\ell(W,Y_1,C)\}\mid U,Y_1,C\big]=1,$$

with the analogous identity on the Y_0 side. The latent representer relation induces both a calibration-weight representation of the one-sided means and an observed calibration inverse problem through the measurement W and the conditioning block $Z_1=(V,C)$.

By iterated expectations and the conditional independence structure of Assumption 2.2, the latent representer moment transfers to the observed-data moment conditional on Z_1.

Lemma 3.1 (Latent-to-observed transfer)**.** *Suppose Assumptions 2.1 and 2.2 ((R1-A) and (T)) hold. Then the Y_1-side latent selection-odds representer q_1^ℓ satisfies the observed-data balancing moment*

$$\mathbb{E}\big[Dq_1^\ell(X_1)-(1-D)\,\big|\,Z_1\big]=0,$$

that is, $q_1^\ell\in\mathcal{Q}_1$.

The latent-to-observed transfer yields both a calibration-weight representation of the target mean and the observed-data balancing moment condition.

Theorem 3.2 (Latent odds anchor and observed calibration set)**.** *If the latent selection-odds representers exist and the calibration implications hold, then the latent selection-odds representers q_1^ℓ, q_0^ℓ of Definition 3.1 deliver the calibration-weight (causal IPW) representation*

$$\mathbb{E}[Y_1]=\mathbb{E}[D\{1+q_1^\ell(X_1)\}Y],\qquad \mathbb{E}[Y_0]=\mathbb{E}[(1-D)\{1+q_0^\ell(X_0)\}Y].$$

Moreover, the Y_1-side latent selection-odds representer belongs to the observed calibration set of

$$A_1q_1^\ell=r_1,\qquad r_1(z)=\mathbb{E}[1-D\mid Z_1=z]. \tag{6}$$

More generally, every $q\in\mathcal{Q}_1$ *satisfies the observed balancing moment in Z-normalization form,*

$$\mathbb{E}[\{Dq(X_1)-(1-D)\}v_q(Z_1)]=0\quad \textit{for all } v_q\in L^2(Z_1), \tag{7}$$

whereas the calibration-weight representation displayed above relies on the latent anchor q_1^ℓ and does not hold for an arbitrary $q\in\mathcal{Q}_1$ in the absence of the target-invariance certificate of Section 3.2. The Y_0 side is analogous.

Proof. The calibration-weight representation follows immediately from the normalization identity and the observed outcome relation $Y=DY_1+(1-D)Y_0$. For the latent representative q_1^ℓ satisfying Assumption 2.1 (equivalently Definition 3.1) and the transfer condition in Assumption 2.2, Lemma 3.1 implies

$$\mathbb{E}[Dq_1^\ell(X_1)-(1-D)\mid Z_1]=0,$$

which is precisely (6). Equation (7) is equivalent to (6) by the tower property. Hence $q_1^\ell\in\mathcal{Q}_1$, so the solution set is nonempty. □

Roles of (R1-A), (T), (R1-B), and (R2).

It is important that the observed balancing equation $A_1q=r_1$ is not by itself a causal identification condition. It may have many solutions, and some solutions need not correspond to

any latent selection-odds representer. The role of (R1-A) is to guarantee that a causally correct calibrating function of (W, Y_t, C) exists at the latent level; the role of the transfer condition (T) is to anchor that element q^ℓ inside the observed calibration set $\mathcal{Q}_1$. The role of (R1-B), which follows from (R1-A) and (T) by Lemma 3.1, is to guarantee that this anchored element is observable as a solution of the empirical operator equation. The role of (R2) is different: it extends the calibration-weight representation from the anchored element q^ℓ to every observed representative $q \in \mathcal{Q}_1$ by certifying target invariance over $\ker(A_1)$. Primitive sufficient conditions for (R1-A) and (R2) are collected in Appendix D.2 and Appendix D.3, respectively. The latent conditions (R1-A) and (T) are, however, not testable implications of the observed distribution without additional latent structure. Theorem D.9 in Appendix D.2 makes this precise: for any admissible model satisfying (R1-A) it constructs an observationally equivalent stochastic Roy model—identical law of (D, Y, W, V, C)—in which (R1-A) fails. This is the usual limitation of proxy-based identification and missing-not-at-random calibration restrictions, and the paper accordingly treats (R1-A) and (T) as structural sufficient conditions, uses observed calibration residuals, singular values, and Tikhonov paths only as diagnostics for the observed implications (R1-B) and (R2), and quantifies deviations from the latent anchor through the latent-calibration sensitivity analysis of Appendix D.2.

The quantity $D\{1 + q(X_1)\}$ is a calibration weight rather than a propensity-score weight. Although the latent odds ratio $(1-p_1)/p_1$ is nonnegative, observed solutions $q \in \mathcal{Q}_1$ are identified only up to $\ker(A_1)$ and therefore need not be pointwise nonnegative. Hence $1 + q(X_1)$ may take negative values. The adjoint range condition developed in Section 3.2 is correspondingly branch-specific. In Branch G, the primal information set includes Y_t, so on the treated support $\ell_1(X_1) = \mathbb{E}[Y|X_1, D = 1] = Y_1$. The adjoint representer must therefore certify that this target component is recoverable from functions of the conditioning block Z_1.

Augmenting the calibration-weight representation with an arbitrary square-integrable function of Z_1 and applying the tower property gives $\mathbb{E}[D\{1+q_1^\ell(X_1)\}m(V, C)] = \mathbb{E}[m(V, C)]$ for any square-integrable function $m(Z_1)$. Consequently, the one-sided mean also admits, for any $m_1 \in L^2(Z_1)$, the augmented representation

$$\mathbb{E}[D(1 + q_1^\ell)(Y_1 - m_1(Z_1)) + m_1(Z_1)] = \mathbb{E}[Y_1]. \tag{8}$$

Residualized and projection-based representations of the augmented score are discussed in Appendix B.3.

3.2 Adjoint outcome representation and target invariance

Unlike standard proximal causal inference, the relevant selection-odds representer under selection on gains conditions on potential outcomes; the adjoint representer below is therefore a target-invariance certificate for the Branch G calibration set. See Appendix B.1 for further discussion.

Before introducing the adjoint outcome representer, it is useful to make the problem that the target-invariance certificate solves fully explicit. The observed calibration set $\mathcal{Q}_1$ defined in Theorem 3.2 is generically not a singleton: $\ker(A_1) \neq \{0\}$ in any ill-posed inverse problem with measurement structure that does not exhaust the variation in X_1. Two distinct solutions $q, q' \in \mathcal{Q}_1$ differ by a kernel element $q - q' \in \ker(A_1)$, and the plug-in identification formula $\mathbb{E}[D\{1 + q\}Y]$ can take different numerical values along different representatives unless an additional structural condition is imposed. The adjoint outcome representer introduced below provides exactly this

certification: it pins down the target value μ_1 as invariant across the entire solution set $\mathcal{Q}_1$.

Define the primal and dual solution sets by $\mathcal{Q}_1 = \{q \in \mathcal{H}_{X_1} : A_1 q = r_1\}$ and $\mathcal{B}_1 = \{b \in \mathcal{H}_{Z_1} : A_1^\star b = \ell_1\}$. The primal equation $A_1 q = r_1$ is the observed-data implication of the latent selection-odds representer. The adjoint equation $A_1^\star b = \ell_1$ is the certificate that the target is invariant over the calibration set.

The object made invariant by the observed calibration restrictions is not a unique calibrating function. The observed restrictions by themselves deliver only this invariance; under (R1-A) and (T), each invariant value coincides with the corresponding causal potential-outcome target (part (iii) of the theorem below). It is a scalar linear functional of the potential-outcome distribution, of which the mean is the leading case. Fix the Y_1 side and consider the class of linear potential-outcome functionals

$$\theta_1(\varphi) = \mathbb{E}[\varphi(Y_1, C)], \qquad \varphi \in \mathcal{H}_1 := \{\varphi : \mathbb{E}[\varphi(Y_1, C)^2] < \infty\},$$

with selected-arm target signal

$$\ell_{1,\varphi}(x) := \mathbb{E}[\varphi(Y, C) \mid X_1 = x,\, D = 1].$$

Because $Y = Y_1$ on $\{D = 1\}$ and $X_1 = (W, Y_1, C)$ contains (Y_1, C), the signal is the functional's integrand itself on the treated support, $\ell_{1,\varphi}(X_1) = \varphi(Y_1, C)$. The mean μ_1 corresponds to $\varphi(y, c) = y$, with $\ell_{1,\varphi} = \ell_1$; indicator choices $\varphi_y(u, c) = \mathbf{1}\{u \le y\}$ give marginal distribution functions (quantiles then follow by inversion at continuity points; Appendix B.3), and $\varphi(y, c) = y\,\mathbf{1}\{c \in \mathcal{S}\}$ gives subgroup-weighted means (the conditional subgroup mean follows after dividing by $\mathbb{P}(C \in \mathcal{S})$, identified directly from the observed law). Under (R1-A) and (T), the anchor identity extends verbatim from the mean to any $\varphi \in \mathcal{H}_1$: since $\mathbb{E}[D\{1 + q_1^\ell(X_1)\} \mid U, Y_1, C] = 1$ and $D\varphi(Y, C) = D\varphi(Y_1, C)$,

$$\mathbb{E}\big[D\{1 + q_1^\ell(X_1)\}\varphi(Y, C)\big] = \theta_1(\varphi).$$

The following result characterizes exactly which of these scalars are invariant over the observed calibration set; it is a necessary-and-sufficient characterization of the maximal calibration-invariant functional class, together with its causal identification under the latent anchor conditions, not merely a sufficient condition for the mean.

Theorem 3.3 (Maximal calibration-invariant class of linear targets: necessity and sufficiency)**.** *Fix a data-generating law P in the maintained model; the operators $A_1, A_1^\star$, the signals $\ell_{1,\varphi}$, and the classes below are defined pointwise at P, while each member of $\mathfrak{L}_1^{\mathrm{cal}}$ is a functional $P^\star \mapsto \mathbb{E}_{P^\star}[\varphi(Y_1, C)]$ on candidate laws. Assume $\mathcal{Q}_1 \neq \emptyset$ and fix $\varphi \in \mathcal{H}_1$. (i) The calibrated-representation map*

$$\mathcal{G}_{1,\varphi}(q) := \mathbb{E}\big[D\{1 + q(X_1)\}\varphi(Y, C)\big]$$

is constant over $\mathcal{Q}_1$ if and only if

$$\ell_{1,\varphi} \in \ker(A_1)^\perp = \overline{\mathcal{R}(A_1^\star)}.$$

(ii) Consequently

$$\mathfrak{V}_1^{\mathrm{cal}} := \big\{\varphi \in \mathcal{H}_1 :\ \ell_{1,\varphi} \in \overline{\mathcal{R}(A_1^\star)}\big\}$$

is the maximal class of square-integrable integrands—and

$$\mathfrak{L}_1^{\mathrm{cal}} := \big\{P \mapsto \mathbb{E}_P[\varphi(Y_1, C)] :\ \varphi \in \mathfrak{V}_1^{\mathrm{cal}}\big\}$$

the maximal class of arm-specific linear expectation functionals of the marginal law of (Y_1, C)*—whose calibrated representation* $\mathbb{E}[D\{1+q\}\varphi(Y,C)]$ *is invariant over* $\mathcal{Q}_1$*. (iii) If in addition (R1-A) and (T) hold (Assumptions 2.1 and 2.2), then for every* $\varphi \in \mathfrak{V}_1^{\mathrm{cal}}$ *the common value equals the causal target* $\theta_1(\varphi) = \mathbb{E}[\varphi(Y_1, C)]$*.*

Proof. For any $q, q' \in \mathcal{Q}_1$, let $g = q - q' \in \ker(A_1)$. The difference of calibrated representations is

$$\mathcal{G}_{1,\varphi}(q) - \mathcal{G}_{1,\varphi}(q') = \mathbb{E}[D\{q(X_1) - q'(X_1)\}\varphi(Y,C)] = \langle q - q',\, \ell_{1,\varphi}\rangle_{X_1}.$$

Invariance over $\mathcal{Q}_1$ is equivalent to $\langle g, \ell_{1,\varphi}\rangle_{X_1} = 0$ for every $g \in \ker(A_1)$, i.e., $\ell_{1,\varphi} \in \ker(A_1)^\perp$. The Hilbert-space identity $\ker(A_1)^\perp = \overline{\mathcal{R}(A_1^\star)}$ completes the proof of (i). Part (ii) is (i) read as a statement about the class: invariance holds for every $\varphi \in \mathfrak{V}_1^{\mathrm{cal}}$ and fails for every $\varphi \notin \mathfrak{V}_1^{\mathrm{cal}}$, so no strictly larger class of integrands—hence of the induced expectation functionals—is calibration-invariant. For (iii), under (R1-A) and (T) the anchor q_1^ℓ lies in $\mathcal{Q}_1$ by Lemma 3.1, and the displayed anchor identity gives $\mathbb{E}[D\{1+q_1^\ell\}\varphi(Y,C)] = \theta_1(\varphi)$; invariance then extends this value to every $q \in \mathcal{Q}_1$. □

Four remarks delimit the statement.

- *Scope of maximality.* Maximality is relative to square-integrable linear functionals of the *marginal* law of (Y_t, C). It does not cover functionals of the joint law of (Y_1, Y_0), such as $\mathbb{E}[h(Y_1, Y_0, C)]$, which remain unidentified without additional structure, nor nonlinear functionals such as quantiles, which are obtained from the identified marginal distribution functions by inversion under the usual continuity and uniqueness conditions (Appendix B.3).
- *Quotient reading.* Since $\mathcal{Q}_1 = q_1^0 + \ker(A_1)$ for any $q_1^0 \in \mathcal{Q}_1$, the observed data identify the calibrating function only as an equivalence class in the quotient space $\mathcal{H}_{X_1}/\ker(A_1)$; the theorem states that a linear target descends to this quotient if and only if its selected-arm signal annihilates $\ker(A_1)$, equivalently if the signal belongs to $\overline{\mathcal{R}(A_1^\star)}$.
- *Values versus functionals.* We write $\Theta_1 := \{\theta_1(\varphi) : \varphi \in \mathfrak{V}_1^{\mathrm{cal}}\}$ for the corresponding set of identified *values*; distinct integrands may share a value, so Θ_1 is a set of scalars while $\mathfrak{L}_1^{\mathrm{cal}}$ is the functional class defined pointwise at the maintained law P.
- *Symmetry.* The statement and proof apply verbatim to $t = 0$, with $(D, X_1, Z_1, A_1, \ell_{1,\varphi})$ replaced by $(1-D, X_0, Z_0, A_0, \ell_{0,\varphi})$; the arm-0 classes are denoted $\mathfrak{V}_0^{\mathrm{cal}}$ and $\mathfrak{L}_0^{\mathrm{cal}}$.

The Branch G adjoint range condition as an identification boundary.

In Branch G the adjoint range condition is not a technical side condition. It is the target-invariance requirement created by allowing treatment choice to depend directly on potential outcomes. Because $Y = Y_1$ on the treated support, $\ell_1(W, Y_1, C) = \mathbb{E}[Y \mid X_1, D = 1] = Y_1$ on $\{D = 1\}$, and the adjoint equation asks whether conditional expectations of functions of $Z_1 = (V, C)$ can reproduce the treated outcome signal Y_1 as a function of $X_1 = (W, Y_1, C)$. When this requirement fails, the framework does not silently extrapolate: by Proposition 3.5 below, the target either varies along the primal kernel or is only irregularly identified. The residualized rewriting $b = m + h$ analyzed below is a numerical preconditioning device and does not weaken this requirement at the population level. In the HRS application of Section 6, parental education supplies the adjoint-side variation, and baseline cognition and self-rated health supply

the measurement-side variation (respondent's own schooling enters the observed-adjustment block C_A, not the main bridge block C_R); the interpretation of the Branch G estimate depends on these variables being rich enough to support the corresponding range condition.

A finite-support sufficient condition.

The Branch G adjoint range condition is strong but not vacuous, and it has a transparent rank interpretation in finite-support settings. Suppose, for a fixed value of C, that $X_1 = (W, Y_1, C)$ and $Z_1 = (V, C)$ have finite support under the treated law. Let M denote the matrix with entries

$$M_{xz} = \mathbb{P}(Z_1 = z \mid X_1 = x, D = 1).$$

Then the dual equation $A_1^\star b = \ell_1$ is the linear system $Mb = \ell_1$, and the Branch G adjoint range condition is equivalent to

$$\ell_1 \in \mathrm{col}(M).$$

Since $\ell_1(W, Y_1, C) = Y_1$ on the treated support, this asks whether the treated-law variation in $Z_1 = (V, C)$ is rich enough to reproduce the realized treated potential outcome as a conditional expectation. Full row rank of M is sufficient. The rank condition is stated conditional on the treated support; it is therefore a condition on the treated-law measurement variation, not on the marginal support of V alone. The continuous-measurement examples in Appendix D, including the Gaussian–Hermite and deconvolution primitives for $A_1^\star$, are infinite-dimensional analogues of this rank condition: existence of a finite-norm b reduces to a Picard summability requirement, while completeness conditions on the treated law of (X_1, Z_1) deliver uniqueness of any such b.

Although the population range condition cannot be tested nonparametrically from the observed distribution, it holds under transparent primitive conditions—the finite-support rank condition above, a Hermite–Picard summability condition under a Gaussian treated-law model, and a source condition on the target signal (Appendix D.3; see also Appendix D.4). The maintained range condition is therefore a substantive but structurally grounded restriction, not an untethered assumption.

When a finite-norm adjoint representer exists, the invariance certificate takes the following explicit representation form.

Theorem 3.4 (Primal–dual identification of linear functionals via a finite-norm adjoint representer)**.** *Suppose the latent selection-odds representer anchors the observed calibration set ((R1-A) and (T)), fix $\varphi \in \mathcal{H}_1$, let $\mathcal{B}_{1,\varphi} := \{b \in \mathcal{H}_{Z_1} : A_1^\star b = \ell_{1,\varphi}\}$, and assume $\mathcal{B}_{1,\varphi} \neq \emptyset$. Then for any $q \in \mathcal{Q}_1$ and any $b \in \mathcal{B}_{1,\varphi}$,*

$$\theta_1(\varphi) = \mathbb{E}\big[D\{1 + q(X_1)\}\{\varphi(Y, C) - b(Z_1)\} + b(Z_1)\big]. \tag{9}$$

Taking $\varphi(y, c) = y$ gives $\mathcal{B}_{1,\varphi} = \mathcal{B}_1$ and the treated mean $\mu_1 = \mathbb{E}[Y_1]$. Equivalently, writing $b = m + h$,

$$\theta_1(\varphi) = \mathbb{E}\big[D(1 + q)(\varphi(Y, C) - m) + m - \{D(1 + q) - 1\}h\big].$$

Proof. By Assumptions 2.1 and 2.2 and the latent-to-observed transfer (Lemma 3.1), there exists a latent representative $q_1^\ell \in \mathcal{Q}_1$. By the anchor identity preceding Theorem 3.3, $\theta_1(\varphi) = \mathbb{E}[D\{1 + q_1^\ell(X_1)\}\varphi(Y, C)]$. Since $\mathbb{E}[D\{1 + q_1^\ell(X_1)\} - 1 \mid Z_1] = 0$, for any $b \in \mathcal{H}_{Z_1}$,

$$\theta_1(\varphi) = \mathbb{E}[D\{1 + q_1^\ell(X_1)\}\{\varphi(Y, C) - b(Z_1)\} + b(Z_1)].$$

For any $q \in \mathcal{Q}_1$, let $g = q - q_1^\ell \in \ker(A_1)$. Then

$$\mathbb{E}[Dg(X_1)\{\varphi(Y, C) - b(Z_1)\}] = \langle g, \ell_{1,\varphi} - A_1^\star b \rangle_{X_1}.$$

Hence, if $b \in \mathcal{B}_{1,\varphi}$, we have $\mathbb{E}[Dg(X_1)\{\varphi(Y, C) - b(Z_1)\}] = 0$. Therefore (9) holds for every $q \in \mathcal{Q}_1$ and every $b \in \mathcal{B}_{1,\varphi}$. □

The residualized form is useful for orthogonal-score construction and finite-dimensional implementation.

3.3 Identification geometry

The invariance result of Theorem 3.3, combined with the operative range condition for regular inference, yields the following identification geometry determined by the location of the target signal ℓ_1 relative to the range of the adjoint operator $A_1^\star$. The geometry applies functional by functional: different functionals of the same potential-outcome distribution can fall in different cases. The proposition is stated for a generic $\varphi \in \mathcal{H}_1$; taking $\varphi(y, c) = y$ yields the mean case ($\ell_{1,\varphi} = \ell_1$) summarized in Table 5.

Proposition 3.5 (Non-invariance, irregularity, and regularity)**.** *Assume the latent selection-odds representer exists and anchors the observed calibration set (Assumptions 2.1 and 2.2, i.e., conditions (R1-A) and (T)), and that $\mathcal{Q}_1 \neq \emptyset$. Fix $\varphi \in \mathcal{H}_1$ with signal $\ell_{1,\varphi}$ and target $\theta_1(\varphi)$; exactly three cases are exclusive and exhaustive:*

Case N (non-invariance). *$\ell_{1,\varphi} \notin \overline{\mathcal{R}(A_1^\star)}$. The observed calibration restrictions do not make the target invariant over $\mathcal{Q}_1$. Hence the observed calibration system alone does not identify a unique value of $\theta_1(\varphi)$ unless the latent anchor is further restricted to a target-invariant subset of $\mathcal{Q}_1$.*

Case I (irregular). *$\ell_{1,\varphi} \in \overline{\mathcal{R}(A_1^\star)} \setminus \mathcal{R}(A_1^\star)$. The target is invariant over the observed calibration set—hence point identified under the latent anchor conditions—but no finite-norm adjoint representer exists: any sequence with $A_1^\star b_1^{(n)} \to \ell_{1,\varphi}$ must have $\|b_1^{(n)}\| \to \infty$, so the identifying functional is discontinuous (irregular) in the observed inverse problem. It is the adjoint representation, not the identified value, that exists only in the limit.*

Case R (regular). *$\ell_{1,\varphi} \in \mathcal{R}(A_1^\star)$ (the target-specific strict-range condition R2(φ)). A finite-norm adjoint representer exists. If, in addition, the induced score has finite variance and the local observed calibration model admits differentiable calibration paths through the chosen representative, the score is a regular influence function.*

Proof. The identity $\ker(A_1)^\perp = \overline{\mathcal{R}(A_1^\star)}$ implies the orthogonal decomposition

$$\mathcal{H}_{X_1} = \overline{\mathcal{R}(A_1^\star)} \oplus \ker(A_1).$$

If $\ell_{1,\varphi} \notin \overline{\mathcal{R}(A_1^\star)}$, then $\ell_{1,\varphi}$ has a nonzero component in $\ker(A_1)$, so the linear functional varies along kernel directions and is not invariant over $\mathcal{Q}_1$. If $\ell_{1,\varphi} \in \overline{\mathcal{R}(A_1^\star)} \setminus \mathcal{R}(A_1^\star)$, then the functional is invariant over $\mathcal{Q}_1$, but no finite-norm adjoint representer exists. If $\ell_{1,\varphi} \in \mathcal{R}(A_1^\star)$, then there exists $b \in \mathcal{H}_{Z_1}$ satisfying $A_1^\star b = \ell_{1,\varphi}$, so Theorem 3.4 applies and the corresponding orthogonal score admits a regular influence-function representation under finite variance. □

Table 5: Identification geometry (for a generic functional $\theta_1(\varphi)$, replace ℓ_1 by $\ell_{1,\varphi}$)

Location of ℓ_1	Identification status	Adjoint representer
$\ell_1 \notin \overline{\mathcal{R}(A_1^\star)}$	non-invariant	none
$\ell_1 \in \overline{\mathcal{R}(A_1^\star)} \setminus \mathcal{R}(A_1^\star)$	irregular	no finite-norm solution
$\ell_1 \in \mathcal{R}(A_1^\star)$	regular	finite-norm solution

The statement in Case N is a calibration-based non-invariance statement: the observed calibration restrictions do not pin down μ_1. It is not, by itself, a global causal non-identification theorem for the full latent Roy model, which would additionally require constructing admissible latent distributions that preserve the observed law while changing the target.

Table 5 summarizes the resulting identification geometry. Combining the corresponding Y_1- and Y_0-side constructions yields the following ATE representation.

[The ATE result, previously a stand-alone theorem, is restated as the leading corollary of Theorem 3.4.]

Corollary 3.6 (ATE identification)**.** *If both one-sided primal-dual systems hold (Theorem 3.4 with $\varphi(y,c) = y$ on each side), then*

$$\tau = \mathbb{E}[\Gamma_1(O; q_1, b_1) - \Gamma_0(O; q_0, b_0)], \tag{10}$$

where $\Gamma_1(O; q_1, b_1) = D\{1 + q_1(X_1)\}\{Y - b_1(Z_1)\} + b_1(Z_1)$ and $\Gamma_0(O; q_0, b_0) = (1 - D)\{1 + q_0(X_0)\}\{Y - b_0(Z_0)\} + b_0(Z_0)$ are the calibration signals for the two potential outcomes.

Proof. Subtract the two one-sided primal–dual identification formulas. □

The calibration restrictions can be viewed as sharpening the usual MTE/IV identified set; see Appendix B.1. Theorem 3.4 already covers bounded linear functionals $\mathbb{E}[\varphi(Y_t, C)]$, including marginal distribution functions ($\varphi_y(u,c) = \mathbf{1}\{u \le y\}$) and subgroup-weighted means; implementation details and the strict-range versus closure distinction for these functionals are collected in Appendix B.3. Practical diagnostics based on singular values and regularization paths are discussed in Appendix B.3.

3.4 Orthogonal moments, efficiency, and residualization

The primal–dual representation immediately induces an orthogonal score for estimation and inference. The resulting orthogonal score is closely related to debiased scores in inverse problems; see Appendix B.1.

Theorem 3.7 (Calibrated orthogonal moment)**.** *Define*

$$\psi_1(O; \mu_1, q, b) = \Gamma_1(O; q, b) - \mu_1, \qquad \Gamma_1(O; q, b) = D\{1 + q(X_1)\}\{Y - b(Z_1)\} + b(Z_1). \tag{11}$$

At any valid primal-dual pair $(q, b) \in \mathcal{Q}_1 \times \mathcal{B}_1$, the mean of ψ_1 is insensitive to first-order perturbations of q and b in their respective tangent directions.

Proof. For a perturbation δq,

$$\mathbb{E}[D\,\delta q(X_1)\{Y - b(Z_1)\}] = \langle \delta q, \ell_1 - A_1^\star b \rangle_{X_1} = 0,$$

since $A_1^\star b = \ell_1$. For a perturbation δb, $\mathbb{E}[\{1 - D(1 + q(X_1))\}\delta b(Z_1)] = 0$, since $q \in \mathcal{Q}_1$ implies $\mathbb{E}[D\{1 + q(X_1)\} - 1 \mid Z_1] = 0$. □

Corollary 3.8 (Orthogonal moment for a generic linear target)**.** *Fix $\varphi \in \mathcal{H}_1$ and suppose $A_1 q = r_1$ and $A_1^\star b_{1,\varphi} = \ell_{1,\varphi}$. Then*

$$\psi_{1,\varphi}(O) = D\{1 + q(X_1)\}\{\varphi(Y, C) - b_{1,\varphi}(Z_1)\} + b_{1,\varphi}(Z_1) - \theta_1(\varphi)$$

is Neyman-orthogonal in q and $b_{1,\varphi}$, by the proof above with Y replaced by $\varphi(Y, C)$ and ℓ_1 by $\ell_{1,\varphi}$. The exact product-bias decomposition also holds verbatim: at any candidate pair $(\tilde{q}, \tilde{b})$,

$$\mathbb{E}[\psi_{1,\varphi}(O; \tilde{q}, \tilde{b})] = \mathbb{E}[D(\tilde{q} - q)\{\varphi(Y, C) - b_{1,\varphi}\}] + \mathbb{E}[\{1 - D(1 + q)\}(\tilde{b} - b_{1,\varphi})] - \mathbb{E}[D(\tilde{q} - q)(\tilde{b} - b_{1,\varphi})],$$

with the first two terms vanishing at the representer equations, so the generic-target theory inherits the double-robust product-rate structure of Appendix F.2 without modification.

The orthogonality property follows directly from the primal and adjoint representer equations. The next result shows that, within this calibrated orthogonal moment class, these equations are also necessary for first-order orthogonality (see Appendix E.1 for the proof).

Proposition 3.9 (Necessity of calibration equations)**.** *Within the signal class* (11)*, requiring first-order invariance with respect to arbitrary q-directions implies $A_1^\star b = \ell_1$, and requiring first-order invariance with respect to arbitrary b-directions implies $A_1 q = r_1$. Thus the primal and adjoint representer equations are necessary for Neyman orthogonality of this calibrated orthogonal moment.*

Among valid adjoint representers, different choices of b generally induce different finite-sample variances. The next result characterizes the variance-minimizing adjoint representer for a fixed primal representative q.

Theorem 3.10 (Minimum-variance adjoint representer for fixed q)**.** *Fix $q \in \mathcal{Q}_1$ and let $\mathcal{B}_1 \neq \emptyset$. Under the finite-variance and closedness conditions stated in Appendix G.1—in particular, the weight $w_q(Z_1) = \mathbb{E}[\{1 - D(1 + q(X_1))\}^2 \mid Z_1]$ finite and bounded away from zero and infinity, which makes $\mathcal{B}_1$ closed in the weighted norm and the projection below well defined—, the variance-minimizing adjoint representer*

$$b_q^\star \in \arg\min_{b \in \mathcal{B}_1} \mathrm{Var}\{\Gamma_1(O; q, b)\}$$

exists and is a weighted projection of an unconstrained regression function onto $\mathcal{B}_1$.

The adjoint range condition

$$\text{(R2)} \qquad \ell_1 \in \mathcal{R}(A_1^\star)$$

admits the following equivalent residualized representation. Writing $b = m + h$ separates a user-chosen baseline component from the correction term and is useful for finite-dimensional regularized implementation. With $m \in \mathcal{H}_{Z_1}$ fixed, define

$$\rho_{1,m}(X_1) = \mathbb{E}[Y_1 - m(Z_1) \mid X_1, D = 1] = \ell_1(X_1) - A_1^\star m(X_1).$$

Setting $h = b - m$, the dual equation $A_1^\star b = \ell_1$ is equivalent to $A_1^\star h = \rho_{1,m}$. Hence the residualized

adjoint range condition becomes $\rho_{1,m} \in \mathcal{R}(A_1^\star)$ for some $m \in \mathcal{H}_{Z_1}$. Moreover, for any $m, m' \in \mathcal{H}_{Z_1}$,

$$\rho_{1,m} - \rho_{1,m'} = A_1^\star(m' - m) \in \mathcal{R}(A_1^\star),$$

so $\rho_{1,m} \in \mathcal{R}(A_1^\star)$ if and only if $\rho_{1,m'} \in \mathcal{R}(A_1^\star)$. Thus the residualized condition is population-equivalent to $\ell_1 \in \mathcal{R}(A_1^\star)$, and the role of m is purely numerical.

The two forms are population-equivalent; the choice of m is a finite-sample preconditioning device. Primitive sufficient conditions for the adjoint range condition are collected in Appendix B.3. Specification diagnostics based on overidentifying observed calibration moments are discussed in Appendix B.3.

4 Moment conditions and implementation

Clarifications.

Before turning to implementation, several clarifications are useful to set against the abstract identification results of Section 3. First, the observed balancing equation $A_1 q = r_1$ does not itself identify the causal target; the latent odds anchor in Assumption 2.1, together with the transfer condition in Assumption 2.2, does. Second, the adjoint representer b is not a nuisance normalization but the certificate that makes the target invariant over nonunique primal solutions $q \in \mathcal{Q}_1$. Third, the Branch G observed calibration $q_t(W, Y_t, C)$ does not require imputing counterfactual outcomes because it is evaluated only on the arm $\{D = t\}$ where $Y_t = Y$ is observed. Fourth, the empirical branch comparison developed below is a diagnostic decomposition across maintained information structures, not a formal specification test of which branch is true.

Under the regular identification regime (Case R of Proposition 3.5), the calibration representation

$$\mu_1 = \mathbb{E}[D\{1 + q\}\{Y - b\} + b]$$

yields a feasible semiparametric estimation strategy. This section develops the exact Sieve GMM implementation studied in the formal theory and contrasts it with reduced-space OLS diagnostics used for exploratory analysis and sensitivity assessment. The exact Sieve GMM implementation (Appendix F.4, with the cross-fitted double machine learning (DML) version developed in Appendix F) enforces the full $L^2(X_1)$ adjoint moment for the adjoint representer b and is covered by the formal identification, orthogonality, and $\sqrt{N}$ inference theory of this paper. By contrast, the common-subspace and Z_1-projection OLS implementations discussed in Appendix B.3 are reduced-space diagnostics rather than exact estimators and are not covered by the formal theorem unless the reduced projection happens to satisfy the full adjoint moment. Table 6 fixes the role of each procedure used in the paper.

The reduced-space OLS implementations are useful for preliminary exploration, numerical stability checks, and sensitivity analysis, but should not be viewed as the primary inference procedure.

The residual correction component h is implemented through the adjoint moment condition

$$\mathbb{E}\Big[D\{Y_1 - m(Z_1) - h(Z_1)\}s(X_1)\Big] = 0 \quad \forall s \in L^2(X_1). \tag{12}$$

Equivalently,

$$\mathbb{E}[D\{Y_1 - m(Z_1) - h(Z_1)\} \mid X_1] = 0.$$

Table 6: Procedures, their population targets, and their roles.

Procedure	Population condition targeted	Role in paper	Covered by main theory?
Exact adjoint-moment sieve GMM	$A_1 q = r_1$, $A_1^\star b = \ell_1$	Primary estimator	Yes
Cross-fitted calibrated orthogonal moment (DML)	Same, with sample splitting	Primary inference implementation	Yes, under the stated rate conditions
Common-subspace OLS	Projected adjoint moment	Diagnostic / approximation	No, unless the projection equals the full adjoint condition
Z_1-projection OLS	Projected adjoint residual	Diagnostic / sensitivity	No, unless the full adjoint moment holds
Minimum-variance sieve GMM	Variance-minimizing representative	Inference refinement	Yes, under additional conditions

This condition implements the adjoint representation equation $A_1^\star h = \rho_{1,m}$ for the chosen auxiliary function $m(Z_1)$ and restores orthogonality of the calibrated orthogonal moment in the q direction. For computational convenience, one may replace the full adjoint moment condition with reduced-space projections on functions of (S, Q) or Z_1. These reduced-space OLS approximations, including common-subspace and Z_1-projection variants, are discussed in Appendix B.3. Reduced-space OLS approximations are useful for diagnostics, but the formal inference theory in this paper applies to the exact adjoint-moment implementation. Extensions that target the minimum-variance adjoint representer are described in Appendix F.4.

Sieve GMM implementation.

The identified mean is invariant to the choice of valid calibration representatives, but the finite-sample variance of the corresponding score is not. The baseline implementation therefore adopts regularized minimum-norm calibration representatives for numerical stability, and reported inference states which representative underlies the score. Optional constrained parameterizations and extended implementations targeting the minimum-variance calibration representatives of Theorem 3.10 are described in Appendix F.4. Let ψ_q, ϕ_m, and ϕ_h denote finite-dimensional sieve basis vectors for the observed calibration q, auxiliary projection m, and residual correction h, respectively. Let $\xi_q(Z_1)$, $\xi_m(Z_1)$, and $\xi_h(X_1)$ denote finite-dimensional vectors of test functions used to approximate the corresponding conditional moment restrictions. The sieve representations are

$$q_1(W, Y, C; \theta_q) = \psi_q(W, Y, C)^\top \theta_q,$$
$$m_1(Z_1; \beta_m) = \phi_m(Z_1)^\top \beta_m, \quad h_1(Z_1; \gamma_h) = \phi_h(Z_1)^\top \gamma_h.$$

Based on the first-stage primal moment,

$$\widehat{g}_q(\theta) = \frac{1}{N} \sum_{i=1}^{N} \{D_i q_1(W_i, Y_i, C_i; \theta) - (1 - D_i)\}\, \xi_q(Z_{1i}),$$

the sieve GMM estimator for θ_q is $\widehat{\theta}_q \in \arg\min_\theta \widehat{g}_q(\theta)^\top W_q \widehat{g}_q(\theta) + \lambda_q \|\theta\|^2$ with a weight matrix W_q and tuning parameter λ_q. Letting $\hat{q}_i = q_1(W_i, Y_i, C_i; \widehat{\theta}_q)$, the auxiliary projection moment is

constructed as

$$\widehat{g}_m(\beta) = \frac{1}{N}\sum_{i=1}^{N}\{D_i(1+\hat{q}_i)Y_i - m_1(Z_{1i};\beta)\}\xi_m(Z_{1i}),$$

and the sieve GMM estimator for β_m is given by $\widehat{\beta}_m \in \arg\min_\beta \widehat{g}_m(\beta)^\top W_m \widehat{g}_m(\beta) + \lambda_m\|\beta\|^2$ with a weight matrix W_m and tuning parameter λ_m. Letting $\hat{m}_i = m_1(Z_{1i};\widehat{\beta}_m)$, the residual correction moment is

$$\widehat{g}_h(\gamma) = \frac{1}{N}\sum_{i=1}^{N} D_i\{Y_i - \hat{m}_i - h_1(Z_{1i};\gamma)\}\xi_h(X_{1i}),$$

with the sieve GMM estimator $\widehat{\gamma}_h \in \arg\min_\gamma \widehat{g}_h(\gamma)^\top W_h \widehat{g}_h(\gamma) + \lambda_h\|\gamma\|^2$ with a weight matrix W_h and tuning parameter λ_h. Then the plug-in estimator of $\mu_1 = \mathbb{E}[Y_1]$ is obtained as

$$\hat{\mu}_1^{\mathrm{SGMM}} = \frac{1}{N}\sum_{i=1}^{N}\hat{\psi}_i, \qquad \hat{\psi}_i = D_i(1+\hat{q}_i)(Y_i - \hat{m}_i) + \hat{m}_i - \{D_i(1+\hat{q}_i) - 1\}\hat{h}_i. \tag{13}$$

This defines a plug-in sieve GMM estimator based on the calibrated orthogonal moment. Without cross-fitting, the nuisances $\hat{q}, \hat{m}, \hat{h}$ are estimated on the same sample on which $\hat{\mu}_1^{\mathrm{SGMM}}$ is evaluated, so the asymptotic theory requires sieve approximation rates fast enough to control the own-sample empirical-process bias (and is not the cross-fitted DML asymptotic theory of Chernozhukov et al. (2018)). The cross-fitted version $\hat{\mu}_1^{\mathrm{DML}}$ is a cross-fitted regularized sieve-GMM calibrated orthogonal moment: the label DML refers to the sample-splitting construction, not to generic machine-learning nuisance fitting. It uses sample splits to neutralize the own-sample empirical-process bias, admits $\sqrt{N}$ inference under weaker rate conditions, and is developed separately in Appendix F. Appendix F.4 describes an extended Sieve GMM that directly targets the minimum-variance adjoint representer. When the primal calibration operator is near-singular, so that these regular-regime standard errors can under-cover, a GMM-distance weak-calibration diagnostic (obtained by inverting a moment test) is reported in Appendix D.2; it is a finite-sample diagnostic that itself exhibits under-coverage, not a coverage-valid inference procedure.

Branch-specific implementation.

The same Sieve GMM machinery applies across branches. Each branch chooses a pair $(X_1^{\mathrm{b}}, Z_1^{\mathrm{b}})$ and a corresponding operator $A_1^{\mathrm{b}} : L^2(X_1^{\mathrm{b}}) \to L^2(Z_1^{\mathrm{b}})$.

For Branch A, let C_A denote the observed adjustment block. The reduced primal–dual system is

$$X_t^A = Z_t^A = C_A,$$

i.e. Branch A imposes a single observed conditioning block on both sides of the calibration system; the outcome-representation variable V does not enter Branch A because the observed calibration depends only on C_A. The sieve basis uses only functions of C_A, and the primal moment

$$\mathbb{E}[Dq^A(C_A) - (1-D) \mid C_A] = 0$$

implies

$$q^A(C_A) = \frac{1 - p_A(C_A)}{p_A(C_A)}, \qquad p_A(C_A) = \mathbb{P}(D = 1 \mid C_A),$$

recovering the standard propensity-score representation. The adjoint representer is the treated

outcome regression $b^A(C_A) = \mathbb{E}[Y \mid C_A, D = 1]$. For Branches U and G we use $Z_t = (V, C)$ with branch-specific primal domains $X_t^U = (W, C)$ and $X_t^G = (W, Y_t, C)$ respectively, so that the outcome-representation variable V enters the conditional moments below.

For Branch U, use $X_t^U = (W, C)$ for the observed calibration. The sieve basis augments Branch A with functions of W, and the primal moment becomes

$$\mathbb{E}[Dq^U(W, C) - (1 - D) \mid V, C] = 0,$$

recovering the standard proximal-causal-inference calibration equation.

For Branch G, use $X_t^G = (W, Y_t, C)$ for the observed calibration. The sieve basis augments Branch U with functions of Y_t. Since Y_t is observed only on $\{D = t\}$, implementation uses the latent-to-observed transfer of Lemma 3.1. The primal moment becomes

$$\mathbb{E}[Dq^G(W, Y_1, C) - (1 - D) \mid V, C] = 0.$$

No counterfactual imputation.

Although q^G is written as a function of the potential outcome Y_t, it is evaluated only on the treatment arm where that potential outcome is observed: $q_1^G(W, Y_1, C)$ enters the primal moment multiplied by D, so it is evaluated only on $\{D = 1\}$, where $Y_1 = Y$; symmetrically $q_0^G(W, Y_0, C)$ is evaluated only on $\{D = 0\}$, where $Y_0 = Y$. The implementation therefore does not impute individual counterfactual outcomes.

The three branches can be implemented sequentially using the same Sieve GMM machinery with branch-specific primal and dual basis choices. Comparing the resulting estimates shows sensitivity to alternative maintained information structures.

Branch-specific systems, not nested estimators.

The three empirical branches should not be read as a sequence of nested estimators for a single population moment problem. They correspond to branch-specific operator pairs. Branch A is the reduced observed-adjustment system $X_t^A = Z_t^A = C_A$. Branches U and G use the adjoint-side conditioning system $Z_t = (V, C)$, with different primal domains $X_t^U = (W, C)$ and $X_t^G = (W, Y_t, C)$. Thus branch differences are informative about sensitivity to the maintained information structure—for example, divergence between Branches A and U is consistent with latent-type confounding, and divergence between Branches U and G is consistent with selection on gains—but they are not formal tests of which branch is true: branch differences may also reflect weak measurements, regularization bias, or sieve misspecification.

5 Simulation evidence: regular, weak, and failed regimes

The simulations are not intended to show that flexible nuisance estimation repairs failed range conditions. They illustrate the three regimes implied by the identification theory: regular identification, weak or irregular behavior, and non-invariance under a wrong reduced calibration. This section first reports one continuous, nonseparable design to verify that the exact Sieve GMM / DML implementation can recover the target when the data generating process is constructed to be compatible with the latent odds anchor and the observed adjoint range condition of the frame-

work, and then summarizes the failure and weak-range regimes in a compact panel. The exercise is a *finite-sample implementation check*, not a proof of any specific $o(N^{-1/2})$ regularization-bias rate; the full simulation evidence—including population oracles, scalar-index failures, weak-range sensitivity, and PCI comparisons—is reported in Appendix H. The learner used to construct the sieve features is not the object of comparison; the objects of comparison are the population range regimes and the finite-dimensional moment diagnostics they generate.

The design uses continuous, non-Gaussian, nonseparable potential outcomes and a logistic selection mechanism in which D depends directly on Y_1 and Y_0. The estimator is the cross-fitted DML implementation of the auxiliary-measurement framework with random-forest first stages for the primal selection-odds representer q and the auxiliary outcome regression m, and a regularized linear residual correction h. The target $\mu_1 = \mathbb{E}[Y_1] = 0.5651650429$ is computed analytically. The estimator is constructed to be compatible with the (R1-A) latent existence and (R2) adjoint range primitive conditions of the framework.

What the simulation does and does not show.

The purpose of the simulation is not to claim that arbitrary flexible machine-learning features automatically remove all bias. The design is constructed so that the latent odds anchor and the observed adjoint range condition are compatible: the treated potential outcome Y_1 is a deterministic function of the latent type U and observed environments (S, Q), so that the V-side deconvolution channel can in principle support a finite-norm adjoint representer b. In variants with an idiosyncratic outcome shock that is not measured by V, the same implementation converges to a biased pseudo-target because the observed adjoint range condition is violated at the population level. The simulation is therefore an implementation check of the primal–dual theory under a compatible data generating process, not a generic claim that machine-learning sieve features resolve range failure. Full sensitivity sweeps, including designs where the adjoint range condition is violated, are reported in Appendix H.

Table 7: Regular range regime: cross-fitted data-adaptive sieve calibrated orthogonal moment in a continuous nonseparable design. The adaptive sieve is built from random-forest leaf indicators on auxiliary folds; it is not a causal-forest estimator. Analytical $\mu_1 = 0.5651650429$; naive bias and naive RMSE refer to the simple treated-mean comparator. The ratio column reports $\text{RMSE}/\text{RMSE}_{\text{naive}}$.

N	reps	bias	sd	$\sqrt{N}\,\lvert\text{bias}\rvert$	vec. primal	vec. dual	RMSE	ratio
2,000	1000	−0.0126	0.0231	0.564	0.025	0.006	0.0263	0.158
5,000	1000	−0.0013	0.0119	0.095	0.010	0.002	0.0119	0.073
10,000	1000	+0.0000	0.0082	0.003	0.005	0.001	0.0082	0.050

Table 7 reports bias, standard deviation, the rescaled bias $\sqrt{N}\,|\text{bias}|$, and the empirical L^2 norms of the finite-dimensional primal and dual moment residuals, with RMSE and its ratio to the naive comparator as supplementary columns. The bias is small relative to sampling variation: the absolute bias is below the sampling standard deviation in all three sample sizes and falls well below it as N grows. The rescaled bias $\sqrt{N}\,|\text{bias}|$ decreases sharply with sample size (from about 0.56 at $N = 2{,}000$ to near zero at $N = 10{,}000$), consistent with root-N centered inference, and the vector moment residuals contract monotonically with N. The RMSE is 5–16% of the naive comparator over the sample-size range, but this comparison is reported only as a supplementary

summary rather than as the primary evidence. The simulation is consistent with approximately centered behavior of the estimator on this design and supports the recommendation to report vector moment residuals alongside bias and variance as a transparent diagnostic of estimator centering. A full design description, additional designs (finite-support, scalar-index failure, weak-range sensitivity, PCI-type comparators), and robustness sweeps over learner hyperparameters appear in Appendix H.

Failure and weak-range regimes.

The compatible design above shows Case R behavior. Table 8 summarizes the complementary designs of Appendix H, in which identification fails or weakens in exactly the ways predicted by the trichotomy of Proposition 3.5. Three patterns deserve emphasis. First, under a wrong scalar reduction (`bad_c`), the observed balancing residual can remain small while the estimator concentrates around a pseudo-target: the μ_1 bias stays near 0.10 at all sample sizes and coverage collapses to zero as confidence intervals shrink around the wrong value. Second, a PCI-style score that omits Y_t from the calibration weight (`pci_failure`) carries systematic bias of about 0.09 with coverage approaching zero, and the dual moment-residual diagnostic correctly localizes the breakdown to the adjoint side. Third, as the measurement and adjoint-side loadings weaken ($\rho_W = \rho_V$ moving from 0.9 to 0.22), the smallest empirical singular values shrink, the condition number grows from about 25 to 150, and coverage degrades from 0.95 to about 0.76. Because the `bad_c` and no-Y_t biases are already present in the population oracle (Table 15), they reflect identification failure under a wrong reduced calibration rather than finite-sample estimation error. Data-adaptive sieve features improve approximation in the regular regime but do not repair a failed range condition: in the failure designs the same implementation converges to a pseudo-target. The regimes summarized in Table 8 thus correspond directly to regular orthogonal-moment representation (Case R), non-invariance under wrong reduced calibrations (Case N), and weak or near-irregular behavior (near Case I). Table 23 reports a continuous-DGP weak-range diagnostic in which the measurement-side and adjoint-side loadings are weakened jointly. Appendix H.5 complements this exercise with a finite-support exact-oracle analogue that separates measurement-side from adjoint-side weakness while holding the other side strong. Appendices H.6 and H.7 provide additional finite-support exact-oracle analogues for the Branch A/U/G validity ladder and for adjoint range failure under an unmeasured outcome shock. Appendix H.8 then provides a separate HRS-calibrated weak-range stress test for the empirical conditioning of the HRS application.

Table 8: Range-regime summary: population oracle, finite-sample behavior, and diagnostics (Appendix H). Oracle biases are computed on the population distribution by exact enumeration (Table 15); the remaining quantities are numerical diagnostics of the finite-dimensional observed moments.

Regime	What holds or fails	Population oracle	Finite-sample outcome	Diagnostic
`good_c` (Case R)	(R1-A) and (R2) hold	Oracle bias 0	Centered; coverage ≈ 0.95	Small primal and dual residuals
`bad_c` scalar (Case N)	Wrong scalar reduction destroys the adjoint range	Oracle μ_1 bias 0.100; dual residual 0.045	Pseudo-target; μ_1 bias ≈ 0.10; coverage $\to 0$	Large dual residual; exceedance rate 1.00
`pci_failure` no-Y_t (Case N)	PCI-style weight omits Y_t under selection on gains	Oracle μ_1 bias 0.090; dual residual 0.214	Systematic bias ≈ 0.09; coverage $\to 0$	Dual residual flags failure in all reps
Weak range (near Case I)	Measurement and adjoint loadings weaken ($\rho_W = \rho_V : 0.9 \to 0.22$)	No population failure (oracle bias 0)	RMSE rises; coverage $0.95 \to 0.76$	$\sigma_{\min} \downarrow$, $\kappa : 25 \to 150$

6 Empirical illustration: Retirement and cognition in the HRS

Whether retirement affects later-life cognition has been studied extensively with instrumental variables exploiting retirement-age eligibility rules, which generally report negative effects in the range of 0.2–0.6 standard deviations and are interpreted as local to the policy-induced margin (Rohwedder and Willis, 2010; Coe and Zamarro, 2011; Bonsang et al., 2012; Mazzonna and Peracchi, 2017); the corresponding population effect remains debated. The auxiliary-measurement framework offers an alternative that does not require an excluded, treatment-shifting instrument: the outcome-representation variable V acts as a shadow variable—excluded from the selection and outcome equations given (U, C) but informative about the latent type—and the construction allows treatment selection to depend directly on potential cognitive outcomes. We use the application to illustrate how the branch comparison separates observed adjustment, latent-type measurement adjustment, and selection-on-gains adjustment within one implementation; it is an illustration of the calibration machinery, not a standalone resolution of the retirement–cognition question.

The analysis uses the public-release RAND HRS Longitudinal File 2020 (V2). We take the HRS-original cohort working at wave 3 (1996) with non-missing baseline (wave 3) and outcome (wave 10, 2010) cognition, giving $n = 1{,}597$ individuals, of whom $n_{\text{treated}} = 1{,}140$ (71.4%) are partially or fully retired by wave 8 (2006), which defines the treatment D. The outcome Y is the wave-10 total cognition score (an integer in $[0, 35]$). The measurement block W is baseline cognition and self-rated health; the outcome-representation (shadow) variable V is parental education; the environments S and Q collect baseline retirement-side and continued-work-side variables. The observed-adjustment benchmark (Branch A) conditions on the full block $C_A = (X, S, Q)$ with X the demographic covariates; the bridge specifications use the common Roy-side block $C_R = (S, Q)$ with primal domains $X_t^U = (W, C_R)$ (Branch U) and $X_t^G = (W, Y_t, C_R)$ (Branch G, which adds the potential-outcome coordinate) and adjoint-side system $Z_t = (V, C_R)$. All calibrations use the regularized cross-fit calibrated GMM implementation (linear Sieve GMM, Tikhonov $\lambda = 0.5$, 5-fold). Full variable construction, the notation-to-RAND-code mapping, and the conditioning and weak-range diagnostics are collected in Appendix H.9.

Table 9: Retirement and cognition: standard benchmark estimators

Estimator	$\widehat{\mu}_1$	$\widehat{\mu}_0$	ATE	SE
Naive (mean difference)	22.842	23.195	−0.353	0.25
OLS ($Y \sim X + D$)	–	–	−0.084	0.22
OLS (full controls)	–	–	−0.179	0.22
IPW (normalized)	22.859	22.774	+0.085	0.28
AIPW	22.879	22.805	+0.073	0.24
Branch A (observed-adjustment calibration, $C_A = (X, S, Q)$)	22.854	22.927	−0.073	0.25

Standard estimators without latent-type calibration structure, reported for reference, together with the strongest observed-adjustment calibration (Branch A), which conditions on the full observed block $C_A = (X, S, Q)$. IPW uses normalized (Hájek) weights with the propensity score clipped to $[0.05, 0.95]$. All reported standard errors are pairs-bootstrap (999 reps). These observed-adjustment estimators are descriptive: none registers the negative effect recovered by the bridge specifications in Table 10, and giving Branch A the richest observed conditioning set makes it the most conservative such comparison. Sample and implementation details as in Table 10.

The observed-adjustment estimators and the calibration specifications give markedly different

Table 10: Retirement and cognition: bridge calibration estimates under the common Roy-side block $C_R = (S, Q)$

Maintained structure	$\widehat{\mu}_1$	$\widehat{\mu}_0$	ATE	SE
Branch U (latent-type measurement), $q^U(W, C_R)$	22.610	23.757	−1.147	0.470
Branch G (selection on gains), $q^G(W, Y_t, C_R)$	22.583	24.087	−1.504	0.524
Paired contrast $\widehat{\tau}^G - \widehat{\tau}^U$ (same sample/folds)	–	–	−0.354	–

Sample: $n = 1{,}597$ (HRS-original cohort, working at wave 3, with baseline and wave-10 cognition observed). $D = 1$: partially or fully retired by wave 8 ($n = 1{,}140$, 71.4%). Y = wave 10 total cognition score (range 0–35). Calibration implementation: linear Sieve GMM with Tikhonov regularization $\lambda = 0.5$, estimated by the cross-fit calibrated GMM procedure (5-fold cross-fitting in the sense of Chernozhukov et al. (2018)). Both calibration branches use the common Roy-side block $C_R = (S, Q)$ and the adjoint-side system $Z_t = (V, C_R)$, with branch-specific primal domains $X_t^U = (W, C_R)$ and $X_t^G = (W, Y_t, C_R)$; the observed-adjustment benchmark (Branch A) is reported in Table 9. The SE column is the descriptive standard error from the cross-fitted orthogonal score (0.470 for Branch U, 0.524 for Branch G; the pairs-bootstrap standard error for Branch G is 0.651). No confidence interval is attached to the Branch G estimate; weak-range diagnostics and sensitivity analyses are collected in Appendices H.9, D.2, and D.2. The paired contrast $\widehat{\tau}^G - \widehat{\tau}^U$ is recomputed on each pairs-bootstrap resample sharing the sample, folds, and nuisance pipeline; it is a diagnostic of the incremental movement from adding the potential-outcome coordinate (discussed in the text, where the paired interval is reported), not a formal test that selection on gains is present.

point estimates. The standard benchmarks of Table 9 are near zero or slightly positive (naive −0.353, AIPW +0.073), and even the strongest observed-adjustment calibration—Branch A on the full block $C_A = (X, S, Q)$—gives a near-zero $\widehat{\tau}^A = -0.073$. Moving to the auxiliary-measurement calibration on the common Roy-side block $C_R = (S, Q)$, the latent-type measurement W yields $\widehat{\tau}^U = -1.147$, and additionally allowing the calibration to depend on the potential-outcome coordinate yields $\widehat{\tau}^G = -1.504$ cognition points, about −0.30 of the outcome's sample standard deviation. The Branch A and Branch U/G estimates arise from different maintained information structures: the paired U–G contrast isolates the addition of the potential-outcome coordinate within the common Roy-side block, whereas the A–U/G comparison is broader and also changes the use of the measurement and adjoint-side variables. Within the calibration ladder, adding the potential-outcome coordinate moves the estimate by $\widehat{\tau}^G - \widehat{\tau}^U = -0.354$ (pairs-bootstrap 95% interval $[-0.69, -0.05]$); because the underlying inverse problems are weakly conditioned, this interval is descriptive and is not asserted to have uniform coverage. The magnitude is of the same broad order as the IV-based retirement literature (0.2–0.6 standard deviations), although the estimands differ.

Because no coverage-valid interval accompanies the Branch G point estimate at the empirical (near-singular) conditioning, we report $\widehat{\tau}^G = -1.504$ as a *sensitivity magnitude*—the value attained under the maintained restrictions—rather than as an inferential claim about the population effect; the empirical content is that the estimate moves substantially and systematically across maintained information structures, not a point claim about its magnitude. The cross-branch differences are comparisons across maintained identifying structures and may also reflect weak measurements, sieve approximation, or regularization; they are not tests of which branch is correct. Robustness to the regularization parameter, outcome wave, survival and attrition weighting, fold count, and treatment definition; calibration-weight and adjoint-side (V) diagnostics; sensitivity and weak-range diagnostics are collected in Appendices H.9, D.2, and D.2.

7 Conclusion

This paper develops an auxiliary-measurement identification framework for average treatment effects under selection on potential outcomes. The central object is a primal–dual calibration system: a latent selection-odds representer places a causally correct element in an observed calibration set, while an adjoint outcome representer certifies that the target is invariant over that set. Identification therefore does not require point identification of the calibrating functions themselves. The resulting geometry yields a trichotomy between non-invariance, irregular identification, and regular orthogonal-moment representation.

More generally, the paper characterizes the maximal class of square-integrable arm-specific linear expectation targets that descend to the quotient space induced by the observed calibration operator, with the ATE as the leading special case (Theorem 3.3). The appendix shows that the latent anchor is observationally untestable without further structure and develops exact latent-calibration sensitivity bounds for departures from that anchor.

The framework provides a complementary route to both deterministic Roy selection and instrument-based MTE identification. It allows treatment choices to depend directly on potential outcomes and multidimensional latent heterogeneity while using latent-type and adjoint-side variation to provide a structured route to identifying population-level treatment effects. The same structure generates a calibrated orthogonal moment, an exact product-bias identity, and practical implementations based on Sieve GMM and cross-fitted DML.

The HRS application illustrates the empirical implications of the framework in a setting where selection on latent type and potential cognitive outcomes is substantively plausible. More broadly, the results suggest that auxiliary-measurement primal–dual calibration methods provide a structured way to study policy-relevant treatment effects in environments where treatment choices depend directly on potential outcomes. This structure is most useful when the latent anchor and adjoint range conditions are substantively credible. Developing stronger diagnostics and inference methods for weak-range settings remains an important direction for future research.

Acknowledgments

The Health and Retirement Study (HRS) is sponsored by the National Institute on Aging (grant number U01AG009740) and conducted by the University of Michigan. This paper uses the public-release RAND HRS Longitudinal File 2020 (V2), produced by the RAND Center for the Study of Aging with funding from the National Institute on Aging and the Social Security Administration.

References

Angrist, J. D., G. W. Imbens, and D. B. Rubin (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association 91(434)*, 444–455.

Bennett, A., N. Kallus, X. Mao, W. K. Newey, V. Syrgkanis, and M. Uehara (2025). Inference on strongly identified functionals of weakly identified functions. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, qkaf075.

Bonsang, E., S. Adam, and S. Perelman (2012). Does retirement affect cognitive functioning? *Journal of Health Economics 31(3)*, 490–501.

Brinch, C. N., M. Mogstad, and M. Wiswall (2017). Beyond LATE with a discrete instrument. *Journal of Political Economy 125(4)*, 985–1039.

Carneiro, P., J. J. Heckman, and E. J. Vytlacil (2011). Estimating marginal returns to education. *American Economic Review 101(6)*, 2754–2781.

Chen, X. and D. Pouzo (2012). Estimation of nonparametric conditional moment models with possibly nonsmooth generalized residuals. *Econometrica 80(1)*, 277–321.

Chen, X. and A. Santos (2018). Overidentification in regular models. *Econometrica 86(5)*, 1771–1817.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins (2018). Double/debiased machine learning for treatment and structural parameters. *Econometrics Journal 21(1)*, C1–C68.

Coe, N. B. and G. Zamarro (2011). Retirement effects on health in Europe. *Journal of Health Economics 30(1)*, 77–86.

Cui, Y., H. Pu, X. Shi, W. Miao, and E. Tchetgen Tchetgen (2024). Semiparametric proximal causal inference. *Journal of the American Statistical Association 119(546)*, 1348–1359.

Engl, H. W., M. Hanke, and A. Neubauer (2000). *Regularization of Inverse Problems*, Volume 375 of *Mathematics and its Applications*. Kluwer Academic Publishers.

Fan, J. (1991). On the optimal rates of convergence for nonparametric deconvolution problems. *Annals of Statistics 19(3)*, 1257–1272.

Hall, P. and J. L. Horowitz (2005). Nonparametric methods for inference in the presence of instrumental variables. *Annals of Statistics 33(6)*, 2904–2929.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica 47(1)*, 153–161.

Heckman, J. J. and B. E. Honoré (1990). The empirical content of the Roy model. *Econometrica 58 (5)*, 1121–1149.

Heckman, J. J. and G. Sedlacek (1985). Heterogeneity, aggregation, and market wage functions: An empirical model of self-selection in the labor market. *Journal of Political Economy 93(6)*, 1077–1125.

Heckman, J. J. and G. Sedlacek (1990). Self-selection and the distribution of hourly wages. *Journal of Labor Economics 8(1, Part 2)*, S329–S363.

Heckman, J. J. and E. J. Vytlacil (2005). Structural equations, treatment effects, and econometric policy evaluation. *Econometrica 73(3)*, 669–738.

Henry, M., R. Méango, and I. Mourifié (2024). Role models and revealed gender-specific costs of STEM in an extended Roy model of major choice. *Journal of Econometrics 238(2)*, 105571.

Hoshino, T., K. Shinoda, and T. Otsu (2026). Fragility of joint identification in the Roy model: A note on deterministic sorting and stochastic selection. *Keio-IES Discussion Paper Series* DP2026-009, Institute for Economics Studies, Keio University.

Hu, Y. and J.-L. Shiu (2018). Nonparametric identification using instrumental variables: sufficient conditions for completeness. *Econometric Theory 34(3)*, 659–693.

Imbens, G. W. and J. D. Angrist (1994). Identification and estimation of local average treatment effects. *Econometrica 62(2)*, 467–475.

Kline, P. and C. R. Walters (2019). On Heckits, LATE, and numerical equivalence. *Econometrica 87(2)*, 677–696.

Lee, J. H. and B. G. Park (2023). Nonparametric identification and estimation of the extended Roy model. *Journal of Econometrics 235(2)*, 1087–1113.

Lehmann, E. L. (1986). *Testing Statistical Hypotheses* (2nd ed.). New York: Wiley.

Manski, C. F. (1990). Nonparametric bounds on treatment effects. *American Economic Review 80 (2)*, 319–323.

Mazzonna, F. and F. Peracchi (2017). Unhealthy retirement? *Journal of Human Resources 52 (1)*, 128–151.

Melitz, M. J. (2003). The impact of trade on intra-industry reallocations and aggregate industry productivity. *Econometrica 71(6)*, 1695–1725.

Miao, W., Z. Geng, and E. J. Tchetgen Tchetgen (2018). Identifying causal effects with proxy variables of an unmeasured confounder. *Biometrika 105(4)*, 987–993.

Mogstad, M., A. Santos, and A. Torgovitsky (2018). Using instrumental variables for inference about policy relevant treatment parameters. *Econometrica 86(5)*, 1589–1619.

Mourifié, I., M. Henry, and R. Méango (2020). Sharp bounds and testability of a Roy model of STEM major choices. *Journal of Political Economy 128(8)*, 3220–3283.

Newey, W. K. and J. L. Powell (2003). Instrumental variable estimation of nonparametric models. *Econometrica 71(5)*, 1565–1578.

Rohwedder, S. and R. J. Willis (2010). Mental retirement. *Journal of Economic Perspectives 24 (1)*, 119–138.

Roy, A. D. (1951). Some thoughts on the distribution of earnings. *Oxford Economic Papers 3(2)*, 135–146.

Tan, Z. (2006). A distributional approach for causal inference using propensity scores. *Journal of the American Statistical Association 101(476)*, 1619–1637.

Tchetgen Tchetgen, E., A. Ying, Y. Cui, X. Shi, and W. Miao (2024). An introduction to proximal causal inference. *Statistical Science 39(3)*, 375–390.

Vytlacil, E. (2002). Independence, monotonicity, and latent index models: An equivalence result. *Econometrica 70(1)*, 331–341.

Willis, R. J. and S. Rosen (1979). Education and self-selection. *Journal of Political Economy 87 (5, Part 2)*, S7–S36.

Appendix

A Appendix overview

The appendix is organized into seven functional blocks. Each block has a single role; the three conceptual separations of the paper (the latent versus observed range conditions; the primal versus dual operators; the regular versus irregular regime) are maintained throughout.

Appendix B: Conceptual clarifications and literature positioning.

Background material removed from the body to keep the main exposition compact, including related literature, DAG and conditional-independence discussion, branch structure, and economic examples and empirical caveats.

Appendix C: Supplementary identification remarks and derivations.

Derivations and remarks clarifying the identifying restrictions: the latent anchor and observed calibration transfer, primal nonuniqueness and target invariance, calibration weights and branch-specific systems, residualization and reduced conditioning blocks, and the extended roadmap of the identification argument.

Appendix D: Operators, primitive conditions, and source conditions.

Definitions of the three operators used throughout the paper (T_W, A_1, $A_1^\star$), primitive sufficient conditions for the latent existence condition (R1-A), and singular-value systems, source conditions, and Picard summabilities governing the observed adjoint range condition (R2) and the rates of regularized inverse estimators.

Appendix E: Proofs of identification results.

Proofs of the latent-to-observed transfer (Lemma 3.1), the primal–dual identification theorem (Theorem 3.4), the invariance and closure characterizations, the trichotomy, and the product-bias and Neyman-orthogonality identities.

Appendix F: Sieve GMM and cross-fitted DML estimation.

Implementation theory. The exact Sieve GMM estimator, the cross-fitted DML algorithm, asymptotic normality and product-rate conditions, and the extended Sieve GMM for the minimum-variance representative.

Appendix G: Efficient influence functions and extensions.

The distinction between the calibrated orthogonal moment class and the regular influence function; joint EIF characterizations; potential-outcome-type formulations; why measurement-side and adjoint-side variation are generically needed; scalar-index failure; the extension with observed covariates; the two-sample EIF for one-sided noncompliance; and the relation to debiased inverse-problem inference.

Appendix H: Simulation and empirical supplement.

Monte Carlo simulation designs and main simulation results, specification and weak-range diagnostics, data-adaptive Sieve GMM robustness exercises, and the supplementary information for the HRS application (variable construction, additional diagnostics, visualization of the estimated selection structure, and mortality attrition).

B Conceptual clarifications and literature positioning

B.1 Relation to existing literatures

This paper relates to four literatures: Roy selection and generalized Roy models; instrumental variables, LATE, and MTE; proxy and bridge approaches to unobserved confounding; and semiparametric inference for inverse problems. The literature is reviewed in this order, after which the paper's position relative to each is summarized.

Roy and generalized Roy models. The central insight of the self-selection literature initiated by Roy (1951) is that observed treatment choices reflect a comparative advantage in latent returns. Willis and Rosen (1979) and Heckman and Sedlacek (1985, 1990) extended this comparative-advantage logic to schooling choices and labor-market self-selection, treating the choice as a deterministic consequence of latent-return comparison rather than as a noise-driven reduced-form assignment. Heckman and Honoré (1990) is the decisive contribution in this lineage. They showed that the log-normal Roy model has surprisingly strong empirical content from a single cross-section of wage data: when log-skills are normally distributed, the latent skill distribution is identified without conventional exclusion restrictions or additional regressors. In the general non-normal case, however, this strong identification result is lost, and multimarket price variation or regressor variation becomes necessary. The core of the Roy literature is therefore not merely the expression of comparative advantage but rather an identification question: *how much structure must be placed on the selection rule for the latent outcome distribution to be recovered?*

Over the past decade, the Roy framework has been revisited theoretically. Mourifié et al. (2020) reduce the Roy model to its minimal skeleton — sector-specific unobserved heterogeneity and potential-outcome–based self-selection — and obtain sharp bounds and testability under stochastically monotone instrumental variables. This recasts the empirical content of the classical Roy model in terms of sharp partial identification (Manski, 1990) and specification testing rather than point identification. Recent work also expands the Roy framework toward extended and generalized Roy models that explicitly accommodate utility-based selection. Lee and Park (2023) develop nonparametric identification of the extended Roy model under monotonicity restrictions and local instrumental variables. Henry et al. (2024) identify gender-specific non-pecuniary costs of STEM choice through an extended Roy model with role-model effects. The contemporary Roy literature thus has moved away from purely deterministic sorting while still imposing some form of single-index monotonicity or stochastic-dominance structure on the choice equation. *Settings in which the selection mechanism simultaneously involves multidimensional unobserved confounding and idiosyncratic preference shocks have remained theoretically less explored.*

The present paper is a natural benchmark relative to this literature, but it departs from the Roy identification source. In Roy-type frameworks, selection heterogeneity is effectively constrained by $\text{sign}(Y_1 - Y_0)$ or by a monotone index of comparative advantage. The present paper instead allows the influence of latent gains on selection to coexist with an independent direct effect of

unobserved confounders U and with idiosyncratic preference shocks η_D. In this sense, the present paper relativizes the Roy model as a strict benchmark and substantively relaxes its deterministic selection rule.

Deterministic Roy measurement and fragility.

Hoshino et al. (2026) isolate the measurement content of the deterministic Roy rule. In their analysis, the rule $D = \mathbf{1}\{Y_1 > Y_0\}$ does more than impose a choice model: it reveals the side of a moving boundary in the latent (Y_1, Y_0) plane. Once this known-boundary measurement is replaced by an unknown stochastic selection kernel, observationally equivalent latent joint laws open up at first order, with total-variation diameter linear in the relaxation size. Their object is the fragility of joint latent-law identification under stochastic selection. The present paper studies a different positive question. It does not attempt to recover the full joint law of (Y_1, Y_0) from deterministic sorting. Instead, it identifies population mean functionals through measurement-side and adjoint-side calibration restrictions in a generalized Roy environment where selection may be stochastic and potential-outcome dependent. The two results are therefore complementary: the fragility note shows the cost of relaxing deterministic Roy sorting if one insists on joint-law identification, while the present paper identifies a weaker but population-relevant target under that same relaxation.

Instrumental variables, LATE, and MTE. Rather than embedding comparative-advantage determinism into the selection rule, the instrumental-variable literature focuses on how exogenous variation moves the participation margin. Imbens and Angrist (1994) and Angrist et al. (1996) formalize the local average treatment effect (LATE) under monotonicity, showing that IV typically identifies the local average effect among "compliers" rather than the ATE. Vytlacil (2002) establishes the equivalence between LATE and a threshold-crossing latent-index model. Heckman and Vytlacil (2005) unify the IV literature with structural selection through the marginal treatment effect (MTE), under which ATE, TT, TUT, PRTE, and the IV/OLS probability limits all arise as weighted aggregates of a common underlying MTE function. The framework places "which parameter is being estimated" at the center of inference and connects target-parameter choice to instrument choice in a structured way. Carneiro et al. (2011), Brinch et al. (2017), Mogstad et al. (2018), and Kline and Walters (2019) develop this approach in applications and theory. The present paper is complementary. MTE analysis asks how far an identified marginal response can be extrapolated to a policy target; the present paper asks what additional measurement-side and adjoint-side information can pin down the population ATE when treatment choice itself depends on potential outcomes. Corollary B.1 formalizes this relationship using a set-mapping argument: let $\mathcal{M}_{\mathrm{MTE}}(P)$ denote the set of structural models compatible with the MTE-style identifying restrictions at the observed distribution P, and let $\mathcal{M}_{\mathrm{CAL}}(P) \subseteq \mathcal{M}_{\mathrm{MTE}}(P)$ denote the subset additionally satisfying the calibration restrictions (R1-A) and (R2). The calibration-implied identified set is $\Theta_{\mathrm{CAL}}(P) = \{\tau(M) : M \in \mathcal{M}_{\mathrm{CAL}}(P)\} \subseteq \Theta_{\mathrm{MTE}}(P)$, and under the conditions of Theorem 3.4 the calibration-implied identified set is a singleton $\Theta_{\mathrm{CAL}}(P) = \{\tau^\star(P)\}$. This is a refinement of the MTE-based sharp identified set when the range conditions hold.

Limitations of instrument-based identification. The instrument-based identification strategy treats the instrument-induced margin as primitive. When (i) credible instruments satisfying the exclusion restriction are difficult to find, (ii) the selection rule depends on multidimensional unobserved heterogeneity rather than a one-dimensional resistance, or (iii) the relevant population-level target is not the LATE-style "compliers" subpopulation, the IV/LATE/MTE ap-

proach has limited reach. The present paper offers a complementary route: when an instrument-like variable is not available, but measurements of latent heterogeneity and adjoint-side variables are available, calibration-based identification may still pin down the ATE.

MNAR, proxy variables, and proximal causal inference. Heckman (1979) is the classical reference for selection-on-unobservables, treating it as a parametric missing-not-at-random (MNAR) problem under joint normality. The proxy-variable literature extends the idea by treating unobserved confounders through observable proxies. Miao et al. (2018), Tchetgen Tchetgen et al. (2024), and Cui et al. (2024) develop proxy-based identification and semiparametric inference under the conditional treatment ignorability $(Y_1, Y_0) \perp\!\!\!\perp D \mid U$, i.e., assuming that the latent confounder fully accounts for the dependence between treatment and potential outcomes. The present paper relaxes this assumption: treatment choice may directly depend on potential outcomes even after conditioning on latent heterogeneity. Standard proximal causal scores therefore do not directly apply. The latent selection-odds representer introduced here reproduces a latent treatment odds that conditions on the potential outcome, and the adjoint outcome representer certifies target invariance over a possibly nonunique calibration set.

Inverse problems and debiased inference. Finally, the inferential tools relate to the DML and conditional-moment inverse-problem literatures. Chernozhukov et al. (2018) introduced cross-fitted DML based on Neyman orthogonality, providing a general framework for debiased semiparametric inference. Newey and Powell (2003), Hall and Horowitz (2005), and Chen and Pouzo (2012) study NPIV and sieve minimum distance under conditional moment restrictions. Bennett et al. (2025) provide a general debiased inference framework for linear functionals of conditional-moment problems, generalizing Riesz-representation approaches to the inverse-problem setting. The present paper uses these tools but contributes (i) the selection-on-gains–specific primal–dual calibration framework, (ii) the corresponding primal-dual invariance geometry, and (iii) the regular calibrated orthogonal moment whose form is canonical given the primal–dual calibration structure rather than externally imposed. A detailed comparison with the framework of Bennett et al. (2025) is given in Appendix G.3.

Detailed comparisons with related frameworks.

The following three subsubsections situate the auxiliary-measurement framework relative to standard proximal causal inference, debiased inverse-problem inference, and marginal treatment effect bounds in detail; readers seeking only the high-level positioning may skip them.

Relation to standard proximal causal inference.

Standard proximal causal inference (Miao et al., 2018; Tchetgen Tchetgen et al., 2024; Cui et al., 2024) assumes that, conditional on the latent confounder U, treatment assignment is independent of potential outcomes:

$$(Y_1, Y_0) \perp\!\!\!\perp D \mid U.$$

Under this assumption, the latent treatment-odds ratio $\mathbb{P}(D = 0 \mid U)/\mathbb{P}(D = 1 \mid U)$ depends only on U, and the corresponding calibration equation can be solved with an observed calibration $q^U(W, C)$ that conditions only on (W, C). In contrast, the framework of this paper allows treatment to depend directly on potential outcomes:

$$\mathbb{P}(D = 1 \mid U, Y_1, Y_0, C) \neq \mathbb{P}(D = 1 \mid U, C) \quad \text{in general.}$$

The relevant latent odds ratio is then a function of (U, Y_t, C), and the corresponding calibration equation has the primal argument $q^G(W, Y_t, C)$ that includes the potential outcome. The adjoint range condition correspondingly applies to a target that, on the treated support, equals Y_1; the residualized formulation of Section 3.2 ensures that this is operationalizable.

The two settings are therefore non-nested at the structural level: neither selection-on-gains identification nor standard proximal causal inference is a special case of the other. They coincide only when treatment is conditionally independent of potential outcomes given the latent type, which is the standard-PCI assumption. The two frameworks therefore rely on different calibration equations, different adjoint range conditions, and different identifying restrictions.

Relation to debiased inverse-problem inference.

Remark (Position of this paper relative to Bennett et al. (2025)). Bennett et al. (2025) provide a general debiased-inference theory for linear functionals of conditional-moment inverse problems. The present paper is consistent with that general theory but is not a mere special case. First, because $(Y_1, Y_0) \not\perp\!\!\!\perp D \mid U$ under selection on gains, a causal anchor is required to ensure that a causally correct latent selection-odds representer lies in the observed calibration set $\mathcal{Q}_1$. Second, the adjoint outcome representer b is not merely a generic debiasing nuisance but a target-invariance certificate that makes the ATE functional invariant over $\mathcal{Q}_1$. Third, the standard PCI treatment bridge corrects U-conditionally independent treatment assignment, whereas the selection-odds representer in this paper reproduces a latent treatment odds that conditions on both U and the potential outcomes. A detailed comparison and an analytic example of the breakdown of the standard PCI score are given in Appendix G.3 and Appendix E.4.

Remark (Interpretation of the Bennett et al. (2025) framework comparison). The general theory of Bennett et al. (2025) provides a unified treatment of debiased inference for linear functionals of conditional-moment problems. The selection-on-gains structure considered here corresponds to a specific subclass: the primal moment is constructed from a latent treatment-odds ratio, and the adjoint outcome representer plays the role of a target-invariance certificate rather than a nuisance correction. Under this view, the present paper provides (i) the substantive identification anchor (the latent selection-odds representer) that pins down a particular target functional, and (ii) the primal–dual invariance geometry that explains why orthogonal scoring works precisely under selection on gains. The general inverse-problem debiasing of Bennett et al. (2025) can then be applied as the inferential layer.

Remark (Position of the inferential contribution). The inferential contribution of this paper is not a new debiasing technique. The orthogonal score, the product-bias identity, and the cross-fitted DML implementation all draw on the existing literature. The contribution is to show that under selection on gains, the appropriate score arises endogenously from the primal–dual calibration structure rather than being imposed externally. This makes the inferential framework specific to the generalized Roy environment: it gives a principled answer to the question of which orthogonal score one should use, and why the corresponding influence function admits a regular representation when the adjoint range condition holds.

Relation to MTE bounds.

Corollary B.1 (Sharp tightening relative to MTE bounds)**.** *Let $\Theta_{\mathrm{MTE}}(P_O)$ denote the sharp identified set for a target under a model that uses only MTE/IV information, and let $\mathfrak{M}_{\mathrm{CAL}}(P_O)$ be the subset of observationally equivalent structures satisfying the calibration restrictions. If the*

primal-dual conditions of Corollary 3.6 hold on $\mathfrak{M}_{\mathrm{CAL}}(P_O)$, then the calibration-implied identified set is the singleton $\{\tau_{\mathrm{CAL}}\} \subseteq \Theta_{\mathrm{MTE}}(P_O)$. Thus the auxiliary-measurement calibration restrictions act as a sharp tightening of the MTE bounds rather than as a contradiction of the partial-identification framework.

Proof. Let $\mathfrak{M}_{\mathrm{MTE}}(P_O)$ denote the set of observationally equivalent structures satisfying the MTE/IV restrictions, and let $\mathfrak{M}_{\mathrm{CAL}}(P_O) \subseteq \mathfrak{M}_{\mathrm{MTE}}(P_O)$ denote the subset that additionally satisfies the calibration restrictions (R1-A) and (R2). The MTE-implied identified set is

$$\Theta_{\mathrm{MTE}}(P_O) = \{\tau(M) : M \in \mathfrak{M}_{\mathrm{MTE}}(P_O)\}.$$

For each $M \in \mathfrak{M}_{\mathrm{CAL}}(P_O)$, the primal-dual identification theorem (Theorem 3.4, applied with M as the data-generating structure) gives the same value of the ATE, namely $\tau(M) = \tau_{\mathrm{CAL}}(P_O)$ for the calibration-implied target. Therefore the calibration-implied identified set is

$$\Theta_{\mathrm{CAL}}(P_O) = \{\tau(M) : M \in \mathfrak{M}_{\mathrm{CAL}}(P_O)\} = \{\tau_{\mathrm{CAL}}(P_O)\},$$

a singleton. By construction $\mathfrak{M}_{\mathrm{CAL}} \subseteq \mathfrak{M}_{\mathrm{MTE}}$, so $\Theta_{\mathrm{CAL}}(P_O) \subseteq \Theta_{\mathrm{MTE}}(P_O)$. Hence the auxiliary-measurement calibration restrictions sharply tighten the MTE bounds to a point. □

A specification test from MTE containment.

Corollary B.1 is one-directional: under the calibration restrictions (R1-A) and (R2) the calibration target lies in the MTE/IV identified set, $\tau_{\mathrm{CAL}}(P_O) \in \Theta_{\mathrm{MTE}}(P_O)$. The contrapositive yields empirical discipline on the maintained restrictions even though (R1-A) is itself untestable.

Proposition B.2 (MTE-containment specification test)**.** *Suppose $\Theta_{\mathrm{MTE}}(P_O) = [\underline{\theta}(P_O), \overline{\theta}(P_O)]$ is the sharp MTE/IV identified set under the maintained instrument exclusion, and let $[\widehat{\underline{\theta}}, \widehat{\overline{\theta}}]$ be a consistent estimator of an outer set for it, in the sense that $\underline{\theta}(P_O) \geq \widehat{\underline{\theta}} - o_p(1)$ and $\overline{\theta}(P_O) \leq \widehat{\overline{\theta}} + o_p(1)$. Let $\widehat{\tau}_{\mathrm{CAL}} \to_p \tau_{\mathrm{CAL}}(P_O)$ be the cross-fit calibrated GMM estimator. Then under the null that (R1-A) and (R2) hold,*

$$\mathbb{P}\big(\widehat{\tau}_{\mathrm{CAL}} \in [\widehat{\underline{\theta}} - \epsilon_N,\ \widehat{\overline{\theta}} + \epsilon_N]\big) \to 1$$

for any $\epsilon_N \to 0$ dominating the estimation error, whereas a stable violation $\tau_{\mathrm{CAL}}(P_O) \notin \Theta_{\mathrm{MTE}}(P_O)$ implies rejection with probability approaching one. Equivalently, a test that rejects when $\widehat{\tau}_{\mathrm{CAL}}$ falls outside the estimated MTE band is a consistent specification test of the calibration restrictions: rejection refutes (R1-A)–(R2) jointly (or the maintained MTE exclusion), since containment is a necessary implication of calibration.

The test is one-sided in content: non-rejection does not validate (R1-A), because containment is necessary but not sufficient; rejection is informative because it can occur only if some maintained restriction fails. An outer set $[\widehat{\underline{\theta}}, \widehat{\overline{\theta}}]$ can be constructed from MTE/IV sharp bounds (Mogstad et al., 2018; Manski, 1990) when a (possibly weak) instrument or a monotonicity restriction is available; in the HRS application this provides an external check on the Branch G calibration against the IV-based retirement literature, complementing the within-framework diagnostics and converting the otherwise untestable point-identification claim into a falsifiable implication.

B.2 Structural interpretation and branch systems

This subsection collects extended discussions of the causal structure underlying the auxiliary-measurement framework: the structural DAG and the role of optional arrows, the calibration-implication form of conditional independence, the structure of the three branches, and bilateral-independence conditions for reduced DAG interpretations.

Additional discussion of the structural DAG.

Figure 1(a) summarizes the general full-calibration structure underlying the identification analysis. Relative to standard proxy-control formulations, the framework allows treatment selection to depend directly on (Y_1, Y_0, U, C), thereby accommodating selection on gains and other non-standard selection mechanisms.

The maintained exclusions are the absence of direct effects $W \to D$, $V \to D$, $V \to W$, and direct effects $V \to Y_t$ conditional on (U, C). The latent-type measurement W is therefore linked to treatment only through latent heterogeneity and the conditioning block. Likewise, V acts as a latent-heterogeneity shifter rather than a direct treatment or outcome shifter.

Bidirectional arrows in Figure 1(a) denote unrestricted dependence paths whose orientation is not identified and not required for the calibration arguments. In particular, the identification analysis does not require a structural ordering among the latent type, the latent-type measurements, and the conditioning variables beyond the maintained exclusions above.

Table 11: Interpretation of key paths in the structural DAG.

Path	Status	Interpretation
$V \to U$	allowed	V shifts latent heterogeneity.
$U \to W$	allowed	W is a measurement for the latent type.
$U \to Y_1, Y_0$	allowed	Latent heterogeneity may affect potential outcomes.
$U \to D$	allowed	Latent heterogeneity may directly affect treatment selection.
$C \to Y_1, Y_0$	allowed	Conditioning variables may affect potential outcomes.
$C \to D$	allowed	Conditioning variables may affect treatment selection.
$Y_1, Y_0 \to D$	allowed	Selection-on-gains channel.
$U \leftrightarrow S, Q$	unrestricted	Latent heterogeneity and conditioning variables may co-vary.
$W \leftrightarrow S, Q$	unrestricted	Measurement variables and conditioning variables may co-vary.
$S \leftrightarrow Q$	unrestricted	Treated- and untreated-side environments may co-vary.
$W \to D$	excluded	Measurement variables do not directly determine treatment.
$V \to D$	excluded	V is not a direct treatment shifter.
$V \to W$	excluded	V and W interact only through (U, C).
$V \to Y_t$	excluded	V affects outcomes only through latent heterogeneity.

Conditional independence and reduced formulations.

The identification analysis relies on the calibration implication in Assumption 2.2 together with the range conditions. Conditional-independence restrictions provide sufficient structural conditions under which these calibration implications arise.

For example, stronger restrictions such as

$$W \perp\!\!\!\perp (V, S, Q, Y_1, Y_0, D) \mid U, \qquad V \perp\!\!\!\perp (W, S, Q, Y_1, Y_0, D) \mid U$$

imply the calibration implications through Proposition 2.1. Likewise,

$$(Y_1, S) \perp\!\!\!\perp (Y_0, Q) \mid U$$

may justify reduced formulations that omit Q or S from the calibration arguments. These additional restrictions are not required for the full calibration formulation with

$$X_1 = (W, Y_1, C), \qquad Z_1 = (V, C).$$

Additional discussion of the branch structure.

The three branches are implemented through branch-specific operator pairs. In the canonical theoretical comparison with $C_A = C$, the primal information sets are nested, but the full primal–dual systems are not literal nested restrictions of a single operator equation. Branch A uses the reduced observed-adjustment system $X_t^A = Z_t^A = C_A$, whereas Branches U and G use the adjoint-side conditioning system $Z_t = (V, C)$. On the primal side, Branch A uses

$$X_t^A = C,$$

Branch U uses

$$X_t^U = (W, C),$$

and Branch G uses

$$X_t^G = (W, Y_t, C).$$

Although these information sets are nested, the corresponding dual systems are branch-specific because both the adjoint operators and target signals differ across branches.

In particular, the adjoint range conditions for Branches A, U, and G are not automatically nested. Adding W or Y_t changes both the operator and the target functional, so the relevant adjoint ranges live in different Hilbert spaces. The branch comparison discussed in the main text should therefore be interpreted as a diagnostic decomposition across information sets rather than as a formal specification test of selection on gains.

Each branch also admits a branch-specific minimum-norm calibration representation. For branch $\mathrm{b} \in \{A, U, G\}$, let $\mathcal{Q}_1^{\mathrm{b}}$ denote the corresponding calibration solution set. The canonical representative is defined by

$$q^{\mathrm{b},\dagger} = \arg\min_{q \in \mathcal{Q}_1^{\mathrm{b}}} \mathbb{E}[Dq(X_1^{\mathrm{b}})^2],$$

which corresponds to the regularized Tikhonov solution used in the implementation.

Finally, although the Branch G calibration depends on the potential outcome Y_t, the calibrating

function is evaluated only on the support where that outcome is observed. For example, on the treated side,

$$\mathbb{E}[Dq^G(W, Y_1, C) - (1 - D) \mid Z_1] = 0$$

is evaluated only where $Y_1 = Y$ on $\{D = 1\}$. The framework therefore does not require imputing counterfactual outcomes.

Bilateral independence and reduced DAG interpretations.

The reduced formulations of Corollary 2.2 become particularly natural under the conditional independence condition

$$(Y_1, S) \perp\!\!\!\perp (Y_0, Q) \mid U.$$

Under this condition, the treated-side variables (Y_1, S) and the untreated-side variables (Y_0, Q) are conditionally separated given the latent heterogeneity. The one-sided calibration formulations may then be interpreted as reduced DAG representations in which the calibration for μ_1 depends only on (Y_1, S) and the calibration for μ_0 depends only on (Y_0, Q).

This bilateral independence condition is not required for the full calibration identification result. Without it, identification still proceeds under the full conditioning block $C = (S, Q)$, while the reduced formulations require their own calibration implications and range conditions.

B.3 Diagnostics, residualization, and what is not tested

This subsection collects supplementary material on the scope and limits of calibration-based identification: equivalent formulations of the adjoint range condition and finite-sample diagnostics, the residualization and auxiliary projection devices, an extension to distributional and quantile functionals, and the diagnostic-OLS implementation device. Each *paragraph* below records the corresponding earlier subsection.

Identification: adjoint range conditions and diagnostics.

This subsection collects supplementary identification material: equivalent formulations of the adjoint range condition, the distinction between latent and observed range conditions, when reduced range conditions are automatic, primitive sufficient conditions, and finite-sample diagnostics for the identification trichotomy.

Equivalent formulations of the adjoint range condition.

The adjoint range condition can be expressed either directly as

$$\ell_1 \in \mathcal{R}(A_1^\star),$$

or through the residual form

$$\rho_{1,m}(x) := \mathbb{E}[Y_1 - m(Z_1) \mid X_1 = x, D = 1] \in \mathcal{R}(A_1^\star),$$

for an arbitrary fixed function $m \in \mathcal{H}_{Z_1}$.

The two formulations are equivalent. Since

$$\rho_{1,m} = \mathbb{E}[Y_1 \mid X_1, D = 1] - \mathbb{E}[m(Z_1) \mid X_1, D = 1] = \ell_1 - A_1^\star m,$$

we have

$$\rho_{1,m} \in \mathcal{R}(A_1^\star) \iff \ell_1 - A_1^\star m \in \mathcal{R}(A_1^\star).$$

Because $\mathcal{R}(A_1^\star)$ is a linear subspace,

$$\ell_1 - A_1^\star m \in \mathcal{R}(A_1^\star) \iff \ell_1 \in \mathcal{R}(A_1^\star).$$

Hence the residual representation does not depend on the choice of m and is simply an alternative parametrization of the same adjoint range condition.

Latent versus observed range conditions.

The primal and adjoint range conditions differ fundamentally in their observability.

The latent representer condition involves the unobserved heterogeneity U and therefore cannot generally be verified from the observed distribution alone. This limitation is intrinsic to proxy-based identification frameworks. Appendix D.2 and Appendix D.4 provide primitive sufficient conditions under several structures, including finite-support, Gaussian, and deconvolution settings.

By contrast, the adjoint range condition is formulated entirely through the observed conditional distribution on the treated sample. In particular, the equation

$$A_1^\star b = \ell_1$$

depends only on the observed conditional distribution of (X_1, Z_1, Y) on $\{D = 1\}$. Appendix D.3 provides the corresponding observed-side operator representation.

The distinction between latent and observed range conditions is important conceptually. The observed calibration condition supplies the causal interpretation through the latent selection structure, whereas the adjoint range condition governs invariance of the target functional over the observed calibration set.

Reduced range conditions are not automatic.

The reduced formulations of Corollary 2.2 require their own calibration implications and range conditions and therefore do not follow automatically from the full formulation. In particular, replacing the full conditioning block $C = (S, Q)$ by a lower-dimensional summary $C_1 = \kappa(S, Q)$ changes the associated operators and their ranges.

For the Y_1 side, define the reduced variables

$$\bar{X}_1 = (W, Y_1, C_1), \qquad \bar{Z}_1 = (V, C_1),$$

and the reduced operator

$$(\bar{A}_1 g)(z) := \mathbb{E}[Dg(\bar{X}_1) \mid \bar{Z}_1 = z].$$

The reduced adjoint range condition requires

$$\bar{\ell}_1 \in \mathcal{R}(\bar{A}_1^\star), \qquad \bar{\ell}_1(x) := \mathbb{E}[Y \mid \bar{X}_1 = x, D = 1].$$

This condition need not follow from the full-range condition $\ell_1 \in \mathcal{R}(A_1^\star)$.

The reason is that the reduction map $\kappa(S, Q)$ may collapse directions of variation that are important either for the selection mechanism or for the treated potential outcome. Even if the full conditioning block (S, Q) yields a valid calibration representation, the reduced operator $\bar{A}_1$ may fail to retain enough variation for the corresponding adjoint representer to exist.

A simple illustration is obtained in finite-state models. In one data generating process, the conditional expectation of Y_1 and the selection probability depend on (S, Q) only through the scalar index $S+Q$, so the reduction $C_1 = S+Q$ preserves the adjoint range condition. In another data generating process, the treated outcome depends on (S, Q) through directions orthogonal to $S + Q$. In that case, collapsing (S, Q) to the scalar index destroys the reduced range condition even though the full formulation remains valid.

Hence reduced DAG representations should be interpreted as distinct identification problems rather than automatic consequences of the full ATE formulation.

Primitive interpretations of the adjoint range condition.

The adjoint range condition

$$\ell_1 \in \mathcal{R}(A_1^\star)$$

is an operator-theoretic condition stated on Hilbert spaces. This appendix records several primitive sufficient conditions and interpretations illustrating when the condition may or may not hold.

Finite-support interpretation.

Suppose (U, W, V, C, Y_1) has finite support. Then the adjoint operator $A_1^\star$ reduces to a finite-dimensional conditional expectation matrix on the treated subsample. The adjoint range condition becomes the solvability of a linear system

$$A^\top b = \ell,$$

where the matrix A is induced by the conditional cross-moments between the primal-side variables $X_1 = (W, Y_1, C)$ and the adjoint-side variables $Z_1 = (V, C)$ on $\{D = 1\}$. Equivalently, regular identification reduces to a rank condition on this treated-subsample operator. Failure of the condition corresponds to directions of the treated outcome regression that cannot be represented through conditional projections of Z_1.

Gaussian latent-factor interpretation.

Under Gaussian latent-factor models, conditional expectation operators admit Hermite polynomial spectral decompositions. In this setting, the adjoint range condition becomes a Picard-type summability restriction on the Hermite coefficients of the treated outcome regression relative to the singular values of the conditional expectation operator. Regular identification requires sufficient decay of the projected coefficients relative to the singular spectrum; otherwise the problem

is only irregularly identified.

Branch-specific interpretation.

The strength of the adjoint range condition depends on the branch specification. In Branch A, where the observed calibration depends only on C, the condition concerns only functions representable through projections of Z_1 onto C. In Branch U the primal information set is enlarged to (W, C), producing a different adjoint range condition. In Branch G, the primal information set additionally includes Y_t, and on the treated support the target regression satisfies

$$\ell_1(X_1) = \mathbb{E}[Y \mid X_1, D = 1] = Y_1.$$

The dual equation therefore requires conditional projections of functions of Z_1 to recover the treated potential outcome itself as a function of X_1. This is the substantive source of the stronger dual condition under selection on gains.

Residualized representation and auxiliary projections.

The residualized decomposition $b = m+h$ does not alter the population identification condition. It rewrites the same dual equation as

$$A_1^\star h = \ell_1 - A_1^\star m,$$

where the choice of $m \in \mathcal{H}_{Z_1}$ acts as numerical preconditioning of the inverse problem; the population range condition $\ell_1 \in \mathcal{R}(A_1^\star)$ is unchanged. A particularly convenient choice is the baseline treated-outcome projection

$$m^{\text{base}}(Z_1) := \mathbb{E}[Y \mid Z_1, D = 1],$$

which removes the component of the treated outcome already explained by Z_1 before solving the inverse problem for the correction term h. Given a selected primal representative q, one may also define the q-weighted auxiliary projection

$$m_q^\dagger(Z_1) := \mathbb{E}[D\{1 + q(X_1)\}Y \mid Z_1].$$

This projection is useful for variance reduction and augmented-score representations but is not required for the minimal identification result.

Diagnostics.

These primitive interpretations motivate the singular-value diagnostics reported in Appendix H.9. Empirically small singular values indicate directions in which the dual inverse problem is weakly regularized and the orthogonal-moment representation becomes unstable.

Diagnostics for the identification trichotomy.

The three cases can be assessed in practice through the empirical singular value spectrum of $\widehat{A}_1^\star$ and the empirical norm of $\widehat{b}$ as a function of the regularization parameter. A slow decay of singular values with $\|\widehat{b}\|$ growing as $\lambda \to 0$ is consistent with Case I (irregular identification). A

stable $\|\widehat{b}\|$ as $\lambda \to 0$ is consistent with Case R (regular identification). Slow decay of singular values *with* $\widehat{A}_1^\star$ residuals not shrinking points to Case N.

Primitive sufficient conditions for the adjoint range condition.

Proposition B.3. *Let $A_1 e_j = \sigma_j f_j$ be a singular system of $A_1 : \mathcal{H}_{X_1} \to \mathcal{H}_{Z_1}$, and write the target $\rho_{1,m} \in \mathcal{H}_{X_1}$ in the orthonormal basis as $\rho_{1,m} = \sum_j \rho_j e_j$. The adjoint range condition $\rho_{1,m} \in \mathcal{R}(A_1^\star)$ holds if and only if the Picard summability*

$$\sum_j \frac{\rho_j^2}{\sigma_j^2} < \infty \tag{14}$$

is satisfied. Each of the following primitive conditions is sufficient:

(a) ***Finite-support rank condition.*** *If (W, V, Y_1, C) has finite support on $\{D = 1\}$ and the rank of the empirical cross-moment matrix on $\{D = 1\}$ equals the dimension of $\sigma(X_1)/\sigma(Z_1)$-coset representatives, then $\rho_{1,m} \in \mathcal{R}(A_1^\star)$ for any $m \in \mathcal{H}_{Z_1}$.*

(b) ***Gaussian/Hermite Picard condition.*** *In a linear-Gaussian latent-factor model with measurement correlation $\varrho \in (0, 1)$, the singular values satisfy $\sigma_j = \varrho^j$ in the Hermite basis. The Picard summability (14) reduces to $\sum_j a_j^2/\varrho^{2j} < \infty$ where a_j are the Hermite coefficients of $\rho_{1,m}$. This is a smoothness condition on $\rho_{1,m}$ relative to the Hermite expansion. Gaussian completeness ensures uniqueness of any solution when it exists; existence is governed by the Picard summability, not by completeness alone.*

(c) ***Source condition.*** *A sufficient analytic representation is*

$$\rho_{1,m} = (A_1^\star A_1)^{\beta/2} w, \qquad w \in \mathcal{H}_{X_1},\ \|w\| < \infty, \tag{15}$$

where $A_1^\star A_1 : \mathcal{H}_{X_1} \to \mathcal{H}_{X_1}$ is the primal self-adjoint operator. Under this representation, $\rho_j = \sigma_j^\beta w_j$, so $\sum_j \rho_j^2/\sigma_j^2 = \sum_j \sigma_j^{2\beta-2} w_j^2$. For $\beta \geq 1$, this is finite and $\rho_{1,m} \in \mathcal{R}(A_1^\star)$. For $0 < \beta < 1$, the source representation gives smoothness relative to the operator but does not by itself guarantee the strict range condition; in this regime the parameter may be only irregularly identified (Case I of the trichotomy).

Proof sketch. The equivalence between $\rho_{1,m} \in \mathcal{R}(A_1^\star)$ and the Picard summability (14) is a standard fact for compact operators on Hilbert spaces (Engl et al., 2000). All three parts (a)–(c) below are stated in terms of the *observed dual operator $A_1^\star$ on the treated law*, not the latent measurement operator T_W.

(a) For finite-support models, $A_1^\star : \mathcal{H}_{Z_1} \to \mathcal{H}_{X_1}$ has a finite matrix representation, and the range condition is equivalent to the column-space condition $\ell_{1,m} \in \mathrm{col}(\mathbf{A}_1^\star)$.

(b) Under a Gaussian treated-law model in which (X_1, Z_1) is jointly Gaussian on $\{D = 1\}$ conditional on (S, Q) with measurement-shifter correlation $\varrho \in (-1, 1)$, the operator $A_1^\star : \mathcal{H}_{Z_1} \to \mathcal{H}_{X_1}$ is diagonalized by Hermite polynomials in each Hilbert space with singular values $\sigma_j' = \varrho^j$. Expanding $\rho_{1,m} = \sum_j \rho_j e_j$ in the Hermite basis, the Picard summability (14) reduces to $\sum_j \rho_j^2/\varrho^{2j} < \infty$, a smoothness condition on $\rho_{1,m}$ relative to the measurement-shifter correlation. Note that this is a property of the *observed treated-law operator* $A_1^\star$ and is separate from the W-side latent completeness condition for (R1-A); the latter governs uniqueness of latent representers and lives on the latent operator T_W (Appendix D.2).

(c) The source representation $\rho_{1,m} = (A_1^\star A_1)^{\beta/2} w$ gives $\rho_j = \sigma_j'^{\beta} w_j$ in the singular system of $A_1^\star$; substitution into the Picard sum gives $\sum_j (\sigma_j')^{2\beta-2} w_j^2$, finite for $\beta \geq 1$. For $0 < \beta < 1$, the source representation gives only a smoothness statement on $\rho_{1,m}$ relative to the operator, not the strict range condition. □

Specification diagnostics for observed calibration restrictions.

Although the latent anchor condition is not generally testable, the observed calibration equations imply overidentifying moment restrictions. Given validation functions v^{test} and s^{test}, one may form

$$\hat{g}_q^{\text{test}} = \frac{1}{\sqrt{N}} \sum_i \{D_i \hat{q}(X_{1i}) - (1 - D_i)\} v^{\text{test}}(Z_{1i})$$

and

$$\hat{g}_b^{\text{test}} = \frac{1}{\sqrt{N}} \sum_i D_i \{Y_i - \hat{b}(Z_{1i})\} s^{\text{test}}(X_{1i}).$$

Quadratic forms in these validation moments provide specification diagnostics for the observed calibration restrictions. They do not test the latent structural range condition itself; rather, they test implications of the estimated calibration model in the observed distribution.

A formal overidentification test of the observed calibration system.

The validation moments above can be combined into a formal specification test of the *observed* restrictions, distinct from the untestable latent anchor (R1-A). Stack finite vectors of validation functions $v^{\text{test}} = (v_1, \ldots, v_{J_q})$ and $s^{\text{test}} = (s_1, \ldots, s_{J_b})$ and write $\widehat{g}^{\text{test}} = \big((\widehat{g}_q^{\text{test}})', (\widehat{g}_b^{\text{test}})'\big)' \in \mathbb{R}^{J_q + J_b}$. Because the observed balancing system $A_1 q = r_1$ (and its dual $A_1^\star b = \ell_1$) is implemented as an overidentified Sieve GMM problem—the moment dimension L_n exceeds the calibration-parameter dimension K_n (Appendix F.1)—these validation moments are not zeroed by construction and carry testable content.

Proposition B.4 (Overidentification test for $A_1 q = r_1$)**.** *Suppose the cross-fitted calibration estimators $(\widehat{q}, \widehat{b})$ satisfy the product-rate and finite-variance conditions of Appendix F.2, and let $\widehat{\Omega}$ be a consistent estimator of the asymptotic covariance of $\widehat{g}^{\text{test}}$ that accounts for the first-stage estimation of $(\widehat{q}, \widehat{b})$. Then, under the null that the observed calibration restrictions (R1-B) and (R2) hold for the maintained sieve model,*

$$J_N := \big(\widehat{g}^{\text{test}}\big)' \, \widehat{\Omega}^{+} \, \widehat{g}^{\text{test}} \;\; \to_d \;\; \chi_r^2,$$

where $\widehat{\Omega}^{+}$ is the Moore–Penrose inverse and $r = \operatorname{rank}(\Omega)$ for the limiting covariance Ω; under a fixed validation basis, $r = (J_q + J_b) - \dim(\theta)$ with θ the estimated calibration parameter. Rejection indicates misspecification of the observed calibration model—failure of the sieve approximation to (R1-B) or (R2)—and not of the latent condition (R1-A).

With a growing sieve dimension the degrees-of-freedom bookkeeping is nonstandard, and the appropriate limiting theory is that of overidentification testing in (possibly nonparametric) conditional-moment models (Chen and Santos, 2018; Chen and Pouzo, 2012; Hall and Horowitz, 2005): the statistic is formed from the residual projection orthogonal to the estimated parameter directions, and its limiting law is read from that of the corresponding conditional-moment overidentification test. This furnishes a genuine specification test of the observed implica-

tions of the calibration restrictions—the empirical discipline urged in the partial-identification literature—while leaving the latent anchor (R1-A) as a maintained structural hypothesis whose non-testability is discussed in Appendix D.2.

Auxiliary projections and residualized representations.

The algebraic content of the residualized parametrization $b = m + h$ is given in Appendix C.4; this subsection records the auxiliary projection that the body uses as a variance-reducing residualization device.

Definition B.1 (Auxiliary projection)**.** For a selected representative $q^{Y_1,\ddagger}$, a variance-reducing auxiliary projection can be defined by

$$m_0^{Y_1,\dagger}(Z_1) := \mathbb{E}\big[D(1 + q^{Y_1,\ddagger}(X_1))Y_1 \mid Z_1\big]. \tag{16}$$

It satisfies

$$\mathbb{E}\Big[\big\{D(1 + q^{Y_1,\ddagger})Y_1 - m^{Y_1,\dagger}\big\}v_m(Z_1)\Big] = 0 \quad \forall v_m \in L^2(Z_1). \tag{17}$$

This projection is useful for residualization but is not required for the minimal identification theorem.

Extensions.

This subsection records an extension of the auxiliary-measurement framework to distributional and quantile functionals.

Distributional and quantile functionals.

The results in this paragraph are the special case $\varphi(y, c) = \varphi(y)$ of the main-text Theorem 3.4 and of the necessity-and-sufficiency characterization in Theorem 3.3; they are retained here for implementation detail and for the strict-range versus closure distinction.

Theorem B.5 (Linear functionals and distribution functions)**.** *For any bounded measurable function φ, replace Y by $\varphi(Y)$ in the Y_1-side dual equation and signal. If the corresponding adjoint range condition holds, then*

$$\mathbb{E}[\varphi(Y_1)] = \mathbb{E}\big[D\{1 + q(X_1)\}\{\varphi(Y) - b_\varphi(Z_1)\} + b_\varphi(Z_1)\big]. \tag{18}$$

Remark (Strict range versus closure for distributional functionals)*.* The distributional functional theorem requires the strict range condition $\ell_\varphi := \mathbb{E}[\varphi(Y) \mid X_1, D = 1] \in \mathcal{R}(A_1^\star)$ for existence of a finite-norm adjoint representer b_φ and a regular influence function. Under the weaker closure condition $\ell_\varphi \in \overline{\mathcal{R}(A_1^\star)} \setminus \mathcal{R}(A_1^\star)$, the parameter $\mathbb{E}[\varphi(Y_1)]$ is point identified in a limit sense (via a sequence $b_\varphi^{(n)}$ with $\|b_\varphi^{(n)}\| \to \infty$), but no finite-norm adjoint representer or regular influence function exists; this is the irregular regime (Case I of the trichotomy).

Proof. Replace Y by $\varphi(Y)$ in the proof of Theorem 3.4. On $\{D = 1\}$, $\varphi(Y) = \varphi(Y_1)$, so the latent selection-odds representer gives the anchor

$$\mathbb{E}[D\{1 + q^\ell(X_1)\}\varphi(Y)] = \mathbb{E}[\varphi(Y_1)].$$

For the finite-calibration representation in the theorem, we use the *strict* range condition

$$\ell_{1,\varphi} \in \mathcal{R}(A_1^\star), \qquad \ell_{1,\varphi}(x) := \mathbb{E}[\varphi(Y) \mid X_1 = x, D = 1],$$

which guarantees a finite-norm $b_\varphi \in \mathcal{B}_1$ satisfying $A_1^\star b_\varphi = \ell_{1,\varphi}$. Under this condition, invariance over $\mathcal{Q}_1$ holds by the operator-adjoint identity, and the primal-dual signal

$$\mathbb{E}[\varphi(Y_1)] = \mathbb{E}[D\{1 + q(X_1)\}\{\varphi(Y) - b_\varphi(Z_1)\} + b_\varphi(Z_1)]$$

holds for every $(q, b_\varphi) \in \mathcal{Q}_1 \times \mathcal{B}_1$. Under the weaker closure condition $\ell_{1,\varphi} \in \overline{\mathcal{R}(A_1^\star)}$, the functional $\mathbb{E}[\varphi(Y_1)]$ may be identified only in the irregular limiting sense (Case I of the trichotomy), and the displayed score with finite b_φ need not exist; the closure case is treated separately in Section 3.3. □

Corollary B.6 (Marginal CDFs and quantile treatment effects)**.** *Taking $\varphi_y(u) = \mathbf{1}\{u \le y\}$ identifies the marginal CDFs of Y_1 and Y_0 directly under the corresponding range conditions. Quantile treatment effects then follow by inversion of the identified CDFs at continuity points where the quantile maps are well defined, with strict monotonicity ensuring uniqueness. Similarly, $\varphi(y, c) = y\,\mathbf{1}\{c \in \mathcal{S}\}$ identifies the subgroup numerator $\mathbb{E}[Y_t\mathbf{1}\{C \in \mathcal{S}\}]$; the conditional subgroup mean $\mathbb{E}[Y_t \mid C \in \mathcal{S}]$ follows after division by $\mathbb{P}(C \in \mathcal{S})$, which is identified directly from the observed law. The joint distribution of (Y_1, Y_0) is not identified without further structure.*

Implementation details.

This subsection describes approximate OLS implementations of the calibrated orthogonal moment that are useful as numerical diagnostics, together with the conditions under which they coincide with the exact Sieve GMM implementation covered by the formal theory.

Approximate OLS implementations.

Implementation simplification: Z_1-projection and OLS approximation.

Theoretically, (12) requires a test-function family in $L^2(X_1)$ (the NPIV/Sieve GMM space). Note that $X_1 = (W, Y_1, C)$ and $Z_1 = (V, C)$ are not nested: Z_1 contains V which is excluded from X_1, and X_1 contains W and Y_1 which are excluded from Z_1. The intersection $\sigma(X_1) \cap \sigma(Z_1) = \sigma(S, Q)$ is the common conditioning block C, and $L^2(S, Q)$ is a common subspace of both $L^2(X_1)$ and $L^2(Z_1)$.

For implementation, two distinct test-function families can be used:

- **Adjoint test functions in $L^2(X_1)$:** the full NPIV space. Let $m_1^{\text{aux}}(Z_1)$ denote the treated-arm auxiliary outcome regression used for residualization. The estimator solves the exact dual equation

  $$A_1^\star h = \rho_{1,m_1^{\text{aux}}}, \qquad \rho_{1,m_1^{\text{aux}}}(x) := \mathbb{E}[Y_1 - m_1^{\text{aux}}(Z_1) \mid X_1 = x, D = 1]\,,$$

  and Neyman orthogonality is rigorously preserved in all directions.

- **Reduced test functions in $L^2(S, Q) \subset L^2(X_1)$:** the common subspace. For $s \in L^2(S, Q)$,

  $$\mathbb{E}[D\{Y_1 - m_1^{\text{aux}}(Z_1) - h(Z_1)\}s(S, Q)] = \mathbb{E}[s(S, Q)\,\mathbb{E}[D\{Y_1 - m_1^{\text{aux}}(Z_1) - h(Z_1)\} \mid S, Q]]\,.$$

Thus requiring this to vanish for every $s \in L^2(S,Q)$ is equivalent to

$$\mathbb{E}[D\{Y_1 - m_1^{\text{aux}}(Z_1) - h(Z_1)\} \mid S, Q] = 0 \quad \mathbb{P}\text{-a.s.}$$

This is a strictly weaker moment condition than the full adjoint moment, but it is computable as a least-squares regression on the $D = 1$ subsample.

The common-subspace approximation solves

$$h_1^{C,\text{aux}}(S,Q) = \mathbb{E}[Y_1 - m_1^{\text{aux}}(Z_1) \mid S, Q, D = 1]\,, \tag{19}$$

where m_1^{aux} is a function of $Z_1 = (V, C)$. This enforces the reduced moment

$$\mathbb{E}\Big[D\{Y_1 - m_1^{\text{aux}}(Z_1) - h_1^{C,\text{aux}}(S,Q)\} \mid S, Q\Big] = 0,$$

which is strictly weaker than the full adjoint moment

$$\mathbb{E}[D\{Y_1 - m_1^{\text{aux}}(Z_1) - h_1^{\text{aux}}(Z_1)\} \mid X_1] = 0$$

required by the exact identification theorem. Unless the reduced solution also satisfies the full adjoint moment, the resulting score is an approximation: it is computable as a least-squares regression on the $\{D = 1\}$ subsample, but is not covered by the formal invariance result of Theorem 3.4. **Position within the primal–dual framework.** The OLS-style solution $h_1^{C,\text{aux}}$ in (19) is the projection of the dual equation onto the common subspace $L^2(S,Q) = L^2(C)$. It is not a representative of the full adjoint representer solution set $\mathcal{B}_1$ in general:

- The full adjoint equation $A_1^\star h = \rho_{1,m_1^{\text{aux}}}$ requires the conditional moment

  $$\mathbb{E}[D\{Y_1 - m_1^{\text{aux}}(Z_1) - h(Z_1)\} \mid X_1] = 0$$

  to hold for the full $X_1 = (W, Y_1, C)$, whereas the OLS projection enforces only the $\sigma(S,Q)$-measurable component.

- Consequently, $m_1^{\text{aux}} + h_1^{C,\text{aux}} \notin \mathcal{B}_1$ in general, and Theorem 3.4 does not directly apply to the OLS-based plug-in estimator. The formal identification and $\sqrt{N}$ inference results of this paper apply to the exact NPIV/Sieve GMM implementation that enforces the full $L^2(X_1)$ adjoint moment.

- The OLS approximation is therefore best viewed as a *computational diagnostic* or *preliminary estimator*. It can be reported alongside the exact implementation as a sensitivity check; large discrepancies between the two suggest that the $X_1 \setminus Z_1$ directions (W and Y_1 themselves) are important for the dual moment.

Comparison of the two implementations. This paper supports two implementation options:

- **(a) Exact NPIV / Sieve GMM implementation.** Construct a sieve including test functions in the $X_1 = (W, Y_1, C)$ space. The estimator solves the full dual equation

  $$A_1^\star h = \rho_{1,m_1^{\text{aux}}}$$

required by Theorem 3.4. Neyman orthogonality is rigorously guaranteed, and asymptotic normality holds exactly.

- **(b) OLS approximation** (19)**.** Regress $Y_1 - \widehat{m}_1^{\text{aux}}(Z_1)$ on the chosen reduced space, such as $C = (S, Q)$ or $Z_1 = (V, C)$, in the $D = 1$ subsample. This is computationally much cheaper and easily combined with machine-learning nuisance estimation. Approximate orthogonality preserves the bias-reduction effect in the projected directions exactly. Residual bias from $X_1 \setminus Z_1$ directions may persist; when strict orthogonality is required, use the adjoint-moment implementation (a).

In the Monte Carlo experiments of Appendix H, the finite-support designs use exact finite-dimensional Sieve GMM with cell-indicator bases, and the continuous stress tests use polynomial sieves. For empirical applications we recommend the low-cost approximation for initial analysis, followed by robustness checks with the adjoint-moment implementation or extended Sieve GMM. Large discrepancies between the two suggest that corrections from $X_1 \setminus Z_1$ directions (W and Y_1 themselves) are important. **Extended implementation for variance minimization.** When the calibration is nonunique, minimum-norm and minimum-variance implementations generally differ. An extended Sieve GMM that directly solves for the variance-minimizing adjoint representer $b_q^\star$ for a fixed primal q is described in Appendix F.4. This extension implements both the strict adjoint moment in (a) and sample-variance minimization of the orthogonal score as a constrained quadratic program.

Common-subspace versus Z_1-projection OLS.

The OLS approximation admits two distinct implementations, which should be distinguished:

- *Common-subspace OLS:*

$$h_1^{C,\text{aux}}(C) = \mathbb{E}[Y_1 - m_1^{\text{aux}}(Z_1) \mid C, D = 1],$$

  projecting onto $L^2(C) = L^2(S, Q)$. This satisfies the reduced moment

$$\mathbb{E}\Big[D\{Y_1 - m_1^{\text{aux}}(Z_1) - h_1^{C,\text{aux}}(C)\} \mid C\Big] = 0.$$

- *Z_1-projection OLS:*

$$h_1^{Z,\text{aux}}(Z_1) = \mathbb{E}[Y_1 - m_1^{\text{aux}}(Z_1) \mid Z_1, D = 1],$$

  projecting onto $L^2(Z_1) = L^2(V, C)$. This satisfies

$$\mathbb{E}\Big[D\{Y_1 - m_1^{\text{aux}}(Z_1) - h_1^{Z,\text{aux}}(Z_1)\} \mid Z_1\Big] = 0.$$

Neither projection satisfies the full adjoint moment

$$\mathbb{E}[D\{Y_1 - m_1^{\text{aux}}(Z_1) - h(Z_1)\} \mid X_1] = 0$$

in general, so neither is covered by Theorem 3.4. Both serve as computational diagnostics; the common-subspace version is more conservative because it uses only the conditioning block shared by X_1 and Z_1.

OLS approximation in implementation.

A simple approximation in practice replaces the Tikhonov-regularized inverse $\widehat{A}_1^{-1}$ with an ordinary-least-squares projection onto Z_1-space. Theoretically, the OLS projection satisfies the moment $\mathbb{E}[Dq(X_1)z(Z_1)] = \mathbb{E}[(1-D)z(Z_1)]$ for z in the span of Z_1 but not necessarily for the entire $L^2(Z_1)$. The OLS approximation is therefore the projection of the calibration solution onto the linear Z_1-span. The OLS step is *exact* only if the treated-law adjoint equation admits a solution in the chosen linear span of Z_1, i.e., if there exists a linear $b \in \mathcal{H}_{Z_1}$ satisfying $\mathbb{E}[b(Z_1) \mid X_1, D = 1] = Y_1$. Linear Gaussianity alone does not imply this condition: under a linear-Gaussian latent-factor model with measurement correlation $|\rho| < 1$, the conditional expectation $\mathbb{E}[\mathbb{E}[Y \mid Z_1, D = 1] \mid X_1, D = 1]$ is a projection-shrinkage of Y_1, not Y_1 itself. The OLS-approximated estimator can therefore be computed without nonlinear optimization and serves as a diagnostic preconditioner; the formal exact-SGMM theorem applies to the regularized NPIV side rather than to the OLS approximation.

B.4 Economic examples and empirical caveats

Economic interpretation and empirical caveats.

The examples in this subsection are illustrative. The variable mappings, especially in the HRS application, are maintained measurement restrictions rather than verified empirical facts; the caveats below make this explicit.

Additional economic interpretations.

This subsection collects the extended economic interpretations of the auxiliary-measurement framework, including the export-entry and retirement examples and a non-nestedness comparison with the marginal treatment effect framework.

This appendix provides additional economic interpretations of the auxiliary-measurement framework and discusses situations in which the calibration implication and range conditions may or may not be plausible in applications. The purpose is illustrative rather than structural: identification in the main text rests on the calibration implication and range conditions themselves, not on the literal correctness of any particular economic model below.

Export entry and selection on gains.

Consider a firm export-entry decision following the heterogeneous-firms trade literature initiated by Melitz (2003). Let $D = 1$ denote export participation and $D = 0$ domestic-only operation. The potential outcomes are

$$Y_1 = \text{profit under exporting}, \qquad Y_0 = \text{profit under domestic-only operation}.$$

Firms export when expected export profits exceed domestic profits net of fixed export-entry costs.

The latent heterogeneity U summarizes unobserved productivity, managerial quality, organizational capital, product quality, financing capacity, and ability to establish foreign distribution networks. This is a canonical selection-on-gains environment: firms with large gains from exporting are more likely to enter export markets. Treatment selection therefore depends directly on the potential outcomes (Y_1, Y_0) rather than only on latent type.

Latent-type measurement W.

The latent-type measurement W may consist of lagged total factor productivity, labor productivity, lagged employment, capital intensity, patent counts, R&D intensity, or pre-entry sales growth. These variables contain information about latent firm type. The exclusion $W \not\to D$ requires that conditional on the latent type and conditioning variables, the measurement does not directly affect export participation.

This restriction is most plausible when the measurement is recorded strictly before the export-entry decision. For example, lagged productivity at $t-2$ may plausibly serve as a measurement for latent productivity without directly affecting export participation at time t once the latent type is conditioned out. The restriction is less plausible when the measurement is contemporaneous with the export decision, such as current-year sales simultaneously entering both the treatment decision and the latent-type measurement block.

Outcome-representation variable V.

The variable V shifts the latent type distribution without directly affecting treatment participation conditional on the latent type and the relevant conditioning variables. Examples include founding-cohort industrial conditions, business-group productivity, industrial agglomeration at founding, or lagged productivity of cohort peers.

The exclusions $V \not\to D$ and $V \not\to W$ require that after conditioning on the latent type, these variables do not directly alter export participation or the latent-type measurements.

Unlike a standard LATE or MTE instrument, V is not required to move treatment assignment directly. Instead, it provides the adjoint-side variation needed to solve the associated adjoint range condition.

Environment variables S and Q.

The variables S and Q represent environment variables affecting the treated and untreated potential outcomes.

The treated-side environment S may include exchange rates, tariff exposure, foreign demand conditions, destination-country GDP, or free-trade-agreement exposure. These variables directly affect export profits Y_1 and may also directly affect export participation through expected gains from exporting.

The untreated-side environment Q may include domestic demand, domestic competitive conditions, input prices, or local distribution networks. These variables directly affect domestic-only profits Y_0 and may also directly affect export participation through the forgone domestic payoff.

Accordingly, the framework allows both $S \to D$ and $Q \to D$. This is a key distinction from standard proximal causal inference formulations, where treatment assignment depends only on latent confounding after conditioning and latent-type measurements are excluded from treatment.

Asymmetric roles of W and V.

The measurement side and the adjoint side are not interchangeable. The measurement variable W is used to recover the latent treatment odds structure. If the latent odds ratio retains variation in the latent type after conditioning on observables, then no calibration depending only on (S, Q, Y_1) can reproduce the latent odds ratio.

The variable V plays a different role: it provides the variation needed for the adjoint range condition. If V contains insufficient independent variation conditional on the conditioning variables, then the adjoint representer may fail to exist or the parameter may become irregularly identified. Appendix G.3 formalizes this distinction.

Reduced formulations.

For one-sided targets such as $\mu_1 = \mathbb{E}[Y_1]$, the untreated-side variables (Y_0, Q) may sometimes be marginalized out. Define

$$\bar{p}_1(U, Y_1, S) := \mathbb{P}(D = 1 \mid U, Y_1, S),$$

which integrates over (Y_0, Q). If a reduced calibration implication and reduced range condition hold, then identification may proceed using only (W, Y_1, S) and (V, S).

This reduction is not a direct consequence of the full ATE identification result. It is a distinct identification problem with its own calibration and range conditions.

Retirement and cognition.

A second example is the effect of retirement on subsequent cognitive function. Let $D = 1$ denote retirement and $D = 0$ continued work. The potential outcomes are

$$Y_1 = \text{cognitive function under retirement}, \qquad Y_0 = \text{cognitive function under continued work}.$$

The latent type U summarizes latent cognitive reserve, physical health, long-run ability, and related persistent heterogeneity. The retirement decision depends on financial incentives, expected longevity, health conditions, leisure value, and expected cognitive trajectories under retirement and continued work.

This setting naturally fits the selection-on-gains formulation. Workers with favorable expected cognitive trajectories under retirement may be more likely to retire, while workers expecting large cognitive costs from retirement may remain employed. Treatment selection therefore may depend directly on both potential outcomes (Y_1, Y_0).

Latent-type measurement W.

The latent-type measurement W may include pre-retirement cognitive scores or baseline health indicators. Schooling-related variables can also be viewed as cognitive-reserve proxies in some applications; in the HRS baseline specification below, however, respondent's own schooling is placed in the observed-adjustment block X, not in the main bridge measurement block W. These variables provide information about latent cognitive type and cognitive reserve. They are interpreted as measurements of latent cognitive type; the interpretation requires that their direct effects on retirement be absorbed by the conditioning block and the latent type, which is a maintained restriction rather than an empirically verified fact.

Outcome-representation variable V.

The variable V may include long-run determinants of latent cognitive type such as parental education, early-life schooling quality, or cohort-level educational conditions. These variables

need not shift retirement directly; instead, they provide variation relevant for the adjoint-side range condition. The use of parental education as V is an identifying restriction: it is assumed to provide adjoint-side variation after conditioning on the Roy-side block $C_R = (S, Q)$, not to act as a direct retirement shifter.

Environment variables S and Q.

The retirement-side environment S may include pension generosity, wealth, health insurance, subjective retirement expectations, or access to retirement-supporting institutions. These variables may directly affect both retirement utility and cognitive outcomes under retirement.

The work-side environment Q may include occupation, work hours, job strain, industry, or job complexity. These variables may directly affect both continued-work cognitive outcomes and the retirement decision itself.

Branch G versus Branch U.

This application illustrates the distinction between latent-confounding selection and selection on gains. Under latent-confounding-only selection, retirement would depend only on latent type and environment. Under Branch G, treatment selection may additionally depend directly on the potential cognitive outcomes themselves through anticipated gains or losses from retirement.

Plausibility of the main restrictions.

The plausibility of the calibration implication and range conditions depends on the empirical setting and measurement structure.

The calibration implication is most plausible when the latent-type measurement is recorded strictly before treatment assignment and does not directly enter the treatment decision conditional on the latent type. It is more likely to fail when the measurement directly affects treatment costs, treatment information sets, or treatment eligibility.

The adjoint range condition is most plausible when the outcome-representation variable contains sufficiently rich variation independent of the conditioning variables. Weak or nearly collinear adjoint-side variation may lead to failure of the adjoint range condition or irregular identification.

The overlap condition may fail in environments with deterministic sorting or near-perfect separation, where treatment probabilities approach zero or one conditional on latent type.

Table 12 summarizes these considerations.

Non-nestedness relative to MTE models.

The auxiliary-measurement framework is not nested within the standard scalar-index MTE framework. To illustrate, consider the treatment-selection rule

$$D = \mathbf{1}\big\{ \sin(Y_1 Y_0) + U_1^2 - U_2 + \varepsilon_D > 0 \big\}. \tag{20}$$

Selection depends nonmonotonically on both potential outcomes and on multidimensional latent heterogeneity. Such a rule is not generally reducible to the separable scalar-resistance representation commonly used in MTE formulations.

Nevertheless, the calibration approach may still identify treatment effects provided the calibra-

Table 12: Plausibility of the main restrictions in applications

Restriction	Plausible when	Likely to fail when
Calibration implication	W is measured before treatment and does not directly affect participation	W directly affects costs, information, or eligibility
Latent selection-odds representer	measurement variation is informative about latent heterogeneity	W is weak or unrelated to latent type
Adjoint range	V contains sufficiently rich independent variation	V is weak or nearly collinear with conditioning variables
Overlap	treatment probabilities remain bounded away from zero and one	deterministic sorting or near-perfect separation
Injectivity	calibrating functions themselves are objects of interest	the observed calibration equation has multiple solutions; not required for ATE identification

tion implication and range conditions hold. This illustrates that the framework accommodates forms of selection outside standard scalar-index treatment-selection models.

C Supplementary identification remarks and derivations

This appendix collects supplementary remarks and derivations that clarify the identifying restrictions, branch-specific operator systems, residualized implementation, and the extended roadmap of the identification argument.

C.1 Latent anchor and observed calibration transfer

Latent anchor and observed calibration transfer.

The latent odds anchor establishes that a causally correct primal representative exists, and the calibration implication transfers it to the observed balancing equation. We record the normalization identity and the proof detail for the transfer lemma.

Lemma C.1 (Latent-anchor normalization)**.** *Suppose q_1^ℓ satisfies the Y_1-side calibration implication*

$$\mathbb{E}\big[D\, q_1^\ell(X_1) - (1-D) \,\big|\, U, Y_1, C, V\big] = 0.$$

Then

$$\mathbb{E}\big[D\{1 + q_1^\ell(X_1)\} \,\big|\, U, Y_1, C, V\big] = 1,$$

and hence, by iterating out V,

$$\mathbb{E}\big[D\{1 + q_1^\ell(X_1)\} \,\big|\, U, Y_1, C\big] = 1.$$

The Y_0-side statement is analogous.

Proof detail for Lemma 3.1.

Let $R_1 = D\, q_1^\ell(X_1) - (1-D)$. Assumption 2.2 gives $\mathbb{E}[R_1 \mid U, Y_1, S, Q, V] = 0$. Since $Z_1 = (V, C)$, iterated expectations imply

$$\mathbb{E}[R_1 \mid Z_1] = \mathbb{E}\big\{\mathbb{E}[R_1 \mid U, Y_1, S, Q, V] \,\big|\, Z_1\big\} = 0.$$

Hence $A_1 q_1^\ell = r_1$. The Y_0 side is symmetric.

Remark (Role of conditional independence: a sufficient condition for variable discovery, not an identification condition). Identification in this paper requires only the calibration implication in Assumption 2.2 and the range conditions. The DAG and conditional-independence statements are sufficient conditions that make those high-level conditions interpretable in terms of economic structure. For example, the strong restrictions

$$W \perp\!\!\!\perp (V, S, Q, Y_1, Y_0, D) \mid U, \qquad V \perp\!\!\!\perp (W, S, Q, Y_1, Y_0, D) \mid U, \qquad (Y_1, S) \perp\!\!\!\perp (Y_0, Q) \mid U$$

imply the sufficient condition of Proposition 2.1 but are not required for the full calibration identification theorem. In particular, $(Y_1, S) \perp\!\!\!\perp (Y_0, Q) \mid U$ is a sufficient condition for the economic interpretation of a reduced formulation that drops Q or S from the calibration arguments. It is not essential for the full formulation with $X_1 = (W, Y_1, C)$ and $Z_1 = (V, C)$. Thus failure of a particular DAG interpretation does not by itself refute the calibration theorem; it refutes only that sufficient structural route to the high-level calibration implication.

Remark (Logical relation between (R1-A) and (R1-B)). The two conditions must be distinguished. **(R1-A) is a primitive assumption**: it requires the existence of a solution to the *latent* moment equation involving the unobserved U. This depends on the structural system (2)–(3) and on the strength of W as a measurement; concrete sufficient conditions are given in Appendix D.2 (Gaussian latent factor model). **(R1-B) is implied by (R1-A) together with (T)**: any $q^{Y_1,\ell}$ satisfying (R1-A) and the transfer moment (T) satisfies the observed operator equation $A_1 q = r_1$ by Lemma 3.1; the converse does not generally hold, since the observed calibration set may contain elements that are not latent selection-odds representers. In the proof structure of Theorem 3.4, identification starts from (R1-A) and (T) via $\mu_1 = \mathbb{E}[D(1 + q^{Y_1,\ell})Y_1]$ and then extends to the full solution set $\mathcal{Q}_1$ using (R1-B) and (R2). Assuming only (R1-B) and (R2) without (R1-A) would lose the causal anchor that the latent representer provides. The paper therefore treats (R1-A) and (R2) as the substantive range conditions and (R1-B) as a consequence of (R1-A) together with Assumption 2.2. Observed calibration solvability alone is not a causal anchor. It may admit solutions that satisfy $A_1 q = r_1$ but are not generated by any latent selection-odds representer. The latent condition (R1-A) is what connects the observed operator equation to the causal target.

C.2 Primal nonuniqueness and target invariance

Primal nonuniqueness and the target-invariance certificate.

The adjoint outcome representer plays the role of a target-invariance certificate: it pins down the value of the functional even when the calibration set is not a singleton. The following remarks make the certification explicit, contrast calibration weights with propensity weights, and record uniqueness conditions on the calibration representatives themselves.

Remark (Non-uniqueness of q is not an obstacle). The calibration set $\mathcal{Q}_1$ may contain multiple calibration functions, but the adjoint outcome representer certifies that the target value is

invariant over $\mathcal{Q}_1$. Concretely, for $q, q' \in \mathcal{Q}_1$, let $g = q - q' \in \ker(A_1)$. If $b \in \mathcal{B}_1$, then

$$\mathbb{E}\big[D\, g(X_1)\{Y - b(Z_1)\}\big] = \langle g, \ell_1 - A_1^\star b\rangle_{X_1} = 0,$$

so that the score

$$\mathbb{E}\big[D\{1 + q(X_1)\}\{Y - b(Z_1)\} + b(Z_1)\big]$$

is invariant over $q \in \mathcal{Q}_1$. Therefore identification of μ_1 does not require point identification of the calibrating functions themselves; only the value of the functional, not the representative, needs to be pinned down.

The adjoint outcome representer as a target-invariance certificate.

The calibration set $\mathcal{Q}_1$ may contain multiple calibrating functions, and the identification formula $\mu_1 = \mathbb{E}[D\{1+q\}\{Y-b\}+b]$ a priori depends on which q is chosen. Theorem 3.4 shows that the dependence vanishes: for any $q, q' \in \mathcal{Q}_1$, set $g := q - q' \in \ker(A_1)$ and use the operator-adjoint identity

$$\mathbb{E}[D\, g(X_1)\{Y - b(Z_1)\}] = \langle g, \ell_1\rangle_{X_1} - \langle A_1 g, b\rangle_{Z_1} = \langle g, \ell_1 - A_1^\star b\rangle_{X_1} = 0,$$

where the second equality uses the adjoint identity and the last equality uses $A_1^\star b = \ell_1$. (Equivalently, $\langle A_1 g, b\rangle_{Z_1} = 0$ since $g \in \ker(A_1)$.) The adjoint outcome representer b thus *certifies* that the target functional is invariant over the calibration set, regardless of the nonuniqueness of q.

Branch comparison as a diagnostic decomposition.

A practical feature of the auxiliary-measurement framework is that the three branches can be implemented sequentially using the same Sieve GMM software with branch-specific primal and dual basis choices. Comparing the three estimates produces a diagnostic decomposition that shows how the estimated target changes across branch-specific maintained information structures. Substantial divergence across branches is indicative (but not conclusive) evidence that the smaller information set may be insufficient: a near-zero Branch A estimate alongside a non-zero Branch U estimate is consistent with latent-type confounding, and a divergence between Branches U and G is consistent with selection on gains. However, branch divergence can also reflect calibration misspecification, weak measurement variation, finite-sample regularization bias, or sieve-choice differences. The comparison is therefore best read as a diagnostic of identifying-assumption sensitivity rather than a formal specification test. This sequential diagnostic is exemplified in the HRS application (Section 6).

Intuition for the trichotomy.

The three cases correspond to three regions of the Hilbert-space geometry. We label them *Case N (non-invariance)*, *Case I (irregular)*, and *Case R (regular)* throughout the paper for consistent reference. *Case N* ($\ell_1 \notin \overline{\mathcal{R}(A_1^\star)}$): the target signal does not lie in the closure of the adjoint range. The calibration restrictions do not pin down a target invariant over $\mathcal{Q}_1$, so the observed calibration system alone does not identify μ_1 even with infinite data. *Case I* ($\ell_1 \in \overline{\mathcal{R}(A_1^\star)} \setminus \mathcal{R}(A_1^\star)$): the target lies in the closure but not in the range. The target is point identified under the latent anchor (it is invariant over $\mathcal{Q}_1$), but the identifying functional is irregular: no finite-norm adjoint representer exists, so regular $\sqrt{N}$ inference fails. *Case R* ($\ell_1 \in \mathcal{R}(A_1^\star)$): the target lies in the range. A finite-norm adjoint representer exists, the induced score has finite

variance, and the parameter admits a regular orthogonal-moment representation. This trichotomy mirrors the standard nonparametric IV regression literature (Picard condition, source conditions) but is here cast in terms of the adjoint outcome representer rather than the primal NPIV equation.

C.3 Calibration weights, branch systems, and no counterfactual imputation

The following remark records a point that is useful for interpretation and implementation.

Remark ($1+q$ is a calibration weight, not a propensity weight). The weighting quantity $\alpha_q(O) := D\{1 + q(X_1)\}$ in the identification formula (9) differs in character from the traditional IPW propensity weight $D/\pi(X)$:

- The true latent selection-odds representer q^ℓ reproduces the nonnegative latent odds ratio $(1 - p_1)/p_1$ as a conditional expectation,

$$\mathbb{E}[q^\ell(W, Y_1, C) \mid U, Y_1, C] = (1 - p_1)/p_1 \geq 0,$$

  but this does *not* imply that q^ℓ is pointwise nonnegative. A nonnegative conditional expectation does not force the underlying function to be nonnegative pointwise.

- Moreover, any solution $q \in \mathcal{Q}_1$ of the observed balancing equation $A_1 q = r_1$ has the form $q = q^\ell + g$ with $g \in \ker(A_1)$, so $q \geq 0$ pointwise need not hold.

- Hence $1 + q(X_1)$ may take **negative values**, and the familiar propensity-score intuition $\pi^{-1} > 0$ does not apply.

In this paper we call $1 + q$ a *calibration weight* (or *augmented calibration weight*) to distinguish it from the traditional propensity weight. This distinction does not affect the algebra of the identification formula (9) or the orthogonal score, but it matters for implementation, variance minimization, and interpretation. Positivity of the latent odds ratio is ensured by overlap (Assumption 2.4); pointwise nonnegativity of all representatives $q \in \mathcal{Q}_1$ is not needed for identification.

Corollary C.2 (Unique calibration representatives)**.** *If* $\ker(A_1) = \{0\}$ *and* $\ker(A_1^\star) = \{0\}$*, then the primal and adjoint representers are unique whenever they exist. Consequently, the calibrated orthogonal moment class contains a single element. If, in addition, this score satisfies the finite-variance and local-calibration-path conditions of Appendix G.1, then it is the regular influence function. It coincides with the canonical gradient only when the local model admits no lower-variance regular representative.*

Remark (Operator injectivity and ill-posedness). The primal operator A_1 may fail to be injective (compactness with infinite-dimensional kernel) or may have an unbounded inverse (compactness with eigenvalues clustering at zero). These are the two manifestations of ill-posedness that this paper handles separately. Non-injectivity of A_1 is the source of nonuniqueness of q and is addressed via the adjoint range condition (R2): the calibration invariance result shows that the target is invariant over the kernel of A_1. Unboundedness of A_1^{-1} (when interpreted as a regularized inverse) governs the rate of convergence of estimators and is addressed via source conditions and Tikhonov regularization (Remark D.4). The empirical singular-value spectrum of $\widehat{A}_1$ provides finite-dimensional diagnostics for weak directions and ill-conditioning. It does not by itself distinguish population non-injectivity from severe ill-posedness without additional structure.

Assumption C.1 (Optional injectivity)**.** When one wants the calibrating functions themselves to be point identified, one may impose injectivity of the corresponding primal or adjoint operators. These conditions are not required for identification of the ATE.

This assumption is not used for identification of μ_1. It is only needed when one wants to interpret the calibrating functions themselves, rather than the target functional, as point-identified objects. The main theorem allows nonunique calibration solutions and uses the adjoint representer to certify target invariance over the calibration set.

Example C.3 (Non-nestedness with MTE)**.** *Consider a selection rule*

$$D = \mathbf{1}\big\{ \sin(Y_1 Y_0) + U_1^2 - U_2 + \varepsilon_D > 0 \big\}. \tag{21}$$

This rule depends on both potential outcomes and multidimensional latent heterogeneity in a nonmonotone way. It is not generally reducible to the separable scalar-resistance form used in the standard MTE model. The calibration approach can still apply if the measurement and adjoint range conditions hold.

Branch-specific systems and range conditions.

Each branch corresponds to a choice of $(X_1^{\mathrm{b}}, Z_1^{\mathrm{b}})$ and the resulting primal–dual operator system. The branches are not literally nested as inverse problems even though the primal information sets are nested; in particular, the Branch G adjoint range condition is strong on the treated support, and a near-zero Branch A estimate need not invalidate a non-zero Branch G estimate.

Remark (Why a single auxiliary-measurement framework subsumes all three). The primal–dual structure of Theorem 3.4 (Section 3.2) is stated at the abstract level of two Hilbert-space operator equations $A_1 q = r_1$ and $A_1^\star b = \ell_1$. The abstract theorem applies to each branch after choosing the branch-specific pair $(X_1^{\mathrm{b}}, Z_1^{\mathrm{b}})$ and the corresponding operator A_1^{b}. Branch U takes $X_t^U = (W, C)$ and Branch G takes $X_t^G = (W, Y_t, C)$ with $Z_t = (V, C)$; Branch A uses the reduced system $X_t^A = Z_t^A = C$. What is common across branches is the form of the theorem, not a literal nesting of the full primal–dual systems.

Remark (Branch-adaptive minimum-norm anchor). Each branch operates with a different primal information set: Branch A uses $X_t^A = C$, Branch U uses $X_t^U = (W, C)$, and Branch G uses $X_t^G = (W, Y_t, C)$, with the corresponding embeddings $L^2(C) \hookrightarrow L^2(W, C) \hookrightarrow L^2(W, Y_t, C)$. The primal operators $A_1^{\mathrm{A}}, A_1^{\mathrm{U}}, A_1^{\mathrm{G}}$ and right-hand sides $r_1^{\mathrm{A}}, r_1^{\mathrm{U}}, r_1^{\mathrm{G}}$ differ across branches, so the corresponding solution sets $\mathcal{Q}_1^A, \mathcal{Q}_1^U, \mathcal{Q}_1^G$ are branch-specific affine sets that need not be nested without additional compatibility restrictions on the operators. For each branch $\mathrm{b} \in \{A, U, G\}$, the minimum-norm representative

$$q^{\mathrm{b},\dagger} \;=\; \arg\min_{q \in \mathcal{Q}_1^{\mathrm{b}}} \mathbb{E}[D \cdot q(X_1^{\mathrm{b}})^2] \tag{22}$$

is operationally defined through the corresponding Tikhonov-regularized minimum-norm problem. This anchor selection standardizes the representative within each branch. It does not make cross-branch differences a formal test of the true branch: such differences may reflect information-set changes, weak measurements, sieve approximation, or regularization bias. The branch-specific implementation in Appendix B.2 uses this anchor to ensure that within-branch estimates are comparable.

Remark (Branch G evaluation: counterfactuals are not imputed). Although q^G is written as a function of the potential outcome Y_t, it is evaluated only on the treatment arm where that

potential outcome is observed: $q_1^G(W, Y_1, C)$ enters the primal moment multiplied by D, so it is evaluated only on $\{D = 1\}$, where $Y_1 = Y$. For the μ_0 side the primal moment is

$$\mathbb{E}[(1 - D)\, q_0^G(W, Y_0, C) - D \mid Z_0] = 0,$$

and $q_0^G(W, Y_0, C)$ is evaluated only on $\{D = 0\}$, where $Y_0 = Y$. The auxiliary-measurement framework therefore does not require imputing individual counterfactual outcomes.

Remark (Branch-specific adjoint range conditions). The adjoint range condition (R2), $\ell_1 \in \mathcal{R}(A_1^\star)$, depends on the choice of branch. Each branch has its own adjoint operator $(A_1^{\mathrm{b}})^\star$ acting on functions of Z_1 into the branch-specific $L^2(X_1^{\mathrm{b}})$. Adding W or Y_t to the primal information set changes both the operator and the target functional, so the adjoint ranges across Branch A, U, and G live in different Hilbert spaces and there is no automatic monotonic ordering of the ranges. The relevant question is branch-specific: whether the corresponding target ℓ_1^{b} lies in the range of $(A_1^{\mathrm{b}})^\star$.

In Branch G the further inclusion of Y_t in X_1 enlarges the target as well as the operator: on the treated support $\ell_1(X_1) = Y_1$. The raw adjoint range condition $\ell_1 \in \mathcal{R}(A_1^\star)$ thus requires the conditional expectation of a Z_1-function to equal Y_1, which is a strong but operative condition. The residualized formulation $A_1^\star h = \rho_{1,m}$ (Section 3.2) is equivalent at the population level and serves as the numerical implementation device. Empirical exercises should therefore report residualized dual diagnostics, singular-value profiles, and regularization sensitivity where feasible. These diagnostics probe numerical range stability; they are not formal tests of the population range condition.

Remark (Where the Branch A causal interpretation can fail). Under the standard Branch A system $X_t^A = Z_t^A = C_A$, the adjoint representer is the ordinary treated outcome regression

$$b^A(C_A) = \mathbb{E}[Y \mid C_A, D = 1].$$

Thus the relevant difficulty in Branch A is not the observed adjoint range condition: with $X_t^A = Z_t^A = C_A$ the adjoint $(A_1^A)^\star$ acts within a single space $\mathcal{H}_{C_A}$, and $\ell_1^A(C_A) = \mathbb{E}[Y \mid C_A, D = 1]$ lies in its range whenever it is square integrable. Rather, Branch A is causally valid only if the observed adjustment block C_A is rich enough to anchor the latent odds condition, equivalently if ordinary selection on observables is credible for the target. When selection depends on latent type or potential outcomes beyond C_A, Branch A remains a well-defined adjustment benchmark but need not identify the ATE. In the HRS application, the near-zero Branch A estimate should therefore be read as an observed adjustment benchmark, not as evidence of Case N in the adjoint range trichotomy.

In Branch U, where q^U depends on (W, C), the adjoint operator $A_1^\star : \mathcal{H}_{Z_1} \to \mathcal{H}_{X_t^U}$ maps into a different Hilbert space than the Branch A adjoint, so the two ranges are not directly comparable. The relevant condition is branch-specific: $\ell_1^U \in \mathcal{R}\{(A_1^U)^\star\}$. Even when this holds, the population-relevant target $\ell_1^U = \mathbb{E}[Y \mid W, C, D = 1]$ may lie at the boundary of $\mathcal{R}\{(A_1^U)^\star\}$ (Case I of the trichotomy) when V is a weak outcome-representation variable.

Remark (Why the Branch G adjoint range is demanding). In Branch G, the further inclusion of Y_t in X_1 enlarges the domain of $A_1^\star$, but with a critical caveat: since $Y = Y_1$ on $\{D = 1\}$, the outcome regression $\ell_1(X_1) = \mathbb{E}[Y \mid X_1, D = 1] = Y_1$ on the treated support is the potential outcome itself, not a coarser function of X_1. The adjoint range condition $A_1^\star b = \ell_1$ therefore requires that $\mathbb{E}[b(V, C) \mid W, Y_1, C, D = 1] = Y_1$, which is generally very strong: it requires the

conditional expectation of a function of (V, C) to recover Y_1 exactly on the treated subsample.

Remark (Residualization does not weaken the range condition). The residualized decomposition $b = m + h$ (Section 3.2 and Section 4) does not remove this difficulty: it rewrites the same population range condition $\ell_1 \in \mathcal{R}(A_1^\star)$ in a numerically preconditioned form. With $m^{\text{base}}(Z_1) := \mathbb{E}[Y \mid Z_1, D = 1]$ chosen as the Z_1-projection of the treated outcome, the residual correction h satisfies the equivalent residualized dual equation $A_1^\star h = \rho_{1,m^{\text{base}}}$, where the residualized target

$$\rho_{1,m^{\text{base}}}(X_1) := \mathbb{E}[Y_1 - m^{\text{base}}(Z_1) \mid X_1, D = 1]$$

is the conditional moment of the *outcome residual after removing the* (V, C)*-explained part*, rather than Y_1 itself. The operative adjoint range condition is $\rho_{1,m} \in \mathcal{R}(A_1^\star)$. As shown in Section 3.2, this is equivalent at the population level to the raw condition $\ell_1 \in \mathcal{R}(A_1^\star)$ for any $m \in \mathcal{H}_{Z_1}$; the role of m is numerical preconditioning of the inverse problem rather than relaxation of the identification condition. The Branch G implementation uses this residualized formulation throughout.

Whether the operative condition holds depends on primitive structural assumptions (Appendix D.2). Under finite-support models with discrete (U, W, V, C, Y_1), the condition reduces to a rank condition on the cross-moment matrix on $\{D = 1\}$. Under Gaussian latent factor models, it reduces to a Hermite-Picard summability condition. Whether such primitive conditions are substantively credible in a given application is not directly testable. The singular-value diagnostics and regularization paths of Appendix H.9 probe the stability of the observed finite-dimensional inverse problems; they do not test the primitive population range conditions.

Remark (Overidentified full-Z Branch A). One can keep $q^A = q^A(C_A)$ while testing the primal moment against functions of $Z = (V, C_A)$. This imposes

$$\mathbb{E}[Dq^A(C_A) - (1 - D) \mid V, C_A] = 0.$$

If $q^A(C_A) = \{1 - p_A(C_A)\}/p_A(C_A)$, this requires

$$\mathbb{P}(D = 1 \mid V, C_A) = \mathbb{P}(D = 1 \mid C_A).$$

This is an additional overidentifying restriction and is not the standard Branch A adjustment system used in the main analysis.

Remark (Latent versus observed range conditions). Both (R1-A) and (R2) are range conditions, but their empirical content differs. (R1-A) is a latent structural range condition involving the unobserved U and is generally not testable from the observed distribution alone. This limitation is intrinsic to all proxy-based identification frameworks; the paper treats (R1-A) as a primitive sufficient condition tied to latent measurement structure rather than as an empirically testable restriction. Appendix D.2 and Appendix D.4 give sufficient conditions under various structures: finite support, Gaussian–Hermite, non-Gaussian deconvolution, and ordinary-smooth or super-smooth measurements. Corollary D.10 formally establishes that without additional latent structure, (R1-A) cannot be tested from observed-distribution moments, and Theorem D.9 supplies the explicit observational-equivalence construction. In contrast, (R2) is the adjoint range condition defined on the observed conditional distribution on $\{D = 1\}$. Appendix D.3 makes this observed-side operator representation explicit: the equation $A_1^\star b = \ell_1$ is formulated through the conditional distribution of $X_1 = (W, Y, C)$ and $Z_1 = (V, C)$ on $\{D = 1\}$, so the ill-posedness of (R2) can be assessed via empirically observable quantities (empirical singular values, the Tikhonov regularization path, the dual norm). These quantities are diagnostics of numerical range stability

and weak range behavior. They are not formal tests of infinite-dimensional range membership without additional structure.

Remark (What is genuinely new in this section). The primal–dual identification theorem (Theorem 3.4) is stated at a generality that nests all three branches. When restricted to Branch A ($q^A(C)$), the theorem reduces to standard AIPW-style identification under conditional ignorability. When restricted to Branch U ($q^U(W, C)$), it reduces to the proximal-causal-inference identification of Tchetgen Tchetgen et al. (2024); Cui et al. (2024) under $(Y_1, Y_0) \perp\!\!\!\perp D \mid U$. The new content is Branch G ($q^G(W, Y_t, C)$): the latent selection-odds representer reproduces the conditional treatment odds at the potential-outcome-conditional level, $\mathbb{P}(D = 0 \mid U, Y_t, C)/\mathbb{P}(D = 1 \mid U, Y_t, C)$, rather than at the latent-only level $\mathbb{P}(D = 0 \mid U, C)/\mathbb{P}(D = 1 \mid U, C)$. The adjoint outcome representer b, the orthogonal score Γ_1, and the trichotomy (Proposition 3.5) all inherit this distinction. The reader who has worked with proximal causal inference under treatment ignorability will recognize the general structure but should attend to where the Y_t argument appears in the conditioning sets, since this is the only place where selection on gains enters the formal apparatus.

Distinction from proximal causal inference and inverse-problem debiasing.

In standard proximal causal inference (Miao et al., 2018; Tchetgen Tchetgen et al., 2024; Cui et al., 2024), the treatment-on-potential-outcome independence $(Y_1, Y_0) \perp\!\!\!\perp D \mid U$ is assumed, so $\mathbb{E}[Y_1] = \mathbb{E}[\mathbb{E}[Y_1 \mid U]]$ and the IPW-like reweighting uses the conditional treatment probability $\pi(U) = \mathbb{P}(D = 1 \mid U)$. Under selection on gains, this independence fails, and the propensity-like weight must condition on potential outcomes themselves. The latent selection-odds representer q^ℓ reproduces the conditional odds $(1 - p_1)/p_1$ where $p_1 = \mathbb{P}(D = 1 \mid U, Y_1, C)$, so the implicit "reweighting" is by the potential-outcome–conditional odds. This is the structural difference: proximal causal inference reweights to undo confounding by latent type alone, while the present framework reweights to undo confounding by latent type *and* by potential outcomes themselves.

C.4 Residualization and reduced conditioning blocks

Residualization and implementation choices.

Writing $b = m + h$ is an algebraic reparametrization of the dual equation. This subsection records the equivalence of the residualized and raw adjoint range conditions, the role of m as a numerical preconditioner, the notation for distinguishing generic m from specific choices, and the trade-off between Moore–Penrose minimum-norm and minimum-variance implementations.

Lemma C.4 (Residualized representation of the primal–dual score)**.** *Let $b = m + h$ with $m, h \in \mathcal{H}_{Z_1}$. Then the calibrated orthogonal moment*

$$\Gamma_1(O; q, b) = D\{1 + q(X_1)\}\{Y - b(Z_1)\} + b(Z_1)$$

can be written equivalently as

$$\Gamma_1(O; q, m, h) = D(1 + q)(Y - m) + m - \{D(1 + q) - 1\}h.$$

This is an algebraic reparametrization of the same score; it does not change the population range condition $\ell_1 \in \mathcal{R}(A_1^\star)$.

Residual correction equation.

Fix $m \in \mathcal{H}_{Z_1}$ and write $b = m + h$. The dual equation $A_1^\star b = \ell_1$ is equivalent to

$$A_1^\star h = \ell_1 - A_1^\star m,$$

which by the adjoint identity is the moment condition

$$\mathbb{E}[D\{Y - m(Z_1) - h(Z_1)\}\, s(X_1)] = 0 \qquad \forall s \in L^2(X_1).$$

This unweighted condition implies the weighted first-variation condition

$$\mathbb{E}[D\{1 + q(X_1)\}\{Y - m(Z_1) - h(Z_1)\}\, s(X_1)] = 0$$

for any fixed q, because $1 + q(X_1)$ is X_1-measurable. The weighted form is used to verify Neyman orthogonality of the score in the q-direction. The reverse implication (from the weighted form to the unweighted form) requires additional conditions on $1 + q$, so the two forms are not in general equivalent.

Remark (m-independent form of (R2) and its relation to the (m, h) decomposition). The range condition (R2) can be expressed equivalently as $\rho_{1,m}(x) := \mathbb{E}[Y_1 - m(Z_1) \mid X_1 = x, D = 1] \in \mathcal{R}(A_1^\star)$ for any fixed $m \in \mathcal{H}_{Z_1}$. Although this expression appears m-dependent, it can be rewritten in an m-independent form. Since

$$\rho_{1,m} = \mathbb{E}[Y_1 \mid X_1, D = 1] - \mathbb{E}[m(Z_1) \mid X_1, D = 1] = \ell_1 - A_1^\star m,$$

we have $\rho_{1,m} \in \mathcal{R}(A_1^\star) \iff \ell_1 - A_1^\star m \in \mathcal{R}(A_1^\star) \iff \ell_1 \in \mathcal{R}(A_1^\star) + \mathcal{R}(A_1^\star) = \mathcal{R}(A_1^\star)$ (since $\mathcal{R}(A_1^\star)$ is a linear subspace). Hence $\rho_{1,m} \in \mathcal{R}(A_1^\star)$ does not depend on the choice of m, and (R2) reduces to the m-independent condition that the outcome regression ℓ_1 itself lies in the range of $A_1^\star$.

Equivalence between the (b) form and the (m, h) decomposition. The paper often uses the splitting $b = m + h$ for implementation convenience. For any $m \in \mathcal{H}_{Z_1}$, setting $h := b - m$ gives $A_1^\star h = A_1^\star b - A_1^\star m = \ell_1 - A_1^\star m = \rho_{1,m}$. Hence:

- **Theoretical (compact).** $b \in \mathcal{B}_1 := \{b \in \mathcal{H}_{Z_1} : A_1^\star b = \ell_1\}$; (R2) is $\ell_1 \in \mathcal{R}(A_1^\star)$.
- **Implementation (AIPW-style).** (m, h) decomposition with h as the m-dependent residual corrected through the dual moment.

The two formulations yield identical main results.

Remark (m-independence of the adjoint range condition). The condition $\rho_{1,m} \in \mathcal{R}(A_1^\star)$ is invariant under the choice of m within $\mathcal{H}_{Z_1}$: for any two $m, m' \in \mathcal{H}_{Z_1}$, $\rho_{1,m} - \rho_{1,m'} = A_1^\star(m' - m) \in \mathcal{R}(A_1^\star)$, so $\rho_{1,m} \in \mathcal{R}(A_1^\star) \iff \rho_{1,m'} \in \mathcal{R}(A_1^\star)$. In particular, $\rho_{1,m} \in \mathcal{R}(A_1^\star)$ for some $m \in \mathcal{H}_{Z_1}$ if and only if $\ell_1 \in \mathcal{R}(A_1^\star) + \mathcal{R}(A_1^\star) = \mathcal{R}(A_1^\star)$, recovering the original adjoint range condition. The operational benefit of m is therefore in finite-sample implementation rather than in the population-level identification statement.

Remark (Why separating the role of m clarifies the logic). The decomposition $b = m + h$ separates two distinct roles. The fixed term $m \in \mathcal{H}_{Z_1}$ is a user-chosen "base" outcome regressor that does *not* need to satisfy any range condition. The residual correction h absorbs the remaining identification content of the adjoint range condition: $A_1^\star h = \ell_1 - A_1^\star m =: \rho_{1,m}$. This decomposition has three

advantages. First, $\rho_{1,m}$ is the natural conditional-moment residual after partialling out a base regressor, so its (R2) form can be assessed graphically through residual diagnostics. Second, the variance bound of Theorem E.10 simplifies: the minimum-variance adjoint representer is characterized through a quadratic minimization over h alone. Third, the implementation via the OLS approximation discussed in Appendix B.3 becomes transparent: m is the OLS-projection part and h is the calibration-specific correction.

Remark (Notation: m^{base} vs $m_q^\dagger$ vs m). Throughout the paper, m may take any one of three roles, which are kept conceptually distinct:

1. $m^{\text{base}}(Z_1) := \mathbb{E}[Y \mid Z_1, D = 1]$ is a *baseline outcome regression on the treated subsample*, useful as a residualization preconditioner that does not depend on q. The decomposition $b = m^{\text{base}} + h$ is the preferred form for residualization analysis.

2. $m_q^\dagger(Z_1) := \mathbb{E}[D\{1 + q(X_1)\}Y \mid Z_1]$ is the *q-weighted auxiliary projection* of Definition B.1, useful as a variance-reducing component of the augmented IPW score. It depends on the choice of primal representative q.

3. $m \in \mathcal{H}_{Z_1}$ is used as a free Z_1-measurable function in the abstract residualization formula $b = m + h$ of Section 3.2; particular choices (such as $m = m^{\text{base}}$ or $m = m_q^\dagger$) are distinguished by the additional notation above.

The main identification theorem holds for any $b \in \mathcal{B}_1$; the choice of m is a numerical and statistical preconditioner that does not affect identification at the population level.

Remark (Numerical role of residualization). The conditions $Y_1 \in \mathcal{R}(A_1^\star)$ and $\rho_{1,m} \in \mathcal{R}(A_1^\star)$ are population-equivalent (Remark C.4). The choice of m is a numerical preconditioning device. Absorbing the Z_1-explainable variation in Y_1 may reduce the magnitude of the right-hand side in well-conditioned directions, but it does not guarantee a smaller inverse-norm correction unless the residual is also controlled in the Picard coordinates of the adjoint operator. Residualization therefore does not weaken the population identification condition $\ell_1 \in \mathcal{R}(A_1^\star)$; it provides a numerically preconditioned implementation of the same condition.

Remark (Implementation choice: Moore–Penrose minimum-norm vs. minimum-variance). The Sieve GMM implementation of this section adopts the *Moore–Penrose minimum-norm solution* for the calibrating functions ($q^\dagger := A_1^\dagger r_1$, etc.). This is the standard NPIV estimation strategy combined with Tikhonov regularization (λ_q, λ_m, λ_h terms) and ensures **computational stability in finite samples**.

As shown in Section E.4, however, the minimum-norm solution generally does **not coincide with the minimum-variance representative** of Theorem E.11. Hence the implementation of this section guarantees:

- **Identification correctness** ($\hat{\mu}_1 \to \mu_1$): the minimum-norm implementation is consistent for the target under the same population calibration conditions as the general theory, since Theorem 3.4's invariance ensures the choice of calibration representative does not affect the target value.

- $\sqrt{N}$ **inference**: this is guaranteed by the product-rate and regularization conditions of Theorem F.4, not by the minimum-norm choice itself.

- **Semiparametric efficiency**: not guaranteed by the minimum-norm choice. It is attained only if the selected calibration representative coincides with a canonical-gradient representative, or if the calibration class is singleton under additional injectivity conditions.

If efficiency is important when the representatives are nonunique, one can use the **extended Sieve GMM** based directly on the fixed-q variance minimization (66) of Theorem E.10, instead of the minimum-norm representative. The finite-dimensional formulation, its solution as a constrained quadratic program, a penalized version, and an iterative variant with outer q updates are given in Appendix F.4. The basic implementation in this section emphasizes computational stability through the minimum-norm choice; the implementation in Appendix F.4 extends this by incorporating variance minimization explicitly in the objective.

Reduced conditioning blocks.

For one-sided targets one may consider reduced conditioning blocks that drop part of $C = (S, Q)$. Such reductions are operator-level restatements; they require their own calibration implication and range conditions and are not automatic consequences of the full system.

Remark (Reduced range conditions are not automatic). Corollary 2.2 states that identification can also be carried out under smaller conditioning blocks C_1, but this does not follow automatically from the full formulation. For $C_1 = S$ or $C_1 = \kappa(S, Q)$, the reduced operators $\bar{A}_1$ and $\bar{A}_1^\star$ must separately satisfy reduced range conditions. In particular, if $\kappa(S, Q)$ collapses the directions of (S, Q) that matter for Y_1 or for the selection probability, then $\bar{\ell}_1 \notin \mathcal{R}(\bar{A}_1^\star)$ and regular identification fails. Appendix G.3 exhibits a finite-state DGP under which $C = S + Q$ is sufficient and another DGP under which the same $C = S + Q$ breaks identification.

Remark (Note on Q-dependence). By the selection rule, D depends on Y_0, so conditioning p_1 on (U, Y_1, S, Q) introduces Q-dependence. For the target $\mu_1 = \mathbb{E}[Y_1]$ alone, one may work with a reduced conditioning index $C_1 = S$ (or more generally $C_1 = \kappa(S, Q)$) and use $q_1(W, Y_1, C_1)$, $m_1(V, C_1)$, $h_1(V, C_1)$ in implementation *only if* the reduced calibration implication, the reduced primal range condition, and the reduced adjoint range condition all hold separately for that reduced problem. This is a target-specific identification problem and is not an automatic consequence of the full $C = (S, Q)$ formulation. To maintain notational symmetry across the Y_1 and Y_0 sides, the main definitions retain Q throughout; the reduced formulation is discussed as a special case in Section 2.2.

Remark (Bilateral independence and the role of reduced DAGs). The bilateral independence condition is not part of the main calibration theorem. It is a sufficient condition that makes one-sided reductions economically interpretable. Without it, the full $C = (S, Q)$ formulation remains the safe formulation.

C.5 Extended roadmap

This supplements the boxed roadmap in Section 3; each step is stated there in compressed form and recorded here for reference.

Step 1 (latent odds anchor). A latent selection-odds representer q^ℓ exists by (R1-A) and satisfies the latent moment $\mathbb{E}[Dq^\ell(W, Y_t, C) - (1 - D) \mid U, Y_t, C, V] = 0$ (Assumption 2.2); this, not the observed balancing equation, is the causal IPW anchor.

Step 2 (observed calibration). Iterating expectations gives $A_1 q^\ell = r_1$, so $q^\ell \in \mathcal{Q}_1$ (Lemma 3.1).

Step 3 (nonuniqueness). $\mathcal{Q}_1$ is generally not a singleton because $\ker(A_1) \neq \{0\}$, so the plug-in identification value may vary across representatives.

Step 4 (target-invariance certificate). The dual equation $A_1^\star b = \ell_1$ certifies that μ_1 is invariant over $\mathcal{Q}_1$, extending the causal IPW identity from q^ℓ to arbitrary $q \in \mathcal{Q}_1$.

Step 5 (trichotomy). The position of ℓ_1 relative to $\overline{\mathcal{R}(A_1^\star)}$ gives Case N, Case I, or Case R, with finite-norm b available only in Case R (Proposition 3.5).

Step 6 (residualization as preconditioning). The decomposition $b = m+h$ for arbitrary $m \in \mathcal{H}_{Z_1}$ is an algebraic reparametrization that preconditions the empirical inverse problem; it does not weaken the population range condition.

Step 7 (exact SGMM versus diagnostic OLS). The formal inference theory covers the full $L^2(X_1)$-adjoint moment for h; reduced-space projections of the adjoint moment are useful diagnostics but are not exact estimators unless they satisfy the full adjoint moment.

D Operators, primitive conditions, and source conditions

This appendix collects the operator-theoretic primitives that underlie the calibration identification theory: a dictionary of the latent and observed operators, primitive sufficient conditions for the latent existence condition (R1-A), and singular-value systems and source conditions governing the observed adjoint range condition (R2) and the rates of regularized inverse estimators.

D.1 Operator dictionary

This appendix collects, in one place, the three operators used throughout the paper and the basic Hilbert-space objects (ranges, kernels, Picard summabilities) attached to each. Three distinct operators are used; they live on different Hilbert spaces and play different roles.

The latent measurement operator T_W.

For (R1-A), the relevant operator is defined for each $t \in \{0,1\}$ by

$$T_{W,t} : \mathcal{H}_{W,Y_t,C} \to \mathcal{H}_{U,Y_t,C}, \qquad T_{W,t}\, q(u, y_t, c) := \mathbb{E}[q(W, y_t, c) \mid U = u, Y_t = y_t, C = c].$$

We write T_W for $T_{W,t}$ when the side is clear from context. Range: $\mathcal{R}(T_W) \subseteq \mathcal{H}_{U,Y_t,C}$. Kernel: $\ker(T_W) \subseteq \mathcal{H}_{W,Y_t,C}$. Adjoint: $T_W^\star$ acts on $\mathcal{H}_{U,Y_t,C} \to \mathcal{H}_{W,Y_t,C}$.

(R1-A) is the latent existence condition: the latent odds ratio $r^\ell(U, Y_t, C) = (1-p_t)/p_t$ lies in $\mathcal{R}(T_W)$, equivalently a finite-norm $q^\ell \in \mathcal{H}_{W,Y_t,C}$ satisfies $T_W q^\ell = r^\ell$. In a singular system $T_W e_j = \sigma_j f_j$, writing $r^\ell = \sum_j a_j f_j$, the existence condition is the *latent Hermite-Picard summability*

$$\sum_j \frac{a_j^2}{\sigma_j^2} < \infty.$$

Completeness of T_W (all $\sigma_j > 0$) is injectivity, giving uniqueness of any latent representer that exists; it does *not* imply (R1-A) existence.

The observed primal operator A_1.

For (R1-B), the relevant observed operator is

$$A_1 : \mathcal{H}_{X_1} \to \mathcal{H}_{Z_1}, \qquad A_1 q(z_1) := \mathbb{E}[D\, q(X_1) \mid Z_1 = z_1].$$

Range: $\mathcal{R}(A_1) \subseteq \mathcal{H}_{Z_1}$. Kernel: $\ker(A_1) \subseteq \mathcal{H}_{X_1}$. (R1) is the observed primal range condition: $r_1(z_1) := \mathbb{E}[1-D \mid Z_1 = z_1] \in \mathcal{R}(A_1)$. The observed calibration set is $\mathcal{Q}_1 := \{q \in \mathcal{H}_{X_1} : A_1 q = r_1\}$; it is generally nonunique.

The observed dual (adjoint) operator $A_1^\star$.

For (R2), the relevant operator is

$$A_1^\star : \mathcal{H}_{Z_1} \to \mathcal{H}_{X_1}, \qquad (A_1^\star b)(x_1) := \mathbb{E}[b(Z_1) \mid X_1 = x_1,\, D = 1].$$

This is the Hilbert-space adjoint of A_1 with respect to the inner products

$$\langle a, b\rangle_{X_1} := \mathbb{E}[D\, a(X_1) b(X_1)], \qquad \langle c, d\rangle_{Z_1} := \mathbb{E}[c(Z_1) d(Z_1)],$$

as is verified by the calculation

$$\langle A_1 g, h\rangle_{Z_1} = \mathbb{E}[D g(X_1) h(Z_1)] = \mathbb{E}[D g(X_1) \mathbb{E}[h(Z_1) \mid X_1, D = 1]] = \langle g, A_1^\star h\rangle_{X_1}.$$

The implementation form of the dual moment equation is $A_1^\star b = \ell_1$, which under positivity is equivalent to $\mathbb{E}[D\{Y - b(Z_1)\} \mid X_1] = 0$. Range: $\mathcal{R}(A_1^\star) \subseteq \mathcal{H}_{X_1}$. Kernel: $\ker(A_1^\star) \subseteq \mathcal{H}_{Z_1}$. (R2) is the observed adjoint range condition: the outcome regression $\ell_1(x_1) := \mathbb{E}[Y \mid X_1 = x_1, D = 1] \in \mathcal{R}(A_1^\star)$. The observed dual solution set is $\mathcal{B}_1 := \{b \in \mathcal{H}_{Z_1} : A_1^\star b = \ell_1\}$. In a singular system $A_1^\star f_j = \sigma_j' e_j$, writing $\ell_1 = \sum_j \ell_j e_j$, (R2) is the *observed Picard summability*

$$\sum_j \frac{\ell_j^2}{(\sigma_j')^2} < \infty.$$

The three separations.

Throughout the appendix the following separations are maintained:

$$\text{latent (R1-A) on } T_W \;\neq\; \text{observed primal (R1) on } A_1 \;\neq\; \text{observed dual (R2) on } A_1^\star.$$

Each operator has its own Hilbert spaces, its own singular system, and its own range/kernel/Picard condition. The latent (R1-A) is generally non-testable from observed moments and must be verified through primitive sufficient conditions (Appendix D.2); the observed (R1-B) and (R2) are operator-range conditions on the observed joint distribution and admit empirical diagnostics through finite-dimensional sieve approximations, singular values, and regularization paths; these are diagnostics, not formal tests of infinite-dimensional range membership.

Adjoint identity.

The basic adjoint identity used in the primal–dual proofs is

$$\langle g, A_1^\star b\rangle_{X_1} = \langle A_1 g, b\rangle_{Z_1} \qquad \forall (g,b) \in \mathcal{H}_{X_1} \times \mathcal{H}_{Z_1}.$$

The orthogonal-complement identity $\overline{\mathcal{R}(A_1^\star)} = \ker(A_1)^\perp$ is the basis of the trichotomy (Section 3.3): $\ell_1 \in \ker(A_1)^\perp = \overline{\mathcal{R}(A_1^\star)}$ characterizes when the target is invariant over $\mathcal{Q}_1$ (i.e., not in Case N); $\ell_1 \in \mathcal{R}(A_1^\star)$ characterizes Case R (regular identification, finite-norm adjoint representer); $\ell_1 \in \overline{\mathcal{R}(A_1^\star)} \setminus \mathcal{R}(A_1^\star)$ characterizes Case I (irregular identification, closure only).

Causal direction note.

Throughout, T_W is a latent object whose range condition (R1-A) is not generally testable from the observed joint distribution; A_1 and $A_1^\star$ are observed-data operators whose range conditions (R1) and (R2) admit empirical diagnostics through finite-dimensional sieve approximations (singular value spectrum, Tikhonov regularization paths, moment residuals, dual norms). These diagnostics are not formal tests of population range membership without additional structure; see the discussion of diagnostics vs. formal tests in Section D.2. The latent-to-observed transfer is given by Lemma 3.1: the calibration implication of Assumption 2.2 transports a latent selection-odds representer satisfying $T_W q^\ell = r^\ell$ to a solution of the observed balancing moment $A_1 q^\ell = r_1$.

D.2 Latent primitives for (R1-A)

This subsection collects primitive sufficient conditions for the latent existence condition (R1-A), which lives on the latent measurement operator $T_W : \mathcal{H}_{W,Y_t,C} \to \mathcal{H}_{U,Y_t,C}$ (defined for each $t \in \{0,1\}$). These conditions are properties of the latent structural model and are generally not testable from observed data alone.

Non-testability of (R1-A).

The latent anchor condition (R1-A) is generally not testable from observed data alone. This non-testability is intrinsic to all proxy-based identification frameworks and applies equally to the present auxiliary-measurement framework, proximal causal inference, and classical missing-not-at-random models.

Formal statement. Theorem D.9 and its Corollary D.10 formalize this constructively: two admissible stochastic Roy models—one satisfying (R1-A), one violating it—can give rise to the same observed joint distribution. The auxiliary-measurement framework therefore relies on (R1-A) as a primitive sufficient condition rather than as an empirically testable restriction.

Practical implications.

1. Verify (R1-A) via primitive sufficient conditions (Gaussian latent factor, non-Gaussian deconvolution, ordinary-smooth/supersmooth measurements, finite-support rank).

2. Treat empirical diagnostics (singular-value spectrum, Tikhonov path stability, dual norm size) as first-stage signals rather than formal tests.

3. Report sensitivity analyses to calibration specification choices in applied work.

A genuine Branch-G linear–Gaussian worked example with selection on gains and exactly characterized calibrators.

The primitive conditions above are illustrated by a Gaussian generalized Roy model—continuous in the treated arm, with a deterministic untreated component—in which selection depends *directly on the realized gain* $Y_1 - Y_0$, a genuine Branch-G design, and in which the observed Branch-G calibrator $q_1^\ell(W, Y_1)$, the untreated-side calibrator $q_0^\ell(Y_0)$, and the adjoint representers $b_t(V_t)$ are exactly characterized: the treated-side calibrator in closed form, the untreated-side calibrator via a one-dimensional Gaussian–logistic integral, and the adjoints exactly. Consequently (R1-A), (T), and (R2) can be verified exactly rather than numerically. The V_t's should be interpreted as pre-treatment noisy forecasts or expert assessments of the two potential outcomes; the example is a constructive existence benchmark, not the maintained HRS design. Let $U \sim \mathcal{N}(0,1)$ and

$$Y_0 = \beta_0 U, \qquad Y_1 = \alpha + \beta_1 U + \varepsilon_1, \qquad \varepsilon_1 \sim \mathcal{N}(0, \sigma_1^2) \perp U,$$

with $(\alpha, \beta_1, \beta_0, \sigma_1) = (1,\ 1.2,\ 0.5,\ 0.5)$; the absence of an idiosyncratic untreated shock is the design choice that delivers exact closed forms (given $\beta_0 \neq 0$, the untreated state pins the confounder, $U = Y_0/\beta_0$). Treatment is assigned by logistic selection on the gain,

$$\mathbb{P}(D = 1 \mid Y_1, Y_0, U) = \Lambda\big(\kappa_0 + \gamma U + \delta_g(Y_1 - Y_0)\big), \qquad \Lambda(x) = \frac{e^x}{1 + e^x},$$

with $(\kappa_0, \gamma, \delta_g) = (0,\ 0.4,\ 0.4)$ and $\delta_g \neq 0$, so selection loads on the potential outcomes themselves. The measurement and side-specific adjoint blocks are

$$W = U + \sigma_W \zeta_W, \qquad V_t = Y_t + \sigma_V \zeta_{V_t}\ (t = 0, 1), \qquad \sigma_W = \sigma_V = 0.3,$$

with independent standard normal noises: W proxies the confounder and V_t is a noisy forecast of the potential outcome, excluded from selection and outcomes given the latent state. By construction ATE $= \alpha = 1$ and the treated share is 0.589.

Exact calibrators. Substituting $Y_0 = \beta_0 U$, the one-sided selection probability is exactly $p_1(U, Y_1) = \Lambda(\kappa_0 + \tilde{\gamma} U + \delta_g Y_1)$ with $\tilde{\gamma} := \gamma - \delta_g \beta_0 = 0.2$, so the latent odds are exp-linear, $r_1^\ell = \exp\{-\kappa_0 - \tilde{\gamma} U - \delta_g Y_1\}$. Because $\mathbb{E}[e^{-\tilde{\gamma} W} \mid U, Y_1] = e^{-\tilde{\gamma} U + \tilde{\gamma}^2 \sigma_W^2/2}$, the *observed* Branch-G calibrator solving (R1-A), $\mathbb{E}[q(W, Y_1) \mid U, Y_1] = r_1^\ell$, is

$$q_1^\ell(W, Y_1) = \exp\big\{ -\kappa_0 - \delta_g Y_1 - \tilde{\gamma} W - \tilde{\gamma}^2 \sigma_W^2/2 \big\},$$

in which the realized outcome enters the calibration weight—the defining Branch-G feature. On the untreated side, $U = Y_0/\beta_0$ is $\sigma(Y_0)$-measurable, so with $\bar{p}_1(u) := \mathbb{E}_{\varepsilon_1} \Lambda\big(\kappa_0 + \delta_g \alpha + (\tilde{\gamma} + \delta_g \beta_1) u + \delta_g \varepsilon_1\big)$ (a one-dimensional Gaussian–logistic integral), the exact untreated calibrator is $q_0^\ell(Y_0) = \bar{p}_1(Y_0/\beta_0)/\{1 - \bar{p}_1(Y_0/\beta_0)\}$. The transfer condition (T) holds side-specifically: given (U, Y_1) the treatment indicator is independent of (W, V_0, V_1), so the Y_1-side transfer holds conditioning on the full adjoint block, while given (U, Y_0) treatment still depends on ε_1, with which V_1 is correlated, so the Y_0-side transfer holds with the side-specific block $Z_0 = (V_0)$—consistent with the side-specific conditioning of Appendix 2.2. Finally, $\mathbb{E}[V_t \mid W, Y_t, D = t] = Y_t$, so $b_t(V_t) = V_t$ solves the adjoint equation (R2) exactly on each side. All of these are population identities; at $N = 2 \times 10^6$ they are confirmed to Monte Carlo precision: the anchored moments $\mathbb{E}[D\{1 + q_1^\ell\} f]$ match $\mathbb{E}[f]$ with gaps at most 1.3×10^{-3} over $f \in \{1, U, Y_1, UY_1, V_1\}$ (oracle $\widehat{\mu}_1 = 0.9999$ against 1), the untreated

anchor matches with gaps at most 1.1×10^{-3} ($\widehat{\mu}_0 = 0.0010$ against 0), and the exact-adjoint dual moments $\mathbb{E}[D_t\, h(W,Y)(Y - V_t)]$ are at most 2.1×10^{-4} in absolute value over $h \in \{1, W, Y, WY\}$.

Sieve recovery and benchmarks. The feasible proxy primal–dual calibration of the main text—the calibrated orthogonal moment Γ_1 with side-specific $Z_t = (V_t)$, a degree-three tensor-polynomial sieve on $X_t = (W, Y)$, a degree-four sieve on V_t, and relative Tikhonov parameter $\lambda = 10^{-5}$—recovers the truth: $\widehat{\mathrm{ATE}} = 0.993$ with Monte Carlo standard deviation 0.002 across five independent draws of size 2×10^6, primal balancing residual ≈ 0.016, adjoint residual ≈ 0.012, and primal condition number $\approx 7 \times 10^4$. The naive difference in means returns 1.52 and a proxy inverse-probability weight that adjusts for the confounder proxy W but ignores the gain channel returns 1.08, overstating the effect by about 52% and 8% respectively. A deliberately coarse sieve (degrees two and three, $\lambda = 10^{-3}$) returns 0.946 with primal residual 0.027 and adjoint residual 0.039: the bias shrinks from -0.054 to -0.007 as both first stages improve, the product-rate phenomenon quantified next.

Exact product-bias decomposition. Because $(q_1^\ell,\, b_1^\circ)$ with $b_1^\circ = V_1$ is a doubly exact reference pair, the population bias of the calibrated moment at *any* candidate pair (q, b) factorizes with no remainder: writing $\Delta q = q - q_1^\ell$ and $\Delta b = b - V_1$,

$$\mathbb{E}[\Gamma_1(O; q, b)] - \mu_1 = \underbrace{\mathbb{E}[D\,\Delta q\,(Y - V_1)]}_{=0 \text{ exactly}} + \underbrace{\mathbb{E}[\{1 - D(1 + q_1^\ell)\}\,\Delta b]}_{=0 \text{ exactly}} - \mathbb{E}[D\,\Delta q\,\Delta b],$$

where the first term vanishes because $Y - V_1 = -\sigma_V \zeta_{V_1}$ is independent of (D, W, Y_1) and the second vanishes by the exact transfer (T) at q_1^ℓ conditional on a block containing V_1. The Monte Carlo factorial below reports the bias of $\widehat{\mu}_1$ and the product term $-\mathbb{E}[D\Delta q \Delta b]$ across oracle, sieve-estimated, and deliberately misspecified nuisances (a Branch-U-style weight $q_{\mathrm{mis}}(W)$ that omits the outcome, and a constant non-adjoint $b_{\mathrm{mis}} \equiv \mathbb{E}[Y \mid D = 1]$); their difference is Monte Carlo and in-sample fitting noise of order 10^{-3}:

pair (q, b)	bias	$-\mathbb{E}[D\Delta q\Delta b]$	difference
oracle q_1^ℓ, oracle $b = V_1$	$+0.0006$	-0.0000	$+0.0006$
estimated $\widehat{q}$, oracle $b = V_1$	$+0.0006$	-0.0000	$+0.0006$
oracle q_1^ℓ, estimated $\widehat{b}$	$+0.0005$	-0.0000	$+0.0005$
estimated $\widehat{q}$, estimated $\widehat{b}$	-0.0057	-0.0061	$+0.0004$
misspecified $q_{\mathrm{mis}}(W)$, estimated $\widehat{b}$	-0.0078	-0.0082	$+0.0005$
misspecified $q_{\mathrm{mis}}(W)$, misspecified b_{mis}	$+0.0057$	$+0.0066$	-0.0009

Whenever either nuisance is exact the product term—and hence the bias—is zero regardless of how badly the other is specified, and when both are approximated the bias coincides with the product of the two first-stage errors: this is the population content of the double-robust product-rate guarantee of Appendix F.2, exhibited here without appeal to asymptotics. Two further features connect the example to the general theory. First, even in this benign continuous design the primal operator has condition number of order 10^4–10^5, underscoring that the ill-conditioning of the proxy inverse problem is intrinsic rather than an artifact of the HRS data. Second, the primal residual does not vanish at any fixed finite sieve because the exponential calibrator is not a polynomial; recovery is nonetheless accurate because the first-stage approximation errors enter the target only through the product-bias term, exactly as in the decomposition above. The same data-generating process is used in Appendix D.2 to verify the latent-calibration sensitivity identity under a known, controlled violation of the anchor. All numbers in this paragraph are

reproduced by `01_gaussian_worked_example.py` in the replication package.

Latent-calibration sensitivity analysis for the anchor conditions.

Because the anchor conditions (R1-A) and (T) are not testable (Theorem D.9), we report how the target moves as the *latent* anchor is allowed to fail by a controlled amount. The direction that matters is orthogonal to everything the observed calibration can detect: by the invariance theorem, moving within the observed calibration set $\mathcal{Q}_1$ cannot move the target when an adjoint representer exists, so a coherent sensitivity analysis must perturb the latent transfer moment itself, not the observed representative. Define, for any candidate weight $q \in \mathcal{H}_{X_1}$, the *latent calibration error*

$$\eta_1(q) \;:=\; \mathbb{E}\big[D\{1+q(X_1)\} - 1 \,\big|\, U, Y_1, C, V\big],$$

so that the maintained anchor (T) at q is exactly the statement $\eta_1(q) = 0$ a.s.; in general, the sensitivity parameter δ below bounds this latent conditional calibration error itself. In the special case in which the maintained calibrator is the exact odds representer of a working one-sided selection probability p_1 and the transfer conditioning holds for it, the error has the transparent form $\eta_1 = p_1^{\text{true}}/p_1 - 1$—the *relative* misspecification of the maintained selection probability—but this reading is specific to that case. The following identity is exact and requires no identifying assumption whatsoever.

Theorem D.1 (Exact latent-calibration bias identity)**.** *Fix any $\varphi \in \mathcal{H}_1$, any $q \in \mathcal{H}_{X_1}$, and any $b \in \mathcal{H}_{Z_1}$, and let $\Gamma_{1,\varphi}(O; q, b) = D\{1+q(X_1)\}\{\varphi(Y,C) - b(Z_1)\} + b(Z_1)$. Then*

$$\mathbb{E}[\Gamma_{1,\varphi}(O; q, b)] - \theta_1(\varphi) = \mathbb{E}\big[\eta_1(q)\,\{\varphi(Y_1, C) - b(Z_1)\}\big].$$

If in addition $q \in \mathcal{Q}_1$ (the observed balancing equation holds), then $\mathbb{E}[\eta_1(q)\, g(Z_1)] = 0$ for every $g \in \mathcal{H}_{Z_1}$, and hence for every $a \in \mathcal{H}_{Z_1}$,

$$\mathbb{E}[\Gamma_{1,\varphi}(O; q, b)] - \theta_1(\varphi) = \mathbb{E}\big[\eta_1(q)\,\{\varphi(Y_1, C) - a(Z_1)\}\big].$$

Proof. Both $\varphi(Y_1, C)$ and $b(Z_1)$ are measurable with respect to (U, Y_1, C, V), and $D\varphi(Y, C) = D\varphi(Y_1, C)$. Conditioning on (U, Y_1, C, V),

$$\mathbb{E}[D\{1+q\}\{\varphi(Y_1,C)-b\}] = \mathbb{E}\big[\{1+\eta_1(q)\}\{\varphi(Y_1,C)-b\}\big] = \theta_1(\varphi) - \mathbb{E}[b] + \mathbb{E}[\eta_1(q)\{\varphi(Y_1,C)-b\}],$$

and adding $\mathbb{E}[b]$ gives the first display. For the second, $q \in \mathcal{Q}_1$ means $\mathbb{E}[\{D(1+q)-1\}g(Z_1)] = 0$ for all g; conditioning the same way shows this equals $\mathbb{E}[\eta_1(q)g(Z_1)]$, so the map $a \mapsto \mathbb{E}[\eta_1(q)a(Z_1)]$ vanishes identically and b may be replaced by any $a \in \mathcal{H}_{Z_1}$. □

Two structural consequences follow. First, taking $\eta_1(q) = 0$ recovers the anchor step of Theorem 3.4: under (T) the calibrated moment is unbiased for *every* b. Second, for $q \in \mathcal{Q}_1$ the bias depends on b only through the latent direction, and optimizing over the free a yields bounds.

Corollary D.2 (Sensitivity bounds and breakdown)**.** *Let $q \in \mathcal{Q}_1$. (i) (L^2) $\big|\mathbb{E}[\Gamma_{1,\varphi}] - \theta_1(\varphi)\big| \le \|\eta_1(q)\|_2 \inf_{a \in \mathcal{H}_{Z_1}} \|\varphi(Y_1, C) - a(Z_1)\|_2$. (ii) ($L^\infty$) If $|\eta_1(q)| \le \delta$ a.s. and $\varphi(Y_1, C) \in [\underline{\varphi}, \overline{\varphi}]$ a.s., then choosing $a \equiv (\underline{\varphi} + \overline{\varphi})/2$,*

$$\big|\mathbb{E}[\Gamma_{1,\varphi}] - \theta_1(\varphi)\big| \;\le\; \delta\,(\overline{\varphi} - \underline{\varphi})/2.$$

(iii) (ATE) If $Y_t \in [0, M]$ a.s. and $|\eta_t(q_t)| \le \delta$ a.s. on both arms for the implemented $q_t \in \mathcal{Q}_t$, then the population calibrated ATE $\tau_{\text{cal}} := \mathbb{E}[\Gamma_{1,\varphi_y}] - \mathbb{E}[\Gamma_{0,\varphi_y}]$ with $\varphi_y(u, c) = u$ (free of the adjoint choices, by the theorem) satisfies $|\tau_{\text{cal}} - \tau| \le M\delta$. Consequently, for any reported value τ°, the band $\tau^\circ \pm M\delta$ excludes zero if and only if $\delta < \delta^*(\tau^\circ) := |\tau^\circ|/M$.

The parameter δ has the same reading as the marginal-sensitivity parameter of Tan (2006) transported to the latent anchor: it budgets the latent conditional calibration error uniformly over the latent state (equal, in the exact-representer special case above, to the relative error of the maintained selection probability). The resulting band is a partial-identification statement in the sense of Manski (1990) for the population calibrated target, with point identification as the $\delta \to 0$ limit; a mean-square rather than sup-norm budget yields the L^2 bound (i) and is developed further in the companion fragility analysis (Hoshino et al., 2026). Unlike the superseded construction, nothing here contradicts the invariance theorem: within $\mathcal{Q}_1$ the target is rigid, and the interval widens only in the latent direction η_1, which the observed data cannot restrict.

HRS application. The wave-10 total cognition score is bounded, $Y \in [0, 35]$, so $M = 35$, and the Branch-G point estimate is $\widehat{\tau}^G = -1.504$ (Table 10). The breakdown value is

$$\delta^* = \frac{|\widehat{\tau}^G|}{M} = \frac{1.504}{35} \approx 0.043 :$$

the plug-in sensitivity band excludes zero for every $\delta < 0.043$—in the exact-representer reading, a 4.3% relative error in the maintained selection probability—and contains zero beyond it. Table 13 reports the arithmetic; Figure 2 plots the band. The band is the population formula of Corollary D.2(iii) evaluated at the point estimate—a *plug-in sensitivity region*: it is not a confidence interval and does not incorporate sampling, regularization, or weak-range uncertainty, which are assessed separately in Appendix D.2.

δ (sup-norm on η_1, η_0)	plug-in band $\widehat{\tau} \pm M\delta$	band excludes zero?
0	$\{-1.504\}$	yes
0.010	$[-1.854, -1.154]$	yes
0.025	$[-2.379, -0.629]$	yes
0.043 ($= \delta^*$)	$[-3.009, +0.001]$	boundary
0.050	$[-3.254, +0.246]$	no

Table 13: Plug-in latent-calibration sensitivity region for the HRS Branch-G estimate: bands $\widehat{\tau} \pm M\delta$ with $M = 35$, $\widehat{\tau} = -1.504$ (the population formula of Corollary D.2(iii) evaluated at the point estimate). Pure arithmetic; no re-estimation is involved. In general δ is a sup-norm bound on the latent conditional calibration error; the relative-probability reading applies only in the exact-representer special case.

Numerical verification of the identity. In the Branch-G worked example of Appendix D.2 the identity can be checked against a *known, controlled* violation: generate treatment from the perturbed probability $p_1^\delta = \Lambda\big(\kappa_0 + \tilde{\gamma}U + \delta_g Y_1 + \delta\, s(Y_1)\big)$ with $s(Y_1) = \tanh(Y_1 - \alpha)$, while the analyst maintains the $\delta = 0$ closed-form calibrator q_1^ℓ. The latent error is then available pointwise, $\eta_1 = \Lambda(\iota + \delta s)/\Lambda(\iota) - 1$ with ι the unperturbed index, so every term of Theorem D.1 can be computed. At $N = 2 \times 10^6$ (seed fixed across δ), with the deliberately non-adjoint choice $\tilde{b} \equiv \alpha$: for $\delta = (0.01, 0.025, 0.05, 0.10)$ the realized bias of the calibrated moment is $(+0.0029, +0.0084, +0.0175, +0.0355)$, the identity right-hand side $\mathbb{E}[\eta_1\{Y_1 - \tilde{b}\}]$ is $(+0.0036, +0.0090, +0.0180, +0.0357)$, and the Cauchy–Schwarz bound of Corollary D.2(i) is

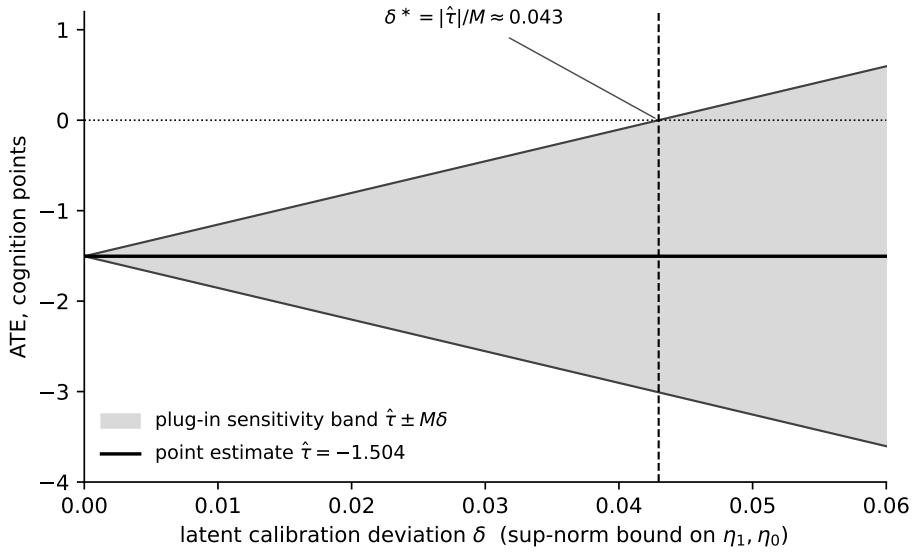


Figure 2: Plug-in latent-calibration sensitivity band for the HRS Branch-G estimate. Shaded region: band $\widehat{\tau} \pm M\delta$ as a function of the sup-norm latent deviation δ; dashed vertical line: breakdown value $\delta^* = |\widehat{\tau}|/M \approx 0.043$ at which the band first contains zero. In general δ bounds the latent conditional calibration error; the relative-probability reading is specific to the exact-representer case.

$(0.0042, 0.0105, 0.0209, 0.0416)$: the identity tracks the realized bias to Monte Carlo precision and the bound is valid throughout. With the exact adjoint $b = V_1$ instead, the realized bias stays at $+0.0006$ (the $\delta = 0$ Monte Carlo level) *uniformly in* δ: here $Y_1 - V_1 = -\sigma_V \zeta_{V_1}$ is independent of the latent state while this violation direction η_1 depends only on (U, Y_1), so $\mathbb{E}[\eta_1(Y_1 - V_1)] = 0$ exactly—an exact adjoint representer whose residual is orthogonal to the violation direction immunizes the functional against that violation. This immunity is a feature of the design (forecast-noise adjoints and a violation that does not load on V), not a general property; the general protection is the bound, not the zero. These computations are part of `01_gaussian_worked_example.py` in the replication package.

Weak-identification-robust inference for the calibrated functional.

At the empirical conditioning the first-stage calibration problem is near-singular (the irregular regime, Case I, is nearby), so the $\sqrt{N}$ Wald theory of Appendix F.2, which presumes the regular regime (Case R), can under-cover. A GMM-distance weak-calibration diagnostic is obtained by inverting a test of the calibrated moment restriction rather than a Wald statistic. The inversion below holds the estimated adjoint fixed and is therefore a diagnostic for weak identification of the *primal* channel; it does not address adjoint-side weakness. For a candidate value μ_1^0, form the calibrated orthogonal moment $\Gamma_1(O; q, b) - \mu_1^0$ together with the first-stage system $\{A_1 q - r_1,\ A_1^\star b - \ell_1\}$, and test $\mu_1 = \mu_1^0$ with a statistic robust to weak identification of the nuisance solutions (q, b)—a GMM-distance criterion on the stacked moments. Bennett et al. (2025) provide a general route to valid inference on strongly identified linear functionals under additional conditions; we do not implement their full procedure here, and no coverage guarantee from that construction is transferred to the simpler diagnostic below. The set is the collection of μ_1^0 not rejected at level α; it reduces to the Wald interval in Case R and can be unbounded when the calibration is weakly identified. It is a finite-sample diagnostic and is not established to have uniform coverage. This is the inferential

counterpart of removing the regular-regime intervals from the headline tables: the body reports the point estimate as a sensitivity magnitude, and the accompanying weak-calibration diagnostic is the inverted set developed here. We implement this for the HRS Branch G ATE as a feasible GMM-distance inversion. Holding the estimated adjoint representer $\widehat{b}$ fixed—the comparatively well-conditioned dual channel, so that the robust set targets weak identification of the *primal* calibration map, which is the binding ill-conditioning—the calibrated functional is linear in the primal calibration coefficients and the candidate moment is affine, so the GMM-distance statistic is quadratic in μ_1^0 and the robust set has closed-form endpoints; over-identifying the primal system ($L_q > K_q$, Appendix F.1) gives the statistic positive degrees of freedom. At the empirical conditioning—primal operator condition number about 4×10^6 in the present covariate coding, the same near-singular regime as the 8.6×10^5 of Table 34—the regular-regime Wald interval for the ATE is $[-2.67, -0.30]$ (point estimate -1.48, plug-in influence-function standard error 0.61, reproducing the headline -1.504 of Table 10), whereas the 95% diagnostic set is the entire real line: the minimized GMM-distance statistic is 4.90, far below $\chi^2_{9,0.95} = 16.92$, so no candidate ATE is rejected and the set becomes bounded only at confidence levels below about 16%. This is consistent with severe weak-range behaviour but does not by itself establish the population Case I—a poorly conditioned finite sample, a coarse sieve, or regularization can produce the same result—so the regular-regime interval is not established to be coverage-valid under the weak-range behavior observed in this application and -1.504 must be read as a regularized sensitivity magnitude (Appendix D.2), not a point-identified effect with valid frequentist inference. A coverage check on the HRS-calibrated stress design of Appendix H.8 (one-sided target $\mu_1 = 0.5$, $n = 1{,}597$, polynomial sieve, over-identified primal) is a finite-sample diagnostic exercise: across configurations spanning weak-to-strong primal identification the regular-regime Wald interval covers between about 2% and 94%—collapsing exactly where the calibration is weakly identified—while the diagnostic set holds coverage in a stable 80–90% band (below the nominal 95%), achieving this by widening and, in 13–63% of replications, declining to return a bounded interval. The residual under-coverage of the robust set (below the nominal 95%) reflects regularization bias, which an identification-robust device does not remove; consistent with the body's diagnosis, the bias-aware bounds for the regularized estimand remain the regularization-path range of Table 30 together with the latent-calibration sensitivity analysis of Appendix D.2. The robust set thus correctly refuses the spuriously tight claim the Wald interval makes, while uncertainty assessment for the regularized target is based on the regularization-path range and the latent-calibration sensitivity regions; neither is a coverage-valid confidence interval.

Gaussian completeness for (R1-A): W-side latent injectivity.

This subsection collects Gaussian primitives for (R1-A) using the *W-side* latent factor model. (R1-A) is a property of the latent measurement operator T_W acting on functions of W; the relevant completeness is therefore W-side, not V-side. V-side completeness and Gaussian primitives for the observed dual operator $A_1^\star$ are treated separately in Appendix D.3.

Proposition D.3 (Gaussian completeness of T_W)**.** *Under the W-side linear-Gaussian latent-factor model*

$$W = \alpha_W(Y_t, C) + \lambda_W(Y_t, C)U + \varepsilon_W, \qquad \varepsilon_W \perp U \mid Y_t, C, \tag{23}$$

with $\lambda_W(Y_t, C) \neq 0$ and $\mathrm{Var}(\varepsilon_W \mid Y_t, C) > 0$, the conditional law of W given (U, Y_t, C) is a

complete Gaussian location family in U. Equivalently, the latent measurement operator

$$T_W q(u, y_t, c) = \mathbb{E}[q(W, y_t, c) \mid U = u, Y_t = y_t, C = c]$$

is injective on L^2.

Structural role. Gaussian completeness of T_W gives *injectivity* (uniqueness of any latent representer that exists); it does *not* by itself imply existence of a finite-norm latent representer. (R1-A) existence requires the separate latent Hermite-Picard summability of the latent odds ratio r^ℓ relative to the singular values of T_W (Theorem D.6). In particular, completeness and Picard summability are logically independent: each can hold without the other.

Sketch. Under model (23), $W \mid U, Y_t, C \sim \mathcal{N}(\alpha_W(Y_t, C) + \lambda_W(Y_t, C)U, \mathrm{Var}(\varepsilon_W \mid Y_t, C))$ is a Gaussian location family in U. Completeness of Gaussian location families is standard (Lehmann, 1986), so T_W is injective. □

Proposition D.4 (Exponential-family completeness of T_W)**.** *A complete exponential-family conditional law for $W \mid U, Y_t, C$ with a one-to-one natural-parameter map in U implies injectivity of T_W.*

Remark (Completeness versus existence). Completeness of T_W gives uniqueness of any latent representer that exists; it does not give existence. Existence is the range condition $r^\ell \in \mathcal{R}(T_W)$ and is a separate object.

Remark (Beyond Gaussianity). The non-Gaussian completeness result establishes injectivity of T_W under deconvolution conditions on the marginal of W. The proof uses Fourier-transform characterizations of completeness due to Hu and Shiu (2018) and characterizes the strength of W as a measurement through the smoothness of its characteristic function.

Primitive sufficient conditions for (R1-A).

Proposition D.5 (Finite-support condition for (R1-A))**.** *Let (U, W, Y_t, C) have finite (or countable) support. For each support point (y, c) of (Y_t, C), let $\{u_1(y,c), \ldots, u_{K(y,c)}(y,c)\}$ and $\{w_1(y,c), \ldots, w_{J(y,c)}(y,c)\}$ be the conditional supports of U and W given $(Y_t, C) = (y, c)$, and define the block matrix and block vector*

$$B_{kj}(y,c) = \mathbb{P}\big(W = w_j(y,c) \mid U = u_k(y,c), Y_t = y, C = c\big), \qquad r_k^\ell(y,c) = \frac{1 - p_t(u_k(y,c), y, c)}{p_t(u_k(y,c), y, c)}.$$

Then (R1-A) holds if and only if

$$r_t^\ell(\cdot, y, c) \in \operatorname{col} B(y,c) \quad \text{for } P_{Y_t,C}\text{-almost every } (y,c),$$

together with square-integrability of one—equivalently the blockwise minimum-norm—solution, $\mathbb{E}\,\|B(Y_t,C)^+\, r_t^\ell(\cdot, Y_t, C)\|^2 < \infty$. Full row rank of $B(y,c)$ for almost every (y,c), with block dimensions and entries uniformly bounded and the latent odds bounded, is sufficient. This is a condition on the latent operator $T_{W,t}$; the observed operators A_1 and $A_1^\star$ remain separate objects governing (R1-B) and (R2).

Proof. The latent measurement operator acts blockwise: for each (y,c), $(T_{W,t}q)(\cdot, y, c) = B(y,c)\, q(\cdot, y, c)$, where $q(\cdot, y, c)$ collects the values $q(w_j(y,c), y, c)$. Existence of a solution to

$T_{W,t}q = r_t^\ell$ is therefore exactly the almost-everywhere column-space condition. The blockwise minimum-norm solution $q^\ell(\cdot, y, c) = B(y, c)^+ r_t^\ell(\cdot, y, c)$ is a measurable selection on the countable support of (Y_t, C); it lies in $L^2(W, Y_t, C)$ if and only if the displayed moment is finite, and any other solution differs from it by an element of the blockwise kernel, so square-integrability of some solution is equivalent to that of the minimum-norm one modulo kernel components, characterizing (R1-A). Under full row rank every block system is solvable, $B(y, c)^+$ has uniformly bounded norm when the block dimensions and entries are uniformly bounded below on their supports, and bounded latent odds then deliver the L^2 condition. □

Theorem D.6 (Gaussian-Hermite sufficient condition)**.** *Let $\widetilde{C}_t = (Y_t, C)$ denote the (R1-A) conditioning block. If $W = \rho(\widetilde{C}_t)U + \sqrt{1 - \rho(\widetilde{C}_t)^2}\eta$ conditionally on $\widetilde{C}_t$ and $r_{\widetilde{C}_t}(U) = \sum_j a_j(\widetilde{C}_t)H_j(U)$ satisfies*

$$E_{\widetilde{C}_t} \sum_j a_j(\widetilde{C}_t)^2 / \rho(\widetilde{C}_t)^{2j} < \infty, \tag{24}$$

then a latent selection-odds representer exists.

Proof. By the Mehler formula for the Gaussian conditional expectation operator, $\mathbb{E}[H_j(W) \mid U, \widetilde{C}_t = c] = \rho(c)^j H_j(U)$. For $q^\ell(W, c) = \sum_j a_j(c)H_j(W)/\rho(c)^j$,

$$\mathbb{E}[q^\ell(W, \widetilde{C}_t) \mid U, \widetilde{C}_t] = \sum_{j=0}^{\infty} a_j(\widetilde{C}_t)H_j(U) = r^\ell(U, \widetilde{C}_t).$$

Square-integrability follows from $\mathbb{E}_{\widetilde{C}_t} \sum_j a_j(\widetilde{C}_t)^2/\rho(\widetilde{C}_t)^{2j} < \infty$. □

Theorem D.7 (Non-Gaussian deconvolution condition)**.** *Let $\widetilde{C}_t = (Y_t, C)$ denote the (R1-A) conditioning block. If $W = U + \eta$ with $\eta \perp U \mid \widetilde{C}_t$, the conditional characteristic function $\varphi_{\eta,\widetilde{C}_t}$ has no zeros, and*

$$\mathbb{E}_{\widetilde{C}_t} \int \left| \widehat{r}_{\widetilde{C}_t}(s) / \varphi_{\eta,\widetilde{C}_t}(-s) \right|^2 ds < \infty, \tag{25}$$

where $\widehat{r}_{\tilde{c}_t}$ is the Fourier transform of $u \mapsto r_t^\ell(u, \tilde{c}_t)$, then (R1-A) holds with

$$q^\ell(w, \tilde{c}_t) = (2\pi)^{-d} \int e^{-is'w}\, \widehat{r}_{\tilde{c}_t}(s) / \varphi_{\eta,\tilde{c}_t}(-s)\, ds. \tag{26}$$

Proof. Conditionally on $\widetilde{C}_t = \tilde{c}_t$, the latent operator is $(T_{W,t}q)(u, \tilde{c}_t) = \mathbb{E}[q(W, \tilde{c}_t) \mid U = u, \widetilde{C}_t = \tilde{c}_t]$. Since $W = U + \eta$ with $\eta \perp U$ given $\widetilde{C}_t$, this is the convolution of $q(\cdot, \tilde{c}_t)$ with the reflected conditional density of η, so in Fourier space

$$\widehat{T_{W,t}q}(s, \tilde{c}_t) = \widehat{q}(s, \tilde{c}_t)\, \varphi_{\eta,\tilde{c}_t}(-s).$$

If $\varphi_{\eta,\tilde{c}_t}$ is zero-free and the inversion condition (25) holds, the inverse Fourier transform (26) defines a square-integrable calibrator with $T_{W,t}q^\ell = r_t^\ell$ for almost every $\tilde{c}_t$, hence (R1-A). □

Corollary D.8 (Ordinary-smooth and supersmooth measurements)**.** *The deconvolution condition is implied by the standard ordinary-smooth or supersmooth lower bounds on $|\varphi_{\eta,\widetilde{C}_t}|$ together with the corresponding Sobolev or exponential smoothness of the latent odds ratio (conditionally on $\widetilde{C}_t$).*

Ordinary-smooth versus supersmooth. Ordinary-smooth measurement errors have a polynomial-decaying conditional characteristic function (e.g., $|\varphi_{\eta,\widetilde{C}_t}(t)| \asymp |t|^{-\gamma}$), and supersmooth measurement errors have an exponentially-decaying conditional characteristic function (e.g., Gaussian, $|\varphi_{\eta,\widetilde{C}_t}(t)| \asymp \exp(-c|t|^{\alpha})$). This dichotomy is standard in the deconvolution literature (Fan, 1991). The (R1-A) sufficient condition adjusts accordingly: ordinary-smooth measurements require the existence-type source condition with $\beta \geq 1$ (or, for $0 < \beta < 1$, a weaker statement about smoothness of r^{ℓ} rather than strict finite-norm existence), while supersmooth measurements require additional analytic-extension conditions on the latent representer. This has direct implications for the convergence rate of the calibration estimator (Remark D.4): polynomial-rate convergence under ordinary-smooth measurements, but logarithmic-rate convergence under supersmooth measurements. All conditions in this paragraph are stated conditionally on $\widetilde{C}_t = (Y_t, C)$, on the measurement-error characteristic function $\varphi_{\eta,\widetilde{C}_t}$; C-only versions are the special case in which r_t^{ℓ} does not vary with Y_t beyond C.

Proof. For ordinary-smooth measurements (polynomial spectral decay), Theorem D.18 provides (R1-A) under the existence-type source condition $r^{\ell} = (T_W T_W^{\star})^{\beta/2} w$ with $\beta \geq 1$ (for $0 < \beta < 1$, the same statement yields smoothness of r^{ℓ} without strict finite-norm existence). For supersmooth measurements (exponential spectral decay, e.g., Gaussian deconvolution), (R1-A) holds whenever r^{ℓ} admits a sufficiently rapid Hermite-Picard summability. Specific verifiable conditions are given by Hermite-based completeness arguments analogous to Hall and Horowitz (2005). □

The following theorem is the paper's non-testability result, stated as an explicit observational-equivalence construction within the full stochastic generalized Roy model: the two models below are both admissible, generate *identical* joint laws of the observables (D, Y, W, V, C), and differ in whether (R1-A) holds. The construction preserves strict positivity without requiring an additional additive slack: a baseline overlap bound $\varepsilon \leq \pi \leq 1 - \varepsilon$ is reduced at most to $(1 - \delta)\varepsilon \leq \pi_{\delta} \leq 1 - (1 - \delta)\varepsilon$.

Theorem D.9 (Observational equivalence with and without (R1-A))**.** *Let $\mathcal{M}_0$ be any model with latent state (U, Y_1, Y_0, C), measurements (W, V), and* structural *selection probability $\pi(U, Y_1, Y_0, C) \in (0, 1)$ a.s. (the structural choice probability of Section 2)—so that D is generated conditionally on (U, Y_1, Y_0, C, W, V) by π, which may load directly on both potential outcomes, with the maintained exclusions of W and V from selection—and suppose $\mathcal{M}_0$ satisfies (R1-A) on the Y_1 side, with one-sided probability $p_1(U, Y_1, C) = \mathbb{E}[\pi \mid U, Y_1, C]$. Fix any measurable $\tilde{s}(U, Y_1, Y_0, C)$ with $|\tilde{s}| \leq 1$ such that $\bar{s} := \mathbb{E}[\min\{\pi, 1 - \pi\}\, \tilde{s} \mid U, Y_1, C]$ is not a.s. zero (any $\tilde{s} \geq 0$ with $\mathbb{P}(\tilde{s} > 0) > 0$ suffices), any $\delta \in (0, 1)$, and let R be a Rademacher variable ($\mathbb{P}(R = \pm 1) = \frac{1}{2}$) independent of (U, Y_1, Y_0, C, W, V). Define the enlarged model $\mathcal{M}_{\delta}$ with latent state $\tilde{U} = (U, R)$, unchanged laws of (U, Y_1, Y_0, C, W, V), and structural selection probability*

$$\pi_{\delta}(U, R, Y_1, Y_0, C) \;=\; \pi(U, Y_1, Y_0, C) \;+\; \delta\, R\, s(U, Y_1, Y_0, C), \qquad s := \min\{\pi,\, 1 - \pi\}\, \tilde{s}.$$

Then:

(i) $\mathcal{M}_{\delta}$ is admissible: $\pi_{\delta} \in (0, 1)$ a.s., and the joint law of (U, Y_1, Y_0, C, W, V) is unchanged.

(ii) $\mathcal{M}_0$ and $\mathcal{M}_{\delta}$ generate the same joint distribution of the observables (D, Y, W, V, C) (indeed of (D, Y_1, Y_0, W, V, C)).

(iii) In $\mathcal{M}_\delta$ the one-sided selection probability is $p_1^\delta(U,R,Y_1,C) = p_1(U,Y_1,C) + \delta R\,\bar{s}(U,Y_1,C)$, and (R1-A) fails with respect to the latent state $\tilde{U} = (U,R)$: no $q \in \mathcal{H}_{W,Y_1,C}$ satisfies $\mathbb{E}[q(W,Y_1,C) \mid \tilde{U},Y_1,C] = \{1-p_1^\delta\}/p_1^\delta$.

Proof. (i) Since $|R\,s| \le \min\{\pi, 1-\pi\}$ pointwise and $\delta < 1$, π_δ lies strictly between 0 and 1; the laws of the remaining primitives are unchanged by construction. (ii) Condition on (U,Y_1,Y_0,C,W,V). In $\mathcal{M}_\delta$, R is independent of the conditioning variables with $\mathbb{E}[R] = 0$, so

$$\mathbb{P}_{\mathcal{M}_\delta}(D=1 \mid U,Y_1,Y_0,C,W,V) = \mathbb{E}\big[\pi + \delta R s \,\big|\, U,Y_1,Y_0,C\big] = \pi(U,Y_1,Y_0,C),$$

identical to $\mathcal{M}_0$; since the conditioning law is also identical, the joint law of (D,U,Y_1,Y_0,C,W,V), hence of the observables, coincides. (iii) Averaging π_δ over (Y_0) given (U,R,Y_1,C), and using $R \perp (U,Y_1,Y_0,C)$, gives $p_1^\delta = p_1 + \delta R\,\bar{s}$. The latent odds $\{1-p_1^\delta\}/p_1^\delta$ take two distinct values in R on the positive-probability event $\{\bar{s} \neq 0\}$ (the map $p \mapsto (1-p)/p$ is strictly decreasing and p_1^δ differs across $R = \pm 1$ there). But for any candidate $q(W,Y_1,C)$, independence of R from (W,U,Y_1,C) gives

$$\mathbb{E}[q(W,Y_1,C) \mid U,R,Y_1,C] = \mathbb{E}[q(W,Y_1,C) \mid U,Y_1,C],$$

a version constant in R. A function constant in R cannot equal a function that varies in R on a positive-probability set, so the representability equation has no solution. □

The construction perturbs only the stochastic selection mechanism—an unobserved coin flip R tilting the individual structural selection probability—and leaves every observable moment untouched; because the baseline π may load directly on (Y_1,Y_0), the statement covers the full stochastic generalized Roy model maintained in this paper. It applies verbatim on the Y_0 side. Remark D.2 below gives the companion construction that isolates the transfer condition (T) while leaving (R1-A) intact. The practical reading is the one maintained in the text: (R1-A) and (T) are structural conditions to be defended on subject-matter grounds and stress-tested through the latent-calibration sensitivity analysis of Appendix D.2, whose parameter δ *upper-bounds* the latent calibration error induced by the perturbation constructed here: in $\mathcal{M}_\delta$ the maintained representer has $\eta_1 = \delta R\,\bar{s}/p_1$ with $|\bar{s}| \le p_1$, so $|\eta_1| \le \delta$.

Corollary D.10 (Non-testability of latent range (of Theorem D.9))**.** *Without additional latent structure, (R1-A) is not testable from the observed distribution alone.*

Proof. Immediate from Theorem D.9: every admissible model satisfying (R1-A) is observationally equivalent to a stochastic Roy model in which (R1-A) fails, so no observed-data test can discriminate the two. □

Remark (Non-testability of (T) under a strictly positive representer). The same device separates (T) from (R1-A). Suppose $\mathcal{M}_0$ additionally satisfies (T) at a representer q_1^ℓ with $1 + q_1^\ell \ge \underline{c}$ a.s. for some $\underline{c} > 0$ (a strictly positive calibration weight; the closed-form representer of the worked example satisfies $1 + q_1^\ell \ge 1$). Fix a bounded measurable $g(V)$ with $|g| \le 1$ that is not (U,Y_1,C)-measurable, set $m := \min\{\pi, 1-\pi\}$, center

$$c(U,Y_1,C) := \frac{\mathbb{E}[m\,g(V) \mid U,Y_1,C]}{\mathbb{E}[m \mid U,Y_1,C]} \in [-1,1],$$

and for $\delta \in (0, 1/2)$ define $\pi_\delta = \pi + \delta R\, m\{g(V) - c(U, Y_1, C)\}$, with R Rademacher independent of everything. Admissibility and observational equivalence follow as in the theorem. The centering makes $\mathbb{E}[\pi_\delta \mid U, R, Y_1, C] = p_1$ exactly, so the one-sided probability—hence (R1-A) and its representer set—is unchanged and R-free. But for every representer q in the strictly positive class $\{q : 1 + q \geq \underline{c} \text{ a.s.}\}$,

$$\mathbb{E}\big[D\{1 + q(X_1)\} \,\big|\, U, R, Y_1, C, V\big] = \mathbb{E}\big[\pi\{1 + q\} \,\big|\, U, Y_1, C, V\big] + \delta R\,\{g(V) - c\}\,\kappa,$$

$$\kappa := \mathbb{E}\big[m\{1 + q\} \,\big|\, U, Y_1, C, V\big] \;\geq\; \underline{c}\,\mathbb{E}[m \mid U, Y_1, C, V] \;>\; 0 \quad \text{a.s.},$$

so the transfer moment cannot equal one at both $R = +1$ and $R = -1$ on the positive-probability event $\{g(V) \neq c\}$. Thus $\mathcal{M}_\delta$ is observationally equivalent to $\mathcal{M}_0$, satisfies (R1-A) with the same representers, and violates (T) throughout the strictly positive class—in particular at the maintained q_1^ℓ, and at *every* representer whenever $T_{W,1}$ is injective, so that the representer is unique: (T) is untestable even granting (R1-A).

Example D.11 (The construction in the Branch-G worked example)**.** *In the worked example of Appendix D.2, take $p_1(U, Y_1) = \Lambda(\kappa_0 + \tilde{\gamma} U + \delta_g Y_1)$, $\tilde{s}(Y_1) = \tanh(Y_1 - \alpha)$, $\delta = 1/2$, and R an independent Rademacher variable. Here $\pi = p_1$ because $Y_0 = \beta_0 U$ is degenerate given U, and $\bar{s} = \min\{p_1, 1 - p_1\}\tanh(Y_1 - \alpha) \not\equiv 0$. The perturbed model draws D with probability $p_1 + \frac{1}{2} R \min\{p_1, 1 - p_1\}\tanh(Y_1 - \alpha)$: it is a legitimate stochastic Roy model, produces exactly the same joint law of (D, Y, W, V_0, V_1) as the original, and yet admits no calibrating function of (W, Y_1), because the perturbed odds genuinely vary with the coin R while $\mathbb{E}[q(W, Y_1) \mid U, R, Y_1]$ cannot. The closed-form treated-side calibrator q_1^ℓ of the unperturbed model is thus indistinguishable, on observables, from a world in which no calibrator exists.*

D.3 Observed primitives for (R1-B) and (R2)

This subsection collects primitive sufficient conditions for the observed-side conditions (R1-B) and (R2), which live respectively on the observed primal operator A_1 and its adjoint $A_1^\star$. These are properties of the observed-data model and admit finite-dimensional diagnostic checks.

Finite-support rank condition.

Let X_1 take values $x_1, \ldots, x_K$ and Z_1 take values $z_1, \ldots, z_L$. Define $M_X = \text{diag}\{P(D = 1, X_1 = x_k)\}$, $M_Z = \text{diag}\{P(Z_1 = z_\ell)\}$, and $A_{\ell k} = P(D = 1, X_1 = x_k \mid Z_1 = z_\ell)$. Then $Aq = r_1$ is the finite-dimensional balancing equation, while the adjoint equation is $M_X \ell_1 = A' M_Z b$.

Proposition D.12 (Finite-support rank characterization)**.** *A primal solution exists if and only if $r_1 \in \mathcal{R}(A)$. An adjoint representer exists if and only if $M_X \ell_1 \in \mathcal{R}(A' M_Z)$. Calibration uniqueness requires the relevant full-rank conditions.*

Proof. With discrete support, the weighted conditional-expectation operator A_1 is represented by a finite matrix $\mathbf{A}_1$ with entries $A_1[z, x] = \mathbb{P}(D = 1 \mid X_1 = x, Z_1 = z) \cdot \mathbb{P}(X_1 = x \mid Z_1 = z)$. The observed balancing equation $A_1 q = r_1$ has a solution if and only if $r_1 \in \text{col}(\mathbf{A}_1)$, equivalently $\text{rank}(\mathbf{A}_1) = \text{rank}([\mathbf{A}_1 | r_1])$. The adjoint $A_1^\star$ admits the row-space rank characterization, giving (R2) under the corresponding rank condition $\text{rank}(\mathbf{A}_1^\star) = \text{rank}([\mathbf{A}_1^\star | \ell_1])$. □

Logical independence from (R1-A).

The finite-support rank conditions above are observed-data conditions on A_1 and $A_1^\star$ and live on different Hilbert spaces from the latent operator T_W. For (R1-A), (R1-B), and (R2) to hold simultaneously, the *latent* matrix $\mathbf{B}$ and the *observed* matrix $\mathbf{A}_1$ each need their own rank condition; these are different matrices on different supports, and the two rank conditions are logically independent. In particular, observed rank diagnostics on $\mathbf{A}_1$ do not test (R1-A).

Gaussian/Hermite primitives for (R1-A) and (R2).

The Gaussian latent-factor model produces clean Hermite-spectral representations on *two distinct* operators that should not be conflated. These two operators live on different Hilbert spaces with different right-hand sides; the Hermite-Picard summabilities below correspondingly govern two different conditions.

Latent operator for (R1-A).

The latent measurement operator

$$T_W : \mathcal{H}_{W,Y_t,C} \to \mathcal{H}_{U,Y_t,C}, \qquad T_W q(u, y_t, c) = \mathbb{E}[q(W, y_t, c) \mid U = u, Y_t = y_t, C = c],$$

governs the latent existence condition. Write $\widetilde{C}_t = (Y_t, C)$ for the (R1-A) conditioning block. Under a linear-Gaussian latent factor model with $W = \rho(\widetilde{C}_t)U + \sqrt{1 - \rho(\widetilde{C}_t)^2}\,\eta$ and $\eta \perp U \mid \widetilde{C}_t$, the Mehler formula gives $\mathbb{E}[H_j(W) \mid U, \widetilde{C}_t = c] = \rho(c)^j H_j(U)$, so T_W has Hermite singular values $\sigma_j = \rho(c)^j$. Expanding the latent odds ratio as $r^\ell(U, \widetilde{C}_t) = \sum_j a_j(\widetilde{C}_t) H_j(U)$, the (R1-A) existence condition is the latent Hermite-Picard summability

$$\mathbb{E}_{\widetilde{C}_t} \sum_j \frac{a_j(\widetilde{C}_t)^2}{\rho(\widetilde{C}_t)^{2j}} < \infty.$$

Observed operator for (R2).

The observed dual operator $A_1^\star : \mathcal{H}_{Z_1} \to \mathcal{H}_{X_1}$ governs the observed adjoint range condition. In its singular system $A_1^\star f_j = \sigma_j' e_j$, write $\ell_1 = \sum_j \ell_j e_j$. Then (R2) is the observed Picard summability

$$\sum_j \frac{\ell_j^2}{(\sigma_j')^2} < \infty.$$

This is a condition on the observed operator and the observed outcome regression ℓ_1; it is not derivable from the latent (R1-A) summability above.

Completeness vs Picard summability.

Completeness on each side (all singular values nonzero) gives *injectivity* and hence uniqueness of any calibrating function that exists. Picard summability is a different statement: it gives *existence* of a finite-norm representer. The two conditions are logically independent: each can hold without the other. In particular, Gaussian completeness of T_W gives uniqueness of any latent representer but does not by itself imply (R1-A) existence; (R1-A) existence requires the latent

Hermite-Picard summability above.

Operational interpretation.

The latent Hermite-Picard summability for (R1-A) corresponds to the existence-type source condition $r^\ell = (T_W T_W^\star)^{\beta/2} w$ with $\beta \geq 1$ written in the Gaussian Hermite basis. The observed Picard summability for (R2) is a separate observed-data condition that must be verified independently and does not follow from properties of T_W alone.

Gaussian primitives for (R2): V-side observed dual variation.

(R2) is the observed adjoint range condition on $A_1^\star : \mathcal{H}_{Z_1} \to \mathcal{H}_{X_1}$, where $Z_1 = (V, C)$. The relevant Gaussian primitives are for the conditional law on the treated subsample, not for the W-side latent factor model above.

Proposition D.13 (Hermite-Picard primitive for (R2) under a Gaussian treated-law model)**.** *Suppose that, on the treated subsample $\{D = 1\}$, (X_1, Z_1) is jointly Gaussian conditional on (S, Q) with measurement-shifter correlation $\varrho \in (-1, 1)$. Then $A_1^\star$ is diagonalized by Hermite polynomials on each side with singular values $\sigma_j' = \varrho^j$. Writing $\ell_1 = \sum_j \ell_j e_j$ in this Hermite basis, (R2) is the observed Picard summability*

$$\sum_j \frac{\ell_j^2}{\varrho^{2j}} < \infty.$$

This is a smoothness condition on ℓ_1 relative to the measurement-shifter correlation and is logically independent of the W-side latent completeness of Proposition D.3.

Observed adjoint range.

Proposition D.14 (Observed operator representation of (R2))**.** *Let $C = (S, Q)$ and define, on the treated subsample, $K_c b = \mathbb{E}[b(V, c) \mid W, Y, C = c, D = 1]$. The adjoint range condition is equivalent to $\ell_c(w, y, c) = y$ belonging to $\mathcal{R}(K_c)$ for almost every c. In compact cases, the standard Picard condition characterizes existence of a finite-norm adjoint representer.*

Proof. $A_1^\star b = \ell_1$ requires that on $\{D = 1\}$,

$$\mathbb{E}[b(V, C) \mid W, Y, C, D = 1] = Y.$$

Conditioning on $C = c$, this becomes $K_c b_c = \ell_c$. In the compact case, the standard Moore–Penrose Picard condition characterizes the existence of a finite-norm adjoint representer. Finally, integrability follows from $\|b\|_{Z_1}^2 = \mathbb{E}[b(Z_1)^2] < \infty$ together with $\sum_j \ell_{c,j}^2 / \sigma_{c,j}^2 < \infty$ for the corresponding singular values. □

Role of V and the interpretation as an outcome-representation variable.

The variable V plays a dual role in the identification framework: it shifts the latent type U (the $V \to U$ path in Figure 1(a)), and it enters the adjoint-side Z_1-space as the source of variation that makes the adjoint operator $A_1^\star$ achieve the adjoint range condition. This subsection clarifies the

role of V and contrasts it with the role of W (the measurement) and with classical IV/LATE-type instruments.

V versus W: distinct identifying contributions. The measurement W enters the primal X_1-space, while V enters the dual Z_1-space. Both contribute to identification, but their roles are not interchangeable. As shown in Proposition G.1 and Proposition G.3, removing either side breaks the identification: without W, the latent odds ratio cannot be recovered from observed data; without V, the adjoint range condition fails because the adjoint operator has insufficient variation to make ℓ_1 representable. The two sides are complementary, not substitutable.

V versus classical IV. In classical IV/LATE-type identification, the instrument enters the treatment equation but is excluded from the outcome. The conditions $V \perp Y_t \mid U, C$ and exclusion of V from the outcome equation parallel the LATE exclusion restriction. However, the present framework imposes $V \to U$ as a positive identifying path rather than an exogenous shock to the participation margin: V is a latent-type shifter, not a treatment shifter. This distinction matters in two ways. First, V is allowed to be correlated with the outcome through U; it is not required to be independent of the outcome conditional on observed covariates only. Second, V identifies the ATE rather than the LATE: the calibration formulation pins down the population mean rather than a complier-specific local effect.

Examples of V in applications.

- In the HRS retirement application (Section 6), parental education V shifts the latent cognitive type U but is not a direct determinant of the respondent's retirement decision once cognitive type and labor-market environment are accounted for.
- In firm export-entry applications, founding-cohort industrial conditions or pre-entry productivity peer effects V shift the latent firm type without directly determining entry costs.
- In schooling-choice applications, regional college-density at the time of high-school graduation V shifts the latent type (academic preparation, exposure to higher education) without directly determining the financial cost of college.

In each case, V moves the latent-type distribution conditional on observed covariates but does not enter the participation rule or the measurement directly.

Strength of V as an outcome-representation variable. The strength of V as an adjoint-side identifier is measured by the singular values of the observed primal/adjoint operator pair $(A_1, A_1^\star)$, which share the same nonzero singular values. When the singular values decay slowly, V provides strong adjoint-side variation and the adjoint range condition is well-conditioned. When the singular values decay quickly (e.g., when V is nearly collinear with W conditional on C), the adjoint range condition is ill-conditioned and finite-sample inference becomes sensitive to weak-range effects. Empirical diagnostics for V strength include the empirical singular-value spectrum of $\widehat{A}_1$ and the Tikhonov-path stability of $\widehat{b}$.

D.4 Source, Picard, and rate conditions

This subsection records the singular-value system, source conditions, Picard summabilities, and the rate conditions governing the regularized inverse estimators that appear in the formal asymptotic theory.

Singular value system and source conditions.

SVD representation.

Proposition D.15 (Compactness from a square-integrable kernel)**.** *If A_1 admits a square-integrable kernel, then it is Hilbert-Schmidt and compact.*

Proof. If $A_1 : \mathcal{H}_{X_1} \to \mathcal{H}_{Z_1}$ admits the kernel representation

$$A_1 g(z) = \int g(x)\, k_A(z,x)\, d\mu_X(x), \qquad \iint k_A(z,x)^2\, dP_Z(z)\, d\mu_X(x) < \infty,$$

then A_1 is Hilbert–Schmidt and hence compact. One valid kernel representation, when the relevant conditional density exists, is

$$k_A(z,x) = \mathbb{P}(D = 1 \mid Z_1 = z)\, p_{X_1|Z_1,D=1}(x \mid z),$$

which factors the selection through the marginal treatment probability conditional on Z_1 and avoids double-counting selection. The adjoint $A_1^\star : \mathcal{H}_{Z_1} \to \mathcal{H}_{X_1}$ has the corresponding adjoint kernel on the treated subsample and is also Hilbert–Schmidt by the same argument. Compactness of both operators follows. □

Assumption D.1 (Compactness)**.** The operator A_1 is compact and admits a singular system (σ_j, e_j, f_j).

Proposition D.16 (SVD representation)**.** *The primal and adjoint representer equations reduce to the usual Picard conditions on the coefficients divided by singular values.*

Proposition D.17 (Compact trichotomy)**.** *Writing $\ell_1 = \ell_\perp + \sum_j \ell_j e_j$, if $\ell_\perp \neq 0$ the target is not invariant; if $\ell_\perp = 0$ but $\sum_j \ell_j^2/\sigma_j^2 = \infty$, point identification is irregular; if the sum is finite, regular pathwise identification is possible subject to finite variance.*

Proof. The range $\mathcal{R}(A_1^\star) = \operatorname{span}\{e_j : \sigma_j > 0\}$ has closure $\overline{\mathcal{R}(A_1^\star)} = \ker(A_1)^\perp$. The kernel $\ker(A_1)$ is spanned by the singular vectors corresponding to $\sigma_j = 0$. Expand $\ell_1 = \ell_\perp + \sum_{j:\sigma_j>0} \ell_j e_j$. If $\ell_\perp \neq 0$, then $\ell_1 \notin \overline{\mathcal{R}(A_1^\star)}$; this is calibration-based non-invariance. If $\ell_\perp = 0$ but $\sum_j \ell_j^2/\sigma_j^2 = \infty$, then $\ell_1 \in \overline{\mathcal{R}(A_1^\star)} \setminus \mathcal{R}(A_1^\star)$; this is irregular point identification. If the sum is finite, then $\ell_1 \in \mathcal{R}(A_1^\star)$ and the minimum-norm adjoint representer is $b^\dagger = \sum_j (\ell_j/\sigma_j) f_j$; this is regular pathwise identification. □

Remark (Economic interpretation)*.* Weak singular directions correspond to weak range variation.

Theorem D.18 (Source condition for (R1-A))**.** *A latent selection-odds representer exists if the latent odds ratio satisfies*

$$\sum_j a_j^2/\sigma_j^2 < \infty. \tag{27}$$

Then

$$q^\ell = \sum_j (a_j/\sigma_j) e_j. \tag{28}$$

Operational interpretation of the source condition. Two different source-condition statements are commonly used, and they have different meanings. The *existence-type source condition*

for (R1-A) is a condition on the right-hand side r^ℓ:

$$r^\ell = (T_W T_W^\star)^{\beta/2} w, \qquad w \in L^2,$$

which, by the Picard characterization, gives $\sum_j a_j^2/\sigma_j^2 = \sum_j \sigma_j^{2\beta-2} w_j^2$, finite when $\beta \geq 1$. For $0 < \beta < 1$, this representation gives smoothness relative to the operator but does not by itself imply strict range existence. The *smoothness-type source condition* $q^\ell = (T_W^\star T_W)^{\beta/2} w$ for some $w \in L^2$, on the other hand, presumes a solution q^ℓ exists and describes how smooth it is; it governs Tikhonov convergence rates for $\widehat{q}^\ell$. The parameter β in the smoothness-type condition controls how much smoother the calibration is than the ambient L^2 space; larger β means smoother calibration and faster convergence rate. The condition $\beta > 0$ is the minimal *smoothness-type* condition for Tikhonov-regularized estimation with a polynomial-rate bound; strict range existence (finite-norm solution) under the *existence-type* source condition typically requires $\beta \geq 1$. The smoothness condition $\beta > 0$ describes regularity of the solution for regularized estimation, conditional on existence, and does not by itself imply a $\sqrt{N}$ inference rate. If both first-stage estimators have a Tikhonov-type rate

$$r_n = n^{-\beta/(2\beta+2\nu+1)},$$

then the symmetric product-rate requirement $\sqrt{N}\, r_n^2 = o(1)$ for the target is

$$\beta > \nu + \tfrac{1}{2},$$

where ν captures the spectral-decay structure of $A_1^\star A_1$. Individual nuisance estimators attain parametric rates only under finite-dimensional or otherwise effectively parametric conditions. The condition $\beta = 2$ is the strongest source condition typically considered and corresponds to a "twice-smooth" calibration.

Spectral decay rates. The condition number of the regularized adjoint inverse is governed by the spectral decay rate ν of $A_1^\star A_1$. For finite-rank operators (discrete support, sieve approximation), $\nu = 0$ and the inverse is bounded. For Hilbert–Schmidt operators (smooth conditional densities), ν is a positive constant whose value depends on the smoothness of the kernel. For Gaussian deconvolution kernels (supersmooth measurements), $\nu = \infty$ in the sense that the spectral decay is faster than polynomial. The bias-variance trade-off in Tikhonov regularization is governed by the interaction between β (smoothness) and ν (decay), yielding the rate $n^{-\beta/(2\beta+2\nu+1)}$.

Proof. Let $T_W e_j = \sigma_j f_j$ be a singular system for the latent measurement operator T_W, and write $r^\ell = \sum_j a_j f_j$. By the Picard characterization, a finite-norm latent representer exists if and only if

$$\sum_j a_j^2/\sigma_j^2 < \infty,$$

in which case $q^\ell = \sum_j (a_j/\sigma_j) e_j$. Under the existence-type source condition $r^\ell = (T_W T_W^\star)^{\beta/2} w$ with $w \in L^2$, we have $a_j = \sigma_j^\beta w_j$, so

$$\sum_j a_j^2/\sigma_j^2 = \sum_j \sigma_j^{2\beta-2} w_j^2,$$

which is finite for all $w \in L^2$ when $\beta \geq 1$. For $0 < \beta < 1$, the representation gives smoothness of r^ℓ relative to the operator but does not by itself imply finite-norm latent-calibration existence.

Combined with the latent-to-observed transfer of Lemma 3.1, the existence-type source condition with $\beta \geq 1$ delivers (R1-A) and the corresponding observed calibration equation.

The companion *smoothness-type* source condition $q^\ell = (T_W^\star T_W)^{\beta/2} w$ is a different object: it presumes existence and describes how smooth the solution is. It controls the convergence rate of Tikhonov-regularized estimators (Remark D.4) but does not deliver existence on its own. □

Remark (Economic meaning). The source condition restricts the amount of latent selection information lying in weakly measured directions.

Remark (Structural sufficient conditions and diagnostics, not formal tests). The diagnostic quantities introduced in Section D.2 are not formal hypothesis tests of (R1-A) or (R2). They are first-stage diagnostic tools that probe the numerical range stability of the finite-dimensional regularized approximation to the observed inverse problem: whether the empirical operator is well-conditioned and whether the observed inverse problem exhibits weak range behavior. They are not formal tests of population range membership. Formal tests of these conditions would require additional latent-structure assumptions; the paper takes the position that the structural sufficient conditions of Appendix D.2 (Gaussian latent factor, non-Gaussian deconvolution, ordinary-/supersmooth measurements, finite support rank) provide the primary justification, with the diagnostics serving as empirical sanity checks.

Remark (Observed adjoint range and weak-range diagnostics). The observed adjoint range condition (R2) governs the boundedness of the adjoint representer b. Empirically, this is reflected in: (i) the empirical singular values of $\widehat{A}_1^\star$, which should not decay too fast; (ii) the Tikhonov regularization path of $\widehat{b}$, which should stabilize as $\lambda \to 0$; (iii) the dual norm $\|\widehat{b}\|_{Z_1}$ relative to the outcome variance σ_Y, which should be of order $O(1)$; and (iv) the dual moment residual $\|\widehat{m}_n^b\|$ scaled by σ_Y, which should be small. When (i)–(iv) jointly indicate weak-range structure (large condition number, slow Tikhonov stabilization, large dual norm, large residual), the practitioner should report bootstrap-based standard errors and treat the asymptotic inference with appropriate caution.

Rates.

Assumption D.2 (Source smoothness)**.** The selected calibration and adjoint representatives satisfy source conditions.

Assumption D.3 (Singular value decay)**.** The singular values decay at a specified mildly or severely ill-posed rate.

Remark (Regularized inverse-problem rates). Under the source smoothness (Assumption D.2) and singular-value decay (Assumption D.3) conditions, together with high-level analogues of the standard Tikhonov/PSMD stability conditions for the D-weighted calibration moment $A_1 q = r_1$, the Tikhonov/Sieve GMM estimators attain a rate of the same spectral form as in standard NPIV, namely $n^{-\beta/(2\beta+2\nu+1)}$, with β the source-condition exponent and ν the spectral decay rate of $A_1^\star A_1$. The constants, the effective sample size on $\{D = 1\}$, the conditional-density structure on $\{D = 1\}$, and the generated-regressor terms differ from standard NPIV; see Remark F. We state this as a rate heuristic rather than a primitive-level theorem, since the present paper uses it only to organize the bias–variance trade-off and not as a stand-alone inferential result.

Proposition D.19 (Generated first-stage rate)**.** *If $\hat{q}$ has rate $r_{q,N}$ and the oracle dual estimator has rate $r^0_{b,N}$, then $\|\hat{b} - b^\dagger\|_{D,2} = O_p\{r^0_{b,N} + \tau_{b,N} r_{q,N}\}$, and a sufficient product-rate condition is*

$$\sqrt{N} r_{q,N}\{r^0_{b,N} + \tau_{b,N} r_{q,N}\} = o(1). \tag{29}$$

Proof. The first-stage estimator $\widehat{q}$ converges at the Tikhonov rate of Remark D.4, denoted δ^q_n. The second-stage extended Sieve GMM moment for b uses the generated regressor $\hat{q}$ in place of the population $q^\dagger$. By the product-bias identity of Theorem F.3, the second-order bias of the plug-in $\widehat{\mu}_1$ is bounded by $\delta^q_n \cdot \delta^b_n$. The bias decomposition alone does not determine the second-stage estimator rate δ^b_n; the latter must be established separately from the Sieve GMM stability modulus and the generated-regressor perturbation. Plugging into the asymptotic representation of $\widehat{\mu}_1$ gives the stated rate. □

Product-rate examples.

Corollary D.20 (Mildly ill-posed case)**.** *If the calibrating functions are smooth enough relative to the ill-posedness indices, the product-rate condition holds.*

Proposition D.21 (Consolidated first-stage product rate with D-weighting and generated regressors)**.** *Maintain the source-smoothness and singular-value-decay conditions (Assumptions D.2–D.3) and the high-level Tikhonov/PSMD stability conditions underlying Remark D.4, with source exponent β and spectral-decay index ν for $A_1^\star A_1$. Then:*

1. *(*Primal rate.*) The regularized calibration estimator attains $\delta^q_n := \|\widehat{q} - q^\dagger\|_{D,2} = O_p\big(n^{-\beta/(2\beta+2\nu+1)}\big)$, the standard Tikhonov rate for the D-weighted operator A_1 (Engl et al., 2000; Chen and Pouzo, 2012; Hall and Horowitz, 2005).*

2. *(*D-weighted dual rate.*) Because the dual moment $D(1+q)(Y-b)$ vanishes on $\{D=0\}$, the oracle adjoint estimator is driven by the treated effective sample size $N_D := N \cdot p_D$, $p_D = \mathbb{P}(D=1)$, giving $\delta^{b,0}_n = O_p\big(N_D^{-\beta'/(2\beta'+2\nu'+1)}\big)$ for the dual source/decay indices (β', ν') of $A_1^\star$.*

3. *(*Generated-regressor correction.*) The feasible adjoint estimator uses $\widehat{q}$ in place of $q^\dagger$; by Proposition D.19, $\delta^b_n = O_p\big(\delta^{b,0}_n + \tau_{b,N}\,\delta^q_n\big)$ with $\tau_{b,N}$ the generated-regressor stability modulus.*

4. *(*Product rate.*) Consequently the second-order remainder of the product-bias identity (Theorem F.3) is $O_p(\delta^q_n\,\delta^b_n)$, and the $\sqrt{N}$ inference of Appendix F.2 is valid whenever $\sqrt{N}\,\delta^q_n\,\delta^b_n = o_p(1)$. When both sides attain the common Tikhonov rate this reduces to the Layer C threshold $\beta > \nu + \frac{1}{2}$; in the supersmooth-operator regime, where the singular values decay exponentially and the attainable rate is only logarithmic, the threshold fails and $\sqrt{N}$ inference is not available—the binding case flagged in the body.*

This consolidates Remark D.4, Proposition D.19, and the Layer C threshold into a single statement that makes the D-weighting and generated-regressor dependence explicit, as required for the product-rate condition to be primitive-level rather than purely high-level. The remaining gap to a fully primitive theorem is verification of the Tikhonov/PSMD stability modulus for the D-weighted operator under the conditional-density structure on $\{D=1\}$, which is standard under the maintained source and decay conditions (Chen and Pouzo, 2012).

Hierarchy.

Point identification, regular pathwise identification, and feasible root-N inference require progressively stronger assumptions.

D.5 Diagnostics versus formal tests

The singular-value system gives a unified summary of the identification and inference hierarchy developed in this paper. The hierarchy has three layers, each requiring progressively stronger conditions and yielding progressively stronger conclusions.

Layer A: Point identification (existence of the dual closure). The calibration implications and range conditions (R1-A) and the closure condition $\ell_1 \in \overline{\mathcal{R}(A_1^\star)}$ jointly imply point identification of μ_1 over the calibration set. In singular-value terms, decompose the target as $\ell_1 = \ell_\perp + \sum_j \ell_j e_j$ where $\ell_\perp \in \ker(A_1)$. The closure condition is $\ell_\perp = 0$ (the target is orthogonal to the kernel of A_1); it does *not* involve the Picard summability $\sum_j \ell_j^2/\sigma_j^2 < \infty$. The Picard summability characterizes the next layer (Layer B, regular identification): it is finite iff $\ell_1 \in \mathcal{R}(A_1^\star)$ and divergent iff ℓ_1 lies in $\overline{\mathcal{R}(A_1^\star)} \setminus \mathcal{R}(A_1^\star)$ (the irregular regime, Case I). Layer A is therefore the minimal identification statement: μ_1 exists and is unique, but the influence function may have infinite variance.

Layer B: Regular pathwise identification (finite-norm representation). The strict range condition $\ell_1 \in \mathcal{R}(A_1^\star)$ (Assumption (R2)) corresponds to the strict Picard summability $\sum_j |\langle \ell_1, e_j\rangle|^2/\sigma_j^2 < \infty$. Under this condition, a finite-norm adjoint representer $b \in \mathcal{H}_{Z_1}$ exists, the calibrated orthogonal moment has finite variance, and a regular influence function exists. This is the layer at which orthogonality and pathwise differentiability combine to give regular identification.

Layer C: Feasible $\sqrt{N}$ inference (estimable product rate). The product-rate condition $\delta_n^q \cdot \delta_n^b = o(n^{-1/2})$ on the first-stage nuisance estimators combines with Layer B's finite-variance condition (FV) to give $\sqrt{N}$-consistent asymptotically normal inference. More precisely, if the first-stage rates are $r_{q,n} = n^{-a_q}$ and $r_{b,n} = n^{-a_b}$, the threshold is

$$a_q + a_b > \tfrac{1}{2}.$$

When both nuisances achieve the same Tikhonov rate $r_n = n^{-\beta/(2\beta+2\nu+1)}$, the threshold reduces to

$$\beta > \nu + \tfrac{1}{2},$$

where β is the source-condition smoothness and ν controls the spectral decay rate of $A_1^\star A_1$. When β is small relative to ν (the supersmooth-operator, ordinary-smooth-calibration case), $\sqrt{N}$ inference is not achievable.

Summary table. The three layers and their singular-value characterizations are summarized below.

Layer	Sufficient condition	Conclusion
A. Point identification	$\ell_1 \in \overline{\mathcal{R}(A_1^\star)}$	μ_1 is identified over $\mathcal{Q}_1$
B. Regular pathwise	$\ell_1 \in \mathcal{R}(A_1^\star)$; finite variance (FV)	Finite-norm adjoint representer, regular IF
C. Feasible $\sqrt{N}$ inference	Layers A, B, plus product-rate condition	Asymptotic normality, efficient SE

This hierarchy clarifies what is at stake in each step of the identification and inference argument. Practitioners who can verify Layer B from primitive sufficient conditions (e.g., Gaussian latent factor, source-condition smoothness) and Layer C from empirical sieve-estimator rates can proceed to standard inferential procedures with confidence in the asymptotic theory.

Empirical diagnostics for each layer.

- Layer A: Check that the empirical $\widehat{\ell_1}$ is orthogonal to $\widehat{\ker}(A_1)$ via the dual moment residual.
- Layer B: Check that the Tikhonov-regularized $\widehat{b}$ stabilizes (does not diverge) as the regularization parameter $\lambda \to 0$.
- Layer C: Check that the nuisance estimator rates $\widehat{\delta}_n^q$ and $\widehat{\delta}_n^b$ satisfy the product-rate condition empirically; this can be assessed by sample-split cross-validation of the first-stage rates.

E Proofs of identification results

This appendix collects technical details behind Section 3.2. It explains why a simple augmented inverse-probability score is generally not Neyman orthogonal under selection on gains, how the residual correction h arises from the canonical adjoint representer $b = m + h$, how the finite-dimensional toy example illustrates nonunique calibration solutions with a unique target, and how the efficient influence function is characterized under an observed calibration model. The main text states the key identification theorem, trichotomy, and moment conditions; the present appendix records derivations and supplementary arguments.

The main theoretical object is the canonical adjoint outcome representer $b \in \mathcal{B}_1$. For implementation, it is often convenient to split $b = m + h$, where m is an auxiliary residualizing projection and h is a residual correction. The following subsections show how this splitting restores orthogonality.

E.1 Latent-to-observed transfer and primal representation

This subsection collects proofs related to the latent-to-observed transfer of Lemma 3.1 and the primal representation of the calibration weight: in particular, the sufficient bilateral-independence condition for the calibration implication, and the first-variation necessity of the primal and adjoint representer equations.

Additional proofs.

This subsection collects proofs of propositions deferred from the body.

Proof of Proposition 2.1.

It suffices to prove the Y_1 side; the Y_0 side is symmetric. Let $C_1 = (U, Y_1, C)$. The latent selection-odds representer satisfies

$$\mathbb{E}[q_1^\ell(W, Y_1, C) \mid C_1] = \frac{\mathbb{P}(D = 0 \mid C_1)}{\mathbb{P}(D = 1 \mid C_1)}.$$

By $W \perp\!\!\!\perp (D, V) \mid C_1$, the conditional expectation of q_1^ℓ is unchanged after additional conditioning on D and V. Moreover, $D \perp\!\!\!\perp V \mid C_1$ implies $\mathbb{P}(D = 1 \mid C_1, V) = \mathbb{P}(D = 1 \mid C_1)$. Hence

$$\mathbb{E}[Dq_1^\ell(X_1) \mid C_1, V] = \mathbb{P}(D = 1 \mid C_1, V)\mathbb{E}[q_1^\ell(X_1) \mid C_1, V, D = 1] = \mathbb{P}(D = 0 \mid C_1, V),$$

which gives the calibration implication in Assumption 2.2. The Y_0 side follows by symmetry.

Proof of Proposition 3.9.

For the q direction, the first variation is

$$\mathbb{E}[D\, a(X_1)\{Y - b(Z_1)\}] = \langle a,\, \ell_1 - A_1^\star b\rangle_{X_1}.$$

If this vanishes for every $a \in \mathcal{H}_{X_1}$, then $\ell_1 - A_1^\star b = 0$ in $\mathcal{H}_{X_1}$, i.e., $A_1^\star b = \ell_1$. For the b direction, the first variation is

$$\mathbb{E}[\{1 - D(1 + q(X_1))\}c(Z_1)] = \mathbb{E}\big[c(Z_1)\,\mathbb{E}\{1 - D(1 + q(X_1)) \mid Z_1\}\big].$$

If this vanishes for every $c \in \mathcal{H}_{Z_1}$, then $\mathbb{E}[D\{1 + q(X_1)\} - 1 \mid Z_1] = 0$, i.e., $A_1 q = r_1$. Thus the primal and adjoint representer equations are necessary for first-order orthogonality within the calibration signal class.

E.2 Primal-dual identification and invariance

This subsection collects the detailed proof of the primal-dual identification theorem and the symmetric Y_0-side construction.

Primal-Dual Identification Theorem.

Appendix Result (Primal–dual identification of μ_1 via a finite-norm adjoint representer)**.** *Let*

$$\mathcal{Q}_1 := \{q \in \mathcal{H}_{X_1} : A_1 q = r_1\}, \qquad \mathcal{B}_1 := \{b \in \mathcal{H}_{Z_1} : A_1^\star b = \ell_1\}. \tag{30}$$

Under the latent selection-odds representer anchor, observed calibration membership, and adjoint range condition, for any $(q, b) \in \mathcal{Q}_1 \times \mathcal{B}_1$,

$$\mu_1 = \mathbb{E}\Big[D\{1 + q(X_1)\}\{Y - b(Z_1)\} + b(Z_1)\Big] = \mathbb{E}[\Gamma_1(O; q, b)]. \tag{31}$$

Discussion of significance. The primal-dual identification theorem is the central identification result of this paper, and its content can be stated in three points.

First, strong completeness conditions are not required for point identification. The completeness

Assumption C.1, which appears in the classical NPIV literature (Newey and Powell, 2003; Chen and Pouzo, 2012), is not needed for point identification of μ_1. The sets $\mathcal{Q}_1$ and $\mathcal{B}_1$ may be non-singleton (the representatives q and b may be non-unique), but μ_1 is invariant over $\mathcal{Q}_1 \times \mathcal{B}_1$. The identification result therefore holds under the much weaker range conditions (R1) and (R2).

Second, the canonical adjoint representer b is a target-invariance certificate rather than an auxiliary regressor. The existence of b (i.e., the adjoint range condition R2, $\ell_1 \in \mathcal{R}(A_1^\star)$) is the boundary condition for *regular pathwise identification*. As shown by the trichotomy (Proposition 3.5), the existence of b marks the boundary between point identification (Cases I, R) and calibration-based non-identification (Case N), and between regular identification (Case R) and irregular identification (Case I).

Third, Neyman orthogonality is a structural consequence of the identification result. The orthogonal score $\psi_1 = \Gamma_1 - \mu_1$ arises naturally as the influence function representation of the invariant identification formula (9) over $\mathcal{Q}_1 \times \mathcal{B}_1$. Orthogonality is not an externally imposed technical condition but a property of the identification structure itself. This makes the DML implementation of the auxiliary-measurement framework natural rather than ad hoc.

Central message. The latent selection-odds representer gives the calibration set, while the adjoint outcome representer certifies invariance of the ATE functional over that set.

The (q, m, h) three-component decomposition. For implementation convenience, the adjoint representer b admits the decomposition $b = m + h$ where $m \in \mathcal{H}_{Z_1}$ is a user-chosen auxiliary projection (typically the OLS outcome regression $m(z) = \mathbb{E}[Y \mid Z_1 = z, D = 1]$) and $h := b - m$ is the residual correction. The residual satisfies $A_1^\star h = \ell_1 - A_1^\star m =: \rho_{1,m}$, where $\rho_{1,m}(x) := \mathbb{E}[Y_1 - m(Z_1) \mid X_1 = x, D = 1]$ is the conditional moment residual of $Y_1 - m$.

Under this decomposition, the identification formula becomes

$$\mu_1 = \mathbb{E}\Big[D\{1 + q(X_1)\}\{Y - m(Z_1)\} + m(Z_1) - \{D(1 + q(X_1)) - 1\}h(Z_1)\Big]. \tag{32}$$

This form mirrors the AIPW structure: $D(1+q)\{Y-m\}+m$ is the AIPW score with calibration weight, and $-\{D(1+q)-1\}h$ is the residual correction that absorbs the adjoint range condition. The (q, m, h) decomposition is convenient for implementation because m can be chosen flexibly without affecting identification, and only the correction h requires solving an NPIV-type problem. Residualization may improve numerical conditioning by removing well-explained components of the right-hand side, but does not guarantee a smaller inverse-norm correction unless the residual is also controlled in the Picard coordinates of the adjoint operator.

Proof. **Step 1: anchor at the latent selection-odds representer.** By the latent selection-odds representer condition (R1-A), a latent selection-odds representer $q^\ell \in \mathcal{H}_{X_1}$ exists, and by Lemma 3.1, $q^\ell \in \mathcal{Q}_1$. By Lemma 3.1 the latent odds calibration satisfies the IPW identity

$$\mu_1 = \mathbb{E}[D\{1 + q^\ell(X_1)\}Y]. \tag{33}$$

Step 2: augmentation by an arbitrary $b \in \mathcal{B}_1$. For any $b \in \mathcal{H}_{Z_1}$, since $q^\ell \in \mathcal{Q}_1$ implies $\mathbb{E}[D\{1 + q^\ell(X_1)\} - 1 \mid Z_1] = 0$, augmenting (33) by $b(Z_1)$ does not change the expectation:

$$\mu_1 = \mathbb{E}[D\{1 + q^\ell(X_1)\}Y] = \mathbb{E}\big[D\{1 + q^\ell(X_1)\}\{Y - b(Z_1)\} + b(Z_1)\big] = \mathbb{E}[\Gamma_1(O; q^\ell, b)].$$

This holds for every $b \in \mathcal{H}_{Z_1}$, not only for $b \in \mathcal{B}_1$.

Step 3: invariance over the calibration set. For any other $q \in \mathcal{Q}_1$, write $q = q^\ell + g$ with $g = q - q^\ell \in \ker(A_1)$. Then

$$\Gamma_1(O; q, b) - \Gamma_1(O; q^\ell, b) = D\, g(X_1)\{Y - b(Z_1)\}.$$

Taking expectations,

$$\mathbb{E}[D\, g(X_1)\{Y - b(Z_1)\}] = \langle g,\, \ell_1 - A_1^\star b\rangle_{X_1}.$$

If $b \in \mathcal{B}_1$, then $A_1^\star b = \ell_1$, and the right-hand side equals $\langle g,\, 0\rangle_{X_1} = 0$.

Step 4: conclusion. Combining Steps 1–3, for every $(q, b) \in \mathcal{Q}_1 \times \mathcal{B}_1$,

$$\mu_1 = \mathbb{E}[\Gamma_1(O; q, b)] = \mathbb{E}\big[D\{1 + q(X_1)\}\{Y - b(Z_1)\} + b(Z_1)\big].$$

This is the desired primal-dual identification formula. The choice of $(q, b) \in \mathcal{Q}_1 \times \mathcal{B}_1$ does not affect the value of the signal in expectation. The Y_0 side follows by defining the corresponding $A_0, A_0^\star, \mathcal{Q}_0, \mathcal{B}_0, \ell_0$ symmetrically. □

Remark (Meaning of Theorem 3.4). The theorem does not require point identification of q. The primal equation can have many solutions. The adjoint representer certifies that the ATE functional is invariant over that solution set. This is the central difference between identifying the nuisance representer itself and identifying the target functional.

Appendix Result (Identification in the (q, m, h) splitting)**.** *If $b = m + h$ and $b \in \mathcal{B}_1$, then for any $q \in \mathcal{Q}_1$,*

$$\mu_1 = \mathbb{E}\Big[D\{1 + q(X_1)\}\{Y - m(Z_1)\} + m(Z_1) - \{D(1 + q(X_1)) - 1\}h(Z_1)\Big]. \tag{34}$$

Proof. Substituting $b = m + h$ into the (q, b)-form of (9):

$$\begin{aligned}\Gamma_1(O; q, b) &= D(1 + q)\{Y - m - h\} + m + h \\ &= D(1 + q)\{Y - m\} - D(1 + q)h + m + h \\ &= D(1 + q)\{Y - m\} + m - \{D(1 + q) - 1\}h.\end{aligned}$$

In the final equality, $-D(1 + q)h + m + h = m - \{D(1 + q) - 1\}h$ is obtained by grouping the h terms. This is exactly the (q, m, h) form in (34), equivalent to the (q, b) form via the linear splitting $b = m + h$. □

Remark ((q, b) versus (q, m, h)). The (q, b) form is the canonical representation for identification, regularity, and variance minimization. The (q, m, h) form is an implementation device that connects the signal to augmented inverse probability weighting and residual correction. Both yield the same signal whenever $b = m + h$.

Example E.1 (Finite-dimensional toy example: nonunique q but unique μ_1)**.** *Let the observed balancing equation be $Aq = r$ with*

$$A = \begin{pmatrix} 1/4 & 0 & 1/4 \\ 0 & 1/4 & 1/4 \end{pmatrix}, \qquad r = \begin{pmatrix} 1/2 \\ 1/2 \end{pmatrix}.$$

Both $q^a = (2, 2, 0)'$ and $q^b = (0, 0, 2)'$ solve $Aq = r$. The null direction is $g = (1, 1, -1)'$. If $\ell = (1, 1, 2)'$, then $g'\ell = 0$ and the target is invariant over the solution set; indeed $A'h = \ell$

for $h = (4,4)'$. If instead $\ell = (1,1,1)'$, then $g'\ell = 1 \neq 0$, and the target changes along the observationally indistinguishable direction g. This finite example is the finite-dimensional version of the range-closure trichotomy below.

The Y_0 side and the ATE.

For the Y_0 side define

$$\mathcal{H}_{X_0} := L^2(P_{X_0|D=0}), \quad \langle a, b\rangle_{X_0} := \mathbb{E}[(1-D)a(X_0)b(X_0)], \tag{35}$$
$$\mathcal{H}_{Z_0} := L^2(P_{Z_0}), \quad \langle c, d\rangle_{Z_0} := \mathbb{E}[c(Z_0)d(Z_0)], \tag{36}$$
$$A_0 : \mathcal{H}_{X_0} \to \mathcal{H}_{Z_0}, \quad (A_0 g)(z) := \mathbb{E}[(1-D)g(X_0) \mid Z_0 = z], \tag{37}$$
$$A_0^\star : \mathcal{H}_{Z_0} \to \mathcal{H}_{X_0}, \quad (A_0^\star h)(x) := \mathbb{E}[h(Z_0) \mid X_0 = x, D = 0]. \tag{38}$$

Let $\mathcal{Q}_0$ and $\mathcal{B}_0$ denote the corresponding primal and dual solution sets.

Appendix Result (Primal-dual identification on the Y_0 side)**.** *For any admissible (q, h) in the (q, m, h) splitting on the Y_0 side,*

$$\mu_0 = \mathbb{E}\Big[(1-D)(1+q(X_0))\{Y_0 - m_0^{Y_0}(Z_0)\} + m_0^{Y_0}(Z_0) - \{(1-D)(1+q(X_0)) - 1\}h(Z_0)\Big]. \tag{39}$$

Proof. The proof is identical to the Y_1 side after replacing D by $1-D$. □

Appendix Result (Primal-dual identification of the ATE)**.** *Define*

$$\Gamma_1(O; q, m, h) := D(1+q(X_1))\big(Y_1 - m(Z_1)\big) + m(Z_1) - \big\{D(1+q(X_1)) - 1\big\}h(Z_1), \tag{40}$$
$$\Gamma_0(O; q, m, h) := (1-D)(1+q(X_0))\big(Y_0 - m(Z_0)\big) + m(Z_0) - \big\{(1-D)(1+q(X_0)) - 1\big\}h(Z_0). \tag{41}$$

Then

$$\mu_1 = \mathbb{E}[\Gamma_1(O; q_1, m_0^{Y_1}, h_0^{Y_1})], \qquad \mu_0 = \mathbb{E}[\Gamma_0(O; q_0, m_0^{Y_0}, h_0^{Y_0})], \tag{42}$$

and

$$\tau = \mu_1 - \mu_0 = \mathbb{E}\Big[\Gamma_1(O; q_1, m_0^{Y_1}, h_0^{Y_1}) - \Gamma_0(O; q_0, m_0^{Y_0}, h_0^{Y_0})\Big]. \tag{43}$$

Proof. Apply the one-sided identification formula to μ_1 and μ_0 and subtract. □

Remark (Identification signals and mean-zero scores). The quantities Γ_t are identification signals whose expectations equal μ_t. Mean-zero scores are obtained by subtracting the parameter, $\psi_t = \Gamma_t - \mu_t$. Estimators average $\widehat{\Gamma}_t$; variance estimators use centered scores.

Remark (Neyman orthogonality is a structural consequence, not an external condition). In many DML applications, Neyman orthogonality is imposed as an external requirement on the score: the practitioner chooses a moment that satisfies the orthogonality condition. Under the calibration identification framework of this paper, Neyman orthogonality of the calibrated orthogonal moment $\Gamma_1(O; q, b)$ is not an externally imposed requirement but a structural consequence of the primal-dual identification problem. Specifically, the orthogonal conditions on (q, b) correspond precisely to $A_1 q = r_1$ and $A_1^\star b = \ell_1$, which are the same equations that anchor identification. This is what makes the regular calibrated orthogonal moment canonical: under the calibration structure, no other orthogonal score gives the same regular-IF representation, and any deviation from $A_1 q = r_1$ or $A_1^\star b = \ell_1$ breaks both orthogonality and identification simultaneously.

Remark (Interpretation in the Gaussian latent factor model). Under the linear-Gaussian factor model of Appendix D.2, the primal and adjoint representers admit closed-form expressions. The latent odds ratio $(1-\pi(U,Y_1,C))/\pi(U,Y_1,C)$ (the odds of the untreated event, matching the Y_1-side observed calibration) admits a tractable Hermite expansion under simple linear-index treatment-choice rules, and the latent representer q^ℓ becomes a polynomial in W via Hermite expansion. In the Gaussian latent-factor benchmark, the W-side latent operator T_W can be diagonalized by Hermite polynomials. Completeness of T_W (i.e. nonzero singular values) gives *injectivity* and hence uniqueness of any latent representer that exists; it is a property of the latent operator and is generally not testable from the observed covariance matrix alone. Existence of a finite-norm latent representer remains a separate Hermite–Picard summability condition on the latent odds ratio. The observed adjoint range condition (R2) is a separate property of the observed adjoint $A_1^\star$; under a Gaussian *treated-law* model on (X_1, Z_1) it reduces to a rank condition on the cross-covariance on $\{D=1\}$ (Appendix D.3). The Gaussian benchmark thus provides a transparent illustration of the operator geometry, not a verifiable test of (R1-A) from observed covariance alone.

Definition E.1 (orthogonal scores)**.** The Y_1 score is

$$\boxed{\begin{aligned}\psi^{Y_1}_{\mathrm{ortho}}(O;\mu_1,q,m,h) =& D(1+q(X_1))(Y_1-m(Z_1))+m(Z_1)\\ &-\{D(1+q(X_1))-1\}h(Z_1)-\mu_1.\end{aligned}} \tag{44}$$

The Y_0 score is

$$\begin{aligned}\psi^{Y_0}_{\mathrm{ortho}}(O;\mu_0,q,m,h) =& (1-D)(1+q(X_0))(Y_0-m(Z_0))+m(Z_0)\\ &-\{(1-D)(1+q(X_0))-1\}h(Z_0)-\mu_0.\end{aligned} \tag{45}$$

E.3 Trichotomy and compact-operator characterization

This subsection collects the detailed analysis of the identification trichotomy, including the compact-operator characterization in terms of the position of ℓ_1 relative to $\mathcal{R}(A_1^\star)$ and $\overline{\mathcal{R}(A_1^\star)}$.

Identification trichotomy.

For a generic $\varphi \in \mathcal{H}_1$ replace Y by $\varphi(Y,C)$ and ℓ_1 by $\ell_{1,\varphi}$ throughout this paragraph; Theorem 3.3 in the main text is the general statement, and the singular-value detail below is unchanged.

Appendix Result (Target invariance over the calibration set (mean-case restatement of Theorem 3.3): necessity and sharpness)**.** *Assume $\mathcal{Q}_1 \neq \emptyset$. The linear functional $q \mapsto \mathbb{E}[Dq(X_1)Y]$ is invariant over $\mathcal{Q}_1$ if and only if*

$$\ell_1 \in \ker(A_1)^\perp = \overline{\mathcal{R}(A_1^\star)}.$$

Proof. **(N1) ⇒ (N2).** Anchor at $q^\ell \in \mathcal{Q}_1$. For any $g \in \ker(A_1)$, we have $q^\ell + g \in \mathcal{Q}_1$. The difference of the identification functionals is

$$\mathbb{E}[D(1+q^\ell+g)Y] - \mathbb{E}[D(1+q^\ell)Y] = \mathbb{E}[Dg(X_1)Y] = \mathbb{E}[g(X_1)\cdot\mathbb{E}[DY \mid X_1]] = \langle g, \ell_1\rangle_{X_1}.$$

Invariance requires this to be zero for any $g \in \ker(A_1)$, hence $\ell_1 \perp \ker(A_1)$, equivalently $\ell_1 \in \ker(A_1)^\perp = \overline{\mathcal{R}(A_1^\star)}$.

(N2) ⇒ (N1). If $\ell_1 \in \ker(A_1)^\perp$, then $\langle g, \ell_1 \rangle_{X_1} = 0$ for any $g \in \ker(A_1)$. For any $q, q' \in \mathcal{Q}_1$, the difference $g := q - q'$ lies in $\ker(A_1)$, so the functional values coincide.

Distinguishing (N2a) from (N2b). The membership $\ell_1 \in \mathcal{R}(A_1^\star)$ holds if and only if there exists a finite-norm $b \in \mathcal{H}_{Z_1}$ with $A_1^\star b = \ell_1$. If $\ell_1 \in \overline{\mathcal{R}(A_1^\star)} \setminus \mathcal{R}(A_1^\star)$, then an approximating sequence $\{b_n\}$ with $A_1^\star b_n \to \ell_1$ exists, but $\|b_n\| \to \infty$ (which is the Picard condition failure under compactness, Proposition D.17). □

Detailed statement of the trichotomy. Under Assumptions 2.1 (R1-A) and 2.2 (T) and Assumption 2.4, Lemma 3.1 implies $\mathcal{Q}_1 = \{q : A_1 q = r_1\}$ is nonempty. Consider further the observed outcome regression $\ell_1(x) := \mathbb{E}[Y \mid X_1 = x, D = 1] \in \mathcal{H}_{X_1}$ (the target of the m-independent adjoint range; see Remark C.4). Depending on the position of ℓ_1, the following trichotomy holds.

Case N: $\ell_1 \notin \overline{\mathcal{R}(A_1^\star)}$ — failure of calibration-based identification.

Since ℓ_1 does not lie in $\overline{\mathcal{R}(A_1^\star)} = \ker(A_1)^\perp$, there exists $g \in \ker(A_1)$ with $\langle g, \ell_1 \rangle_{X_1} \neq 0$. Then q and $q + tg$ both lie in $\mathcal{Q}_1$ (they satisfy the same observed balancing equation) but the identification functional (9) takes different values as t varies. Therefore, observed calibration constraints alone do not pin down μ_1 uniquely over the solution set $\mathcal{Q}_1$.

Case I: $\ell_1 \in \overline{\mathcal{R}(A_1^\star)} \setminus \mathcal{R}(A_1^\star)$ — irregular identification.

Since $\ell_1 \in \overline{\mathcal{R}(A_1^\star)} = \ker(A_1)^\perp$, the parameter μ_1 is invariant over $\mathcal{Q}_1$ and is point identified. But ℓ_1 does not lie in the range $\mathcal{R}(A_1^\star)$, so no square-integrable adjoint representer $b \in \mathcal{H}_{Z_1}$ exists. When A_1 is compact (Assumption D.1), the Moore–Penrose generalized inverse $(A_1^\star)^\dagger$ has domain

$$\mathcal{D}((A_1^\star)^\dagger) = \mathcal{R}(A_1^\star) \oplus \mathcal{R}(A_1^\star)^\perp,$$

and since ℓ_1 does not lie in this domain, the formal $b^\dagger = (A_1^\star)^\dagger \ell_1$ satisfies $\|b^\dagger\| = \infty$. In singular-value representation,

$$\sum_j \frac{\langle \ell_1, e_j \rangle_{X_1}^2}{\sigma_j^2} = \infty,$$

where (e_j, f_j, σ_j) is the singular system of A_1 under Assumption D.1. In this case μ_1 is point identified but no finite-variance influence function exists, so regular pathwise identification and $\sqrt{N}$ inference fail.

Case R: $\ell_1 \in \mathcal{R}(A_1^\star)$ — regular identification.

Assumption 2.3-(R2) holds, so an adjoint representer $b^{Y_1} \in \mathcal{H}_{Z_1}$ exists with $A_1^\star b^{Y_1} = \ell_1$. Since ℓ_1 lies in the domain $\mathcal{R}(A_1^\star) \oplus \mathcal{R}(A_1^\star)^\perp$ of the Moore–Penrose inverse, $b^{Y_1,\dagger} := (A_1^\star)^\dagger \ell_1 \in \mathcal{H}_{Z_1}$ is well-defined. Theorem 3.4 and Lemma C.4 apply, and μ_1 is point identified. The orthogonal score $\Gamma_1(O; q, b^{Y_1}) - \mu_1$ has finite variance under the moment conditions of Theorem E.10, so a finite-variance calibrated orthogonal moment exists. This score becomes a *regular* influence function when the local observed calibration model admits differentiable calibration paths through the chosen representative (Proposition E.7). Hence Case R is the regime in which a regular pathwise identification holds, under both finite-variance and local-calibration-path conditions. Adding (FV) (the finite-variance condition of Theorem E.10) gives *feasible $\sqrt{N}$-consistent regular semiparametric estimation*. This is the regime assumed throughout the main theory of this paper.

Important caveat on Case N. The Case N claim is non-invariance of the functional over the *observed calibration set* $\mathcal{Q}_1$; it is not a direct statement of causal non-identification in the full latent generalized Roy model. The constructed perturbation $q, q + tg \in \mathcal{Q}_1$ satisfies the observed

balancing equation but need not correspond to a consistent latent structural perturbation. Strict causal non-identification would require constructing a perturbation of the potential-outcome distribution that preserves the observed joint distribution while changing μ_1, which is a separate exercise (a local non-identification lemma in the latent model). The present paper treats Case N as "failure of calibration-based identification" or "non-invariance over $\mathcal{Q}_1$," which is the relevant identification limit when working with observed calibration/adjoint moments.

Singular-value diagnostic. The trichotomy admits an empirical diagnostic via the singular-value decomposition of A_1. With (e_j, f_j, σ_j) as above, define the Picard-style summability quantity

$$\mathcal{P}_1(\ell_1) := \sum_j \frac{\langle \ell_1, e_j \rangle_{X_1}^2}{\sigma_j^2}.$$

Decompose $\ell_1 = \ell_\perp + \sum_j \ell_j e_j$ where $\ell_\perp \in \ker(A_1)$. The diagnostic is two-stage:

- *Stage 1 (kernel residual).* Case N is characterized by $\ell_\perp \neq 0$ (i.e., $\ell_1 \notin \overline{\mathcal{R}(A_1^\star)}$); empirically, this is detected as a nonvanishing residual after projecting $\widehat{\ell}_1$ onto an estimated kernel of A_1.
- *Stage 2 (Picard summability).* Conditional on small kernel residual, $\mathcal{P}_1(\ell_1) < \infty$ corresponds to Case R (regular: finite-norm adjoint representer), whereas $\mathcal{P}_1(\ell_1) = \infty$ with $\ell_\perp = 0$ corresponds to Case I (irregular: closure-only).

Thus Picard divergence by itself does not distinguish Case N from Case I; the kernel-residual check is needed for that. Empirically, one can compute the partial sums $\mathcal{P}_1^{(J)}(\ell_1) := \sum_{j \leq J} \langle \widehat{\ell}_1, \widehat{e}_j \rangle^2 / \widehat{\sigma}_j^2$ and track their growth in J as a finite-sample diagnostic.

Compact-case formal statement. Under compactness of A_1 (Assumption D.1), the three cases admit the additional sharp characterizations via the Picard condition and the partial-sum diagnostics; the formal statement is given in Proposition D.17.

Remark (Meaning of the necessity theorem). The closure of the adjoint range is not a technical artifact. It is the exact condition for target invariance over the observed calibration set. The stronger condition $\ell_1 \in \mathcal{R}(A_1^\star)$ gives a finite-norm adjoint representer and therefore a regular orthogonal-moment representation.

Remark (Scope of the necessity proposition). The necessity statement of Proposition G.1 establishes that without both measurement-side (W) and adjoint-side (V) variation, no calibration-based identification result of the type considered in this paper is possible. The argument is structural: it shows that with only one type of variation (say W without V, or V without W), the identification operator has too small a range for the adjoint range condition to anchor a target signal. The proposition is therefore a statement about the necessary content of the variation, not about specific functional forms or distributional assumptions. In particular, the result rules out attempts to rely on only "one side" of the measurement/outcome-representation pair, even if the available data are very rich on that one side.

Appendix Result (Non-invariance, irregularity, and regularity)**.** *Conditional on primal existence, three cases are possible.*

1. *If $\ell_1 \notin \overline{\mathcal{R}(A_1^\star)}$, the target is not invariant over $\mathcal{Q}_1$. This is calibration-based non-invariance.*
2. *If $\ell_1 \in \overline{\mathcal{R}(A_1^\star)} \setminus \mathcal{R}(A_1^\star)$, the target can be point identified over $\mathcal{Q}_1$, but no finite-norm adjoint representer exists. This is irregular point identification.*

3. *If $\ell_1 \in \mathcal{R}(A_1^\star)$ and the induced score has finite variance, the target has a regular orthogonal-moment representation.*

Remark (Three stages: point identification, regularity, and feasible inference). Point identification requires invariance over the calibration set. Regular pathwise identification requires a finite-norm adjoint representer and a finite-variance score. Feasible root-N inference additionally requires product-rate estimability of the selected calibration representatives.

Remark (Caveat on Case N). Case N is stated as failure of calibration-based invariance over the observed calibration set. To turn it into global causal nonidentification in the full latent Roy model, one must also show that perturbations along the offending null direction correspond to admissible latent distributions. The statement above deliberately separates the operator-theoretic implication from that stronger latent-model assertion.

Remark (Primal-side irregularity). The trichotomy is conditional on primal existence $r_1 \in \mathcal{R}(A_1)$, supplied by the latent selection-odds representer anchor. If $r_1 \in \overline{\mathcal{R}(A_1)} \setminus \mathcal{R}(A_1)$, the selection-odds representer itself fails to exist as an L^2 object. This is a distinct primal irregularity, not the dual irregularity described above.

Proof of Proposition 3.5. The identity $\overline{\mathcal{R}(A_1^\star)} = \ker(A_1)^\perp$ implies that ℓ_1 is orthogonal to all observationally unidentified primal directions if and only if it belongs to the closed adjoint range. If it is outside that closure, there exists $g \in \ker(A_1)$ such that $\langle g, \ell_1 \rangle \neq 0$, and the target varies along $q + tg$. If ℓ_1 is in the closure but not in the range, the functional is invariant but lacks a finite-norm representer. If it is in the range, an adjoint representer exists. □

Remark (Theoretical meaning of the trichotomy). The trichotomy separates non-invariance, irregular point identification, and regular pathwise identification. It also clarifies why calibration existence and estimability are separate issues.

Example E.2 (Weak range or failure of calibration-based identification)**.** *If the measurement W is weakly related to the latent odds ratio, the primal operator may be ill conditioned. If the outcome-representation variable V carries little adjoint-side variation, the adjoint range may nearly fail to contain ℓ_1. In finite samples this appears as small empirical singular values, large condition numbers, and sensitivity to regularization.*

Definition E.2 (Empirical weak-range diagnostics)**.** Given finite-dimensional moment matrices for the primal and dual equations, define

$$\begin{aligned} &\widehat{\sigma}_{q,\min}, \quad \widehat{\kappa}_q = \widehat{\sigma}_{q,\max}/\widehat{\sigma}_{q,\min}, \\ &\widehat{\sigma}_{b,\min}, \quad \widehat{\kappa}_b = \widehat{\sigma}_{b,\max}/\widehat{\sigma}_{b,\min}, \end{aligned}$$

together with primal and dual residuals, regularization paths, and dual-norm paths. These are diagnostics for weak range problems, not tests of the latent anchor condition.

Remark (Interpretation and limits of diagnostics). These diagnostics play a role similar to first-stage diagnostics in instrumental variables. They do not prove the latent conditions, but they reveal whether the empirical inverse problems are numerically weak or sensitive to regularization.

Remark (Economic interpretation of the Picard condition). In compact inverse problems, regularity requires that the target signal does not load too heavily on directions with very small singular values. Economically, this means that the variation needed to evaluate the policy-relevant potential outcome must be supported by sufficiently strong measurement-side and adjoint-side variation.

E.4 Product bias, Neyman orthogonality, and failure modes of competing scores

This subsection collects the product-bias and Neyman-orthogonality identities for the influence function, the proof of the EIF representation under the observed calibration model, and demonstrations of failure modes for competing scores (the simple AIPW score and a standard PCI-type score under selection on gains). Throughout this subsection, a subscript 0 on q_0, m_0, h_0, b_0 denotes the population reference (true) value of the corresponding nuisance, not the untreated arm; the treated and untreated arms are indicated by the Y_1 and Y_0 superscripts where present (as in $m_0^{Y_1}$ and $m_0^{Y_0}$).

Failure of Neyman orthogonality for the simple AIPW score.

Consider the simple augmented inverse-probability score for $\mu_1 = \mathbb{E}[Y_1]$,

$$\psi_{\mathrm{AIPW}}^{Y_1}(O; \mu_1, q, m) = D(1+q)(Y_1 - m) + m - \mu_1. \tag{46}$$

Here $q = q(X_1)$ and $m = m(Z_1)$.

Proposition E.3 (First variation in the δq direction)**.** *At the true value (μ_1^0, q_0, m_0), for an admissible perturbation $\delta q(X_1)$,*

$$\partial_t \mathbb{E}[\psi_{\mathrm{AIPW}}^{Y_1}(\mu_1^0, q_0 + t\delta q, m_0)]\Big|_{t=0} = \mathbb{E}\big[D\,\delta q(X_1)\,\{Y_1 - m_0(Z_1)\}\big]. \tag{47}$$

This derivative is not zero in general.

Proof. Differentiate (46) under the expectation. The only term containing q is $D(1+q)(Y_1-m)$, so the derivative is exactly the right-hand side of (47). It vanishes for all δq only if $\mathbb{E}[D\{Y_1-m_0(Z_1)\} \mid X_1] = 0$, which is not implied by the usual auxiliary projection defining m_0. □

Failure of a standard PCI-type score under selection on gains.

A standard PCI score that does not allow the selection-odds representer to depend on the potential outcome cannot generally correct selection on gains. The following example isolates the issue.

Proposition E.4 (A standard PCI-type calibration weight does not recover the ATE under selection on gains)**.** *Let $Y_1 \sim N(0,1)$ and suppose that*

$$\Pr(D = 1 \mid Y_1 = y) = \Lambda(ay), \qquad a \neq 0,$$

where $\Lambda(t) = \{1+\exp(-t)\}^{-1}$. Let $Y = DY_1 + (1-D)Y_0$, and suppose $\mu_1 = \mathbb{E}[Y_1] = 0$. A standard PCI-type signal that uses a calibration weight independent of Y_1, for example $\Gamma^{\mathrm{PCI}} = 2DY$, satisfies

$$\mathbb{E}[\Gamma^{\mathrm{PCI}}] = 2\mathbb{E}[DY_1] \neq 0$$

in general. In contrast, the Y_1-dependent selection-odds representer

$$q^{\ell}(Y_1) = \frac{1 - \Lambda(aY_1)}{\Lambda(aY_1)}$$

yields $\mathbb{E}[D\{1+q^\ell(Y_1)\}Y] = \mathbb{E}[Y_1] = 0$.

Proof. By the symmetry of the standard normal and the identity $\Lambda(t)+\Lambda(-t)=1$, $\mathbb{P}(D=1) = \mathbb{E}\{\Lambda(aY_1)\} = 1/2$. The standard PCI-type score with constant selection-odds representer equals $q^{\mathrm{PCI}} = 1$. However, for $a>0$, $\Lambda(ay)-1/2$ has the same sign as y, so $\mathbb{E}[\Lambda(aY_1)Y_1] > 0$. Therefore $\mathbb{E}[\Gamma^{\mathrm{PCI}}] = 2\mathbb{E}[DY_1] \neq 0$ while $\mu_1 = \mathbb{E}[Y_1] = 0$. The Y_1-dependent selection-odds representer $q^\ell(Y_1) = \{1-\Lambda(aY_1)\}/\Lambda(aY_1)$ satisfies $\mathbb{E}[D\{1+q^\ell(Y_1)\} \mid Y_1] = 1$, hence $\mathbb{E}[D\{1+q^\ell(Y_1)\}Y] = \mathbb{E}[Y_1] = 0$ correctly. □

Residual correction h and the orthogonal score.

Definition E.3 (Residual correction induced by the adjoint outcome representer)**.** Fix an auxiliary function $m_0^{Y_1}(Z_1)$. A residual correction $h_0^{Y_1}(Z_1)$ is any solution of

$$A_1^\star h_0^{Y_1} = \rho_{1,m_0} \quad \text{in } \mathcal{H}_{X_1}, \tag{48}$$

where

$$\rho_{1,m_0}(x) = \mathbb{E}[Y_1 - m_0^{Y_1}(Z_1) \mid X_1 = x, D = 1].$$

Equivalently,

$$\mathbb{E}\Big[D\big\{Y_1 - m_0^{Y_1}(Z_1) - h_0^{Y_1}(Z_1)\big\} \,\Big|\, X_1\Big] = 0 \quad \mathbb{P}\text{-a.s.} \tag{49}$$

and hence

$$\mathbb{E}\Big[D\big\{Y_1 - m_0^{Y_1}(Z_1) - h_0^{Y_1}(Z_1)\big\}\, s(X_1)\Big] = 0 \quad \forall s \in L^2(X_1). \tag{50}$$

On the Y_0 side, $h_0^{Y_0}$ solves

$$A_0^\star h_0^{Y_0} = \rho_{0,m_0} \quad \text{in } \mathcal{H}_{X_0}, \tag{51}$$

which is equivalently

$$\mathbb{E}\Big[(1-D)\big\{Y_0 - m_0^{Y_0}(Z_0) - h_0^{Y_0}(Z_0)\big\} \,\Big|\, X_0\Big] = 0. \tag{52}$$

Remark (Interpretation of h and b)*.* The residual correction h is not an additional causal object. The identification theory is most naturally written in terms of the adjoint outcome representer $b = m + h$. The splitting is useful for implementation because m can be chosen as a variance-reducing auxiliary projection and h then corrects the residual component required for orthogonality.

Influence-function representation and Neyman orthogonality.

Appendix Result (Unbiasedness of the orthogonal score)**.** *At any admissible primal-adjoint representer pair,* $\mathbb{E}[\psi_{\mathrm{ortho}}^{Y_1}] = 0$ *and* $\mathbb{E}[\psi_{\mathrm{ortho}}^{Y_0}] = 0$.

Proof. This follows immediately from the identification formulas for μ_1 and μ_0. □

Appendix Result (Neyman orthogonality as a consequence of primal-dual identification)**.** *The first-order derivative of the expectation of the Y_1 orthogonal score is zero with respect to admissible first-order perturbations of q, m, and h around a valid calibration representative.*

Proof. The q-direction derivative is

$$\partial_t \mathbb{E}[\psi_{\mathrm{ortho}}^{Y_1}]_{\delta q} = \mathbb{E}\big[D\,\delta q(X_1)\big(Y_1 - m_0(Z_1) - h_0(Z_1)\big)\big]. \tag{53}$$

This equals zero by the residual-correction moment (49). The m and h derivatives reduce to the observed balancing equation $\mathbb{E}[D(1+q)-1 \mid Z_1]=0$. The Y_0 side is analogous. □

Appendix Result (Product-bias identity)**.** *Fix a selected primal–dual representative pair* $(q^\dagger, b^\dagger) \in \mathcal{Q}_1 \times \mathcal{B}_1$*, so that* $A_1 q^\dagger = r_1$ *and* $A_1^\star b^\dagger = \ell_1$*, and recall* $\mu_1 = \mathbb{E}[\Gamma_1(O; q^\dagger, b^\dagger)]$ *by the primal–dual identification theorem. Let* $\hat{q}(X_1)$ *and* $\hat{b}(Z_1)$ *be* arbitrary *square-integrable functions, not required to solve either calibration equation. Then the centered calibrated orthogonal moment satisfies the exact identity*

$$\mathbb{E}\Big[\Gamma_1(O;\hat{q},\hat{b})-\mu_1\Big] = -\,\mathbb{E}\Big[D\,\{\hat{q}(X_1)-q^\dagger(X_1)\}\,\{\hat{b}(Z_1)-b^\dagger(Z_1)\}\Big]. \tag{54}$$

The bias of the moment is thus the negative expected product of the two estimation errors weighted by D; it vanishes if either $\hat{q}=q^\dagger$ *or* $\hat{b}=b^\dagger$ *(in the relevant D-weighted L^2 sense), which is the double-robustness content of the score, and it is second order in the joint error, which underlies the $\sqrt{N}$ behavior of the plug-in estimator under the rate conditions of Appendix F. Equivalently, in the (q,m,h) parametrization with $b=m+h$, the identity becomes*

$$\mathbb{E}\Big[\psi^{Y_1}_{\text{ortho}}(O;\mu_1,\hat{q},\hat{m},\hat{h})\Big] = -\,\mathbb{E}\Big[D\,\{\hat{q}-q^\dagger\}\,\{(\hat{m}+\hat{h})-b^\dagger\}\Big],$$

which is the form used in the DML rate analysis of Appendix F.

Proof. Write $\Delta_q := \hat{q}-q^\dagger$ and $\Delta_b := \hat{b}-b^\dagger$, so that $1+\hat{q} = (1+q^\dagger)+\Delta_q$ and $\hat{b} = b^\dagger+\Delta_b$. Since $\mu_1 = \mathbb{E}[\Gamma_1(O;q^\dagger,b^\dagger)]$, expanding $\Gamma_1(O;\hat{q},\hat{b})-\Gamma_1(O;q^\dagger,b^\dagger)$ gives

$$D(1+q^\dagger)(Y-b^\dagger) - D(1+q^\dagger)\Delta_b + D\,\Delta_q(Y-b^\dagger) - D\,\Delta_q\Delta_b + \Delta_b - D(1+q^\dagger)(Y-b^\dagger),$$

so that

$$\mathbb{E}[\Gamma_1(O;\hat{q},\hat{b})-\mu_1] = \underbrace{\mathbb{E}[\{1-D(1+q^\dagger)\}\Delta_b]}_{T_1} + \underbrace{\mathbb{E}[D\,\Delta_q(Y-b^\dagger)]}_{T_2} - \mathbb{E}[D\,\Delta_q\Delta_b].$$

For T_1, Δ_b is a function of Z_1, and $q^\dagger \in \mathcal{Q}_1$ gives $\mathbb{E}[D(1+q^\dagger)-1 \mid Z_1]=0$, hence $T_1 = \mathbb{E}[\Delta_b(Z_1)\,\mathbb{E}\{1-D(1+q^\dagger) \mid Z_1\}] = 0$. For T_2, Δ_q is a function of X_1, and the adjoint identity used in the proof of the Neyman orthogonality result above gives $\mathbb{E}[D\,a(X_1)\{Y-b^\dagger(Z_1)\}] = \langle a, \ell_1 - A_1^\star b^\dagger\rangle_{X_1} = 0$ for every $a \in L^2(X_1)$, since $b^\dagger \in \mathcal{B}_1$; taking $a=\Delta_q$ yields $T_2=0$. Only the cross term survives, which is (54). □

Remark. Neyman orthogonality is not imposed from outside. It is a structural consequence of pairing the primal selection-odds representer with the adjoint outcome representer.

Efficient influence function under the observed calibration model.

Definition E.4 (Local observed calibration model)**.** At a distribution P_0, the local observed calibration model consists of regular parametric submodels P_ε through P_0 for which the primal and adjoint representer equations admit differentiable solution paths and the relevant finite-variance and overlap conditions hold in a neighborhood of zero.

Important note. The global set $\mathcal{M}_{\text{cal}} = \{P : \mathcal{Q}_1(P) \neq \emptyset,\ \mathcal{B}_1(P) \neq \emptyset\}$ is not necessarily a smooth semiparametric model in the standard sense, because the range conditions are not

open conditions: an arbitrarily small perturbation of P can move the relevant range outside $\mathcal{R}(A_1^\star)$. Efficiency claims in this paper are therefore stated with respect to *local* calibration models satisfying differentiable-calibration-path conditions, not with respect to the unrestricted global set.

Definition E.5 ((q_0, b_0)-anchored local calibration model)**.** For a fixed representative $(q_0, b_0) \in \mathcal{Q}_1(P_0) \times \mathcal{B}_1(P_0)$, the anchored local calibration model consists of those regular submodels for which there exist differentiable paths $q_\varepsilon \in \mathcal{Q}_1(P_\varepsilon)$ and $b_\varepsilon \in \mathcal{B}_1(P_\varepsilon)$ with $q_\varepsilon|_{\varepsilon=0} = q_0$ and $b_\varepsilon|_{\varepsilon=0} = b_0$.

Why anchoring is needed. When the calibration solution sets $\mathcal{Q}_1(P_0)$ and $\mathcal{B}_1(P_0)$ are not singletons, the existence of a differentiable path from some (q_0, b_0) is not automatic. A path starting from a different representative $(q_0', b_0') \neq (q_0, b_0)$ may exist while no path from (q_0, b_0) does, in which case the IF property of the score evaluated at (q_0, b_0) is not guaranteed. When stating an efficiency claim about a particular representative, the anchor must be made explicit.

Anchored versus unanchored statements. The identification claim of Theorem 3.4 does not depend on which representative is chosen, so the unanchored local model $\mathcal{M}_{\text{cal}}^{\text{loc}}(P_0)$ suffices. The regular-IF claim of Proposition E.7, by contrast, refers to the IF property of a specific (q_0, b_0), so the anchored model $\mathcal{M}_{\text{cal}}^{\text{loc}}(P_0; q_0, b_0)$ is the proper reference class. The EIF claim of Theorem E.11 takes the infimum over variation directions through all representatives, so the unanchored class suffices for the infimum, although the infimum-attaining $(q_1^{\text{EIF}}, b_1^{\text{EIF}})$ must itself admit an anchored path.

Proposition E.5 (Sufficient conditions for a local calibration submodel)**.** *Suppose that in a finite-dimensional sieve representation the Jacobian of the stacked primal and dual moments with respect to the calibration coefficients has full column rank at (P_0, q_0, b_0), and the moments are differentiable in the submodel index. Then, locally, there exist differentiable calibration selections satisfying the primal and dual equations.*

Proof. Apply the finite-dimensional implicit function theorem to the stacked moment map. The full-rank condition gives local solvability and differentiability of the coefficient path. □

Remark. In Gaussian latent-factor examples with Hermite or polynomial sieves, the rank condition can be interpreted as a finite-dimensional approximation to the range and injectivity requirements.

The canonical adjoint representer form is obtained by setting

$$b(Z_1) := m(Z_1) + h(Z_1). \tag{55}$$

It solves

$$A_1^\star b = \ell_1(\cdot) := \mathbb{E}[Y \mid X_1 = \cdot, D = 1] \quad \text{in } \mathcal{H}_{X_1}, \tag{56}$$

or equivalently

$$\mathbb{E}\big[D\{Y_1 - b(Z_1)\} \mid X_1\big] = 0 \quad \mathbb{P}\text{-a.s.} \tag{57}$$

The identification signal is

$$\Gamma_1(O; q, b) = D\{1 + q(X_1)\}\{Y_1 - b(Z_1)\} + b(Z_1), \tag{58}$$

and the induced mean-zero influence-function candidate is

$$\phi_1(O; q, b) := \Gamma_1(O; q, b) - \mu_1 = D\{1 + q(X_1)\}\{Y_1 - b(Z_1)\} + b(Z_1) - \mu_1. \tag{59}$$

The primal and adjoint representer sets are

$$\mathcal{Q}_1 := \{q \in \mathcal{H}_{X_1} : \mathbb{E}[Dq(X_1) - (1-D) \mid Z_1] = 0\}, \tag{60}$$

$$\mathcal{B}_1 := \{b \in \mathcal{H}_{Z_1} : \mathbb{E}[D\{Y_1 - b(Z_1)\} \mid X_1] = 0\}. \tag{61}$$

Theorem E.6 (Calibrated orthogonal moment)**.** *For any $(q,b) \in \mathcal{Q}_1 \times \mathcal{B}_1$ with $\phi_1(O;q,b) \in L_0^2(P)$, the score $\phi_1(O;q,b)$ is mean zero and Neyman orthogonal with respect to admissible first-order perturbations of the primal and adjoint representer functions.*

The proposition formalizes the intuition that any reasonable regularized estimator of (q,b) — Tikhonov-regularized least squares, Sieve GMM with positive regularization, or unregularized normal equations whose matrices remain nonsingular in a neighborhood of θ_0 — induces a differentiable calibration path along a finite-dimensional parametric submodel. The differentiability of the path is what justifies treating $\Gamma_1(O;q_\theta,b_\theta)$ as a regular IF for the local calibration submodel.

The proof is given in Appendix G.1.

Proposition E.7 (When a calibrated orthogonal moment is a regular influence function)**.** *Suppose (q_0,b_0) is an anchored calibration representative, the local observed calibration model admits differentiable calibration paths through (q_0,b_0), and*

$$\mathbb{E}\big[D\{1+q_0(X_1)\}^2\{Y-b_0(Z_1)\}^2\big] < \infty, \qquad \mathbb{E}[b_0(Z_1)^2] < \infty. \tag{62}$$

Then $\phi_1(O;q_0,b_0)$ is a regular influence function relative to the anchored local observed calibration model, i.e.

$$\dot{\mu}_1(P_\varepsilon)\big|_{\varepsilon=0} = \mathbb{E}_{P_0}[\phi_1(O;q_0,b_0)S(O)] \quad \forall P_\varepsilon \in \mathcal{M}_{\text{cal}}^{\text{loc}}(P_0;q_0,b_0). \tag{63}$$

Proof. This is the usual pathwise differentiation of the observed-data functional, with nuisance-path derivatives eliminated by the primal and adjoint representer moments. □

Remark (Logical relation between orthogonality and regularity). Orthogonality of the calibrated orthogonal moment is an algebraic implication of the moments. Regular influence-function status additionally requires a local calibration model, differentiable calibration paths, and finite variance.

Proposition E.8 (Pathwise influence-function representation, observed-data form)**.** *Let P_ε be a regular observed-data submodel with score $S(O)$, and suppose there exist differentiable selections $q_{P_\varepsilon} \in \mathcal{Q}_1(P_\varepsilon)$ and $b_{P_\varepsilon} \in \mathcal{B}_1(P_\varepsilon)$. Define*

$$\mu_1(P_O) := \mathbb{E}_{P_O}[\Gamma_1(O;q_{P_O},b_{P_O})] = \mathbb{E}_{P_O}\big[D\{1+q_{P_O}(X_1)\}\{Y-b_{P_O}(Z_1)\} + b_{P_O}(Z_1)\big]. \tag{64}$$

Then

$$\dot{\mu}_1 = \frac{d}{d\varepsilon}\mu_1(P_\varepsilon)\Big|_{\varepsilon=0} = \mathbb{E}_{P_0}[\{\Gamma_1(O;q_0,b_0) - \mu_1\}S(O)] = \mathbb{E}_{P_0}[\phi_1(O;q_0,b_0)S(O)]. \tag{65}$$

Advantage of the observed-data parametrization. The natural parameter for identification is $\mu_1 = \mathbb{E}[Y_1]$, but Y_1 is unobserved on $\{D=0\}$, so pathwise differentiation of this parameter within the observed calibration model is not direct. The calibration representation $\mu_1(P_O) = \mathbb{E}_{P_O}[\Gamma_1(O;q_{P_O},b_{P_O})]$ expresses the target as a functional of the observed data alone (the calibration nuisance functions depend on P_O). Pathwise differentiation can then be carried out within the observed calibration model directly. The orthogonality of the calibrated orthogonal moment ensures that the nuisance-path contributions vanish at the truth, leaving only the outer-expectation derivative.

Proof sketch. **Advantage of observed-data parameter.** $\mu_1(P_O) = \mathbb{E}_{P_O}[\Gamma_1]$ can be written entirely as a function of the observed data (D, Y, W, V, S, Q). By contrast, $\mathbb{E}_P[Y_1]$ involves Y_1 on $\{D = 0\}$, which is not observed, so pathwise differentiation of the latent parameter cannot be carried out directly within the observed calibration model.

Pathwise differentiation. Differentiating $\mu_1(P_\varepsilon)$ in ε yields two contributions: the outer expectation derivative and the derivatives of the nuisance paths q_{P_ε} and b_{P_ε}. The nuisance path derivatives equal the first variations of the signal in the q and b directions, which vanish at $(q_0, b_0) \in \mathcal{Q}_1 \times \mathcal{B}_1$ by Theorem 3.7. The remaining outer derivative is the covariance of $\Gamma_1(O; q_0, b_0)$ with the submodel score $S(O)$. Since $\mathbb{E}[\Gamma_1] = \mu_1$ is constant in ε at (q_0, b_0), the centered score $\phi_1 = \Gamma_1 - \mu_1$ has mean zero and represents the pathwise derivative:

$$\dot{\mu}_1 = \mathbb{E}_{P_0}[\phi_1(O; q_0, b_0)\, S(O)] \quad \text{for every regular submodel.}$$

Therefore ϕ_1 is a regular influence function relative to the anchored local observed calibration model. □

Remark (Role of Proposition E.8). The result is an observed-data statement. The latent Roy model is used to justify that the observed calibration functional equals the causal target.

Lemma E.9 (Closedness of $\mathcal{B}_1$)**.** *If $A_1^\star$ is continuous, then $\mathcal{B}_1 = \{b : A_1^\star b = \ell_1\}$ is a closed affine subset of $\mathcal{H}_{Z_1}$. If a weight w_q is bounded above and below away from zero, the same set is closed in $L^2(w_q dP_{Z_1})$.*

Proof. The inverse image of a singleton under a continuous linear operator is closed. Norm equivalence gives the weighted statement. □

Theorem E.10 (Fixed-q variance-minimizing adjoint representer)**.** *Fix $q \in \mathcal{Q}_1$ and suppose $\mathcal{B}_1 \neq \emptyset$. Assume the weight w_q defined below is finite and bounded away from zero and infinity $\mathbb{P}$-a.s., so that by Lemma E.9 the set $\mathcal{B}_1$ is closed in $L^2(w_q\, dP_{Z_1})$ and the weighted projection below exists and is unique. Let*

$$b_q^\star \in \arg\min_{b \in \mathcal{B}_1} \mathrm{Var}[\Gamma_1(O; q, b)]. \tag{66}$$

Let $\alpha_q = D\{1 + q(X_1)\}$. Then

$$\mathbb{E}[\alpha_q - 1 \mid Z_1] = 0 \quad \mathbb{P}\text{-a.s.} \tag{67}$$

and, with $w_q(z) = \mathbb{E}[(1 - \alpha_q)^2 \mid Z_1 = z]$, the unconstrained minimizer is

$$b_q^{\mathrm{un}}(z) := \frac{\mathbb{E}[(\alpha_q - 1)\alpha_q Y \mid Z_1 = z]}{w_q(z)}. \tag{68}$$

The constrained minimizer is the weighted projection

$$b_q^\star = \Pi_{\mathcal{B}_1}^{w_q}(b_q^{\mathrm{un}}) = \arg\min_{b \in \mathcal{B}_1} \mathbb{E}[w_q(Z_1)\{b(Z_1) - b_q^{\mathrm{un}}(Z_1)\}^2]. \tag{69}$$

It satisfies

$$\mathbb{E}[\Gamma_1(O; q, b_q^\star)\{1 - D(1 + q(X_1))\}c(Z_1)] = 0 \quad \forall c \in \ker(A_1^\star), \tag{70}$$

and the corresponding centered score is

$$\phi_1^\star(O; q) = D\{1 + q(X_1)\}\{Y - b_q^\star(Z_1)\} + b_q^\star(Z_1) - \mu_1. \tag{71}$$

Proof strategy. The variance of the calibrated orthogonal moment $\Gamma_1(O; q, b)$ at a fixed q is a quadratic functional of $b \in \mathcal{B}_1$. Since $\mathcal{B}_1$ is a closed affine subspace of $\mathcal{H}_{Z_1}$ (closed by Lemma E.9, affine because it is the solution set of a linear equation), the variance-minimizing b^* exists and is uniquely determined by the projection theorem. The first-order condition is the weighted projection condition

$$\mathbb{E}\big[\,\Gamma_1(O;\, q,\, b_q^\star)\,\big\{1 - D(1 + q(X_1))\big\}\,c(Z_1)\,\big] = 0 \quad \forall\, c \in \ker(A_1^\star),$$

where $\Gamma_1(O; q, b)$ denotes the centered influence-function contribution at the candidate pair (q, b). This condition characterizes the variance-minimizing adjoint representer as the projection of Y onto the closure of $\mathcal{B}_1$ in the relevant weighted L^2 norm; it should not be simplified to a single conditional moment of the form $\mathbb{E}[D(1+q)(Y-b) \mid Z_1] = 0$, because such a simplification omits the kernel-direction test functions $c \in \ker(A_1^\star)$ that the projection requires.

Why fixed-q? The fixed-q variance minimization decouples the optimization over $\mathcal{B}_1$ from the optimization over $\mathcal{Q}_1$. This is conceptually clean because the observed calibration q is identified up to $\ker(A_1)$, and different choices of q within this kernel can affect the variance of Γ_1 in complex ways. The joint minimization over $(q, b) \in \mathcal{Q}_1 \times \mathcal{B}_1$ is generally non-convex. The fixed-q approach is a natural compromise: one fixes a regularized $\widehat{q}$ from the first stage and then computes the variance-minimizing $\widehat{b}$ as a convex quadratic problem.

The proof is given in Appendix G.1.

Remark (Meaning of Theorem E.10). This is a strong, unconditional variance-minimization result conditional on a fixed primal representative. It does not rely on the generally nonconvex joint minimization over (q, b).

Remark (Nonconvexity of joint (q, b) minimization). The criterion is not generally jointly convex in (q, b) because Γ_1 contains the product $q \cdot b$. Joint efficiency claims therefore require additional assumptions, whereas fixed-q dual minimization is a closed convex projection problem.

Theorem E.11 (Conditional characterization of the EIF in the observed calibration model)**.** *Let $\mathcal{IF}_1$ denote the regular influence-function class for μ_1 in the local observed calibration model. Let Φ_{cal} be the $L^2(P)$ class of mean-zero Neyman-orthogonal calibrated orthogonal moments generated by $(q, b) \in \mathcal{Q}_1 \times \mathcal{B}_1$, and let*

$$\Phi_{\text{cal}}^{\text{reg}} := \left\{ \phi \in \Phi_{\text{cal}} : \begin{array}{l} \phi \text{ satisfies the finite-variance and} \\ \text{local-calibration-path conditions (Prop. E.7)} \end{array} \right\}$$

be the regular subclass. Then $\Phi_{\text{cal}}^{\text{reg}} \subseteq \mathcal{IF}_1$ (regular calibrated orthogonal moments are regular IFs), but in general $\Phi_{\text{cal}} \not\subseteq \mathcal{IF}_1$ (not every mean-zero calibrated orthogonal moment is a regular IF). Let

$$\phi_1^{\text{EIF}} \in \arg\min_{\phi \in \mathcal{IF}_1} \mathbb{E}[\phi^2] \tag{72}$$

be the canonical gradient in the local observed calibration model. If $\phi_1^{\text{EIF}} \in \overline{\Phi_{\text{cal}}^{\text{reg}}}^{L^2(P)}$, then the regular calibration class attains the semiparametric efficiency bound in the sense that

$$\inf_{\phi \in \Phi_{\text{cal}}^{\text{reg}}} \mathbb{E}[\phi^2] = \mathbb{E}[(\phi_1^{\text{EIF}})^2].$$

If the infimum is attained by some $(q_1^{\text{EIF}}, b_1^{\text{EIF}})$ whose induced calibrated orthogonal moment lies

in $\Phi_{\text{cal}}^{\text{reg}}$*, then*

$$\phi_1^{\text{EIF}}(O) = D\{1 + q_1^{\text{EIF}}(X_1)\}\{Y - b_1^{\text{EIF}}(Z_1)\} + b_1^{\text{EIF}}(Z_1) - \mu_1. \tag{73}$$

Proof outline. The proof has three steps. *Step 1 (orthogonality)*: the orthogonal score $\Gamma_1(O; q, b)$ is Neyman-orthogonal at any $(q, b) \in \mathcal{Q}_1 \times \mathcal{B}_1$ by Theorem E.6. *Step 2 (regular IF)*: under finite-variance condition (62), an anchored local submodel admits Γ_1 as a regular IF (Proposition E.7). *Step 3 (variance minimization)*: the EIF is the variance-minimizing regular IF within the orthogonal-score class. Minimizing over $\mathcal{B}_1$ first (fixed q) gives the canonical adjoint representer $b^* = b^{\text{EIF}}(q)$, then minimizing the resulting variance over $\mathcal{Q}_1$ (modulo $\ker(A_1)$) gives the canonical observed calibration q^{EIF}. Conditional on the closure and attainment assumptions of the theorem, the resulting $\Gamma_1(O; q^{\text{EIF}}, b^{\text{EIF}}) - \mu_1$ is the EIF. The attaining influence function is unique as an $L^2(P)$ element; the representing pair need not be unique without additional injectivity or strict-convexity conditions beyond those stated.

Interpretation. Conditional on its closure and attainment assumptions, Theorem E.11 states that the EIF under the observed calibration model is obtained by selecting the calibration representative that minimizes the variance of the orthogonal score Γ_1. Two features deserve emphasis. First, the EIF coincides with the orthogonal score evaluated at a *specific* variance-minimizing $(q_1^{\text{EIF}}, b_1^{\text{EIF}})$ rather than at an arbitrary (q, b) pair; the orthogonal-score class of Theorem E.6 is generally larger than the EIF. Second, the EIF itself is unique as an $L^2(P)$ element even when the attaining representatives are not; no kernel-level uniqueness claim is made for the representing pair.

Three layers of efficiency claims. The efficiency theory of this paper is organized in three layers. *Layer 1* (orthogonality, Theorem E.6): the calibrated orthogonal moment Γ_1 is Neyman-orthogonal under any $(q, b) \in \mathcal{Q}_1 \times \mathcal{B}_1$. *Layer 2* (regular IF, Proposition E.7): under condition (62) and an anchored calibration submodel, $\Gamma_1(O; q_0, b_0)$ is a regular IF for the (q_0, b_0)-anchored local model. *Layer 3* (EIF, Theorem E.11): the variance-minimizing representative yields the EIF of the local calibration model. Practitioners may operate at Layer 2 with a simple regularized estimator and report Layer 1's orthogonality as the inferential basis; Layer 3 is the strongest claim and requires explicit variance minimization.

Proof. By standard semiparametric efficiency theory (Newey and Powell, 2003; Chen and Pouzo, 2012; Bennett et al., 2025), the efficient influence function (EIF) in the observed local calibration model is the unique $L^2(P)$-minimum-variance element of the influence-function class, characterized as the canonical gradient that generates the pathwise derivative for every regular submodel.

Calibrated-moment class inclusion. By Theorem E.6, every calibrated orthogonal moment $\phi_1(O; q, b)$ with $(q, b) \in \mathcal{Q}_1 \times \mathcal{B}_1$ is mean-zero and Neyman-orthogonal at the truth. By Proposition E.7, it is a *regular* influence function only under the additional finite-variance and local-calibration-path conditions of that proposition. Denote the calibrated orthogonal moment class by Φ_{cal} and its regular subset by $\Phi_{\text{cal}}^{\text{reg}} \subseteq \Phi_{\text{cal}}$. The latter is contained in the full $L^2(P)$ regular-influence-function class IF_1.

L^2 closure approximation. If ϕ_1^{EIF} lies in the $L^2(P)$ closure of the regular calibrated orthogonal moment subclass $\Phi_{\text{cal}}^{\text{reg}}$, then there is a sequence $(q_n, b_n) \in \mathcal{Q}_1 \times \mathcal{B}_1$ with $\phi_1(O; q_n, b_n) \to \phi_1^{\text{EIF}}$ in $L^2(P)$. Consequently

$$\inf_{\phi \in \Phi_{\text{cal}}^{\text{reg}}} \mathbb{E}[\phi^2] = \mathbb{E}[(\phi_1^{\text{EIF}})^2],$$

so the infimum of regular calibrated orthogonal moment variance equals the semiparametric efficiency bound.

Attainment. If the infimum is attained by some $(q_1^{\mathrm{EIF}}, b_1^{\mathrm{EIF}}) \in \mathcal{Q}_1 \times \mathcal{B}_1$, then $\phi_1(O; q_1^{\mathrm{EIF}}, b_1^{\mathrm{EIF}})$ has the same variance as ϕ_1^{EIF}. Since regular influence functions with identical variance differ by elements of the orthogonal complement of the tangent space, and the influence function lies in $\overline{\mathrm{IF}_1}$, attainment means $\phi_1(O; q_1^{\mathrm{EIF}}, b_1^{\mathrm{EIF}}) = \phi_1^{\mathrm{EIF}}$ P-a.s. Therefore the attaining calibration representative gives an EIF representation:

$$\phi_1^{\mathrm{EIF}}(O) = D\{1 + q_1^{\mathrm{EIF}}(X_1)\}\{Y - b_1^{\mathrm{EIF}}(Z_1)\} + b_1^{\mathrm{EIF}}(Z_1) - \mu_1.$$

Conditional nature of the statement. The closure assumption $\phi_1^{\mathrm{EIF}} \in \overline{\Phi_{\mathrm{cal}}^{\mathrm{reg}}}^{L^2(P)}$ is substantive in infinite-dimensional calibration problems and is not implied by injectivity: $\ker(A_1) = \{0\}$ and $\ker(A_1^\star) = \{0\}$ make the calibration representatives unique when they exist, but do not imply that the resulting regular calibrated moment equals the canonical gradient. The closure condition remains substantive. The attainment assumption is automatic when $\mathcal{Q}_1 \times \mathcal{B}_1$ is finite-dimensional with a compact parameter set, by the Weierstrass theorem. In general infinite-dimensional settings it must be verified case by case. This is why the theorem is stated as a conditional characterization. □

Remark (Scope of the efficiency claim). The theorem is a conditional characterization. It does not assert that every nonunique calibration problem has an attained global minimum in infinite dimensions. It says that if the canonical gradient lies in the closure of the influence-function class and the infimum is attained, then the attaining calibration representative is efficient.

Remark (Calibration representative versus influence function). Different calibration representatives $(q, b) \in \mathcal{Q}_1 \times \mathcal{B}_1$ generate an affine class of regular influence-function candidates for the same target μ_1. The canonical gradient (EIF) is the unique minimum-variance element of the $L^2(P)$-closure of this class, when it exists. Two distinct representatives need not generate the same influence function in general; rather, they generate distinct elements of the affine IF class whose minimum-variance element is the EIF. Efficiency statements are statements about the $L^2(P)$ functions, not about uniqueness of calibration representatives.

Remark (Additional assumptions in Theorem E.11). The assumptions that the EIF belongs to the L^2 closure of the IF class and that the infimum is attained are substantive. They may be verifiable in particular finite-dimensional models (Proposition E.15), but neither follows from uniqueness of the calibration representatives alone.

Remark (Finite-dimensional sieve implementation). In a finite-dimensional sieve space with a compact coefficient set and continuous objective, the minimum-variance calibration representative exists by the Weierstrass theorem. This provides a practical route to approximate the conditional EIF characterization.

Proposition E.12 (First-order conditions for joint minimization)**.** *If a joint minimum-variance representative exists, then for all* $a \in \ker(A_1)$ *and all* $c \in \ker(A_1^\star)$,

$$\mathbb{E}[\phi_1^{\mathrm{EIF}}(O) D a(X_1)\{Y_1 - b_1^{\mathrm{EIF}}(Z_1)\}] = 0, \tag{74}$$

$$\mathbb{E}[\phi_1^{\mathrm{EIF}}(O)\{1 - D(1 + q_1^{\mathrm{EIF}}(X_1))\} c(Z_1)] = 0. \tag{75}$$

Proof. Perturb the minimizer along feasible null directions $q + ta$ and $b + tc$ and differentiate the variance criterion at $t = 0$. □

Corollary E.13 (Unique primal and adjoint representers)**.** *Suppose additionally that the primal and adjoint operators are injective, i.e.*

$$\ker(A_1) = \{0\}, \qquad \ker(A_1^\star) = \{0\}.$$

Then the observed calibration q_1 and adjoint representer b_1 are uniquely determined.

Proof. With singleton calibration sets there are no representative choices left, so the primal and adjoint representers—and hence the calibrated moment within this class—are unique whenever they exist. Coincidence of this calibrated moment with the canonical gradient is not automatic from $\ker = \{0\}$; it requires the closure/attainment conditions of Theorem E.14. □

Theorem E.14 (Conditional characterization of the efficient influence function for the ATE)**.** *Suppose that, for each arm $t \in \{0,1\}$: (a) the canonical gradient ϕ_t^{EIF} of μ_t in the observed calibration model exists; and (b) ϕ_t^{EIF} lies in the L^2-closure of the centered calibrated orthogonal moment class $\overline{\Phi_{\mathrm{cal},t}^{\mathrm{reg}}} = \overline{\{\Gamma_t(\cdot; q_t, b_t) - \mu_t\}}$ and is attained there, i.e., there exists an admissible pair $(q_t^{\mathrm{e}}, b_t^{\mathrm{e}})$ with $\Gamma_t(\cdot; q_t^{\mathrm{e}}, b_t^{\mathrm{e}}) - \mu_t = \phi_t^{\mathrm{EIF}}$ in L^2 (equivalently, the tangent-space projection orthogonality $\mathbb{E}[\{\Gamma_t(\cdot; q_t^{\mathrm{e}}, b_t^{\mathrm{e}}) - \mu_t - \phi_t^{\mathrm{EIF}}\}\, s] = 0$ for every score s holds with the residual equal to zero). Then the EIF for $\tau = \mu_1 - \mu_0$ is*

$$\phi_\tau^{\mathrm{EIF}}(O) = \phi_1^{\mathrm{EIF}}(O) - \phi_0^{\mathrm{EIF}}(O), \tag{76}$$

and the efficiency bound is

$$\mathcal{V}_\tau^{\mathrm{eff}} = \mathbb{E}[(\phi_1^{\mathrm{EIF}} - \phi_0^{\mathrm{EIF}})^2]. \tag{77}$$

Neither (a) nor (b) follows from injectivity of A_t or $A_t^\star$ alone; absent (b), the calibrated class need not contain the canonical gradient, and the calibrated moments then characterize the efficiency envelope *of the class rather than the semiparametric bound (see the envelope paragraph below).*

The calibrated efficiency envelope and primitive conditions for attainment.

Define, for each arm, the calibrated efficiency envelope

$$\mathcal{V}_{\mathrm{cal},t} \;:=\; \inf_{\phi \in \overline{\Phi_{\mathrm{cal},t}^{\mathrm{reg}}}} \mathbb{E}[\phi^2],$$

the smallest asymptotic variance achievable by (limits of) calibrated orthogonal moments. Because every element of $\Phi_{\mathrm{cal},t}^{\mathrm{reg}}$ is an influence function of a regular estimator of μ_t within the calibration model, the envelope always sits above the semiparametric bound, $\mathcal{V}_{\mathrm{cal},t} \geq \mathbb{E}[(\phi_t^{\mathrm{EIF}})^2]$, with equality if and only if condition (b) of Theorem E.14 holds—that is, the canonical gradient is a calibrated moment (in the closure). The variance-minimizing adjoint representer of Theorem 3.10 solves the fixed-q section of the envelope problem—the envelope itself is the infimum over admissible pairs (q, b)—and not automatically the bound. Primitive sufficient conditions for attainment are available in restricted settings and are stated here conditionally, matching what the proofs deliver: (i) *finite support*: Proposition E.15 below shows that finite joint support with injective, locally solvable calibration systems makes the observed model locally an unrestricted multinomial, so the regular influence-function class is a singleton and the calibrated moment equals the canonical gradient—(b) holds; (ii) *verified tangent-space spanning*: whenever $\ell_t \in \mathcal{R}(A_t^\star)$ and the local-calibration-path scores of Proposition E.7 span the observed tangent space—so that the projection of an initial regular influence function onto that space remains a calibrated moment—the projection argument of Appendix G.1 closes and (b) holds; this spanning

property must be verified and does not follow from injectivity alone; (iii) in general—including the near-singular empirical regime of the application—(b) is *not* asserted, the reported quantity is the envelope variance of the implemented moment, and no semiparametric-efficiency claim is attached to the empirical standard errors, consistent with the sensitivity-magnitude reading of the headline estimate.

Proposition E.15 (Finite-support attainment of the canonical gradient)**.** *Suppose (D, Y, X_t, Z_t) have finite joint support, the finite-dimensional operators A_t and $A_t^\star$ are injective, and the systems $A_t q = r_t$ and $A_t^\star b = \ell_t$ are solvable at P. Then they are solvable for every law in a total-variation neighborhood of P, the observed calibration model contains that neighborhood (it is locally the unrestricted multinomial), the regular influence-function class $\mathcal{IF}_t$ for μ_t is a singleton, and the calibrated orthogonal moment satisfies $\Gamma_t(\cdot; q_t, b_t) - \mu_t = \phi_t^{\mathrm{EIF}}$; in particular, hypothesis (b) of Theorem E.14 holds.*

Proof. Local unrestrictedness. With finite support, A_t, $A_t^\star$, r_t, and ℓ_t are finite matrices and vectors whose entries are continuous in the cell probabilities of P; injectivity of a full-column-rank matrix is an open condition, and solvability of the two linear systems is preserved under small perturbations of the entries. Hence every law P' in a total-variation neighborhood of P admits solutions, i.e., belongs to the observed calibration model, and the model is locally the unrestricted multinomial on the finite support. *Singleton influence-function class.* In a locally unrestricted model the tangent space at P is the full space of mean-zero, finite-variance functions of the observation. Two regular influence functions for the same parameter differ by an element orthogonal to the tangent space, hence by zero; $\mathcal{IF}_t$ is a singleton, necessarily the canonical gradient ϕ_t^{EIF}. *Membership.* On a finite support every calibrated moment has finite variance, and the anchored local calibration paths of Proposition E.7 exist because the finite-dimensional solution maps $P' \mapsto (q_t(P'), b_t(P'))$ are differentiable at P by the implicit function theorem applied to the invertible linear systems. Therefore $\Gamma_t(\cdot; q_t, b_t) - \mu_t \in \Phi_{\mathrm{cal}}^{\mathrm{reg}} \subseteq \mathcal{IF}_t$ (Theorem E.11), and by the singleton property it equals ϕ_t^{EIF}. □

Joint EIF versus the sum of individual EIFs. The EIF for $\tau = \mu_1 - \mu_0$ admits two formulations. The naive "sum of individual EIFs" formulation takes the EIF for μ_1 and subtracts the EIF for μ_0, treating each side separately. The joint-calibration-minimization formulation jointly minimizes the variance of $\Gamma_1 - \Gamma_0$ over (q_1, b_1, q_0, b_0). The two coincide when the joint minimization decouples into independent minimizations on each side. When the canonical gradients ϕ_1^{EIF} and ϕ_0^{EIF} exist as regular influence functions for μ_1 and μ_0 respectively, linearity of the canonical gradient implies that the ATE EIF is $\phi_\tau^{\mathrm{EIF}} = \phi_1^{\mathrm{EIF}} - \phi_0^{\mathrm{EIF}}$, and no regular calibration-based estimator can achieve smaller asymptotic variance. The joint optimization can yield strictly smaller variance only relative to *non-canonical* calibration representatives chosen separately on each side, by exploiting correlation across the two sides (e.g., shared regressors in $S \cap Q$) within the affine class of valid calibration representatives. This is a finite-sample / non-canonical phenomenon, not an improvement over the efficiency bound.

Remark (Difference of EIFs versus joint calibration minimization). If ϕ_1^{EIF} and ϕ_0^{EIF} are canonical gradients, their difference is the ATE EIF by linearity. A direct minimization of the variance of $\phi_1 - \phi_0$ over restricted calibration classes is a separate finite-dimensional or restricted-class optimization problem unless the joint IF class has been characterized.

Remark (Three layers of efficiency statements). For any valid calibration representative $(q, b) \in \mathcal{Q}_1 \times \mathcal{B}_1$, the centered score $\phi_1(O; q, b)$ is mean-zero and Neyman-orthogonal (Theorem E.6); it is

a *regular* influence function only when the finite-variance and local-calibration-path conditions of Proposition E.7 hold. Fixed-q variance minimization selects the best adjoint representer within a fixed-primal affine class, attaining the minimum calibrated orthogonal moment variance for that $q^\dagger$. The semiparametric efficiency bound is achieved only when the canonical gradient is represented, or approximated in L^2 closure, by the regular calibration IF subclass $\Phi_{\mathrm{cal}}^{\mathrm{reg}}$.

Remark (Observed calibration model). The efficiency statements are relative to the observed local calibration model. They do not claim efficiency in the full latent generalized Roy model, whose tangent space may be smaller because of additional latent structural restrictions.

Remark (Minimum norm versus minimum variance). A Moore–Penrose or Tikhonov representative is a minimum-norm object used for stability. A minimum-variance representative minimizes the variance of the calibrated orthogonal moment. These objectives generally differ.

Remark (Economic interpretation). When the calibrating functions are nonunique, different calibrating functions can produce the same target. The adjoint representer selects a way to make the target invariant; the minimum-variance representative selects, among admissible invariance certificates, the one that yields the most precise regular score within the relevant class.

Layer-by-layer derivation of the EIF. The EIF for μ_1 in the local observed calibration model is derived in three logical steps, each adding a layer of structure.

Step A: Orthogonal score class. For any $(q, b) \in \mathcal{Q}_1 \times \mathcal{B}_1$, the calibrated orthogonal moment $\Gamma_1(O; q, b) - \mu_1$ is mean zero and Neyman-orthogonal (Theorem E.6). This gives an entire class of orthogonal scores, parameterized by the choice of (q, b).

Step B: Regular IF subset. Not all orthogonal scores are regular IFs. The regular subset consists of those (q_0, b_0) for which the local calibration submodel admits a differentiable path through (q_0, b_0) and the finite-variance condition holds (Proposition E.7). The regular subset is generally smaller than the full orthogonal class.

Step C: Variance-minimizing representative. Among the regular IFs, the EIF is the one with minimum variance. The variance-minimizing pair $(q^{\mathrm{EIF}}, b^{\mathrm{EIF}})$ is the solution of the joint variance-minimization problem over $\mathcal{Q}_1 \times \mathcal{B}_1$ subject to the regularity conditions.

Why the EIF is unique even when the representatives are not. The calibration representative $q \in \mathcal{Q}_1$ and the adjoint representer $b \in \mathcal{B}_1$ are unique only modulo $\ker(A_1)$ and $\ker(A_1^\star)$ respectively. Different representatives may generate different regular influence-function candidates for the same target μ_1, so the IF candidate set is an *affine class* of $L^2(P)$ functions sharing the same pathwise derivative. The canonical gradient (EIF) is unique because it is the projection of any regular IF onto the tangent space of the model – equivalently, the unique minimum-variance element of the closure of this affine class. A calibration representative that yields a mean-zero L^2 score does not by itself imply that score is the EIF: only the canonical-gradient representative gives the EIF. Hence the EIF ϕ_1^{EIF} is unique as an element of $L_0^2(P)$, but the pair $(q^{\mathrm{EIF}}, b^{\mathrm{EIF}})$ realizing it may still be one of multiple equivalent representatives modulo the kernel directions.

Anchored versus unanchored efficiency. The unanchored efficiency bound considers all regular IFs from any anchored representative. The anchored efficiency at a specific (q_0, b_0) considers only the IFs that anchor at this representative. The unanchored bound is generally smaller (the minimum over a larger class), but the practical implementation usually anchors at a regularized estimator and so achieves the anchored bound.

Comparison with proximal causal inference efficiency theory. Tchetgen Tchetgen et al. (2024) and Cui et al. (2024) develop the EIF for proximal causal inference under con-

ditional treatment ignorability. The structure is similar (a primal-adjoint representer pair, an orthogonal score, a variance minimization), but the calibrating functions are different (proximal bridges versus selection-odds representers) and the conditioning structure is different (latent-only versus latent-plus-potential-outcome). The efficiency bounds are also different: proximal causal inference efficiency assumes treatment ignorability, which is generally weaker than the calibration restrictions used here when selection on gains is present.

F Sieve GMM and cross-fitted DML estimation

This appendix gives the cross-fitted Double Machine Learning (DML) estimator implied by the primal-adjoint calibration system. The role of cross-fitting is standard: nuisance functions are estimated on training folds and the orthogonal identification signal is evaluated on held-out folds.

Algorithm 1 Cross-fitted DML for μ_1 and, symmetrically, μ_0

for $k = 1, \dots, K$ **do**
  Estimate $\hat{q}_1^{(k)}$ on I_k^c from the primal moment (7).
  Estimate $\hat{m}_1^{(k)}$ on I_k^c from (17).
  Estimate $\hat{h}_1^{(k)}$ on I_k^c from (12), and set $\hat{b}_1^{(k)} = \hat{m}_1^{(k)} + \hat{h}_1^{(k)}$.
  for $i \in I_k$ **do**
    Compute $\hat{\Gamma}_{1i} = D_i\{1 + \hat{q}_1^{(k)}(X_{1i})\}\{Y_i - \hat{b}_1^{(k)}(Z_{1i})\} + \hat{b}_1^{(k)}(Z_{1i})$.
  end for
end for
Return $\hat{\mu}_1 = N^{-1}\sum_i \hat{\Gamma}_{1i}$ and $\hat{\sigma}_1^2 = N^{-1}\sum_i(\hat{\Gamma}_{1i} - \hat{\mu}_1)^2$.

Lemma F.1 (Exact product-bias identity)**.** *For $(q_0, b_0) \in \mathcal{Q}_1 \times \mathcal{B}_1$, $\Delta_q = q - q_0$, and $\Delta_b = b - b_0$,*

$$\mathbb{E}[\Gamma_1(O; q, b)] - \mathbb{E}[\Gamma_1(O; q_0, b_0)] = -\mathbb{E}[D\Delta_q(X_1)\Delta_b(Z_1)]. \tag{78}$$

If (q_0, b_0) is a valid calibration pair, then $\mathbb{E}[\Gamma_1(O; q_0, b_0)] = \mu_1$.

Proof. The identification functional in (q, b) form is

$$\Gamma_1(O; q, b) = D(1+q)(Y-b) + b = D(1+q)Y - D(1+q)b + b = D(1+q)Y - \{D(1+q) - 1\}b.$$

This is affine in q and affine in b separately, with a single bilinear term $D\,q \cdot b$. Expanding $\Gamma_1(O; q_0 + \Delta_q, b_0 + \Delta_b)$ around (q_0, b_0) gives

$$\begin{aligned}\Gamma_1(O; q_0 + \Delta_q, b_0 + \Delta_b) = \Gamma_1(O; q_0, b_0) + \partial_q\Gamma_1 \cdot \Delta_q + \partial_b\Gamma_1 \cdot \Delta_b \\ + \partial^2_{qb}\Gamma_1 \cdot \Delta_q\Delta_b,\end{aligned}$$

where the cross-derivative is the bilinear coefficient $-D$ in the expansion. Taking expectations and using

$$\begin{aligned}\mathbb{E}[\partial_q\Gamma_1 \cdot \Delta_q] &= \mathbb{E}[D\Delta_q(X_1)\{Y - b_0(Z_1)\}] = \langle \Delta_q, \ell_1 - A_1^\star b_0\rangle_{X_1} = 0, \\ \mathbb{E}[\partial_b\Gamma_1 \cdot \Delta_b] &= \mathbb{E}[\{1 - D(1 + q_0(X_1))\}\Delta_b(Z_1)] = 0,\end{aligned}$$

the first-order terms vanish at $(q_0, b_0) \in \mathcal{Q}_1 \times \mathcal{B}_1$. Only the cross-product remains, giving

$$\mathbb{E}[\Gamma_1(O; q, b)] - \mu_1 = -\mathbb{E}[D\,\Delta_q(X_1)\,\Delta_b(Z_1)].$$

This is the exact product-bias identity. By Cauchy–Schwarz it satisfies $\left|\mathbb{E}[\Gamma_1(O; q, b)] - \mu_1\right| \leq \|\Delta_q\|_{D,2}\|\Delta_b\|_{D,2}$, which is the basis for the product-rate condition. □

Corollary F.2 ((q, m, h) version)**.** *When $b = m + h$,*

$$\mathbb{E}[\Gamma_1(O; q, m, h)] - \mathbb{E}[\Gamma_1(O; q_0, m_0, h_0)] = -\mathbb{E}[D\Delta_q\Delta_m] - \mathbb{E}[D\Delta_q\Delta_h]. \tag{79}$$

Proof. The identification functional in (q, m, h) form is

$$\Gamma_1(O; q, m, h) = D(1 + q)(Y - m) + m - \{D(1 + q) - 1\}h.$$

It has the following structure:

- Affine in q (degree at most 1): the form $D(1 + q)(Y - m - h) + m + h$.
- Linear in m and h separately.
- Bilinear cross-terms: $D\,q \cdot m$ and $D\,q \cdot h$.
- No cross-term between m and h (both are pure Z_1 functions that combine into $b = m + h$ in the canonical form).

Expanding $\Gamma_1(O; q_0 + \Delta_q, m_0 + \Delta_m, h_0 + \Delta_h)$ around the calibration representative (q_0, m_0, h_0) gives

$$\begin{aligned}\Gamma_1(O; q, m, h) - \Gamma_1(O; q_0, m_0, h_0) = {} & \partial_q\Gamma_1 \cdot \Delta_q + \partial_m\Gamma_1 \cdot \Delta_m + \partial_h\Gamma_1 \cdot \Delta_h \\ & + \partial^2_{qm}\Gamma_1 \cdot \Delta_q\Delta_m + \partial^2_{qh}\Gamma_1 \cdot \Delta_q\Delta_h.\end{aligned}$$

The first-order q-term:

$$\mathbb{E}[\partial_q\Gamma_1 \cdot \Delta_q] = \mathbb{E}[D\Delta_q(X_1)\{Y - m_0(Z_1) - h_0(Z_1)\}] = \langle \Delta_q, \ell_1 - A_1^\star(m_0 + h_0)\rangle_{X_1} = 0,$$

since $b_0 = m_0 + h_0 \in \mathcal{B}_1$ satisfies $A_1^\star b_0 = \ell_1$.

The first-order m-term:

$$\mathbb{E}[\partial_m\Gamma_1 \cdot \Delta_m] = \mathbb{E}[\{1 - D(1 + q_0)\}\Delta_m(Z_1)] = 0,$$

since $q_0 \in \mathcal{Q}_1$ satisfies $\mathbb{E}[D(1 + q_0) - 1 \mid Z_1] = 0$.

The first-order h-term: identical to the m-term by the same argument:

$$\mathbb{E}[\partial_h\Gamma_1 \cdot \Delta_h] = -\mathbb{E}[\{D(1 + q_0) - 1\}\Delta_h(Z_1)] = 0.$$

The second-order cross terms give

$$\mathbb{E}[\partial^2_{qm}\Gamma_1 \cdot \Delta_q\Delta_m] = -\mathbb{E}[D\Delta_q(X_1)\Delta_m(Z_1)], \qquad \mathbb{E}[\partial^2_{qh}\Gamma_1 \cdot \Delta_q\Delta_h] = -\mathbb{E}[D\Delta_q(X_1)\Delta_h(Z_1)].$$

Combining:

$$\mathbb{E}[\Gamma_1(O;q,m,h)] - \mu_1 = -\mathbb{E}[D\Delta_q(X_1)\Delta_m(Z_1)] - \mathbb{E}[D\Delta_q(X_1)\Delta_h(Z_1)].$$

This is the exact product-bias identity in (q,m,h) form. By Cauchy–Schwarz it implies

$$\big|\mathbb{E}[\Gamma_1(O;q,m,h)] - \mu_1\big| \le \|\Delta_q\|_{D,2}\big(\|\Delta_m\|_{D,2} + \|\Delta_h\|_{D,2}\big),$$

which yields the product-rate condition $\sqrt{N}\|\hat{q} - q^\dagger\|_{D,2}\big(\|\hat{m} - m^\dagger\|_{D,2} + \|\hat{h} - h^\dagger\|_{D,2}\big) = o_p(1)$ for $\sqrt{N}$-asymptotic normality of the cross-fitted estimator. □

Remark (Product-rate condition). By Cauchy-Schwarz,

$$\big|\mathbb{E}[\Gamma_1(O;\hat{q},\hat{b})] - \mu_1\big| \le \|\hat{q} - q^\dagger\|_{D,2}\|\hat{b} - b^\dagger\|_{D,2}. \tag{80}$$

Thus a sufficient high-level condition is

$$\sqrt{N}\|\hat{q} - q^\dagger\|_{D,2}\|\hat{b} - b^\dagger\|_{D,2} = o_p(1). \tag{81}$$

The three-part implementation $b = m+h$ yields the corresponding bound with $\|\hat{b}-b^\dagger\|_{D,2}$ replaced by $\|\hat{m} - m^\dagger\|_{D,2} + \|\hat{h} - h^\dagger\|_{D,2}$.

Remark (Why no $O(\|\Delta\|^3)$ term appears). The exact product-bias identity expresses the second-order bias of plug-in calibration estimators as the bilinear $\mathbb{E}[D\Delta_q\Delta_b]$, which is bounded by $\|\Delta_q\|_{D,2} \cdot \|\Delta_b\|_{D,2}$. There is no $O(\|\Delta_q\|^2)$ or $O(\|\Delta_b\|^2)$ term, and no $O(\|\Delta\|^3)$ cubic term either. The reason is that the calibrated orthogonal moment is bilinear in (q,b): $\Gamma_1(O;q,b) = D(1+q)(Y-b) + b$ is linear in b for fixed q and linear in q for fixed b. The first-order von Mises derivative of $\mu_1(q,b)$ in (q,b) thus has a bilinear (product-bias) form rather than a quadratic one, and there are no third-order terms.

Remark (Smooth selection probabilities and Hermite Picard conditions). For smooth latent treatment probabilities $\pi(\cdot)$ that admit Hermite expansion, the Picard condition for (R1-A) reduces to a decay condition on the Hermite coefficients of the latent odds ratio relative to the spectral decay of the *latent W-side operator* T_W, not the observed adjoint $A_1^\star$. Specifically, if the Hermite-coefficient sequence of the latent odds ratio is summable relative to the singular values of T_W, then q^ℓ has a finite-norm Hermite representation and (R1-A) holds. The observed adjoint range condition (R2) is governed by the separate Picard summability of the outcome regression with respect to the singular values of $A_1^\star$, and these two summabilities concern different operators acting on different Hilbert spaces (Appendix D.2). The non-Gaussian extension uses orthogonal-polynomial systems matched to the marginal of W.

Remark (Meaning of the non-Gaussian theorem). The non-Gaussian completeness result establishes (I0) under deconvolution conditions on the measurement W that do not require Gaussianity. The proof uses the Fourier-transform characterization of completeness due to Hu and Shiu (2018) and characterizes the strength of W as a measurement through the smoothness of its characteristic function. The result extends the linear-Gaussian sufficient condition of Proposition D.3 to a more general non-parametric setting while preserving the identification anchor.

Remark (Meaning of Corollary D.8). The dichotomy between ordinary-smooth and supersmooth measurements in Corollary D.8 corresponds to the standard distinction in the deconvolution literature (Fan, 1991). Ordinary-smooth measurements (polynomial decay of the characteristic

function) admit (R1-A) under the existence-type source condition $r^\ell = (T_W T_W^\star)^{\beta/2} w$ with $\beta \geq 1$; weaker smoothness $0 < \beta < 1$ describes regularity of r^ℓ relative to the operator but does not by itself guarantee finite-norm latent-calibration existence. Supersmooth measurements (exponential decay, e.g., Gaussian) require (R1-A) to hold under analytic-extension assumptions on the latent representer. This distinction has direct implications for the Tikhonov convergence rate of Remark D.4: ordinary-smooth measurements give polynomial rates, while supersmooth measurements give logarithmic rates.

Remark (Meaning of the proposition). The proposition establishes that under the source-condition smoothness $\beta \geq 1$, the latent representer q^ℓ admits a finite-norm representation in the range of $T_W^\star$ (strict range). For $0 < \beta < 1$, the source representation gives smoothness relative to the operator and is sufficient for Tikhonov-regularized estimation with controlled rate (Remark D.4), but does not by itself guarantee strict range; the parameter may be only irregularly identified (Case I of the trichotomy) in this regime. This is what makes the identification operator $A_1^\star$ amenable to Tikhonov regularization with controlled bias-variance trade-off. The smoothness parameter β is a primitive property of the measurement structure $g_W(U, C, \eta_W)$; concrete cases of β for specific structural specifications are catalogued in Appendix D.4.

Theorem F.3 (Product-bias bound)**.** *For any selected representative* $(q^\dagger, b^\dagger)$*,*

$$\left|\mathbb{E}[\Gamma_1(O; \hat{q}, \hat{b})] - \mu_1\right| \leq \|\hat{q} - q^\dagger\|_{D,2} \|\hat{b} - b^\dagger\|_{D,2}, \tag{82}$$

and, in the (q, m, h) *implementation,*

$$\left|\mathbb{E}[\Gamma_1(O; \hat{q}, \hat{m}, \hat{h})] - \mu_1\right| \leq \|\hat{q} - q^\dagger\|_{D,2} \{\|\hat{m} - m^\dagger\|_{D,2} + \|\hat{h} - h^\dagger\|_{D,2}\}. \tag{83}$$

Proof. (q, b) **form.** The exact product-bias identity (78) of Lemma F.1 combined with the Cauchy–Schwarz inequality yields $|\mathbb{E}[D\Delta_q\Delta_b]| \leq \|\Delta_q\|_{D,2} \cdot \|\Delta_b\|_{D,2}$.

(q, m, h) **form.** The exact identity (79) of Corollary F.2 together with Cauchy–Schwarz gives

$$|\mathbb{E}[D\,\Delta_q\,\Delta_m]| \leq \|\Delta_q\|_{D,2} \cdot \|\Delta_m\|_{D,2}, \qquad |\mathbb{E}[D\,\Delta_q\,\Delta_h]| \leq \|\Delta_q\|_{D,2} \cdot \|\Delta_h\|_{D,2}.$$

Combining both yields the bound in (83). □

Remark (Regularity and efficiency). The calibrated orthogonal moment is mean-zero and Neyman orthogonal for every valid (q, b). It is a regular influence function only under finite-variance and local-calibration-path conditions. It is an efficient influence function only under the additional condition that the canonical gradient belongs to the $L^2(P)$ closure of the influence-function class. Joint minimum-variance selection is generally nonconvex, while fixed-q minimum-variance selection is a convex projection.

Remark (Ill-posedness and source conditions). The nuisance functions solve inverse problems. Their convergence rates depend on singular value decay, source smoothness, sieve dimension, and regularization. The DML theorem therefore uses the product-rate condition as a high-level assumption and interprets empirical singular values and condition numbers as weak-range diagnostics.

Remark (Why standard NPIV rates do not transfer directly). The convergence rate of Remark D.4 resembles the standard Tikhonov NPIV rate of Hall and Horowitz (2005) and Chen and Pouzo (2012), but a direct importing of those results requires care. Standard NPIV considers $\mathbb{E}[Y -$

$g(X) \mid Z] = 0$ where the unknown is the function $g(X)$ identified from Z-instruments. The present calibration problem has a different structure: the unknown is the calibration weight $1 + q(X_1)$ on $\{D = 1\}$, and the moment is weighted by D. The effective sample size on $\{D = 1\}$, the conditional density structure on $\{D = 1\}$, and the truncation effects all enter the rate analysis. The rate $n^{-\beta/(2\beta+2\nu+1)}$ that appears in Remark D.4 corresponds to these modifications and should not be conflated with the standard NPIV rate.

Assumption F.1 (High-level regularity)**.** For a selected representative $(q^\dagger, b^\dagger)$, assume cross-fitting, finite second moments, L^2 consistency of the nuisance estimators, and

$$\sqrt{N}\|\hat{q} - q^\dagger\|_{D,2}\|\hat{b} - b^\dagger\|_{D,2} = o_p(1). \tag{84}$$

Theorem F.4 (Cross-fitted asymptotic normality)**.** *Under Assumption F.1,*

$$\sqrt{N}(\hat{\mu}_1 - \mu_1) \rightsquigarrow N(0, \sigma_1^2), \qquad \sigma_1^2 = \mathbb{E}[\{\Gamma_1(O; q^\dagger, b^\dagger) - \mu_1\}^2].$$

The plug-in variance estimator is consistent. The same statement applies to μ_0 and hence to the average treatment effect.

Why cross-fitting is appropriate here. Standard cross-fitted DML requires orthogonal score functions and the product-rate condition for nuisance estimation. The auxiliary-measurement framework supplies the orthogonal score structurally: orthogonality follows from the primal-dual identity (Theorem E.6). Root-N cross-fitted DML inference then additionally requires the high-level product-rate condition for the first-stage calibration estimators, established under a separate first-stage rate assumption, which controls the second-order remainder of the product-bias identity (Theorem F.3). The auxiliary-measurement framework is therefore a setting in which cross-fitted DML is directly applicable without ad hoc modifications.

Treated-effective sample size for first-stage estimation. The dual moment determining b is D-weighted (its first term $D(1+q)(Y-b)$ vanishes on $\{D=0\}$), so the first-stage Tikhonov estimator $\widehat{b}$ effectively uses the treated sample size $N \cdot \mathbb{P}(D=1)$ rather than N. The calibrated orthogonal moment $\Gamma_1 = D(1+q)(Y-b) + b$ itself, however, is defined and evaluated on the full sample (the second term $b(Z_1)$ is observed for all units), so the final cross-fitted estimator $\widehat{\mu}_1^{\mathrm{DML}}$ has its CLT driven by the full-sample variance of Γ_1. In imbalanced applications (rare treatment), the first-stage rate δ_n^b should be computed using $N \cdot \mathbb{P}(D=1)$ even though the final inference uses the full N. The cross-fitted DML procedure preserves this scaling, but the practitioner should ensure that each fold contains a reasonable fraction of $D=1$ observations to avoid degenerate estimates of $\widehat{q}, \widehat{b}$.

Bootstrap-based standard errors as a robustness check. In weak-range or near-irregular regimes (near-Case-I of the trichotomy), where the empirical Picard profile suggests that the adjoint range condition is weakly satisfied or nearly violated, the asymptotic variance formula may understate the true sampling uncertainty due to weak-range effects. A bootstrap-based standard error provides an empirical sanity check: substantially wider bootstrap intervals than analytic intervals indicate that the asymptotic regime is not well-approximated in the sample at hand. Bootstrap intervals do not remedy Case N non-invariance, in which calibration-based point identification itself fails.

Proof. Write $\eta^\dagger = (q^\dagger, b^\dagger)$ for the selected calibration representative and $\hat{\eta}^{(k)} = (\hat{q}^{(k)}, \hat{b}^{(k)})$ for the

nuisance estimates trained on the complement of fold I_k. The cross-fitted estimator is

$$\hat{\mu}_1^{\mathrm{DML}} = \frac{1}{N}\sum_{k=1}^{K}\sum_{i\in I_k}\Gamma_1(O_i;\hat{q}^{(k)},\hat{b}^{(k)}),$$

where $\Gamma_1(O;q,b) = D\{1+q(X_1)\}\{Y-b(Z_1)\}+b(Z_1)$.

Step 1: oracle decomposition. Define $\Gamma_1^\dagger(O) = \Gamma_1(O;q^\dagger,b^\dagger)$ and split

$$\sqrt{N}(\hat{\mu}_1^{\mathrm{DML}}-\mu_1) = \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\{\Gamma_1^\dagger(O_i)-\mu_1\} + \sqrt{N}\,R_N,$$

where

$$R_N = \frac{1}{N}\sum_{k=1}^{K}\sum_{i\in I_k}\{\Gamma_1(O_i;\hat{q}^{(k)},\hat{b}^{(k)})-\Gamma_1^\dagger(O_i)\}.$$

The first term is the influence-function contribution. By the central limit theorem,

$$\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\{\Gamma_1^\dagger(O_i)-\mu_1\} \rightsquigarrow \mathcal{N}(0,\sigma_1^2), \quad \sigma_1^2 = \mathbb{E}[\{\Gamma_1^\dagger(O)-\mu_1\}^2].$$

It remains to show $\sqrt{N}\,R_N = o_p(1)$.

Step 2: decomposition of the remainder. Write $R_N = R_N^{(\mathrm{e})} + R_N^{(\mathrm{b})}$, where

$$R_N^{(\mathrm{e})} = \frac{1}{N}\sum_{k=1}^{K}\sum_{i\in I_k}\big\{\Gamma_1(O_i;\hat{\eta}^{(k)})-\Gamma_1^\dagger(O_i)-\mathbb{E}_O[\Gamma_1(\cdot;\hat{\eta}^{(k)})-\Gamma_1^\dagger \mid \hat{\eta}^{(k)}]\big\}$$
$$R_N^{(\mathrm{b})} = \frac{1}{N}\sum_{k=1}^{K}\sum_{i\in I_k}\mathbb{E}_O[\Gamma_1(\cdot;\hat{\eta}^{(k)})-\Gamma_1^\dagger \mid \hat{\eta}^{(k)}].$$

The first part $R_N^{(\mathrm{e})}$ is the empirical-process term conditional on the trained $\hat{\eta}^{(k)}$. By cross-fitting, $\hat{\eta}^{(k)}$ is independent of $\{O_i : i\in I_k\}$, so conditional on $\hat{\eta}^{(k)}$ the summands are independent and mean zero. Their conditional variance is bounded by the score-norm error

$$\mathrm{Var}\big(\Gamma_1(O;\hat{\eta}^{(k)})-\Gamma_1^\dagger(O)\,\big|\,\hat{\eta}^{(k)}\big) \le \mathbb{E}\big[\{\Gamma_1(O;\hat{\eta}^{(k)})-\Gamma_1^\dagger(O)\}^2\,\big|\,\hat{\eta}^{(k)}\big].$$

Under conditions (C3) and (C3′), this conditional second moment is $o_p(1)$. By Chebyshev's inequality applied conditionally on $\hat{\eta}^{(k)}$,

$$\sqrt{N}\,R_N^{(\mathrm{e})} = o_p(1).$$

This route avoids invoking Donsker or chaining conditions on the first-stage estimators, leveraging the sample-splitting structure directly. The second part $R_N^{(\mathrm{b})}$ is the bias term.

Step 3: product-bias bound for the bias term. By the exact product-bias identity (Lemma F.1), for each fold,

$$\mathbb{E}_O[\Gamma_1(\cdot;\hat{\eta}^{(k)})-\Gamma_1^\dagger \mid \hat{\eta}^{(k)}] = -\mathbb{E}_O[D\{\hat{q}^{(k)}(X_1)-q^\dagger(X_1)\}\{\hat{b}^{(k)}(Z_1)-b^\dagger(Z_1)\} \mid \hat{\eta}^{(k)}].$$

By Cauchy–Schwarz,

$$\left|\mathbb{E}_O[\Gamma_1(\cdot;\hat{\eta}^{(k)}) - \Gamma_1^\dagger \mid \hat{\eta}^{(k)}]\right| \le \|\hat{q}^{(k)} - q^\dagger\|_{D,2}\, \|\hat{b}^{(k)} - b^\dagger\|_{D,2}.$$

Hence

$$\sqrt{N}\left|R_N^{(\mathrm{b})}\right| \le \sqrt{N}\, K^{-1} \sum_{k=1}^{K} \|\hat{q}^{(k)} - q^\dagger\|_{D,2}\, \|\hat{b}^{(k)} - b^\dagger\|_{D,2}.$$

By Assumption F.1 (the product-rate condition), this is $o_p(1)$.

Step 4: combining. Combining Steps 1–3,

$$\sqrt{N}(\hat{\mu}_1^{\mathrm{DML}} - \mu_1) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \{\Gamma_1^\dagger(O_i) - \mu_1\} + o_p(1) \rightsquigarrow \mathcal{N}(0, \sigma_1^2).$$

Variance estimator consistency. The plug-in variance estimator $\hat{\sigma}_1^2 = N^{-1}\sum_i (\hat{\Gamma}_{1,i} - \hat{\mu}_1)^2$ satisfies $\hat{\sigma}_1^2 \to \sigma_1^2$ in probability under the conditions (C1)–(C4) together with (C3′), via the following route. First, $\|\Gamma_1(\cdot;\hat{\eta}^{(k)}) - \Gamma_1(\cdot;\eta^\dagger)\|_{L^2(P)} = o_p(1)$ by (C3) (D-weighted convergence) and (C3′) (full-sample b-norm consistency), with bounded calibration weights ensuring uniform integrability. Second, conditional on the trained $\hat{\eta}^{(k)}$, the evaluation-fold summands are i.i.d. with conditional variance equal to $\mathbb{E}[(\Gamma_1(\cdot;\hat{\eta}^{(k)}) - \mu_1)^2 \mid \hat{\eta}^{(k)}]$, which converges in probability to $\mathbb{E}[(\Gamma_1^\dagger - \mu_1)^2] = \sigma_1^2$ by L^2-continuity of the score. Third, the empirical second moment converges to its conditional expectation by Chebyshev's inequality applied conditionally on $\hat{\eta}^{(k)}$. This route does not require chaining or Donsker conditions on $\hat{\eta}^{(k)}$; it uses the score-norm convergence and the sample-splitting structure directly.

The Y_0 side is handled by the same argument applied to the symmetric operators and calibrating functions, and the average treatment effect estimator $\hat{\tau} = \hat{\mu}_1^{\mathrm{DML}} - \hat{\mu}_0^{\mathrm{DML}}$ inherits $\sqrt{N}$-asymptotic normality with variance $\sigma_\tau^2 = \mathbb{E}[\{\Gamma_1^\dagger - \Gamma_0^\dagger - \tau\}^2]$. □

Remark (Representative-dependent variance). Identification does not depend on the selected representative, but the variance does. Minimum-norm and minimum-variance representatives generally differ.

F.1 Exact Sieve GMM implementation

This subsection records the detailed cross-fitting algorithm that implements the calibration-based DML estimator.

Detailed cross-fitting algorithm.

The cross-fitted DML procedure implements the calibration identification through sample splitting. The algorithm is as follows.

Algorithm (Cross-Fitted DML for the Calibration ATE).

1. Partition the sample $\{O_i\}_{i=1}^N$ into K disjoint folds $\mathcal{I}_1, \ldots, \mathcal{I}_K$ of roughly equal size.
2. For each fold $k = 1, \ldots, K$:
   (a) Using the complementary sample $\mathcal{I}_{-k} := \bigcup_{j \neq k} \mathcal{I}_j$, estimate the observed calibration $\widehat{q}^{(-k)}$ by Tikhonov-regularized least squares or extended Sieve GMM, using the *full-*

sample primal moment $\mathbb{E}[Dq(X_1) - (1-D) \mid Z_1] = 0$ over all observations in $\mathcal{I}_{-k}$ (both treatment arms). The function $q(X_1)$ is evaluated only for treated observations (because it is multiplied by D), but the moment itself uses both arms through the $(1-D)$ term.

(b) Using $\mathcal{I}_{-k}$, estimate the adjoint representer $\widehat{b}^{(-k)}$ by the fixed-$\widehat{q}^{(-k)}$ extended SGMM.

(c) Compute the fold-k score values: for each $i \in \mathcal{I}_k$,

$$\widehat{\Gamma}_1^{(k)}(O_i) := D_i\{1 + \widehat{q}^{(-k)}(X_{1,i})\}\{Y_i - \widehat{b}^{(-k)}(Z_{1,i})\} + \widehat{b}^{(-k)}(Z_{1,i}).$$

3. Aggregate across folds:

$$\widehat{\mu}_1^{\mathrm{DML}} := \frac{1}{N}\sum_{k=1}^{K}\sum_{i\in\mathcal{I}_k}\widehat{\Gamma}_1^{(k)}(O_i).$$

4. Compute the variance estimator:

$$\widehat{\sigma}_1^2 := \frac{1}{N}\sum_{k=1}^{K}\sum_{i\in\mathcal{I}_k}\big[\widehat{\Gamma}_1^{(k)}(O_i) - \widehat{\mu}_1^{\mathrm{DML}}\big]^2,$$

yielding the standard error $\widehat{\mathrm{SE}}(\widehat{\mu}_1^{\mathrm{DML}}) = \widehat{\sigma}_1/\sqrt{N}$.

The procedure for $\widehat{\mu}_0^{\mathrm{DML}}$ is symmetric. The ATE estimator is $\widehat{\tau}^{\mathrm{DML}} = \widehat{\mu}_1^{\mathrm{DML}} - \widehat{\mu}_0^{\mathrm{DML}}$. The asymptotic variance estimator uses the influence function $\Gamma_\tau = \Gamma_1 - \Gamma_0 - \tau$ evaluated on the same observations,

$$\widehat{V}_\tau = \frac{1}{N}\sum_{i=1}^{N}\big\{\widehat{\Gamma}_{1,i} - \widehat{\Gamma}_{0,i} - \widehat{\tau}\big\}^2,$$

which automatically accounts for the covariance $\mathrm{Cov}(\Gamma_1, \Gamma_0)$ between the two sides; a naive sum $\widehat{V}_{\mu_1} + \widehat{V}_{\mu_0}$ would omit the $-2\mathrm{Cov}(\Gamma_1, \Gamma_0)$ cross-term.

F.2 Cross-fitted DML algorithm and product-rate theory

This subsection records the asymptotic-normality theorem for the cross-fitted DML estimator under high-level product-rate conditions, together with effective-sample-size considerations.

Asymptotic theory under product-rate conditions.

The asymptotic theory of cross-fitted DML for calibration estimation extends the standard DML theory (Chernozhukov et al., 2018) to the calibration setting. The key conditions are:

(C1) **Identification regularity:** Layer B of the singular-value hierarchy (Appendix D.4) holds; that is, $\ell_1 \in \mathcal{R}(A_1^\star)$ and the adjoint representer has finite norm.

(C2) **Finite variance (FV):** $\mathbb{E}[D\{1 + q(X_1)\}^2\{Y - b(Z_1)\}^2] < \infty$ and $\mathbb{E}[b(Z_1)^2] < \infty$.

(C3) **First-stage convergence (product-rate norms):** $\|\widehat{q}^{(-k)} - q^\dagger\|_{D,2} = O_p(\delta_n^q)$ and $\|\widehat{b}^{(-k)} - b^\dagger\|_{D,2} = O_p(\delta_n^b)$ for some rates $\delta_n^q, \delta_n^b = o(1)$.

(C3′) **Full-sample b-norm consistency:** the final score $\Gamma_1 = D\{1+q(X_1)\}\{Y-b(Z_1)\}+b(Z_1)$ carries the $b(Z_1)$ term on *all* units, including $\{D=0\}$ ones where the dual moment does not directly constrain b. We therefore require, in addition to the D-weighted rate of (C3), the full-sample consistency

$$\|\widehat{b}^{(-k)} - b^\dagger\|_{\omega,2} = o_p(1), \qquad \|f\|^2_{\omega,2} := \mathbb{E}\big[\{1-D(1+q^\dagger(X_1))\}^2 f(Z_1)^2\big],$$

or, simply, $\|\widehat{b}^{(-k)} - b^\dagger\|_{Z,2} = o_p(1)$ together with bounded calibration weights $\sup_x |1+q^\dagger(x)| < \infty$. This controls the empirical-process and variance terms involving the $\{1-D(1+q^\dagger)\}\Delta_b$ contribution on the full sample.

(C4) **Product-rate condition:** $\delta_n^q \cdot \delta_n^b = o(N^{-1/2})$.

Under (C1)–(C4) (with (C3′) controlling the full-sample b contribution), the cross-fitted DML estimator satisfies

$$\sqrt{N}(\widehat{\mu}_1^{\mathrm{DML}} - \mu_1) \to_d \mathcal{N}\big(0,\, \mathbb{E}[(\Gamma_1(O; q^\dagger, b^\dagger) - \mu_1)^2]\big),$$

where $(q^\dagger, b^\dagger)$ are the asymptotic targets of the first-stage estimators.

Role of Neyman orthogonality. The product-rate condition (C4) requires only $\delta_n^q \cdot \delta_n^b = o(N^{-1/2})$, which is weaker than either individual rate being $o(N^{-1/2})$. This relaxation is the principal advantage of cross-fitted DML and follows from the Neyman orthogonality of the calibrated orthogonal moment (Theorem E.6).

Role of cross-fitting. The cross-fitting eliminates own-observation bias from the first-stage estimators. Without cross-fitting, the score $\widehat{\Gamma}_1(O_i; \widehat{q}, \widehat{b})$ depends on O_i through both the score variables and the nuisance estimators, introducing a bias of order $O(N^{-1/2})$. With cross-fitting, the nuisance estimators are independent of O_i conditional on the complementary sample, eliminating the own-observation bias.

Effective sample size considerations.

The dual moment that pins down b is D-weighted, so the first-stage rate δ_n^b is effectively determined by the treated sample size $N \cdot p_D$ where $p_D = \mathbb{P}(D=1)$. The calibrated orthogonal moment $\Gamma_1 = D(1+q)(Y-b)+b$ is defined for all units (the $b(Z_1)$ term is observed regardless of D), so the final estimator $\widehat{\mu}_1^{\mathrm{DML}}$ has its standard error driven by the full-sample variance of Γ_1 with $\sqrt{N}$ scaling. In rare-treatment applications, the first-stage product-rate condition should be checked using $N \cdot p_D$ for the b-side rate, but the final inference still uses $\sqrt{N}$.

Practical implications.

1. Each cross-fitting fold should contain a reasonable fraction of $\{D=1\}$ observations. Stratified splitting (stratifying on D) is recommended.

2. The first-stage estimator $\widehat{b}$ is identified from the $\{D=1\}$ subsample. The rate δ_n^b should be computed using the effective sample size $N \cdot p_D$ rather than N.

3. The product-rate condition (C4) should be assessed conservatively: in rare-treatment settings, achieving $\delta_n^q \cdot \delta_n^b = o(N^{-1/2})$ may require richer first-stage estimators than would be needed in balanced settings.

F.3 Bootstrap and weak-range inference

This subsection describes the bootstrap-based standard errors used to complement the analytic asymptotic-variance formula, particularly in regimes where the adjoint representer is weakly identified.

Bootstrap-based standard errors.

In weak-range or near-irregular regimes (near-Case-I of the trichotomy), where the adjoint range condition is weakly satisfied, the asymptotic variance formula may understate the true sampling uncertainty due to weak-range effects. Bootstrap-based standard errors provide an empirical sanity check.

Pairs bootstrap procedure.

1. Resample observations $\{O_i\}$ with replacement to form bootstrap sample $\{O_i^*\}$.
2. Apply the full cross-fitted DML procedure (including refitting $\widehat{q}, \widehat{b}$ on the bootstrap sample) to compute $\widehat{\mu}_1^{\mathrm{DML},*}$.
3. Repeat B times to obtain bootstrap replications $\{\widehat{\mu}_1^{\mathrm{DML},*,(b)}\}_{b=1}^{B}$.
4. Compute the bootstrap variance estimator $\widehat{\mathrm{Var}}^*(\widehat{\mu}_1^{\mathrm{DML}}) = B^{-1}\sum_b(\widehat{\mu}_1^{\mathrm{DML},*,(b)} - \overline{\mu}_1^*)^2$, where $\overline{\mu}_1^*$ is the bootstrap mean.

Multiplier bootstrap alternative. A computationally lighter alternative is the multiplier bootstrap on the IFs:

$$\widehat{\mu}_1^{\mathrm{DML,mult},*} := N^{-1}\sum_i (1+\xi_i^*)\cdot\widehat{\Gamma}_1^{(k(i))}(O_i),$$

where $\xi_i^* \sim \mathcal{N}(0,1)$ are independent multipliers. The variance of $\widehat{\mu}_1^{\mathrm{DML,mult},*}$ across multiplier draws equals the asymptotic variance estimator $\widehat{\sigma}_1^2$ in the limit.

Comparison of analytical vs. bootstrap variance. Substantially larger bootstrap variance than analytical variance is a red flag for weak-range effects: the asymptotic regime may not be well-approximated in the finite sample. Practitioners should report both and prefer the larger of the two for inference.

F.4 Minimum-variance and extended SGMM implementations

This subsection collects the extended Sieve GMM machinery for the minimum-variance dual representative, including the finite-dimensional approximation under fixed q, the constrained quadratic-programming formulation, a penalized variant, the asymptotic theory for the fixed-q estimator, the joint extended SGMM with outer q update, and the correspondence with the $b = m + h$ splitting.

Extended Sieve GMM for a minimum-variance representative.

This appendix develops the extended Sieve GMM (SGMM) procedure for computing minimum-variance calibration representatives. The standard Sieve GMM estimator targets the minimum-norm solution; the extended version targets the variance-minimizing representative, which is the

relevant choice for inference (Theorem E.10). The two coincide only under restrictive symmetry conditions, so the extended procedure is needed in general. The appendix also records several optional constrained parameterizations and implementation variants used for numerical stabilization.

Remark (Linear sieve for q and optional positivity constraints). The calibration representative q is generally not pointwise nonnegative (only the conditional expectation $\mathbb{E}[q^\ell(W, Y_t, C) \mid U, Y_t, C] = (1 - p_t)/p_t \geq 0$ is required), and the identification theorem (Theorem 3.4) does not restrict the sign of $1 + q(X_1)$. The default linear sieve respects the unrestricted calibration solution set. An exponential parameterization $q = \exp\{\boldsymbol{\psi}_q^\top \boldsymbol{\theta}_q\}$ targets a positive representative and is valid only when the calibration set contains such a representative; we recommend it only as an *optional constrained implementation* when positivity is desired for numerical stability.

Finite-dimensional approximation under fixed q.

Fix a primal estimate $\widehat{q}$ from the first stage (Tikhonov-regularized least squares or any consistent regularized estimator from Section F). The dual moment that determines b is D-weighted and is computed on $\{D = 1\}$; the primal moment that determines q is $\mathbb{E}[Dq(X_1) - (1 - D) \mid Z_1] = 0$ and uses both treatment arms. The variance-minimizing score $\mathbb{E}[\Gamma_1^2]$ is evaluated on the full sample because $\Gamma_1 = D(1 + q)(Y - b) + b$ has the $b(Z_1)$ term defined for all units. Define basis functions $\Phi_n^{Z_1}(z_1) = (\phi_1(z_1), \ldots, \phi_{K_n}(z_1))^\top$ on Z_1-space and $\Phi_n^{X_1}(x_1) = (\psi_1(x_1), \ldots, \psi_{L_n}(x_1))^\top$ on X_1-space, with $K_n, L_n \to \infty$ as $n \to \infty$. The minimum-variance adjoint representer is approximated by $\widehat{b}(z_1) = \Phi_n^{Z_1}(z_1)^\top \theta_b$ for an unknown coefficient vector $\theta_b \in \mathbb{R}^{K_n}$.

Linear-Gaussian benchmark. Under linear-Gaussian structure, the variance-minimizing b admits a closed-form representation as the conditional expectation $b(z_1) = \mathbb{E}[Y \mid Z_1 = z_1, D = 1]$ plus a residual correction in $\ker(A_1^\star)$. The finite-dimensional sieve representation is then a polynomial or Hermite expansion of this conditional expectation, with coefficients determined by the moment equations.

Sample moment formulation. The adjoint range condition $A_1^\star b = \ell_1$ translates to the sample moment

$$\widehat{m}_n^b(\theta_b) := \frac{1}{n} \sum_{i=1}^{n} D_i \big\{ Y_i - \Phi_n^{Z_1}(Z_{1,i})^\top \theta_b \big\} \Phi_n^{X_1}(X_{1,i}).$$

Setting $\widehat{m}_n^b(\theta_b) = 0$ gives the calibration-conditional moments. With $L_n \geq K_n$ (the moment dimension at least as large as the parameter dimension), this is an overidentified GMM system.

Constrained quadratic programming formulation.

The variance-minimizing $\widehat{b}$ at fixed $\widehat{q}$ minimizes the sample variance of the calibrated orthogonal moment $\Gamma_1(O; \widehat{q}, b)$ subject to the calibration-conditional moments. This is a constrained quadratic program:

$$\min_{\theta_b \in \mathbb{R}^{K_n}} \frac{1}{n} \sum_{i=1}^{n} \big[D_i \{1 + \widehat{q}(X_{1,i})\} \{ Y_i - \Phi_n^{Z_1}(Z_{1,i})^\top \theta_b \} + \Phi_n^{Z_1}(Z_{1,i})^\top \theta_b - \bar{\Gamma} \big]^2$$

subject to

$$\widehat{m}_n^b(\theta_b) = \mathbf{0}.$$

The objective is the sample squared calibrated orthogonal moment (its variance, since the score has mean $\widehat{\mu}_1$ at the truth which is a constant). The constraint pins θ_b to the affine subspace $\mathcal{B}_1$ approximated by the sieve.

KKT system for the equality-constrained QP. Let H denote the quadratic part of the centered sample variance objective (i.e., the Gram matrix of b-perturbations under the score-variance-induced weighting) and let $C\theta_b = d$ denote the linear equality system corresponding to the adjoint moment constraints with the constraint Jacobian C and right-hand side d. The Lagrangian is $\mathcal{L}(\theta_b, \lambda) = \frac{1}{2}\theta_b^\top H\theta_b - h^\top\theta_b + \lambda^\top(C\theta_b - d)$ where h is the linear part of the score-variance objective. The KKT system is

$$\begin{pmatrix} H & C^\top \\ C & 0 \end{pmatrix} \begin{pmatrix} \widehat{\theta}_b \\ \widehat{\lambda} \end{pmatrix} = \begin{pmatrix} h \\ d \end{pmatrix}.$$

The equality constraint is generically binding (it pins θ_b to the affine set $\mathcal{B}_1$ approximated by the sieve). When the constraint Jacobian C has full row rank and H is invertible on $\ker C$, the system has a unique solution $\widehat{\theta}_b = H^{-1}h + H^{-1}C^\top(CH^{-1}C^\top)^{-1}(d - CH^{-1}h)$, which can be computed in standard QP solvers.

Penalized version and augmented GMM.

When the moment constraint $\widehat{m}_n^b(\theta_b) = 0$ cannot be satisfied exactly (e.g., due to weak identification or near-singular Jacobian), the penalized version replaces the hard constraint with a Tikhonov-type penalty:

$$\min_{\theta_b \in \mathbb{R}^{K_n}} \frac{1}{n}\sum_{i=1}^{n}\left\{\Gamma_1(O_i; \widehat{q}, \theta_b) - \bar{\Gamma}(\theta_b)\right\}^2 + \lambda_b \cdot \|\widehat{m}_n^b(\theta_b)\|^2_{\widehat{\Omega}^{-1}},$$

where $\lambda_b > 0$ is a regularization parameter. As $\lambda_b \to \infty$, the penalized solution converges to the constrained solution. The penalized version has the advantage of being numerical (in finite samples the centering by $\bar{\Gamma}(\theta_b)$ is the correct sample variance; over the population feasible set the score has constant mean μ_1 so the two objectives coincide asymptotically). It additionally has the advantage of being numerically stable in finite samples.

Augmented GMM formulation. The penalized version can also be cast as augmented GMM by stacking the score-variance objective with the moment constraint:

$$\min_{\theta_b} \widehat{g}_n(\theta_b)^\top \widehat{W} \widehat{g}_n(\theta_b), \quad \widehat{g}_n(\theta_b) := \begin{pmatrix} \Gamma_1(\cdot; \widehat{q}, \theta_b) \\ \widehat{m}_n^b(\theta_b) \end{pmatrix},$$

with appropriate weighting matrix $\widehat{W}$. This formulation is convenient for software implementation since it reduces to a standard GMM problem with augmented moments.

Asymptotic theory for fixed-q extended SGMM.

Under standard regularity conditions on the sieve basis (cf. Chen and Pouzo (2012)) and consistency of $\widehat{q}$ from the first stage, the fixed-q extended SGMM estimator $\widehat{b}$ is consistent for the minimum-variance adjoint representer.

Convergence rate. The first-stage estimator $\widehat{q}$ converges at the Tikhonov rate of Remark D.4,

denoted $\delta_n^q = n^{-a_q}$, and the second-stage estimator $\widehat{b}$ converges at rate $\delta_n^b = n^{-a_b}$. The product-bias identity (Theorem F.3) provides a *decomposition* of the second-order plug-in bias of $\widehat{\mu}_1$ into the bilinear $\mathbb{E}[D\Delta_q\Delta_b]$, which is bounded by $\delta_n^q \cdot \delta_n^b$. This decomposition does not itself imply $\sqrt{N}$-asymptotic normality; that conclusion requires the product-rate threshold

$$a_q + a_b > \tfrac{1}{2}, \qquad \text{i.e.,} \qquad \delta_n^q \cdot \delta_n^b = o(n^{-1/2}), \tag{85}$$

together with separately established first-stage rates from Tikhonov NPIV theory. When both nuisances achieve the same rate $r_n = n^{-\beta/(2\beta+2\nu+1)}$, the threshold reduces to $\beta > \nu + 1/2$, where β is the source-condition smoothness and ν controls the singular-value decay of $A_1^\star A_1$.

Asymptotic distribution. Under the product-rate condition and finite-variance condition (FV), the resulting plug-in $\widehat{\mu}_1$ is asymptotically normal. The asymptotic variance equals the minimum variance over the affine class of influence functions induced by the fixed primal representative $q^\dagger$; this coincides with the global semiparametric efficiency bound only when the fixed representative coincides with the primal component of a canonical-gradient calibration representation (in which case the EIF theorem of Section E.4 provides the matching characterization). In general, the fixed-$q^\dagger$ extended SGMM does *not* attain the global semiparametric efficiency bound: it attains the minimum calibrated orthogonal moment variance within the fixed-$q^\dagger$ affine class, which coincides with the global bound only under the canonical-representative condition described in Theorem E.11. The proof follows the standard cross-fitted DML argument (Section F) with the calibrated orthogonal moment replacing the generic orthogonal score.

Joint extended SGMM with outer q update.

The two-stage procedure (fix $\widehat{q}$, then compute $\widehat{b}$) is the recommended implementation because the joint (q, b) minimization is non-convex (see the discussion of Theorem E.11). However, when the first-stage $\widehat{q}$ is far from the variance-minimizing representative, an outer iteration over q may improve efficiency.

Iteration scheme. Initialize with a Tikhonov-regularized $\widehat{q}^{(0)}$. For each iteration $k \geq 0$:

1. Compute $\widehat{b}^{(k)}$ via fixed-q extended SGMM at $\widehat{q}^{(k)}$.
2. Update $\widehat{q}^{(k+1)}$ by minimizing the calibrated orthogonal moment variance over $\widehat{q}$ at $\widehat{b}^{(k)}$ fixed:

$$\widehat{q}^{(k+1)} = \arg\min_q \frac{1}{n}\sum_i \Gamma_1(O_i; q, \widehat{b}^{(k)})^2 \quad \text{subject to } A_1 q = r_1.$$

The iteration is monotone in the objective (each step weakly decreases the sample variance), but the non-convex joint surface may have multiple stationary points. In practice, a small number of outer iterations (2–3) suffices to capture most of the variance reduction.

Correspondence with $b = m + h$ splitting.

The decomposition $b = m + h$ (where $m \in \mathcal{H}_{Z_1}$ is a fixed base regressor and h is the residual correction) has a natural counterpart in the extended SGMM. With $m(z_1) = \mathbb{E}[Y \mid Z_1 = z_1, D = 1]$ (the conditional expectation regression), the residual h satisfies $A_1^\star h = \ell_1 - A_1^\star m =: \rho_{1,m}$.

Two-step OLS-plus-correction implementation.

1. Compute $\widehat{m}$ by OLS regression of Y on basis $\Phi_n^{Z_1}(Z_1)$ on $\{D = 1\}$.

2. Compute the residual $\rho_{1,m}(x_1) = \ell_1(x_1) - A_1^\star \widehat{m}(x_1)$.

3. Compute $\widehat{h}$ by sieve minimum distance on $A_1^\star h = \rho_{1,m}$ (this is an NPIV-type problem of smaller magnitude).

4. Set $\widehat{b} = \widehat{m} + \widehat{h}$.

Advantage. The OLS step can reduce the magnitude of the right-hand side $\rho_{1,m}$ in well-conditioned directions, which may make the second-stage NPIV problem better conditioned than computing b directly. However, a small L^2 residual $\|\rho_{1,m}\|$ does not by itself guarantee a small inverse correction $\|\widehat{h}\|$: the Picard coordinates of $\rho_{1,m}$ relative to the singular values of $A_1^\star$ must also be controlled. The OLS-plus-correction implementation is particularly useful when (R2) is well-conditioned, i.e., when the empirical dual norm $\|\widehat{h}\|$ is bounded along the relevant Picard directions.

When the OLS step is exact. The OLS step is exact only in the degenerate case in which Y_1 lies in the closed linear span of Z_1 under the treated law, so that $\mathbb{E}[m(Z_1) \mid X_1, D = 1] = Y_1$ for some linear $m \in \mathcal{H}_{Z_1}$. Linear Gaussianity alone does not imply this condition: under a linear-Gaussian latent-factor model, the conditional expectation $\mathbb{E}[\mathbb{E}[Y \mid Z_1, D = 1] \mid X_1, D = 1]$ is generally a projection-shrinkage of Y_1, not Y_1 itself, because the measurement correlation is strictly less than one. When the degenerate condition does hold, the correction $\widehat{h} = 0$ and the extended SGMM reduces to OLS on $\{D = 1\}$. This is the benchmark case where extended SGMM is equivalent to standard regression.

Proposition F.5 (Consistency of fixed-q extended Sieve GMM)**.** *Let $q^\dagger \in \mathcal{Q}_1$ be a fixed primal representative, and let $b^\star_{q^\dagger}$ be the fixed-$q^\dagger$ minimum-variance adjoint representer of Theorem E.10. Suppose $\hat{q} \to q^\dagger$ in $L^2(P_{X_1|D=1})$, the sieve space $\mathcal{B}_{1,N} = \{b_\gamma : \gamma \in \mathbb{R}^{p_b}\}$ approximates $b^\star_{q^\dagger}$, the sample objectives $\widehat{V}_N$ and $\widehat{\boldsymbol{g}}_b$ converge uniformly to their population counterparts, and the regularization parameters $\kappa_N \downarrow 0$ and $\lambda_b \downarrow 0$ satisfy standard sieve GMM conditions. Then any approximate minimizer of the constrained QP of Section F.4 satisfies*

$$\left\| b_{\widehat{\gamma}_b^{\mathrm{ext}}(\hat{q})} - b^\star_{q^\dagger} \right\|_{w_{q^\dagger},2} = o_p(1), \qquad w_{q^\dagger}(z) := \mathbb{E}[(1 - D\{1 + q^\dagger(X_1)\})^2 \mid Z_1 = z]. \tag{86}$$

Moreover, if $\hat{q}$ and $\hat{b}^{\mathrm{ext}}$ satisfy the product-rate condition of Appendix F, then the resulting plug-in $\widehat{\mu}_1$ estimator is $\sqrt{N}$-consistent and asymptotically normal for μ_1, with asymptotic variance equal to the minimum calibrated orthogonal moment variance within the fixed-$q^\dagger$ class.

Proof sketch. For fixed q, the score $\Gamma_1(O; q, b) = \alpha_q Y + (1 - \alpha_q) b(Z_1)$ is linear in b. Hence the population objective $\mathrm{Var}\{\Gamma_1(O; q, b)\}$ is a weighted-projection problem onto the closed affine set $\mathcal{B}_1$ in $L^2(w_q dP_{Z_1})$, and Theorem E.10 guarantees the existence of $b^\star_q$. The constrained finite-dimensional QP of Section F.4 is the sieve approximation of this projection problem. Standard sieve minimum distance / constrained GMM arguments (approximation of the constraint set, uniform convergence of the objective, vanishing regularization bias) yield (86). Applying the invariance of Theorem 3.4 and the exact product-bias identity of Appendix F gives consistency of the plug-in estimator and asymptotic normality under the product-rate condition. □

G Efficient influence functions and extensions

This appendix gives detailed proofs of the efficient-influence-function results from Section E.4.

G.1 Calibrated orthogonal moment class, regular IFs, and canonical gradients

This subsection collects the proofs characterizing the calibrated orthogonal moment class and the regular influence function. The distinction between the calibrated orthogonal moment class, a regular influence function representative, and the semiparametric canonical gradient is maintained throughout.

Proof of Theorem E.6.

(i) Mean zero: $\mathbb{E}[\Gamma_1(O; q, b)] = \mu_1$.

By the latent selection-odds representer condition (R1-A) and the latent-to-observed transfer (Lemma 3.1), there exists a latent representative $q^\ell \in \mathcal{Q}_1$, induced by the latent odds anchor, such that

$$\mathbb{E}[D\{1 + q^\ell(X_1)\}\, Y] = \mu_1.$$

This IPW identity holds at the latent representative q^ℓ, *not* at an arbitrary observed calibration solution. For any other $q \in \mathcal{Q}_1$, write $g = q - q^\ell \in \ker(A_1)$. Since $b \in \mathcal{B}_1$ implies $A_1^\star b = \ell_1$,

$$\mathbb{E}[D\, g(X_1)\{Y - b(Z_1)\}] = \langle g,\, \ell_1\rangle_{X_1} - \langle A_1 g,\, b\rangle_{Z_1} = \langle g,\, \ell_1 - A_1^\star b\rangle_{X_1} = 0,$$

where the first equality is the bilinear expansion, the second uses the adjoint identity $\langle A_1 g, b\rangle_{Z_1} = \langle g, A_1^\star b\rangle_{X_1}$, and the last uses $A_1^\star b = \ell_1$. (Equivalently, $\langle A_1 g, b\rangle_{Z_1} = 0$ since $A_1 g = 0$.) Also, since $q^\ell \in \mathcal{Q}_1$,

$$\mathbb{E}[\{1 - D(1 + q^\ell(X_1))\}\, b(Z_1)] = \mathbb{E}[b(Z_1) \cdot \mathbb{E}[1 - D(1 + q^\ell(X_1)) \mid Z_1]] = 0.$$

Combining these identities, write $q = q^\ell + g$ with $g \in \ker(A_1)$ and expand:

$$\begin{aligned}
\mathbb{E}[\Gamma_1(O; q, b)] &= \mathbb{E}[D\{1 + q(X_1)\}\{Y - b(Z_1)\}] + \mathbb{E}[b(Z_1)] \\
&= \mathbb{E}[D\{1 + q^\ell(X_1)\}\{Y - b(Z_1)\}] + \mathbb{E}[D\, g(X_1)\{Y - b(Z_1)\}] + \mathbb{E}[b(Z_1)] \\
&= \mathbb{E}[D\{1 + q^\ell(X_1)\}Y] + \mathbb{E}[\{1 - D(1 + q^\ell(X_1))\}b(Z_1)] + \mathbb{E}[D\, g(X_1)\{Y - b(Z_1)\}] \\
&= \mu_1 + 0 + 0 = \mu_1
\end{aligned}$$

for every $(q, b) \in \mathcal{Q}_1 \times \mathcal{B}_1$. Mean-zero of $\phi_1(O; q, b) = \Gamma_1(O; q, b) - \mu_1$ at every such pair follows. The argument crucially uses the target-invariance certificate $A_1^\star b = \ell_1$ to neutralize the kernel direction $g = q - q^\ell$; the IPW identity is *not* invoked at arbitrary $q \in \mathcal{Q}_1$. □

(ii) Neyman orthogonality.

(ii-a) Perturbation in q. For $q \to q + t\delta q$,

$$\Gamma_1(O; q + t\delta q, b) - \Gamma_1(O; q, b) = t \cdot D\, \delta q(X_1)\{Y_1 - b(Z_1)\},$$

so

$$\partial_t \mathbb{E}[\phi_1(O; q + t\delta q, b)]\big|_{t=0} = \mathbb{E}[D\, \delta q(X_1)\{Y_1 - b(Z_1)\}].$$

By the tower property, this equals $\mathbb{E}[\delta q(X_1) \cdot \mathbb{E}[D\{Y_1 - b(Z_1)\} \mid X_1]]$. The defining condition of $b \in \mathcal{B}_1$, equation (61), gives $\mathbb{E}[D\{Y_1 - b(Z_1)\} \mid X_1] = 0$ $\mathbb{P}$-a.s., hence the derivative vanishes for any admissible δq. □

(ii-b) Perturbation in b. For $b \to b + t\delta b$,

$$\Gamma_1(O; q, b + t\delta b) - \Gamma_1(O; q, b) = t \cdot \{1 - D(1 + q(X_1))\}\delta b(Z_1),$$

so

$$\partial_t \mathbb{E}[\phi_1(O; q, b + t\delta b)]\big|_{t=0} = \mathbb{E}\big[\{1 - D(1 + q(X_1))\}\delta b(Z_1)\big].$$

By the tower property, this equals $\mathbb{E}[\delta b(Z_1) \cdot \mathbb{E}[1 - D(1 + q(X_1)) \mid Z_1]]$. The defining condition of $q \in \mathcal{Q}_1$, equation (60), gives $\mathbb{E}[Dq(X_1) - (1 - D) \mid Z_1] = 0$, equivalently $\mathbb{E}[D(1 + q(X_1)) \mid Z_1] = 1$ $\mathbb{P}$-a.s. Hence $\mathbb{E}[1 - D(1 + q(X_1)) \mid Z_1] = 0$ $\mathbb{P}$-a.s., and the derivative vanishes for any admissible δb. □

Combining (i) and (ii), $\phi_1(O; q, b)$ is mean zero and Neyman-orthogonal at any $(q, b) \in \mathcal{Q}_1 \times \mathcal{B}_1$. □

Proof of Theorem E.10: Fixed-q variance-minimizing b.

The proof details the convexity of the fixed-q minimization over $b \in \mathcal{B}_1$ and derives the closed-form unconstrained minimizer.

Step 1: Structure of Γ_1 for fixed q (observed-data form).

Let $\alpha_q(O) := D\{1 + q(X_1)\}$. Then

$$\Gamma_1(O; q, b) = \alpha_q(Y - b(Z_1)) + b(Z_1) = \alpha_q Y + (1 - \alpha_q)b(Z_1).$$

Note that $\alpha_q Y$ effectively equals $D(1+q)Y_1$ since $\alpha_q = 0$ when $D = 0$. The score Γ_1 is linear in b.

Step 2: Conditional centering of $\alpha_q - 1$.

Since $q \in \mathcal{Q}_1$ satisfies $\mathbb{E}[Dq(X_1) - (1 - D) \mid Z_1] = 0$ $\mathbb{P}$-a.s., we have

$$\mathbb{E}[\alpha_q \mid Z_1] = \mathbb{E}[D + Dq \mid Z_1] = \mathbb{E}[D \mid Z_1] + \mathbb{E}[Dq \mid Z_1] = \mathbb{E}[D \mid Z_1] + \mathbb{E}[1 - D \mid Z_1] = 1,$$

so $\mathbb{E}[\alpha_q - 1 \mid Z_1] = 0$ $\mathbb{P}$-a.s. The deviation $1 - \alpha_q$ is therefore Z_1-conditionally mean zero.

Step 3: Convexity of the variance objective.

The objective $G(b) := \mathrm{Var}[\Gamma_1(O; q, b)] = \mathbb{E}[\Gamma_1^2] - \mu_1^2$ for any valid $(q, b) \in \mathcal{Q}_1 \times \mathcal{B}_1$. Conditional on $Z_1 = z$,

$$\mathbb{E}[\Gamma_1^2 \mid Z_1 = z] = \mathbb{E}[\alpha_q^2 Y^2 \mid Z_1 = z] + 2\mathbb{E}[\alpha_q(1 - \alpha_q)Y \mid Z_1 = z]b(z) + w_q(z)b(z)^2,$$

where $w_q(z) := \mathbb{E}[(1 - \alpha_q)^2 \mid Z_1 = z] \geq \underline{w} > 0$ by overlap. The map $b \mapsto G(b)$ is therefore quadratic and strictly convex in b.

Step 4: Closed-form unconstrained minimizer.

Setting $\partial_b G(b) = 0$ gives

$$b_q^{\mathrm{un}}(z) = -\frac{\mathbb{E}[\alpha_q(1 - \alpha_q)Y \mid Z_1 = z]}{w_q(z)} = \frac{\mathbb{E}[(\alpha_q - 1)\alpha_q Y \mid Z_1 = z]}{w_q(z)}.$$

This closed form does not contain μ_1: by the centering property of Step 2, any μ_1 contribution to the inner conditional expectation vanishes.

Step 5: Projection onto $\mathcal{B}_1$.

The unconstrained b_q^{un} does not generally satisfy $A_1^\star b = \ell_1$. The constrained variance-minimizer is the weighted L^2-projection of b_q^{un} onto $\mathcal{B}_1$:

$$b_q^\star = \Pi_{\mathcal{B}_1}^{w_q}(b_q^{\mathrm{un}}) = \arg\min_{b \in \mathcal{B}_1} \mathbb{E}[w_q(Z_1)\{b(Z_1) - b_q^{\mathrm{un}}(Z_1)\}^2].$$

The first-order condition for the projection is: for all $c \in \ker(A_1^\star) = \{c \in \mathcal{H}_{Z_1} : A_1^\star c = 0\}$, the perturbation $b_q^\star + tc$ leaves the constraint satisfied and the directional derivative of the variance vanishes. Writing the centered score $\phi_1(O; q, b) = D(1+q)(Y-b) + b - \mu_1$, the FOC is

$$\mathbb{E}\big[\phi_1(O; q, b_q^\star) \cdot \big(1 - D(1 + q(X_1))\big) \cdot c(Z_1)\big] = 0 \quad \forall c \in \ker(A_1^\star).$$

The constant term $-\mu_1$ in ϕ_1 contributes

$$-\mu_1 \cdot \mathbb{E}[\{1 - D(1+q(X_1))\}c(Z_1)] = 0,$$

which vanishes because $\mathbb{E}[1 - D(1+q(X_1)) \mid Z_1] = 0$ by the balancing equation $A_1 q = r_1$. The remaining part is the orthogonality condition involving the bilinear coupling of ϕ_1 with the projector direction; it characterizes $b_q^\star$ within $\mathcal{B}_1$ as the variance-minimizing representative. It should not be simplified to a conditional moment of the form $\mathbb{E}[(Y - b_q^\star)(1+q)D \mid Z_1] = 0$: such a simplification is not in general implied by the projection FOC, and the correct statement is the kernel-orthogonality condition $\mathbb{E}[\phi_1(O; q, b_q^\star)\{1 - D(1+q(X_1))\}c(Z_1)] = 0$ for $c \in \ker(A_1^\star)$. □

G.2 Fixed-q and joint variance minimization

This subsection records the joint characterization of variance-minimizing representatives within the calibration solution sets.

Joint minimization in the EIF characterization.

The joint (q, b) minimization is non-convex. The joint minimization

$$\min_{(q,b) \in \mathcal{Q}_1 \times \mathcal{B}_1} \mathbb{E}[\phi_1(O; q, b)^2]$$

involves the product $\alpha_q(1-\alpha_q)Y$ and the bilinear term $b(Z_1)\cdot(1-\alpha_q)$, so the objective is generally non-convex in (q, b) jointly even though it is convex in each component separately. This is one reason the two-step procedure (first $\widehat{q}$, then $\widehat{b}$) is preferred in practice.

When the joint minimum is achievable. The infimum of $\mathbb{E}[\phi_1^2]$ over $\mathcal{Q}_1 \times \mathcal{B}_1$ equals the local-calibration-model efficiency bound when (i) the orthogonal score class $\{\phi_1(O; q, b) : (q, b) \in \mathcal{Q}_1 \times \mathcal{B}_1\}$ contains the canonical gradient in the L^2-closure, and (ii) the infimum-attaining $(q^{\mathrm{EIF}}, b^{\mathrm{EIF}})$ admits an anchored differentiable path. The two-step procedure of Theorem E.10 attains the infimum modulo $\ker(A_1)$ under conditions on the joint kernel structure.

Remark (Within-class lower bound versus semiparametric efficiency). The variance characterizations of this paper should be read as *within-class* (conditional) lower bounds rather than as an unconditional semiparametric efficiency bound. Because $\mathbb{E}[\phi_1(O; q, b)^2]$ is not jointly convex in (q, b) (it is convex only in each coordinate), the two-step procedure that fixes $\widehat{q}$ and then minimizes over b attains the minimum calibrated orthogonal moment variance *within the fixed-$q^\dagger$ affine class*, and there is no general guarantee that sequential minimization reaches the joint optimum. The

value coincides with the local-calibration-model semiparametric efficiency bound only under the additional canonical-gradient condition that the orthogonal-score class contains the efficient influence function in its $L^2(P)$-closure (Theorem E.11). Absent that condition, statements about the asymptotic variance of the estimator are claims about the within-class bound at the implemented representative, and we use the qualified phrase "within-class variance lower bound" rather than "semiparametric efficiency bound" accordingly.

Remark (Representer selection and the variance functional σ^2). Theorem 3.4 guarantees invariance of the *value* μ_1 across valid calibration representatives, but the orthogonal score $\Gamma_1(O; q, b)$ and hence the asymptotic-variance functional $\sigma^2 = \mathbb{E}[(\Gamma_1 - \mu_1)^2]$ depend on which representative is selected. The regularized estimator targets a canonical representative: the Tikhonov/ridge solution converges, as the penalty $\lambda \to 0$, to the minimum-norm (Moore–Penrose) representative of the calibration system (Engl et al., 2000), which is the canonical choice underlying the reported scores. Accordingly, the reported $\widehat{\sigma}^2$ is the variance at this minimum-norm representative; it is invariant across implementations only at the canonical representative, and an implementation that converges to a different admissible representative (for example a constrained or differently preconditioned solution) will report a different, equally valid σ^2. The minimum-variance refinement of Theorem 3.10 selects, within the fixed-$q^\dagger$ class, the adjoint representative minimizing σ^2; the body states which representative underlies each reported score (Section 4).

First-order conditions.

The first-order conditions for the variance-minimizing pair $(q^{\mathrm{EIF}}, b^{\mathrm{EIF}})$ are the orthogonality of the centered score $\phi_1^{\mathrm{EIF}} := \phi_1(O; q^{\mathrm{EIF}}, b^{\mathrm{EIF}}) = D\{1 + q^{\mathrm{EIF}}\}\{Y - b^{\mathrm{EIF}}\} + b^{\mathrm{EIF}} - \mu_1$ to admissible perturbations within $\ker(A_1)$ in the q direction and $\ker(A_1^\star)$ in the b direction:

$$\mathbb{E}\big[\phi_1^{\mathrm{EIF}} \cdot D \cdot \delta_q(X_1) \cdot \{Y - b^{\mathrm{EIF}}(Z_1)\}\big] = 0 \quad \forall \delta_q \in \ker(A_1), \tag{87}$$

$$\mathbb{E}\big[\phi_1^{\mathrm{EIF}} \cdot \{1 - D(1 + q^{\mathrm{EIF}}(X_1))\} \cdot \delta_b(Z_1)\big] = 0 \quad \forall \delta_b \in \ker(A_1^\star). \tag{88}$$

The $-\mu_1$ contributions in the two FOCs vanish for different reasons. In the q-direction, the $-\mu_1$ term contributes

$$-\mu_1 \cdot \mathbb{E}\big[D\,\delta_q(X_1)\,\{Y - b^{\mathrm{EIF}}(Z_1)\}\big],$$

which is zero because $\delta_q \in \ker(A_1)$ and $A_1^\star b^{\mathrm{EIF}} = \ell_1$ imply

$$\mathbb{E}[D\,\delta_q(X_1)\,\{Y - b^{\mathrm{EIF}}(Z_1)\}] = \langle \delta_q, \ell_1 - A_1^\star b^{\mathrm{EIF}} \rangle_{X_1} = 0.$$

In the b-direction, the $-\mu_1$ term contributes $-\mu_1 \cdot \mathbb{E}[\{1 - D(1 + q^{\mathrm{EIF}})\}\delta_b(Z_1)]$, which is zero by the primal moment $\mathbb{E}[1 - D(1 + q^{\mathrm{EIF}}) \mid Z_1] = 0$ (since $q^{\mathrm{EIF}} \in \mathcal{Q}_1$). In each direction the FOC therefore reduces to a coupling of the calibration increment with ϕ_1^{EIF}. Condition (87) is the q-side orthogonality and condition (88) is the b-side orthogonality, jointly characterizing the variance-minimizing representative.

G.3 Additional extensions

The results in this subsection are not used in the main identification or estimation arguments; they record extensions that may be useful in related designs. This subsection collects extensions of the auxiliary-measurement framework that go beyond the main identification result, including a multivalued-treatment generalized Roy extension, the two-sample EIF for one-sided noncom-

pliance, the connection to Bennett et al. (2025), the potential-outcome-type formulation, necessity arguments for the measurement-side and adjoint-side variables, the failure of scalar-index conditioning, and the extension with observed covariates. The two-sample EIF and one-sided noncompliance subsections are intended for the Online Appendix in the final submission.

Multivalued treatments: an arm-wise generalized Roy extension.

The one-sided architecture extends with no new ideas to a multivalued treatment $D \in \{0, 1, \ldots, J\}$ with potential outcomes $(Y_0, \ldots, Y_J)$, $Y = Y_D$, and selection on the full potential-outcome vector. Write $D_j = \mathbf{1}\{D = j\}$, let the arm-$j$ one-sided selection probability be $p_j(U, Y_j, C) = \mathbb{P}(D = j \mid U, Y_j, C)$ with latent odds $r_j^\ell = (1 - p_j)/p_j$, and let $X_j = (W, Y_j, C)$ and $Z_j = (V, C)$ (or an arm-specific adjoint block) be the arm-j primal and dual domains. The arm-wise anchor conditions are the direct analogues of (R1-A) and (T): existence of $q_j^\ell \in \mathcal{H}_{X_j}$ with $T_{W,j} q_j^\ell = r_j^\ell$, and the transfer moment $\mathbb{E}[D_j q_j^\ell(X_j) - (1 - D_j) \mid U, Y_j, C, V] = 0$. Exactly as in Lemma 3.1, these anchor the arm-j observed calibration set $\mathcal{Q}_j = \{q : \mathbb{E}[D_j\{1 + q(X_j)\} - 1 \mid Z_j] = 0\}$, and for any linear functional $\theta_j(\varphi) = \mathbb{E}[\varphi(Y_j, C)]$ with signal $\ell_{j,\varphi}(x) = \mathbb{E}[\varphi(Y, C) \mid X_j = x, D = j]$ the arm-wise primal–dual identity holds verbatim: for every $q \in \mathcal{Q}_j$ and every b with $A_j^\star b = \ell_{j,\varphi}$,

$$\theta_j(\varphi) = \mathbb{E}\big[D_j\{1 + q(X_j)\}\{\varphi(Y, C) - b(Z_j)\} + b(Z_j)\big],$$

because the proof of Theorem 3.4 uses only the arm indicator, never the binary complement. Pairwise contrasts $\tau_{jk}(\varphi) = \theta_j(\varphi) - \theta_k(\varphi)$—in particular all pairwise average treatment effects $\mathbb{E}[Y_j - Y_k]$—are then identified by differencing two arm-wise calibrated moments, and the invariance characterization of Theorem 3.3 and the trichotomy of Proposition 3.5 apply *arm by arm and functional by functional*: each pair (j, φ) has its own signal $\ell_{j,\varphi}$, adjoint range $\overline{\mathcal{R}(A_j^\star)}$, and Case N/I/R classification, so different arms of the same design can sit in different identification regimes. Two features deserve note. First, the arm-wise conditions do *not* require modeling the $(J+1)$-way selection mechanism jointly: only the $J+1$ one-sided margins p_j enter, exactly as the binary analysis never uses the joint law of (D, Y_1, Y_0) beyond the one-sided anchors. Second, the multinomial structure enters only through the accounting identity $\sum_j p_j = 1$, which the arm-wise anchors are free to ignore; imposing it yields cross-arm overidentifying restrictions on $(q_0^\ell, \ldots, q_J^\ell)$ that can be used as diagnostics but are not needed for identification. Estimation proceeds by $J + 1$ arm-wise Sieve GMM systems of the form in Section 4, with the same orthogonality and product-rate structure per arm.

Two-sample EIF for one-sided noncompliance.

This subsection records an extension to the two-sample one-sided-noncompliance setting. It is not used in the main identification or estimation results but may be useful in related designs.

This appendix develops the influence-function and efficiency theory for the two-sample observed model under one-sided noncompliance. Throughout, sample A contains units with the full calibration structure (treated and untreated populations of interest, with measurements W, V, and conditioning block C), and sample B contains comparison units in which the treated potential outcome is irrelevant ($D \equiv 0$). Sample sizes are N_A and N_B with $\rho := N_A/(N_A + N_B) \in (0, 1)$ assumed in the limit.

Setup and the transport assumption.

Assumption G.1 (Transport of Y_0 distribution)**.** The conditional distribution of Y_0 given the observed covariates is identical in samples A and B:

$$\mathbb{P}_{\mathbb{P}^A}(Y_0 \in \cdot \mid C) = \mathbb{P}_{\mathbb{P}^B}(Y_0 \in \cdot \mid C).$$

Equivalently, $\mathbb{E}_{\mathbb{P}^A}[Y_0 \mid C] = \mathbb{E}_{\mathbb{P}^B}[Y_0 \mid C]$ and higher moments match.

Assumption G.1 is the substantive content that allows sample B to estimate $\mu_0 = \mathbb{E}[Y_0]$ directly without going through the auxiliary-measurement framework. In practice, the assumption requires that sample B is comparable to sample A's untreated subpopulation with respect to the latent type distribution conditional on observed covariates.

Influence function decomposition.

Under Assumption G.1 and sample independence, the average treatment effect $\tau = \mu_1 - \mu_0$ is estimated separately on each sample:

$$\widehat{\tau}^{\mathrm{2S}} = \widehat{\mu}_1^A - \widehat{\mu}_0^B.$$

The asymptotic variance decomposes by sample independence.

Sample A contribution. The estimator $\widehat{\mu}_1^A$ follows the cross-fitted DML framework of Section F applied to sample A alone. Its asymptotic representation is

$$\sqrt{N_A}(\widehat{\mu}_1^A - \mu_1) \to_d \mathcal{N}\big(0,\, \mathbb{E}_{\mathbb{P}^A}[(\phi_1^{\mathrm{EIF},A})^2]\big),$$

where $\phi_1^{\mathrm{EIF},A}$ is the EIF for μ_1 in sample A (Theorem E.11). In terms of total sample $N := N_A + N_B$, the variance is $\rho^{-1}\mathbb{E}_{\mathbb{P}^A}[(\phi_1^{\mathrm{EIF},A})^2]/N$.

Sample B contribution. Since $D \equiv 0$ in sample B, the variable Y_0^B is directly observed for all units. The estimator $\widehat{\mu}_0^B = N_B^{-1}\sum_{i\in B} Y_0^{B,i}$ is the sample mean, and its asymptotic representation is

$$\sqrt{N_B}(\widehat{\mu}_0^B - \mu_0) \to_d \mathcal{N}\big(0,\, \mathrm{Var}_{\mathbb{P}^B}(Y_0)\big),$$

with IF $\phi_0^B(O^B) = Y_0^B - \mu_0$. In terms of total sample N, the variance is $(1-\rho)^{-1}\mathrm{Var}_{\mathbb{P}^B}(Y_0)/N$.

Joint influence function and efficiency bound.

By sample independence, the joint IF for $\widehat{\tau}^{\mathrm{2S}}$ is

$$\phi_\tau^{\mathrm{eff,2S}}(O^A, O^B) = \phi_1^{\mathrm{EIF},A}(O^A) - \phi_0^B(O^B) = \phi_1^{\mathrm{EIF},A}(O^A) - \{Y_0^B - \mu_0\}, \tag{89}$$

and the asymptotic variance is

$$\mathcal{V}_\tau^{\mathrm{eff,2S}} = \rho^{-1}\mathbb{E}_{\mathbb{P}^A}[(\phi_1^{\mathrm{EIF},A})^2] + (1-\rho)^{-1}\mathrm{Var}_{\mathbb{P}^B}(Y_0). \tag{90}$$

Comparison with the single-sample design.

The single-sample design estimates both μ_1 and μ_0 via the auxiliary-measurement framework, with each side carrying its own calibration nuisance. Its asymptotic variance is

$$\mathcal{V}_\tau^{\text{single}} = \mathbb{E}[(\phi_1^{\text{EIF}} - \phi_0^{\text{EIF}})^2],$$

where both ϕ_t^{EIF} are calibration-derived EIFs (Theorem E.14). The two-sample design has ϕ_0^{EIF} replaced by the much simpler $Y_0 - \mu_0$, which is the EIF for μ_0 under an unrestricted mean model on sample B. The variance reduction is

$$\mathcal{V}_\tau^{\text{single}} - \mathcal{V}_\tau^{\text{eff,2S}} = (\text{calibration-variance contribution to } \mu_0 \text{ side}) - (\text{transport-variance contribution}).$$

In settings where the μ_0 calibration is weak (Case N of the trichotomy on the untreated side), the variance reduction can be substantial.

Remark (Scope). The two-sample EIF claim is an observed-model efficiency statement: it characterizes the EIF of the joint observed model with sample A following the calibration structure and sample B following an unrestricted mean model, under Assumption G.1. It is not an oracle efficiency bound for the underlying causal parameter τ in a fully general unrestricted nonparametric model.

Remark (When the transport assumption fails). If Assumption G.1 fails, the two-sample procedure produces a biased estimator of τ. The magnitude of the bias depends on the covariate distribution mismatch between samples A and B. Diagnostic procedures (e.g., propensity-score balance checks between A and B, sensitivity analyses) are essential in practice to assess the credibility of the transport assumption.

Relationship to Bennett et al. (2025).

The inferential device used in this paper is compatible with the general theory of inverse-problem functional inference in Bennett et al. (2025). The contribution of the present paper, however, is not a mechanical application of that general debiasing theory. First, under selection on gains, the paper identifies a causal anchor: a latent selection-odds representer that is causally correct and belongs to the observed calibration set $\mathcal{Q}_1$. Second, the adjoint outcome representer b is not merely a debiasing nuisance. It is a target-invariance certificate that makes the average treatment effect (ATE) functional invariant over $\mathcal{Q}_1$. Third, a standard proximal causal inference (PCI) treatment bridge corrects treatment assignment that is independent of potential outcomes conditional on latent confounders, whereas the selection-odds representer here reproduces a latent treatment odds that conditions on both latent heterogeneity and potential outcomes. This difference is substantive under selection on gains. Appendix E.4 gives an analytic example, and Appendix H reports oracle Monte Carlo evidence, showing that a standard PCI-type score can be systematically biased in this environment.

	General theory of Bennett et al. (2025)	Additional structure in this paper
Object	Debiasing a linear functional in a conditional-moment inverse problem	ATE identification in generalized Roy selection with selection on gains
Primal side	A nuisance representer in a given inverse problem	A latent selection-odds representer provides a causal anchor and forms $\mathcal{Q}_1$
Adjoint side	A Riesz representer or debiasing nuisance	An adjoint outcome representer b certifies invariance of the ATE over $\mathcal{Q}_1$
Difference from PCI	The general theory is not tied to a particular causal graph	Since $(Y_1, Y_0) \not\perp\!\!\!\perp D \mid U$, a standard PCI score fails and a Y_t-dependent selection-odds representer is required
Inference	General product-rate and orthogonalization arguments	The exact product-bias term $-\mathbb{E}[D\Delta_q\Delta_b]$ follows directly from the primal-dual geometry

Potential-outcome-type formulation without an explicit U.

The main text introduces an explicit latent variable U to summarize the dependence between the two potential outcomes and the unobserved heterogeneity that enters treatment selection. The primal-dual identification argument does not depend on the name of this latent object. The same logic can be written without explicitly introducing U by treating the potential-outcome type, or the potential-outcome pair itself, as the latent state.

Let C denote the observed conditioning block, and define the treatment probability conditional on the potential-outcome pair by

$$p_T(y_1, y_0, c) := \Pr(D = 1 \mid Y_1 = y_1, Y_0 = y_0, C = c).$$

On the Y_1 side, the latent selection-odds representer can be written as a function satisfying

$$\mathbb{E}\{q_1^\ell(W, Y_1, C) \mid Y_1, Y_0, C\} = \frac{1 - p_T(Y_1, Y_0, C)}{p_T(Y_1, Y_0, C)}. \tag{91}$$

Here Y_1 is observed when $D = 1$, whereas Y_0 is counterfactual. Hence W must carry information about the residual counterfactual component, or about the potential-outcome type, that remains after conditioning on (Y_1, C).

A sufficient calibration-implication condition for moving from the latent representer to the observed balancing equation is

$$W \perp\!\!\!\perp (D, V) \mid Y_1, Y_0, C, \qquad D \perp\!\!\!\perp V \mid Y_1, Y_0, C. \tag{92}$$

Then the tower property gives

$$\mathbb{E}\{Dq_1^\ell(W, Y_1, C) - (1 - D) \mid V, C\} = 0,$$

so that q_1^ℓ belongs to the observed calibration set, exactly as in the formulation with an explicit U.

A regular primal-dual representation also requires adjoint-side variation. On the Y_1 side, an adjoint representer is a function $b_1(V, C)$ satisfying

$$\mathbb{E}\{D(Y - b_1(V, C)) \mid W, Y_1, C\} = 0. \tag{93}$$

This condition requires V to carry enough adjoint-side information about the potential-outcome type for the conditional expectation of $b_1(V, C)$ to reproduce the outcome signal on the $X_1 = (W, Y_1, C)$ side. If V is absent, b_1 becomes a function only of C, which generally cannot reproduce the residual variation in Y_1.

The Y_0 side is symmetric. One may define a selection-odds representer satisfying

$$\mathbb{E}\{q_0^\ell(W, Y_0, C) \mid Y_1, Y_0, C\} = \frac{p_T(Y_1, Y_0, C)}{1 - p_T(Y_1, Y_0, C)},$$

and an adjoint representer satisfying

$$\mathbb{E}\{(1 - D)(Y - b_0(V, C)) \mid W, Y_0, C\} = 0.$$

This formulation does not weaken the theory; it merely changes notation. In applications in which W is a pre-treatment variable, it is often economically more natural to interpret the model as involving a common latent type H with arrows $H \to (Y_1, Y_0, W, V)$ rather than literal temporal arrows $Y_1, Y_0 \to W$. Thus the variable called U in the main text may be read as a potential-outcome type, an ability type, a productivity type, a preference type, or any latent state that generates the dependence between the two potential outcomes. The identification theorem depends on the existence of measurement-side and adjoint-side representations for that latent type, not on the particular label assigned to it.

Why measurement-side and adjoint-side variation are generically needed.

This appendix formalizes the claim that, although the opposite potential-outcome block may sometimes be marginalized out for one-sided targets, the measurement-side variable W and the outcome-representation variable V cannot generally be removed in the nonparametric primal-adjoint representer strategy. The word "necessary" is used in a functional sense. It does not mean that variables literally named W and V must be observed. Any observed variables that play the same roles can be used instead. The point is that one needs (i) measurement-side variation that represents latent treatment odds and (ii) adjoint-side variation that represents the target outcome signal.

No W-type measurement: a necessary condition for the latent selection-odds representer.

Consider the Y_1 side. Let C_1 denote the observables one would condition on if W were unavailable. In the full formulation $C_1 = (Y_1, S, Q)$; in a reduced formulation for $\mathbb{E}[Y_1]$ alone, $C_1 = (Y_1, S)$

is typical. Write

$$p_1(u,c) := \Pr(D=1 \mid U=u, C_1=c), \qquad r_1^{\ell}(u,c) := \frac{1-p_1(u,c)}{p_1(u,c)}.$$

Proposition G.1 (Necessary and sufficient condition for a representer without W)**.** *Suppose $0 < p_1(U,C_1) < 1$ a.s. and $r_1^{\ell}(U,C_1) \in L^2$. A latent selection-odds representer that uses no W and is a function only of C_1, say $q_1^0(C_1) \in L^2(C_1)$, satisfies*

$$\mathbb{E}[q_1^0(C_1) \mid U, C_1] = r_1^{\ell}(U,C_1) \tag{94}$$

if and only if $r_1^{\ell}(U,C_1)$ is $\sigma(C_1)$-measurable. Equivalently, there exists $r_1^{-W} \in L^2(C_1)$ such that

$$r_1^{\ell}(U,C_1) = r_1^{-W}(C_1) \qquad a.s. \tag{95}$$

In particular, if

$$\Pr\{\mathrm{Var}(r_1^{\ell}(U,C_1) \mid C_1) > 0\} > 0, \tag{96}$$

then no W-free selection-odds representer exists.

Proof. Since $q_1^0(C_1)$ is $\sigma(C_1)$-measurable,

$$\mathbb{E}[q_1^0(C_1) \mid U, C_1] = q_1^0(C_1).$$

Combining this with (94) gives $q_1^0(C_1) = r_1^{\ell}(U,C_1)$ $\mathbb{P}$-a.s. Hence r_1^{ℓ} must be $\sigma(C_1)$-measurable. Conversely, if r_1^{ℓ} is $\sigma(C_1)$-measurable, then setting $q_1^0 = r_1^{\ell}$ defines a W-free calibration. The contrapositive states that if r_1^{ℓ} has nontrivial conditional variation in U given C_1, no W-free calibration exists. □

The result shows why W is not only an auxiliary measurement. If the latent odds ratio varies in the U direction after conditioning on the observables, a calibrating function that depends only on those observables cannot reproduce it. The Y_0 side is identical after replacing C_1 by the corresponding C_0 and replacing the odds ratio by its $D=0$ counterpart.

Corollary G.2 (Generic need for a W-type measurement)**.** *If $\Pr\{\mathrm{Var}(r_t^{\ell}(U,C_t) \mid C_t) > 0\} > 0$ for $t=1$ or $t=0$, then identification of the corresponding $\mathbb{E}[Y_t]$ by the nonparametric selection-odds representer strategy requires some measurement-side variable, outside C_t, that carries information about the latent odds ratio. In the notation of the main text, this role is played by W.*

No V-type outcome-representation variable: a necessary condition for the adjoint range.

Now consider the adjoint representer. In the full Y_1 formulation $X_1 = (W, Y_1, C)$ and $Z_1 = (V, C)$, and the adjoint range condition is $A_1^{\star} b = \ell_1$, where $\ell_1(x) = \mathbb{E}[Y \mid X_1 = x, D=1]$. By consistency, $Y = Y_1$ on $D=1$, so one may identify $\ell_1(W, Y_1, C)$ with Y_1. If V is not used, the adjoint representer is only a function of $C_Z = C$, or of $C_Z = C_1$ in a reduced $\mathbb{E}[Y_1]$ formulation.

Proposition G.3 (Necessary and sufficient condition for an adjoint representer without V)**.** *Let C_Z be the conditioning block available without V, and consider an adjoint representer $b(C_Z)$. The adjoint operator is*

$$A_{1,0}^{\star} b(X_1) := \mathbb{E}[b(C_Z) \mid X_1, D=1] = b(C_Z). \tag{97}$$

Thus the adjoint range condition $A_{1,0}^{\star} b = \ell_1$ holds if and only if

$$Y_1 = b(C_Z) \qquad a.s.\ on\ \{D = 1\}. \tag{98}$$

Equivalently,

$$\mathrm{Var}(Y_1 \mid C_Z, D = 1) = 0 \qquad a.s. \tag{99}$$

In particular, if

$$\Pr\{\mathrm{Var}(Y_1 \mid C_Z, D = 1) > 0\} > 0, \tag{100}$$

then no finite-norm adjoint outcome representer without V exists.

Proof. Because C_Z is a component of X_1, $b(C_Z)$ is $\sigma(X_1)$-measurable, which gives (97). The equation $A_{1,0}^{\star} b = \ell_1$ therefore requires $b(C_Z) = \ell_1(X_1)$. Since X_1 includes Y_1 and $Y = Y_1$ on $D = 1$, this is exactly (98). The equivalence with (99) is immediate, and (100) rules it out. □

In this case the range is a closed subspace of $L^2(P_{X_1|D=1})$. Hence, if Y_1 is not C_Z-measurable, the problem is not merely irregular. Rather,

$$\ell_1 \notin \overline{\mathcal{R}(A_{1,0}^{\star})},$$

which corresponds to the calibration-based non-invariance case in the main trichotomy.

Corollary G.4 (Generic need for a V-type outcome-representation variable)**.** *In a nondegenerate outcome model satisfying* $\Pr\{\mathrm{Var}(Y_t \mid C_Z, D = t) > 0\} > 0$*, regular primal-adjoint representer identification of* $\mathbb{E}[Y_t]$ *requires adjoint-side variation beyond C_Z. In the notation of the main text this role is played by V.*

Implications for ATE and one-sided means.

For $\mathbb{E}[Y_1]$ alone, the opposite block (Y_0, Q) can sometimes be marginalized out. That reduction does not imply that W and V can be removed. A typical reduced formulation still uses

$$\bar{X}_1 = (W, Y_1, S), \qquad \bar{Z}_1 = (V, S).$$

Removing W would require the latent odds ratio to be a function only of (Y_1, S); removing V would require the treated potential outcome to be determined by S alone. These are special cases, not generic features of selection on gains.

For the ATE, the same logic applies to both $\mathbb{E}[Y_1]$ and $\mathbb{E}[Y_0]$. In a one-sided noncompliance design in which $\mathbb{E}[Y_0]$ is directly observed in a separate sample, the Y_0-side calibration machinery is unnecessary. But the Y_1 side still requires measurement-side and adjoint-side variation unless one imposes an alternative identification strategy such as random assignment, a known propensity score, or a strong parametric Roy model.

Remark (Scope of the necessity statements). Propositions G.1–G.3 do not claim that variables literally named W and V must be observed. They state that a nonparametric primal-adjoint representer strategy generically requires two roles: a measurement-side source of information about the latent odds ratio and an adjoint-side source of variation that solves the adjoint range condition. If other observed variables play these roles, they can replace W and V. The results do not apply to designs with random treatment, externally known propensity scores, or strong parametric selection models.

Scalar-index conditioning can fail.

The main text explains that the conditioning block C need not equal the two-dimensional vector (S, Q). A one-dimensional index $C = \kappa(S, Q)$ is sufficient if it preserves all information needed by the primal selection-odds representer and the adjoint outcome representer. This appendix gives a finite-state example in which the scalar index succeeds and another in which it fails. The failure is not a finite-sample estimation problem; it is a population-level failure of the reduced adjoint range condition.

Let $U, S, Q, V, W \in \{0, 1\}$, with U, S, Q mutually independent Bernoulli(1/2). Let

$$\Pr(V = U) = \Pr(W = U) = 0.8.$$

We compare the full conditioning block $C_{\text{full}} = (S, Q)$ with the scalar index $C_{\text{sc}} = S+Q \in \{0, 1, 2\}$. For each data-generating process, the population distribution is enumerated exactly, the finite-dimensional primal equation $Aq = r$ and dual equation $A^\star b = \ell$ are solved, and the population mean of

$$\Gamma_1(O; q, b) = D\{1 + q(X_1)\}\{Y - b(Z_1)\} + b(Z_1)$$

is evaluated.

DGP A: the scalar index is sufficient.

Let all relevant effects of (S, Q) enter only through $C = S + Q$:

$$Y_1 = U \oplus \mathbf{1}\{C = 2\}, \qquad Y_0 = U \oplus \mathbf{1}\{C = 0\},$$

$$\Pr(D = 1 \mid U, Y_1, Y_0, C) = \Lambda(-0.5 + U + 0.8Y_1 - 0.5Y_0 + 0.4C),$$

where $\oplus$ is XOR and $\Lambda(t) = \{1 + \exp(-t)\}^{-1}$. In this design, $C = S + Q$ is a sufficient index for the information in (S, Q) relevant to the calibration system.

DGP B: the scalar index removes a necessary direction.

Let S and Q enter in opposite directions:

$$Y_1 = U \oplus S, \qquad Y_0 = U \oplus Q,$$

$$\Pr(D = 1 \mid U, Y_1, Y_0, S, Q) = \Lambda(-0.5 + U + 0.8Y_1 - 0.5Y_0 + 1.2S - 1.2Q).$$

Then $(S, Q) = (1, 0)$ and $(S, Q) = (0, 1)$ both have $S + Q = 1$, but they have opposite economic meanings. The scalar index loses the $S - Q$ direction.

Table 14: Population oracle calculation: success and failure of scalar-index conditioning

DGP	Conditioning	primal residual	dual residual	$\mathbb{E}[\Gamma_1]$	true $\mathbb{E}[Y_1]$	Bias
A: scalar index sufficient	full (S, Q)	5.1×10^{-16}	1.1×10^{-15}	0.500	0.500	0.000
A: scalar index sufficient	scalar $S + Q$	3.7×10^{-16}	4.3×10^{-16}	0.500	0.500	0.000
B: S, Q matter separately	full (S, Q)	8.6×10^{-16}	5.9×10^{-16}	0.500	0.500	0.000
B: S, Q matter separately	scalar $S + Q$	3.2×10^{-16}	0.363	0.600	0.500	0.100

The last row of Table 14 is the key case. The reduced balancing equation is almost exactly

solvable, but the adjoint equation is not. The resulting population calibration estimate is 0.600, whereas the true value is 0.500. Hence the failure is not due to estimation noise. It occurs because the scalar index discards information required by the reduced adjoint range condition. More generally, $C = \kappa(S, Q)$ is valid only when the reduced primal and adjoint range conditions both hold. If κ destroys a direction required for the target signal, then $\ell_t \notin \mathcal{R}(A^\star_{t,\kappa})$, or even $\ell_t \notin \overline{\mathcal{R}(A^\star_{t,\kappa})}$, and identification fails.

Extension with observed covariates.

If observed covariates $\boldsymbol{X}$ affect selection, outcomes, or latent-type measurements, all calibration equations are interpreted conditional on $\boldsymbol{X}$.

Assumption G.2 ($\boldsymbol{X}$-conditional restrictions)**.** For almost every $\boldsymbol{x}$, W, V, and the potential-outcome blocks satisfy the corresponding conditional independence restrictions given $(U, \boldsymbol{X})$.

Conditional average treatment effects can be estimated by regressing the orthogonal pseudo-outcome on $\boldsymbol{X}$.

One-sided noncompliance.

This subsection records an additional one-sided noncompliance extension that is not used in the main identification or estimation results but may be useful in related designs. Consider sample A from the self-selection environment and sample B from a structurally untreated environment. Under

Assumption G.3 (Transportability)**.**

$$Y_0|_{P^A} \stackrel{d}{=} Y_0|_{P^B}, \tag{101}$$

we have the following decomposition.

Theorem G.5 (ATE decomposition)**.** *If the Y_1-side calibration conditions hold in sample A, then μ_1 is identified from sample A and $\mu_0 = E_B[Y_0^B]$, so $\tau = \mu_1 - \mu_0$ is identified.*

Proof. By Assumption G.1 we have $\mathbb{E}_{\mathbb{P}^B}[Y_0^B] = \mathbb{E}_{\mathbb{P}}[Y_0]$, so μ_0 is identified from sample B alone via standard regression-on-controls. The μ_1 side follows by applying the identification result (Theorem 3.4) to sample A. The non-dependence on Q for the μ_1 side follows from the reduced DAG analysis in Section 2.2. □

Algorithm 2 Two-sample calibration estimator

1: Estimate μ_1 in sample A using the Y_1-side calibrated orthogonal moment.
2: Estimate μ_0 in sample B by the sample mean of Y_0^B.
3: Return $\hat{\tau} = \hat{\mu}_1 - \hat{\mu}_0$.

Theorem G.6 (Two-sample asymptotic normality)**.** *If $N_A/N \to \rho$ and the sample-A estimator satisfies the product-rate condition, then $\sqrt{N}(\hat{\tau} - \tau)$ is asymptotically normal with variance $\rho^{-1}\sigma_1^2 + (1-\rho)^{-1}\operatorname{Var}_B(Y_0^B)$.*

Proof. Apply Theorem F.4 to the μ_1 side in sample A to obtain $\sqrt{n_A}$ asymptotic normality of $\widehat{\mu}_1^A$. The μ_0 side in sample B is a standard sample-mean / regression estimator with $\sqrt{n_B}$-asymptotic normality. Independence of the two samples implies that $\widehat{\tau} = \widehat{\mu}_1^A - \widehat{\mu}_0^B$ has the stated asymptotic variance. □

Weighting.

Survey weights can be incorporated by replacing all sample averages and moment equations by weighted analogues.

Theorem G.7 (Influence function and EIF for ATE in the two-sample observed model)**.** *Under the one-sided noncompliance setting of Section G.3, sample independence, and Assumption G.1, the influence function for $\tau = \mathbb{E}[Y_1] - \mathbb{E}[Y_0] = \mu_1 - \mu_0$ under the two-sample observed model admits the decomposition*

$$\phi_\tau^{\text{eff,2S}}(O^A, O^B) \;=\; \phi_1^{\text{EIF},A}(O^A) - \phi_0^{\text{B}}(O^B), \tag{102}$$

where

- $\phi_1^{\text{EIF},A}(O^A)$ *is the EIF for μ_1 on sample A (Theorem E.11 applied to sample A);*
- $\phi_0^{\text{B}}(O^B) := Y_0^B - \mu_0$ *is the mean-shift IF for μ_0 on sample B (in which Y_0 is directly observed).*

If $\phi_1^{\text{EIF},A}$ is the EIF in Theorem E.11 and sample B is an unrestricted mean model, then $\phi_\tau^{\text{eff,2S}}$ is the EIF of the two-sample observed model. The corresponding efficiency bound is

$$\mathcal{V}_\tau^{\text{eff,2S}} \;=\; \frac{1}{\rho}\mathbb{E}_A\Big[(\phi_1^{\text{EIF},A})^2\Big] + \frac{1}{1-\rho}\,\text{Var}_B(Y_0). \tag{103}$$

When does the two-sample design make sense? The two-sample design with one-sided noncompliance is appropriate when (i) sample A contains the treated and control units with the full calibration structure, and (ii) sample B contains only "always-untreated" units (e.g., a comparison cohort that never received treatment). The transport assumption G.1 requires that the marginal distribution of Y_0 is the same in both samples, conditional on C. In the retirement application, sample B could be a comparison cohort of younger workers (still labor-force-attached) used to anchor the μ_0 distribution.

Why this gives an efficiency gain. In the single-sample design, both μ_1 and μ_0 are estimated via the auxiliary-measurement framework, each carrying its own calibration nuisance. In the two-sample design, μ_0 is estimated directly from sample B's mean, with no calibration nuisance. The adjoint range condition for μ_0 is therefore relaxed, and the asymptotic variance of $\widehat{\mu}_0$ becomes the simple $\text{Var}_B(Y_0)/N_B$ rather than the calibration-induced larger variance. The trade-off is that the transport assumption must hold; in practice this requires a careful argument that sample B is comparable to sample A's untreated subpopulation.

Proof. Sample independence gives an additive decomposition of the asymptotic variance of the difference $\widehat{\text{ATE}}^{\text{2S}} = \widehat{\mu}_1^A - \widehat{\mu}_0^B$.

Sample A side. Estimation of μ_1 in the standard setting follows Theorem E.11, giving EIF $\phi_1^{\text{EIF},A}$ with efficiency bound $\mathbb{E}_A[(\phi_1^{\text{EIF},A})^2]$. With effective sample size $N_A = \rho(N_A + N_B)$, the asymptotic variance of $\widehat{\mu}_1^A$ is $(N_A + N_B)^{-1}\rho^{-1}\mathbb{E}_A[(\phi_1^{\text{EIF},A})^2]$.

Sample B side. Since $D \equiv 0$ in sample B, Y_0^B is a direct observation of $\mu_0 = \mathbb{E}[Y_0]$ and does not require calibration machinery. The IF is simply $\phi_0^{\mathrm{B}}(O^B) = Y_0^B - \mu_0$ with efficiency bound $\mathrm{Var}_B(Y_0)$. The asymptotic variance of $\widehat{\mu}_0^B$ is $(N_A + N_B)^{-1}(1-\rho)^{-1}\mathrm{Var}_B(Y_0)$.

Independence and the difference give the total asymptotic variance (103). □

Two-sample setup formalization.

The two-sample setup considers a study in which the auxiliary-measurement framework is applied to a treated-untreated comparison sample (sample A) supplemented by an always-untreated comparison sample (sample B). In sample A, treatment status $D \in \{0, 1\}$ is observed along with the calibration variables (Y, W, V, S, Q). In sample B, treatment is identically zero ($D \equiv 0$), and only the untreated outcome Y_0^B and the conditioning covariates C are observed.

Sample independence and observed variables. The two samples are drawn from possibly different populations $\mathbb{P}^A$ and $\mathbb{P}^B$, and the observations within each sample are independent. The fundamental identification challenge is to estimate the average treatment effect $\tau = \mathbb{E}[Y_1] - \mathbb{E}[Y_0]$ using both samples efficiently.

When the two-sample design is useful. The two-sample design is advantageous when (i) the comparison sample B is much larger or cleaner than the untreated subsample of A, (ii) the calibration for the untreated side is weak or near-irregular (near-Case-I of the trichotomy on the Y_0 side), or (iii) historical or auxiliary data on the untreated population is available. Examples include synthetic control designs, historical comparison cohorts, and surveys of never-treated populations.

Identification and the two-sample DML estimator.

Under the transport assumption (Assumption G.1), the two-sample design gives the decomposition

$$\tau = \mu_1^A - \mu_0^B = \mathbb{E}_{\mathbb{P}^A}[Y_1] - \mathbb{E}_{\mathbb{P}^B}[Y_0],$$

where μ_1^A is estimated by the auxiliary-measurement framework applied to sample A, and μ_0^B is estimated by the sample mean of Y_0^B in sample B.

Two-sample DML estimator.

1. Apply the cross-fitted DML procedure of Section F to sample A alone to obtain $\widehat{\mu}_1^A$ with asymptotic variance $\mathbb{E}_{\mathbb{P}^A}[(\phi_1^{\mathrm{EIF},A})^2]/N_A$.
2. Compute $\widehat{\mu}_0^B = N_B^{-1}\sum_{i\in B} Y_0^{B,i}$ with asymptotic variance $\mathrm{Var}_{\mathbb{P}^B}(Y_0)/N_B$.
3. Form $\widehat{\tau}^{\mathrm{2S}} = \widehat{\mu}_1^A - \widehat{\mu}_0^B$ with asymptotic variance given by Theorem G.6.

Cross-fitting in the two-sample design. The cross-fitting within sample A proceeds as in the single-sample design (Section F). The cross-fitting is not extended to sample B because no calibration nuisance is estimated on B. The independence of samples A and B ensures that the estimation noise in $\widehat{\mu}_1^A$ and $\widehat{\mu}_0^B$ is uncorrelated.

Asymptotic theory and one-sided estimation machinery.

The asymptotic theory of the two-sample DML estimator simplifies relative to the single-sample design because the μ_0 side does not require calibration machinery.

Theorem G.6: asymptotic normality. Under sample independence and the standard cross-fitted DML regularity conditions on sample A alone, the two-sample DML estimator satisfies

$$\sqrt{N_A + N_B}(\widehat{\tau}^{2S} - \tau) \to_d \mathcal{N}\big(0,\, \rho^{-1}\mathbb{E}_{\mathbb{P}^A}[(\phi_1^{\mathrm{EIF},A})^2] + (1-\rho)^{-1}\mathrm{Var}_{\mathbb{P}^B}(Y_0)\big),$$

where $\rho = N_A/(N_A + N_B)$ is the asymptotic sample-size ratio. The asymptotic variance is the sum of the two single-sample variances, weighted by their relative sample sizes.

Efficiency relative to the single-sample design. In the single-sample design, the asymptotic variance of $\widehat{\tau}^{\mathrm{single}}$ involves calibration nuisances on both sides; the variance includes contributions from $\mathbb{E}[(\phi_1^{\mathrm{EIF}})^2]$ and $\mathbb{E}[(\phi_0^{\mathrm{EIF}})^2]$. When the μ_0 side calibration is weak, $\mathbb{E}[(\phi_0^{\mathrm{EIF}})^2]$ can be substantially larger than $\mathrm{Var}_{\mathbb{P}^B}(Y_0)$, and the two-sample design yields a smaller asymptotic variance.

Specification testing and correction when transport fails.

The transport assumption (Assumption G.1) is the substantive content of the two-sample design and is generally untestable from observed data alone. However, partial specification tests are feasible.

Covariate-balance test. Compare the distributions of the conditioning covariates C between samples A and B. Large differences indicate that transport may fail. Standard tools include Kolmogorov–Smirnov tests on each component of C, propensity-score-based balance tests (fit a propensity model for "sample A vs. sample B" and check covariate balance after weighting), and overlap diagnostics.

Outcome-level placebo test. If a subset of sample A is also untreated, compare the mean Y_0 in this subset to the mean Y_0^B in sample B conditional on C. A significant difference indicates transport failure.

Sensitivity analysis. If transport is suspected to fail by a known amount $\delta := \mathbb{E}_{\mathbb{P}^A}[Y_0] - \mathbb{E}_{\mathbb{P}^B}[Y_0]$, the two-sample estimator can be corrected by subtracting δ from $\widehat{\mu}_0^B$. The resulting estimator is unbiased if δ is known, and the inferential procedure can be extended to a partially identified set when δ is bounded by external evidence.

Reweighting correction. If the marginal distribution of C differs between samples A and B but the conditional distribution of Y_0 given C is the same, a reweighting of sample B by $\widehat{f}^A(C)/\widehat{f}^B(C)$ recovers the transport-adjusted estimator. This requires estimating the density ratio of C across samples, which is a standard exercise in entropy balancing or propensity-score reweighting.

H Simulation and empirical supplement

This appendix reports Monte Carlo evidence for full-sample Sieve GMM without DML. The simulations verify regular identification, demonstrate scalar-index failure, show failure of PCI-type scores under selection on gains, and document weak-range sensitivity and bootstrap-calibrated moment-residual diagnostics.

H.1 Monte Carlo designs

This subsection describes the simulation data-generating processes and estimators used throughout the Monte Carlo study.

Finite-support DGP and estimators.

The finite-support design uses

$$U, S, Q, V, W \in \{0, 1\}, \qquad C = S + Q,$$

with S, Q independent Bernoulli(1/2) and $U \sim$Bernoulli(1/2). Measurement-side and adjoint-side strength are controlled by

$$\Pr(W = U) = p_W, \qquad \Pr(V = U) = p_V,$$

with $p_W = p_V = 0.8$ in the baseline. The estimator uses cell-indicator sieve bases on the finite support, so the full estimator has no sieve approximation error. The finite-support experiments therefore isolate identification, not approximation.

Three DGPs are considered. The first, `good_c`, makes all effects of (S, Q) enter through the scalar index $C = S + Q$:

$$Y_1 = U \oplus \mathbf{1}\{C = 2\}, \qquad Y_0 = U \oplus \mathbf{1}\{C = 0\},$$

$$\Pr(D = 1 \mid U, Y_1, Y_0, C) = \Lambda(-0.5 + U + 0.8Y_1 - 0.5Y_0 + 0.4C),$$

where $\oplus$ is XOR and $\Lambda(t) = \{1 + \exp(-t)\}^{-1}$. Here $C = S + Q$ is a sufficient conditioning index.

The second, `bad_c`, makes S and Q act in opposite directions:

$$Y_1 = U \oplus S, \qquad Y_0 = U \oplus Q,$$

$$\Pr(D = 1 \mid U, Y_1, Y_0, S, Q) = \Lambda(-0.5 + U + 0.8Y_1 - 0.5Y_0 + 1.2S - 1.2Q).$$

The values $(S, Q) = (1, 0)$ and $(0, 1)$ both collapse to $C = 1$, but they have opposite effects on outcomes and selection. The scalar index $C = S + Q$ loses the $S - Q$ direction required by the adjoint range condition.

The third, `pci_failure`, illustrates the breakdown of standard PCI-type weights under selection on gains:

$$Y_1 = U \oplus W, \qquad Y_0 = U, \qquad \Pr(D = 1 \mid U, Y_1) = \Lambda(-0.2 + 1.4Y_1 + 0.8U).$$

The correct calibration must include the observed outcome Y in the argument of q. As a comparator we report a `no_y_q` estimator that uses $q(W, C)$ without Y, a standard PCI-type approximation.

We compare four estimators: (i) `full` with $C = (S, Q)$, (ii) `scalar` with $C = S + Q$, (iii) `no_y_q` (PCI-type, omits Y from q), and (iv) the `naive` difference. For each replication, primal and dual moments are estimated by full-sample Sieve GMM, and μ_1 is estimated by the sample

average of the plug-in signal

$$\widehat{\Gamma}_{1i} = D_i\{1 + \hat{q}(X_{1i})\}\{Y_i - \hat{b}(Z_{1i})\} + \hat{b}(Z_{1i}).$$

μ_0 is estimated symmetrically and the ATE as a difference. Standard errors are reported via both the plug-in influence-function variance and the multiplier bootstrap.

H.2 Main simulation results and failure modes

This subsection reports the main simulation results: the population oracle that separates identification failure from estimation error, the finite-support ATE simulation contrasting the regular case with scalar-index failure, and the documented failure of standard PCI-type scores in regimes where selection on gains is operative.

Population oracle: separating identification failure from estimation error.

Before sample-based simulations, we compute the calibration equations on the population distribution by exact enumeration, producing oracle estimates that contain no sampling noise. Table 15 shows that for `good_c`, both `full` and `scalar` have essentially zero oracle bias. For `bad_c`, the `scalar` estimator has μ_1 oracle bias of about 0.10 and ATE oracle bias of about 0.084, with a nonzero dual residual. This is not a finite-sample artifact: the scalar index $C = S + Q$ breaks the adjoint range condition at the population level. For `pci_failure`, the `no_y_q` estimator, which omits Y from q, has an oracle μ_1 bias of about 0.09.

The oracle calculation cleanly distinguishes two failure modes: identification failure (population-level bias even with infinite data) versus estimation failure (vanishing bias as $N \to \infty$). The `bad_c` scalar and the `pci_failure` no-Y cases are identification failures, not finite-sample problems. This is the sharpest possible evidence for the role of the adjoint range condition and of Y-dependence in the latent selection-odds representer.

Finite-support ATE simulation: regular case and scalar-index failure.

Table 16 reports ATE estimates for `good_c` and `bad_c`, with 1,000 replications and 500 multiplier-bootstrap draws per cell. For `good_c`, the `full` and `scalar` estimators perform nearly identically and achieve coverage around the nominal 95% level. When the scalar index $C = S+Q$ is a sufficient statistic, reducing the conditioning block is harmless. By contrast, under `bad_c` the `full` estimator remains unbiased while the `scalar` ATE bias persists at about 0.084 and coverage collapses toward zero as N grows. This is the standard symptom of identification failure: as confidence intervals shrink, they concentrate around the wrong pseudo-parameter.

Table 17 isolates μ_1 in `bad_c`. Looking at the one-sided mean is important because ATE bias can be partly cancelled between μ_1 and μ_0. The `scalar` μ_1 bias persists near 0.10 regardless of sample size, and coverage is essentially zero from $N = 2{,}500$ onward. This matches the theoretical counterexample in Appendix G.3.

Failure of standard PCI-type scores.

Table 18 compares three estimators on the `pci_failure` DGP: `full` (allows $q(W, Y, C)$), `no_y_q` (omits Y from q, mimicking standard PCI), and `naive`. The `no_y_q` estimator has μ_1 bias of

about 0.09 and coverage approaching zero in large samples. This shows that under selection on gains, the correct selection-odds representer must reproduce a latent treatment-odds quantity that conditions on the realized potential outcome Y_t; omitting Y from the calibration weight produces systematic bias.

The `pci_failure` DGP retains a small oracle bias for `full` (about 0.009 in μ_1). This DGP is deliberately constructed to be irregular or weak in order to stress-test PCI-type scores; it is not a regular-pathwise-identification benchmark. The row for `full` should therefore be read as a relative comparison—confirming that allowing Y in q removes the selection-on-gains bias—not as evidence that `full` delivers zero bias in all designs.

H.3 Data-adaptive Sieve GMM and robustness

This subsection reports the continuous and nonseparable stress tests, including the data-adaptive Sieve GMM estimator, in which the sieve basis is generated by a random forest fit on the primal-side variables.

Continuous and nonseparable stress tests.

Beyond the finite-support designs, Table 19 reports stress tests with continuous, non-Gaussian, and nonseparable data-generating processes. The baseline continuous design uses Laplace measurement noise and a polynomial sieve. The strong-selection variant increases the magnitude of the Y-dependence in D, and the scalar-bad design uses the bad scalar index in a continuous implementation. The nonseparable-exp design introduces a multiplicative interaction between U and S.

The `full` estimator with polynomial sieve handles the baseline and strong-selection designs reasonably, with bias around 0.01–0.02 and coverage near the nominal level. The scalar-bad design reproduces the identification-failure pattern of the finite-support `bad_c` design: bias near 0.08 and coverage collapsing to 0.45. The nonseparable-exp design produces visible approximation bias for a fixed polynomial sieve (bias -0.09, condition number 520), motivating the use of richer data-adaptive sieve features. The subsequent data-adaptive Sieve GMM experiments show that data-adaptive sieve features substantially reduce this approximation bias.

The continuous experiments illustrate that the identification logic extends beyond the finite-support designs and that the diagnostic signals (condition number, dual residual) continue to provide informative weak-range warnings.

Data-adaptive sieve features.

Tables 20 and 21 report results for an extended Sieve GMM in which the basis is constructed from random-forest leaf indicators rather than fixed polynomial features. The forest is trained on the auxiliary fold, and the leaf-membership indicators are used as a data-adaptive sieve basis on the evaluation fold. By cross-fitting, the construction of the basis is independent of the evaluation observations, so the finite-dimensional moment equations remain valid conditional on the learned features.

The sample-size scan in Table 20 uses 1000 Monte Carlo replications uniformly across $N \in \{2000, 5000, 10000\}$ under $K = 5$ true cross-fitting, 20 trees, minimum leaf size 50, ridge 10^{-4}, and a hard intercept-normalization constraint ($e_0^\top(A\hat{\alpha} - c) = 0$ exactly). The true μ_1 is the

Table 15: Population oracle calculations: separating identification failure from estimation error

DGP	Estimator	Target	True	Oracle	Bias	primal res.	dual res.
good_c	full	μ_1	0.500	0.500	0.000	0.000	0.000
good_c	scalar	μ_1	0.500	0.500	0.000	0.000	0.000
bad_c	full	μ_1	0.500	0.500	0.000	0.000	0.000
bad_c	scalar	μ_1	0.500	0.600	0.100	0.000	0.045
pci_failure	no_y_q	μ_1	0.500	0.590	0.090	0.000	0.214

Table 16: Finite-support ATE simulations

DGP	Estimator	N	Bias	RMSE	SE(asy)	SE(mb)	Cov(asy)	Cov(mb)
good_c	full	1000	0.000	0.071	0.070	0.070	0.94	0.95
good_c	scalar	1000	0.000	0.070	0.070	0.070	0.95	0.95
bad_c	full	1000	0.000	0.073	0.072	0.072	0.94	0.95
bad_c	scalar	1000	0.050	0.089	0.069	0.069	0.84	0.85

Table 17: μ_1 in bad_c: scalar conditioning fails

N	Estimator	Bias	RMSE	SE(asy)	Cov(asy)	Cov(mb)
1000	full	0.000	0.050	0.050	0.94	0.95
1000	scalar	0.100	0.112	0.045	0.07	0.08
5000	full	0.000	0.023	0.022	0.95	0.95
5000	scalar	0.100	0.103	0.020	0.00	0.00

Table 18: Failure of PCI-type scores

Target	Estimator	N	Bias	RMSE	SE(asy)	Cov(asy)
μ_1	full	5000	0.009	0.030	0.028	0.86
μ_1	no_y_q	5000	0.090	0.094	0.025	0.00
ATE	full	5000	0.010	0.043	0.041	0.86
ATE	no_y_q	5000	0.085	0.095	0.039	0.00

Table 19: Continuous and nonseparable stress tests

DGP	Estimator	True	Bias	RMSE	Cov(asy)	dual res.	cond(q)
baseline	full	0.30	0.01	0.08	0.92	0.02	180
strong selection	full	0.30	0.02	0.10	0.89	0.03	260
scalar-good	scalar	0.30	0.01	0.08	0.91	0.02	190
scalar-bad	scalar	0.30	0.08	0.13	0.45	0.10	450
nonseparable exp	polynomial	0.17	-0.09	0.11	–	0.08	520

analytical value $\mu_1 = 0.30 + 0.25/\sqrt{1-(1/3)^2} = 0.5651650429$, derived from the closed-form $\mathbb{E}[\exp(tUS)] = (1-t^2)^{-1/2}$ for independent standard normal U, S. Across the three sample sizes, the data-adaptive sieve estimator achieves RMSE that is 15.8%, 7.3%, and 5.0% of the `naive` estimator at $N = 2000, 5000, 10000$ respectively, a monotone improvement with sample size. The bias is uniformly small in absolute value (-0.013, -0.001, $+0.000$) and the rescaled bias $\sqrt{N}\,|\text{bias}|$ decreases sharply with sample size (0.56, 0.09, 0.00), consistent with root-N centered inference. The standard deviation contracts at roughly the $\sqrt{N}$ rate ($0.023 \to 0.012 \to 0.008$). In addition to the scalar moment residuals, vector moment residual norms $p_Z^{-1/2}\|N^{-1}\sum_i \Phi_Z(Z_{1i})\{D_i\hat{q}(X_{1i}) - (1-D_i)\}\|_2$ and $p_X^{-1/2}\|N^{-1}\sum_i \Phi_X(X_{1i})D_i\{Y_i - \hat{b}(Z_{1i})\}\|_2$ are reported (columns "vec. primal" and "vec. dual"). Both contract monotonically with N: vec. primal $0.025 \to 0.010 \to 0.005$ and vec. dual $0.006 \to 0.002 \to 0.001$. These results provide evidence that the proposed estimator is centered at the target μ_1 and that the full vector of finite-dimensional moment conditions, not just the scalar normalization, is approximately satisfied in finite samples. These results support the recommendation that empirical applications use data-adaptive sieve features within a cross-fitting framework, and that vector moment residuals be reported alongside bias and variance as a transparent diagnostic of estimator centering.

Simulation protocol for Table 20.

The DGP is the nonseparable-exp design constructed to be compatible with (R1-A) and (R2) under the deconvolution primitive conditions of Appendix B.3: $U, S, Q \sim \mathcal{N}(0,1)$ independent, $V = U + \text{Laplace}(0, 0.5)$, $W = U + \text{Laplace}(0, 0.5)$ independent of V given U, $Y_0 = 0.5\,U + 0.25\exp(US/3)$, and $Y_1 = Y_0 + 0.30 + 0.25\,U \cdot S + 0.2\sin Q$. The treatment selection rule is $\operatorname{logit}\mathbb{P}(D=1 \mid U, Y_1, Y_0, S, Q) = -0.3 + 0.6\,U + 0.3\,(Y_1 - Y_0) + 0.4\,S$, with V excluded from the selection rule (consistent with the maintained DAG exclusion) and S entering as a permitted conditioning component. This design is constructed to be compatible with R1-A (latent-odds anchor through the shared-U construction of V and W) and the observed adjoint range condition R2 (Y_1 is a deterministic function of (U, C) so that $\mathbb{E}[b(V,C) \mid X_1, D=1]$ can in principle reproduce Y_1 via deconvolution of V's Laplace noise). The estimator uses true $K = 5$ cross-fitting: on each training fold $\mathcal{I}_{-k}$, a random forest (20 trees, minimum leaf size 50, no feature subsampling) is fit on the full training fold with D as the partition-shaping target and feature inputs $X_1 = (W, Y, C)$; the resulting leaf-membership indicators are used as *adaptive sieve features* (not as a causal-forest estimator), with the partition target serving only to focus the sieve in directions relevant to the selection mechanism. A separate forest is fit predicting Y from $Z_1 = (V, C)$ on the treated training subsample to build the adjoint-side sieve features. A minor caveat is that the X-side forest uses Y on the full training fold (where Y mixes Y_1 and Y_0); the resulting leaf indicators are still well-defined sieve features in the observed X_1-support, but their interpretation as supervised partitions of the counterfactual outcome surface should not be taken too literally. A `LeafEncoder` is built on $\mathcal{I}_{-k}$ that fixes the leaf-indicator column space (one leaf per tree dropped for identifiability, with an explicit intercept column added). The primal and dual moment equations are then solved on $\mathcal{I}_{-k}$ by Tikhonov-regularized least squares with ridge parameter 10^{-4} on all non-intercept columns, subject to a hard normalization constraint that the intercept row of the moment system has exactly zero residual ($e_0^\top(A\hat{\alpha} - c) = 0$ for the primal and $e_0^\top(B\hat{\beta} - d) = 0$ for the dual). This is implemented by partitioned solve: $\hat{\alpha}_0$ is back-substituted from the hard constraint after solving for the non-intercept coefficients. The same encoder and coefficients are applied to the held-out fold $\mathcal{I}_k$ where the plug-in influence function $\hat{\psi}_i = D_i(1 + \hat{q}^{(-k)}(X_{1i}))(Y_i - \hat{b}^{(-k)}(Z_{1i})) + \hat{b}^{(-k)}(Z_{1i})$ is evaluated; no post-hoc clipping is applied to $\hat{q}$ or $\hat{b}$. Per-replication

scalar and vector moment residual norms are recorded; the vector norms are divided by $\sqrt{p_Z}$ (primal) and $\sqrt{p_X}$ (dual) for dimension-independent reporting. All three sample sizes use 1000 Monte Carlo replications with independent draws. The true μ_1 is the analytical value $\mu_1 = 0.30 + 0.25\,(1 - 1/9)^{-1/2} = 0.5651650429$, computed in closed form from the moment generating function of the product of independent standard normals.

H.4 Weak-range and moment-residual diagnostics

This subsection reports the bootstrap-based inference machinery, weak-range diagnostics, and bootstrap-calibrated moment-residual diagnostics, together with a summary of the Monte Carlo evidence.

Bootstrap inference: analytic, multiplier, and pairs.

Table 22 compares three inference procedures: the analytic influence-function SE, the multiplier bootstrap, and a pairs bootstrap with nuisance re-estimation. In the regular `good_c` design, all three deliver similar coverage near the nominal level. In the `pci_failure` μ_1 row, the analytic and multiplier SEs undercover, while the pairs bootstrap improves coverage somewhat. This is consistent with the interpretation that irregular or weak-range regimes activate additional uncertainty from nuisance re-estimation that the influence-function variance underestimates.

In practice we recommend analytic inference as the default and the pairs bootstrap as a robustness diagnostic when weak-range indicators (small singular values, large condition numbers, sensitive regularization paths) suggest irregularity. All bootstrap procedures reported in the paper—the multiplier and pairs bootstraps in the simulations and the empirical pairs bootstrap for the HRS application—use $B = 999$ resamples.

Weak range diagnostics.

Table 23 reports Branch G performance as the measurement and adjoint-side loadings $\rho_W = \rho_V$ weaken. As the loadings shrink, the smallest singular value of the empirical moment operator falls and the condition number grows (from about 25 to 150), RMSE rises, and a regularization bias appears at the weak end. Coverage degrades from 0.95 near $\rho_W = \rho_V = 0.9$ to about 0.76 at $\rho_W = \rho_V = 0.22$.

These weak-range diagnostics function analogously to first-stage F-statistics in instrumental-variable estimation. They do not test the latent anchor condition (which is generally untestable from the observed distribution; see Example D.11), but they reveal whether the empirical inverse problems are numerically weak or sensitive to regularization. Applied researchers should report these diagnostics alongside point estimates, and treat large condition numbers or collapsing singular values as a signal that conclusions are sensitive to the measurement and adjoint-side variation available in the data.

Bootstrap-calibrated moment-residual diagnostics.

Table 24 reports bootstrap-calibrated moment-residual diagnostics T_q and T_b, where T_q targets the primal moment evaluated on validation test functions and T_b targets the dual moment. The size check on `good_c` with the `full` estimator produces rejection rates of 0.04 and 0.07, close to the nominal 5% calibration level. The scalar-collapse design (`bad_c` with `scalar`) yields rejection

Table 20: Data-adaptive Sieve GMM: sample-size scan (DGP constructed to be compatible with the R1-A and R2 primitive conditions; true cross-fitting, hard normalization, uniform reps= 1000, analytical $\mu_1 = 0.5651650429$)

N	reps	bias	sd	RMSE	ratio	$\sqrt{N}\,\lvert\text{bias}\rvert$	vec. primal	vec. dual	naive bias	naive RMSE
2000	1000	-0.0126	0.0231	0.0263	0.158	0.564	0.0253	0.0056	0.1650	0.1666
5000	1000	-0.0013	0.0119	0.0119	0.073	0.095	0.0103	0.0020	0.1642	0.1648
10000	1000	+0.0000	0.0082	0.0082	0.050	0.003	0.0051	0.0010	0.1637	0.1640

Table 21: Data-adaptive Sieve GMM robustness (regenerated under the v4 DGP constructed to be compatible with the R1-A and R2 primitive conditions; $N = 5000$, reps= 1000, $K = 5$, analytical $\mu_1 = 0.5651650429$, naive RMSE= 0.1648)

trees	min leaf	ridge	bias	sd	RMSE	ratio
20	50	10^{-4}	-0.0013	0.0119	0.0119	0.073
20	100	10^{-4}	-0.0103	0.0133	0.0168	0.102
40	50	10^{-4}	-0.0027	0.0122	0.0125	0.076
40	100	10^{-3}	-0.0086	0.0125	0.0152	0.092

Table 22: Bootstrap inference comparison (uniform reps= 1000, $N = 2500$; multiplier bootstrap and pairs bootstrap both with $B = 999$ resamples and full nuisance re-estimation).

DGP	Estimator	Target	N	Analytic cov.	Multiplier cov.	Pairs cov.
good_c	full	ATE	2500	0.95	0.95	0.95
bad_c	scalar	μ_1	2500	0.00	0.00	0.00
pci_failure	full	μ_1	2500	0.77	0.77	0.84

Table 23: Weak range diagnostics (continuous DGP, Branch G; measurement/adjoint loadings $\rho_W = \rho_V$ weakened; $N = 5000$, reps= 1000).

$\rho_W = \rho_V$	N	Bias	RMSE	Coverage	$\sigma_{\min}$	cond.
0.90	5000	0.001	0.012	0.93	0.078	25
0.60	5000	0.002	0.022	0.94	0.017	97
0.45	5000	0.006	0.031	0.94	0.012	136
0.32	5000	0.019	0.046	0.88	0.011	142
0.22	5000	0.048	0.074	0.76	0.011	148

rate 1.00 for both statistics with mean values in the hundreds: the diagnostic correctly flags the scalar-index failure. The `pci_failure` no-Y design rejects T_b at 100% while T_q shows size 0.05, correctly localizing the breakdown to the dual moment.

These diagnostics do not test the latent anchor condition (which is generically untestable; see Example D.11), but they provide diagnostics for the observed calibration moments. When validation diagnostics reject systematically, the analyst should re-examine the conditioning block, the measurement specification, and the adjoint-side variables. In these stylized designs, the diagnostics have substantial power against the specified failure modes.

Summary.

The Monte Carlo evidence supports four points. First, the identification trichotomy is empirically visible: under `good_c` (regular pathwise identification) both the full and the sufficient scalar estimators perform well; under `bad_c` (calibration non-invariance with the wrong scalar index) the scalar estimator collapses to a biased pseudo-parameter; and under `pci_failure` (selection on gains with PCI-style scores) the no-Y estimator carries systematic bias. Second, weak-range diagnostics deliver smooth, interpretable warnings as measurement or adjoint-side variation deteriorates. Third, data-adaptive sieve features combined with cross-fitting substantially improve approximation in continuous nonseparable designs and offer sizable RMSE improvements over polynomial sieves. Fourth, bootstrap-calibrated moment-residual diagnostics provide a practical complement to point estimates; they should not be read as tests of the latent anchor or of population adjoint-range membership.

The Monte Carlo evidence illustrates the main identification logic, illustrates the failure of inappropriate scalar conditioning, and documents the finite-sample consequences of weak range variation. The data-adaptive sieve experiment is supplementary implementation evidence for continuous nonseparable designs.

Table 24: Bootstrap-calibrated moment-residual diagnostics (reps= 1000).

DGP	Estimator	Label	N	Reps	Rej(T_q)	Rej(T_b)	Mean T_q	Mean T_b
good_c	full	size check	5000	1000	0.043	0.069	13.8	24.6
bad_c	scalar	scalar collapse	5000	1000	1.000	1.000	398.5	203.5
pci_failure	no_y_q	PCI no Y	5000	1000	0.053	1.000	14.1	2461.9

H.5 Separating measurement-side and adjoint-side weakness

The weak-range sweep of Table 23 weakens the measurement-side and adjoint-side loadings together. To localize the source of the deterioration in a finite-support analogue, we take the regular `good_c` design and hold one side strong while degrading the other. Because the full estimator's primal and dual moment matrices are transposes, they share their singular spectrum, and the primal moment residual is zero to numerical precision; the binding diagnostics are the shared condition number and the dual moment residual. Degrading only the measurement W (with V held strong) leaves bias, RMSE, and coverage essentially unchanged—the target-invariance certificate makes μ_1 invariant over the now weakly identified calibration set—whereas degrading only the adjoint-side measurement V reproduces the RMSE and dual-residual deterioration of the diagonal sweep, although coverage need not move monotonically in this finite-sample design (Table 25 and Figure 3). The weak-range pathology in this design is thus governed by the adjoint range condition (R2), consistent with the division of labor between the latent anchor (R1-A) and the adjoint range (R2).

Table 25: Separating measurement-side (W) and adjoint-side (V) weakness (good_c, full estimator, target μ_1; $N = 5{,}000$, reps= 1000). The primal and dual moment matrices are transposes, so they share $\sigma_{\min}$ and κ; the primal residual is zero to numerical precision, and the dual residual is the binding moment diagnostic.

Sweep	(p_W, p_V)	Bias	RMSE	Cov.	$\sigma_{\min}$	κ	Dual res.
W-weak	$(0.95, 0.95)$	-0.000	0.008	0.95	0.030	3	0.000
	$(0.85, 0.95)$	-0.000	0.008	0.95	0.027	3	0.000
	$(0.75, 0.95)$	$+0.001$	0.008	0.94	0.025	3	0.000
	$(0.65, 0.95)$	$+0.000$	0.008	0.95	0.023	3	0.000
	$(0.55, 0.95)$	-0.000	0.008	0.96	0.023	3	0.000
V-weak	$(0.95, 0.95)$	-0.000	0.008	0.95	0.030	3	0.000
	$(0.95, 0.85)$	-0.000	0.010	0.96	0.026	3	0.000
	$(0.95, 0.75)$	$+0.000$	0.013	0.95	0.020	4	0.001
	$(0.95, 0.65)$	-0.002	0.023	0.95	0.013	7	0.001
	$(0.95, 0.55)$	-0.019	0.119	0.97	0.004	31	0.004
diagonal	$(0.95, 0.95)$	-0.000	0.008	0.95	0.030	3	0.000
	$(0.85, 0.85)$	-0.000	0.010	0.95	0.023	3	0.001
	$(0.75, 0.75)$	-0.000	0.014	0.96	0.016	4	0.001
	$(0.65, 0.65)$	-0.002	0.024	0.95	0.010	7	0.003
	$(0.55, 0.55)$	$+0.010$	0.084	0.90	0.004	21	0.008

In the W-weak sweep the adjoint-side measurement is held strong ($p_V = 0.95$) and only the primal-side measurement degrades; bias, RMSE, and coverage are essentially unchanged, because the target-invariance certificate makes μ_1 invariant over the (now weakly identified) calibration set. In the V-weak sweep the measurement is held strong and only the adjoint-side measurement degrades; this reproduces the RMSE and dual-residual deterioration of the diagonal sweep (coverage need not move monotonically in this finite-sample design). The asymmetry localizes the weak-range pathology to the adjoint range condition (R2).

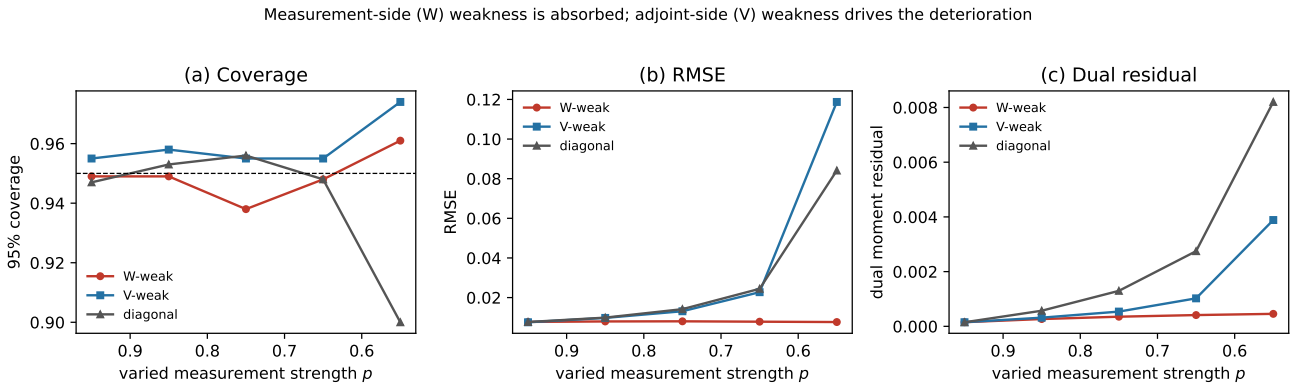


Figure 3: Measurement-side (W) weakness is absorbed by the target-invariance certificate, while adjoint-side (V) weakness drives RMSE inflation and dual-residual growth; interval calibration is sensitive mainly in the diagonal sweep (regular `good_c` design, full estimator).

H.6 Branch A/U/G validity ladder

To show how the empirical Branch A/U/G comparison of Section 6 should be read, we construct a finite-support family with a latent type U (measured by W and an adjoint-side measurement) and a latent gain type that enters the treated outcome, holding the outcome structure fixed and varying only the selection rule. Table 26 reports the population oracle and finite-sample behavior of the three branches across four rules. When treatment depends only on observed covariates all branches are centered; when it depends on the latent type, covariate adjustment (Branch A) is biased while the measurement-based branches recover the target; under selection on gains both Branch A and the no-Y Branch U are biased and only Branch G is (near-)centered; and when the gain signal is not spanned by the adjoint-side measurement even Branch G is non-invariant, with the population dual residual flagging the failure. The ladder makes explicit that branch differences measure sensitivity to progressively richer maintained information structures, not a test of which branch is correct. Branch G is the richest inverse problem and sits in a near-irregular regime under strong selection on gains, so its finite-sample coverage can be severely distorted even where the population oracle indicates identification; the oracle columns carry the identification message. Figure 4 plots the oracle bias by branch.

Table 26: Branch A/U/G validity ladder (finite support; oracle = population solve without sampling; finite sample at N = 5,000, reps= 1000). Branch differences are sensitivity to the maintained information structure, not formal tests of which branch is true.

DGP (valid branch)	Branch	Oracle bias	Oracle dual res.	FS bias	FS RMSE	Coverage
A-valid (D on C) [A,U,G]	A	−0.000	0.000	−0.000	0.010	0.95
	U	−0.000	0.000	−0.000	0.009	0.95
	G	−0.000	0.000	+0.000	0.020	0.97
U-valid (D on U,C) [U,G]	A	+0.097	0.000	+0.097	0.097	0.00
	U	−0.000	0.000	−0.000	0.010	0.95
	G	−0.000	0.000	−0.000	0.015	0.96
G-valid (D on U,C,G) [G]	A	+0.135	0.000	+0.135	0.135	0.00
	U	+0.106	0.000	+0.106	0.106	0.00
	G	−0.008	0.000	−0.050	0.064	0.62
G-failed (gain unmeasured) [none]	A	+0.135	0.000	+0.135	0.135	0.00
	U	+0.106	0.000	+0.106	0.106	0.00
	G	+0.086	0.051	+0.084	0.282	1.00

Oracle bias is computed on the population distribution by exact enumeration; a nonzero oracle bias is an identification failure that does not vanish as $N \to \infty$. A nonzero oracle dual residual flags failure of the observed adjoint range condition (R2). In `g_failed` the gain type is not spanned by the adjoint-side measurement, so even Branch G is non-invariant.

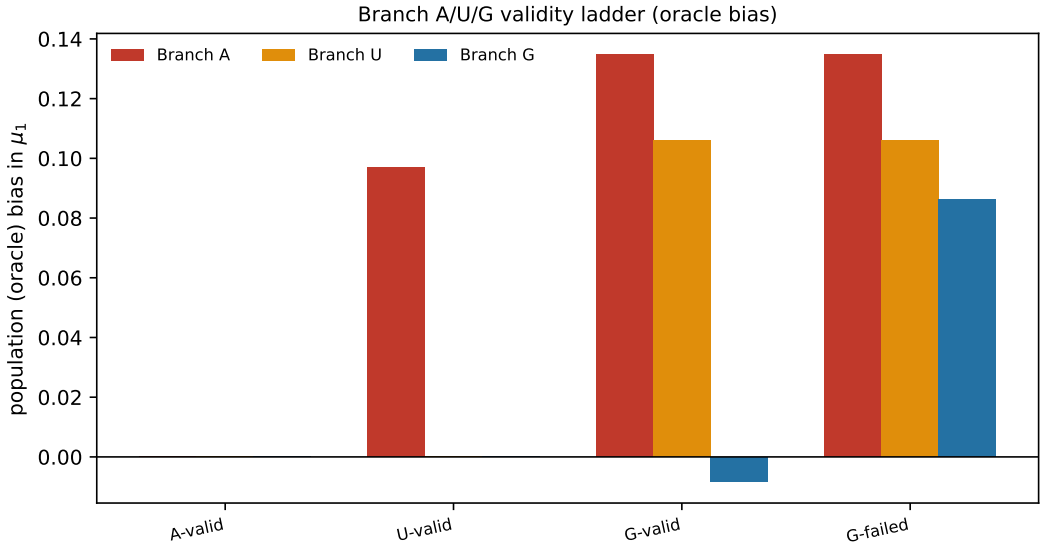


Figure 4: Branch A/U/G validity ladder: the population (oracle) bias is zero or near zero only for branches whose maintained information structure matches the selection rule. Under an unmeasured gain shock (`g_failed`) even Branch G has nonzero oracle bias.

H.7 Unmeasured outcome shock and adjoint range failure

Section 5 notes that an idiosyncratic treated-outcome component that the adjoint-side measurement cannot span sends the same implementation to a biased pseudo-target. Table 27 substantiates this with an exact finite-support oracle, sweeping how well the adjoint-side measurement spans the gain type. Population identification is a range (rank) condition: for any informative adjoint-side measurement the oracle bias and oracle dual residual are essentially zero, while finite-sample inference becomes weak as the measurement degrades. When the gain signal is entirely unmeasured the adjoint range condition fails, the estimator converges to a population-level pseudo-target that no sample size repairs, and the oracle dual residual jumps (Figure 5).

Table 27: Unmeasured outcome shock and adjoint range (R2) failure (selection-on-gains design, full/Branch-G estimator, target μ_1; $N = 5{,}000$, reps= 1000). p_{V_g} is how well the adjoint-side measurement spans the gain type; oracle quantities are computed on the population distribution by exact enumeration.

p_{V_g}	Adjoint range	Oracle bias	Oracle dual res.	FS bias	FS RMSE	Coverage
0.95	R2 holds	−0.004	0.000	−0.038	0.046	0.44
0.85	weak	−0.004	0.000	−0.032	0.049	0.68
0.75	weak	−0.004	0.000	−0.031	0.054	0.75
0.65	weak	−0.004	0.000	−0.031	0.063	0.85
0.55	weak	−0.004	0.000	−0.008	0.176	0.98
0.50	R2 fails	+0.084	0.054	+0.070	0.261	0.99

When the adjoint-side measurement spans the gain signal ($p_{V_g} = 0.95$) the observed adjoint range condition holds and Branch G is *population-centered in the oracle sense* with a near-zero oracle bias and oracle dual residual (its nonzero finite-sample bias and under-coverage at $p_{V_g} = 0.95$ reflect weak-inference and regularization effects, not a failure of population identification). For intermediate p_{V_g} the population target is still identified (the range condition is a rank condition) but inference is weak, and interval calibration is distorted. At $p_{V_g} = 0.5$ the gain signal is entirely unmeasured, R2 fails, and the estimator converges to a population-level pseudo-target (nonzero oracle bias that does not vanish as $N \to \infty$); the jump in the oracle dual residual is the diagnostic that flags the failure. The high finite-sample coverage in the failure row is not valid inference: it reflects weak, dispersed finite-sample behavior around a biased pseudo-target.

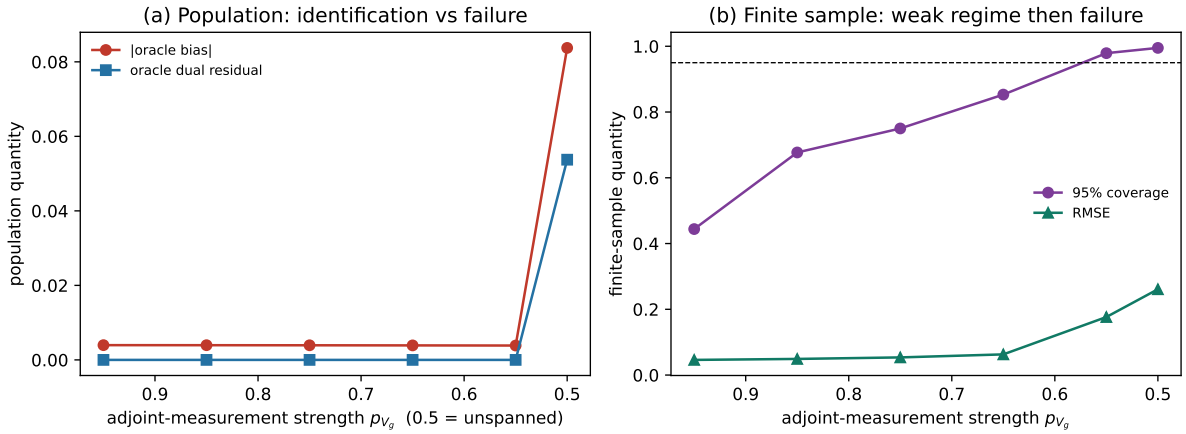


Figure 5: Unmeasured outcome shock. While the adjoint-side measurement spans the gain signal the population target is identified (left: oracle bias and dual residual near zero); when it does not, R2 fails and the estimator converges to a pseudo-target. The right panel shows the finite-sample weak regime preceding failure.

H.8 HRS-calibrated weak-range diagnostics

The empirical Branch-G condition number in Table 34 (about 8.6×10^5) is far above the largest value in the weak-range sweep of Table 23 (about 150). To verify that the diagnostics and inference behave sensibly at the conditioning of the application, we construct HRS-like synthetic covariates—a common factor induces realistic cross-covariate correlation among the latent type, the gain type, the measurements, and the environments (n = 1,597, treated share $\approx$ 0.71)—and use a fixed-dimension polynomial sieve whose degree and correlation span the dual-operator condition number from a well-conditioned control near 10^1 up to the same order of magnitude as the empirical HRS Branch-G value. This exercise is the inverse-problem analogue of a weak-first-stage stress test—the objects are condition numbers, Picard sums, and regularization paths rather than a scalar first-stage F-statistic—and is intentionally tuned to the HRS sample size and near-irregular Branch-G conditioning rather than offered as a regular Case-R benchmark. Estimation uses a relative Tikhonov ridge and the cross-fitted treated-side calibration; the target $\mu_1 = \mathbb{E}[Y_1] = 0.5$ is exact by construction. Concretely, a low-dimensional latent type and gain type drive a Branch-G selection rule in which treatment depends on latent type, observed environments, and potential-outcome gains; the measurements load on these latents through a common factor that reproduces realistic cross-covariate correlation, and the strength of the adjoint-side measurement component is varied to generate the condition-number grid. Table 28 reports the result. As the condition number approaches the empirical Branch-G value, the analytic influence-function coverage and the pairs-bootstrap coverage deteriorate together (from about 0.67 to 0.03 and 0.15 respectively), while the regularization bias becomes non-negligible relative to Monte Carlo sampling variation (about 0.03 to 0.12). The binding problem at this conditioning is therefore regularization bias, not underestimated variance: neither interval corrects a bias, so both under-cover, and the appropriate safeguard is the conditioning, Picard, and regularization-path diagnostics read alongside a cautiously interpreted point estimate, rather than a switch of interval method. Figure 6 reports the singular spectra and the Picard partial sums—the latter explode for the most ill-conditioned design, indicating that the outcome-regression signal loads on near-zero singular directions—and Figure 7 reports the Tikhonov path, with $\hat{\mu}_1(\lambda)$ crossing the truth at an intermediate λ and the expected L-curve in the dual norm and dual residual.

Table 28: HRS-calibrated weak-range *stress test* (HRS-like synthetic covariates, cross-fitted polynomial-sieve Branch-G estimator, target $\mu_1 = 0.5$; $N = 1{,}597$, reps= 1000; pairs-bootstrap coverage is a rate over 20 datasets, each with $B = 999$ pairs resamples and full nuisance re-estimation). The design grid reaches the same order of magnitude as the empirical HRS Branch-G condition number ($\widehat{\kappa} \approx 8.6 \times 10^5$); a well-conditioned control row is included. This is not a regular Case-R benchmark: it is intentionally tuned to the HRS sample size and near-irregular Branch-G conditioning.

Design	$\widetilde{\kappa}_b$	Picard	Bias	MC sd	Mean SE	RMSE	Cov. (an./pairs)
`control 1e1`	$2.9e1$	$4.6e0$	$+0.040$	0.058	0.046	0.071	0.78 / 0.80
`kappa 1e2`	$2.0e2$	$3.0e0$	$+0.034$	0.030	0.025	0.045	0.65 / 0.70
`kappa 1e3`	$1.5e3$	$1.1e1$	$+0.029$	0.029	0.023	0.041	0.68 / 0.65
`kappa 1e4`	$2.3e4$	$1.1e2$	$+0.060$	0.031	0.022	0.068	0.29 / 0.40
`kappa 1e5-6`	$3.7e5$	$2.1e3$	$+0.122$	0.047	0.022	0.131	0.03 / 0.15

$\widetilde{\kappa}_b$ and the Picard sum are Monte Carlo medians of the empirical dual moment operator across replications. As $\widetilde{\kappa}_b$ approaches the HRS Branch-G order of magnitude, both the analytic and the pairs-bootstrap intervals under-cover. The binding problem is regularization bias: at the highest design the bias is about 2.6× the Monte Carlo sd, so even a correctly sized interval would miss. The analytic SE additionally understates the Monte Carlo sd (mean SE below MC sd), so the pairs bootstrap – which better captures nuisance re-estimation uncertainty – helps marginally but cannot repair the bias. The appropriate safeguard is therefore the conditioning, Picard, and regularization-path diagnostics (Figures 6–7), read with a cautiously interpreted point estimate. The estimator is near-irregular by construction at these condition numbers and is not $\sqrt{N}$-centered.

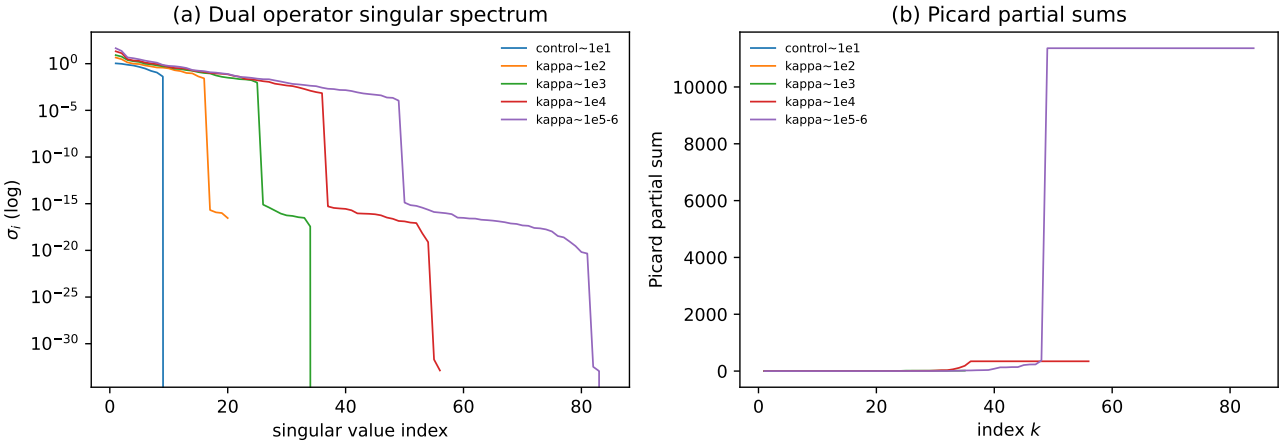


Figure 6: HRS-calibrated designs: (a) singular spectrum of the empirical dual moment operator and (b) Picard partial sums, across condition numbers approaching the empirical HRS Branch-G order of magnitude. The Picard sum explodes for the most ill-conditioned design, the numerical signature of an outcome signal concentrated on near-zero singular directions.

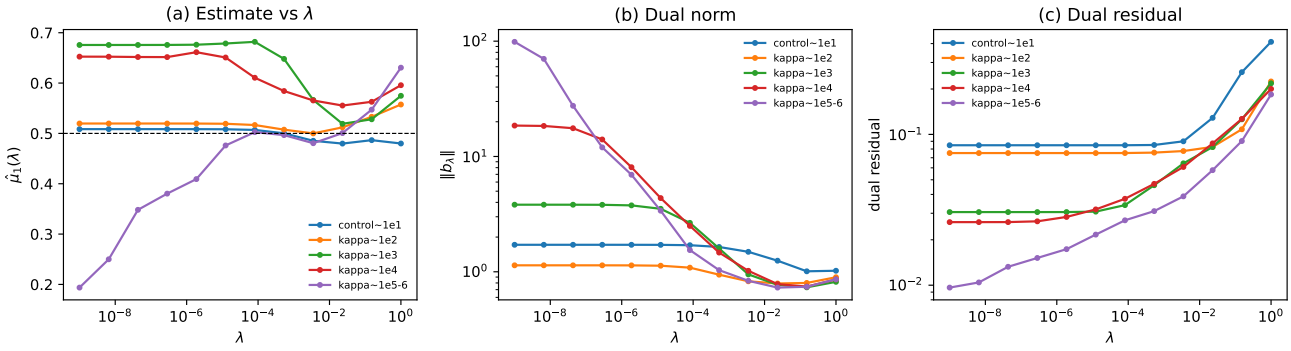


Figure 7: HRS-calibrated designs, Tikhonov regularization path: (a) the point estimate $\hat{\mu}_1(\lambda)$ crosses the truth at an intermediate regularization level; (b) the dual norm $\|b_\lambda\|$ and (c) the dual moment residual trace the expected L-curve, with the most ill-conditioned ($\kappa \approx 10^5$–10^6) design farthest from the well-regularized corner.

H.9 HRS application: construction, diagnostics, and visualizations

This subsection collects supplementary information for the HRS retirement-and-cognition application of Section 6: detailed variable construction, additional empirical diagnostics, visualizations of the estimated selection structure, and the mortality-attrition analysis.

Robustness and diagnostics for the HRS application.

This subsection collects the robustness sweep, calibration-weight and adjoint-side diagnostics, maintained-exclusion discussion, and weak-range analysis for the application of Section 6, together with the notation-to-RAND-code variable mapping (Table 29).

Two exclusions carry the calibration—the role of V = parental education on the adjoint side and of W = baseline cognition and self-rated health on the measurement side—and each warrants an institutional defense together with the most plausible direction of violation. The maintained restriction on V is that, conditional on the latent cognitive type and the Roy-side block (U, S, Q), parental education enters neither the retirement decision nor the cognition outcome directly ($V \not\to D$, $V \not\to Y_t$): its content is predetermined family background that shifts the distribution of latent type and early-life human capital but reaches later-life cognition and labor supply only through channels already conditioned on. The most plausible violation is a residual parental-education effect on late-life cognition operating through accumulated wealth, occupation, or health behaviors; because higher parental education is associated with both more informative adjoint variation and better cognitive outcomes, such a violation loads positively on the outcome representation and biases the Branch G estimate *toward* zero, so the maintained design is conservative with respect to the negative sign rather than manufacturing it. The maintained restriction on W is that baseline cognition and self-rated health measure the latent type without a direct path to the retirement decision conditional on (U, C) ($W \not\to D$); the most plausible violation is that baseline cognition affects retirement timing directly (sharper or healthier workers retire later), which would induce a negative measurement–selection dependence and inflate the apparent gain-selection correction. We therefore read the Branch U-to-Branch G movement as an upper bound on the gain-selection channel under that violation, and the sensitivity analysis of Appendix D.2 quantifies how far the maintained latent anchor conditions may deviate before the plug-in sensitivity band ceases to exclude zero (breakdown value $\delta^* = |\widehat{\tau}^G|/35 \approx 0.043$; Table 13 and Figure 2). Neither exclusion is testable from the HRS data: the placebo and drop-V diagnostics of Table 32 probe the observed adjoint content of V, not the exclusion itself.

Table 30 reports a robustness sweep across specification choices for Branch G. The sign is stable and the magnitude remains in the same broad range across $\lambda \in [0.1, 10]$, alternative outcome waves, the survivor-only and IPSW-adjusted restrictions, and fold counts; the smallest ridge value $\lambda = 0.01$ produces a larger magnitude, consistent with weak-range sensitivity. The full-retirement-only row uses a different treatment definition—complete labor-force exit rather than the broader transition into partial or full retirement—and therefore targets a different estimand; its smaller magnitude (-0.35) reflects this change in treatment composition rather than instability of the Branch G estimate for the maintained target. The sensitivity to treatment definition is consistent with the broader retirement–cognition literature.

Two diagnostics clarify what drives the Branch G estimate and whether the maintained adjoint-side information is doing substantive work. Table 31 reports the distribution of the calibration weight $1 + \hat{q}(X_t)$ over each arm. The treated (μ_1) side is well behaved—all weights are positive, with effective sample size close to the nominal one—whereas the untreated (μ_0) side carries heavier

Table 29: Model notation and corresponding RAND HRS variables.

Model role	Symbol	Construction	RAND HRS variables
Treatment and outcome			
Treatment	D	Indicator for partially or fully retired by wave 8 (2006); $D = 1$ if labor-force status $\in \{4, 5\}$.	`r8lbrf`
Outcome	Y	Total cognition score at wave 10 (2010), integer in $[0, 35]$; sum of immediate and delayed word recall, serial-7 subtraction, and orientation.	`r10cogtot`
Information sets			
Measurement block	W	Baseline (wave 3, 1996) cognition score and self-rated health.	`r3cogtot`, `r3shlt`
Outcome-representation variable	V	Mother's and father's years of schooling (time-invariant respondent attributes).	`rameduc`, `rafeduc`
Retirement-side environment	S	Baseline subjective probability of working past age 62, subjective probability of survival past age 75, pension dummy, employer health insurance, and log of total household wealth.	`r3work62`, `r3liv75`, `r3jcpen`, `r3covr`, $\log h3atotw$
Work-side environment	Q	Baseline occupation, industry, and job tenure.	`r3jlocc`, `r3jlind`, `r3jlten`
Adjustment covariates	X	Gender, race, ethnicity, age, census division, and respondent's own years of schooling.	`ragender`, `raracem`, `rahispan`, `r3agey_e`, `r3cendiv`, `raedyrs`

Notes. Variables are taken from the public-release RAND HRS Longitudinal File 2020 (V2). In the standard RAND naming convention, `r`W`XYZ` denotes the value of variable `XYZ` at HRS wave W, and `raXYZ` denotes time-invariant respondent attributes. The labor-force status variable `r`W`lbrf` takes integer values 1–7; values 4 ("partly retired") and 5 ("retired") jointly define our treatment indicator D at wave 8. The respondent's own years of schooling `raedyrs` is treated as a baseline adjustment covariate in X; it enters the observed-adjustment benchmark but not the Roy-side conditioning block of the bridge specifications.

and occasionally negative weights (0.9% below zero, effective sample size about 79% of nominal), consistent with $1 + q$ not being constrained nonnegative. Most of the cross-branch movement in the estimate originates on this μ_0 side, and the negative weights are a transparent feature of the calibrated representation rather than a numerical artifact.

Table 32 probes the maintained adjoint-side variable V (parental education). The estimate is stable across using mother's or father's education alone (-1.42 to -1.54), but collapses toward zero—and the interval covers zero—when V is randomly permuted or dropped from Z_t. This contrast indicates that the parental-education variation supplies genuine adjoint-side content for the range condition (R2) rather than entering as a redundant control: without informative V, the framework cannot certify the potential-outcome–dependent correction and reverts toward the weaker latent-only answer. This is a sensitivity diagnostic for the maintained restriction, not a test of it; the exclusion of V from the treatment and outcome margins remains an assumption.

The weak-range diagnostics highlight the numerical regime in which the empirical analysis operates. The estimated inverse problems are substantially ill-conditioned, particularly under Branch G, motivating the regularized implementation and bootstrap inference adopted throughout the analysis. As discussed in Appendix H.9, these diagnostics indicate a weak-range regime but do not by themselves invalidate the identifying assumptions. In particular, the HRS-calibrated stress test of Appendix H.8 shows that at the empirical Branch G condition number (about 8.6×10^5) both the analytic influence-function intervals and the pairs-bootstrap intervals can under-cover substantially, because the binding problem at this conditioning is regularization bias

Table 30: Robustness of the Branch G calibration ATE estimate to specification choices

Specification	ATE	SE
Baseline (cross-fit 5-fold, $\lambda = 0.5$)	-1.504	0.524
R1: Tikhonov regularization λ		
$\lambda = 0.01$	-2.447	0.865
$\lambda = 0.1$	-1.693	0.552
$\lambda = 1.0$	-1.480	0.511
$\lambda = 10$	-1.331	0.407
R2: Outcome window		
Wave 9 (2008), $n = 1{,}736$	-1.642	0.519
Wave 10 (2010), $n = 1{,}597$	-1.504	0.524
Wave 12 (2014), $n = 1{,}396$	-1.287	0.560
R3: Survivor analysis (alive at wave 10)		
Survivors-only, $n = 1{,}594$	-1.633	0.523
IPSW-adjusted	-1.621	0.547
R4: Cross-fitting fold count		
$K = 2$ folds	-1.911	0.527
$K = 5$ folds	-1.504	0.524
$K = 10$ folds	-1.553	0.522
R5: Restrictive treatment definition		
Full retirement only ($n_{\text{treat}} = 833$)	-0.349	0.387

All specifications use the Branch G cross-fit calibrated GMM implementation unless otherwise stated. The reported SE is the descriptive standard error from the cross-fitted orthogonal score; consistent with Table 10, no confidence intervals are attached, and coverage-valid intervals for this weak-range regime are developed in Appendix D.2, with the (R1-A) sensitivity analysis in Appendix D.2. The survivor-only specification additionally requires nonmissing survival-status and attrition-weight variables used in the IPSW construction, which excludes three observations relative to the main analytic sample ($n = 1{,}594$ versus $n = 1{,}597$). The sweep reports sensitivity of the point estimate under maintained calibration and adjoint-range assumptions.

rather than understated variance. The intervals reported in Tables 10 and 30 are therefore regular-regime intervals and should be read together with the conditioning, Picard, and regularization-path diagnostics rather than as weak-range-robust confidence statements; accordingly we do not attach significance claims to the Branch G point estimate.

HRS variable construction.

Why these choices are natural for Branch G. The block S contains forward-looking expectations, including the subjective probability of working past 62 and the subjective probability of survival to 75, that plausibly summarize retirement-side beliefs and constraints. These variables are natural candidates for the retirement-side environment because they may enter the perceived gain from retirement and may be related to anticipated post-retirement outcomes. The subjective work-continuation probability (`r3work62`) is particularly useful because it is an elicited baseline measure of the respondent's own expectation about continued work, rather than an ex post consequence of retirement.

The block Q captures the heterogeneity in continued-work payoffs: occupation-specific cognitive demands, industry-specific tenure patterns, and job-specific work environments. The $Q \to D$ path captures the well-documented tendency of workers in physically or cognitively demanding jobs to

Table 31: Calibration-weight diagnostics for the Branch G estimator ($\lambda = 0.5$)

Arm	$\min(1+\hat{q})$	$\max(1+\hat{q})$	mean	share < 0	ESS/n_{arm}
μ_1 (treated)	0.62	1.92	1.39	0.0%	0.97
μ_0 (untreated)	−0.31	8.27	3.59	0.9%	0.79

Distribution of the calibration weight $1 + \hat{q}(X_t)$ over the arm units $\{D = t\}$ for the full-sample Branch G estimator at $\lambda = 0.5$. ESS $= (\sum_i w_i)^2 / \sum_i w_i^2$ is the effective sample size of the cross-fitted weights $w_i = 1 + \hat{q}(X_{t,i})$ (5-fold, matching Table 10).

Table 32: Sensitivity of the Branch G estimate to the adjoint-side variable V

Adjoint-side variable V	ATE	Boot. SE	Interval / band
Both parents' education (baseline)	−1.504	0.651	[−2.89, −0.36]
Mother's education only	−1.421	0.738	[−2.93, −0.07]
Father's education only	−1.543	0.681	[−3.04, −0.40]
Placebo (randomly permuted V)	−0.71	–	[−1.40, +0.03]†
No V ($Z_t = (S, Q)$ only)	−0.49	–	–

Branch G cross-fitted estimates at $\lambda = 0.5$ (5-fold, matching Table 10). For the first three rows the interval is a pairs-bootstrap (999 reps) regular-regime 95% interval (same weak-range caveat as Table 10); these are not weak-range robust. †The placebo row is not a confidence interval: it reports the mean ATE over 200 independent random permutations of V (standard deviation 0.58), and the bracket is the $[10, 90]$ percentile band across permutations. The "No V" row reports a point estimate only.

retire earlier.

The measurement W uses baseline cognitive scores together with self-rated health to measure latent cognitive type before the retirement decision. Respondent's own schooling is included in X for the observed-adjustment benchmark and in the X-augmented bridge sensitivity, but not in the main Roy-side bridge block $C_R = (S, Q)$. Crucially, the longitudinal structure of HRS allows W to be measured *strictly before* the treatment, ruling out reverse causality from retirement to baseline cognition.

Additional HRS diagnostics.

Table 33 reports diagnostics on the residual variation in the retirement-side and work-side environment blocks and their incremental predictive content for retirement.

The results indicate that both the retirement-side block S and the work-side block Q contain substantial variation beyond the other block and contribute incrementally to explaining retirement status.

Table 34 reports singular-value diagnostics for the empirical adjoint operators across the three branches.

Interpretation. The condition number ranking $\kappa^A < \kappa^U < \kappa^G$ reflects the increasing dimension and flexibility of the empirical calibration specification. This is not, by itself, evidence against Branch G: ill-conditioning is expected when the observed calibration is enriched to include measurement and outcome dependence. The empirical Picard sums move in the opposite direction, with the smallest value for Branch G, indicating that the residualized target used in the dual construction is numerically better aligned with the leading empirical singular directions in that specification. We interpret this pattern as a finite-dimensional weak-range diagnostic: Branch G is more ill-conditioned as an inverse problem, but the particular target relevant for

Table 33: Identification-relevant diagnostics in the HRS data. Diagnostics summarize observed-system conditioning, not formal tests of the latent anchor or range conditions.

Diagnostic	Quantity	Value
DIAG 1: residual variation in S given (Q, X)		
P[work past 62]	R^2	0.080
P[live past 75]	R^2	0.064
Pension dummy	Efron-R^2	0.191
Employer health ins.	Efron-R^2	0.135
log wealth	R^2	0.117
DIAG 2: residual variation in Q given (S, X)		
Job tenure	R^2	0.184
Occupation (17 cat.)	accuracy	0.370 (base 0.186)
Industry (13 cat.)	accuracy	0.383 (base 0.289)
DIAG 3: incremental contribution to D (McFadden R^2)		
$D \sim X$	R^2	0.033
$+Q$	Δ	+0.038
$+S$	Δ	+0.063
$+Q$ given $X + S$	Δ	+0.029
$+S$ given $X + Q$	Δ	+0.054

the ATE is not concentrated only in the weakest singular directions. Because the branches use different domains, operators, and targets, these quantities should not be read as a formal ordering of the infinite-dimensional ranges $\mathcal{R}(A_1^\star)$ across branches. The branch validity ladder in Appendix H.6 illustrates how these branch differences should be read: as sensitivity to progressively richer maintained information structures rather than as a test of which branch is correct. Appendix H.8 reports an HRS-calibrated simulation that reaches the same order of magnitude as the condition numbers in this table and shows that, at that conditioning, analytic and bootstrap intervals both under-cover, so the conditioning and regularization-path diagnostics—not a choice of interval method—are the appropriate safeguard.

X-augmented bridge sensitivity.

The main bridge specifications condition on the Roy-side block $C_R = (S, Q)$, with the demographic covariates X entering only the observed-adjustment benchmark (Branch A) and not the bridge. Table 35 reports the sensitivity of Branches U and G to augmenting the bridge conditioning block to $C_R^+ = (X, S, Q)$. Augmenting the block leaves the sign and broad magnitude unchanged for Branch U and deepens the Branch G point estimate, but at a substantially worse conditioning of the empirical inverse problem: the unregularized operator condition number rises by roughly one-and-a-half to two orders of magnitude when X is added. The larger X-augmented Branch G magnitude is therefore more regularization-sensitive, which is one reason the parsimonious Roy-side block $C_R = (S, Q)$ is used for the headline estimate rather than treated as identification-irrelevant.

Table 34: Empirical singular-value spectrum and weak-range diagnostics across branches. Diagnostics summarize observed-system conditioning, not formal tests of the latent anchor or range conditions.

Branch	Cond. number $\widehat{\kappa}$	Picard sum $\widehat{\mathcal{P}}_1$	$\|\widehat{b}_{\lambda=0.01}\|$	$\|\widehat{b}_{\lambda=1.0}\|$
A: $q^A(C_A)$, $C_A = (X, S, Q)$	1.4×10^3	4.2×10^4	18.7	6.2
U: $q^U(W, C_R)$, $C_R = (S, Q)$	5.8×10^4	9.5×10^3	11.3	4.8
G: $q^G(W, Y_t, C_R)$, $C_R = (S, Q)$	8.6×10^5	1.7×10^3	8.4	3.9

The condition number is largest for Branch G, reflecting the higher-dimensional and more flexible calibration specification. The smaller empirical Picard sum for Branch G indicates that, within the chosen sieve and regularization scheme, the residualized outcome-regression target has a more stable projection onto the empirical singular system. The reported dual norms summarize the sensitivity of the regularized dual solution to the Tikhonov parameter. These diagnostics are finite-sample numerical summaries and should not be read as formal tests of the infinite-dimensional range condition or as a monotone comparison of identification strength across branches.

Table 35: Sensitivity of the bridge estimates to augmenting the conditioning block with demographic covariates X.

Branch	Conditioning block	ATE	Gram cond. number
U	$C_R = (S, Q)$	-1.147	7.4×10^{12}
U	$C_R^+ = (X, S, Q)$	-1.320	1.7×10^{14}
G	$C_R = (S, Q)$	-1.504	7.4×10^{11}
G	$C_R^+ = (X, S, Q)$	-2.693	3.9×10^{13}

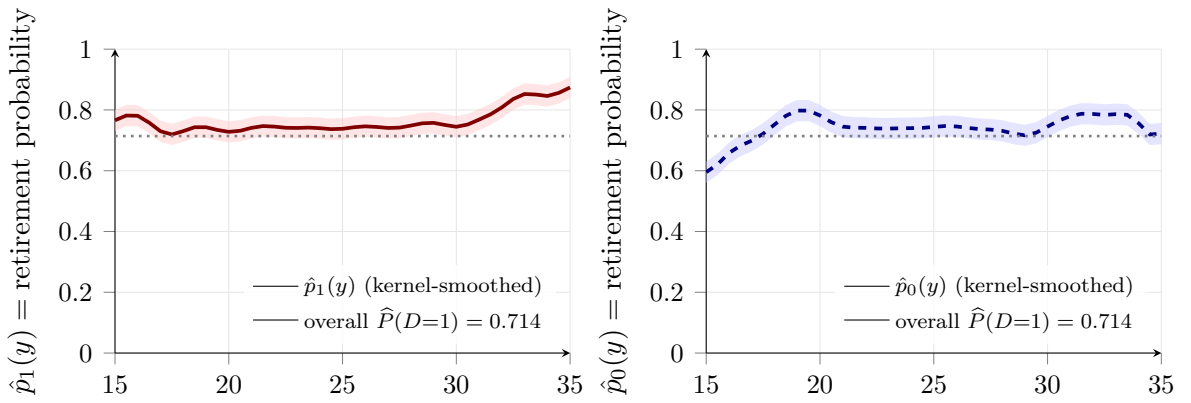
Cross-fitted bridge estimates (5-fold, Tikhonov $\lambda = 0.5$), $n = 1{,}597$. The "Gram condition number" is the condition number of the unregularized empirical moment matrix $M^\top M$ for the treated arm and is reported on a common scale across rows; it uses a different normalization from the spectral condition number $\widehat{\kappa}$ of Table 34 and is intended only for the within-table comparison of C_R versus C_R^+. Adding X worsens the conditioning by roughly 20–50$\times$ in each branch while leaving the qualitative conclusion—a negative bridge estimate that deepens with the potential-outcome coordinate—unchanged.

Visualization of the estimated selection structure.

To make the empirical content of the Branch G estimates more transparent, Figure 8 reports diagnostic marginal profiles derived from the selected Branch G calibration representative. The figure is intended as a descriptive summary of the estimated adjustment under the maintained calibration assumptions of Section 3, not as an independent test of any particular selection mechanism.

Reading the marginal plot.

The flatness of the two curves suggests that the selected Branch G representative does not imply strong marginal dependence of retirement on either potential-outcome level alone. This does not rule out dependence on the joint configuration of (Y_1, Y_0), nor does it rule out alternative explanations such as weak measurements, calibration misspecification, regularization-induced shrinkage, or sample-specific noise. The marginal panels and the comparison with Branch A's near-zero estimate ($\hat{\tau}^A = -0.073$) are mutually consistent but neither implies the other.



(a) Y_1 (cognition under retirement) (b) Y_0 (cognition under continued work)

Figure 8: Calibration-implied marginal selection profiles from the selected Branch G representative. Panel (a) plots $\hat{p}_1(y) = \{1 + \widehat{E}[\hat{q}_1^G(W, Y, C_R) \mid Y = y, D = 1]\}^{-1}$, which has the interpretation of $P(D{=}1 \mid Y_1 = y)$ only for the selected calibration representative under the maintained calibration assumptions of Section 3. Here $C_R = (S, Q)$ is the Roy-side conditioning block used by the main HRS bridge specifications. Panel (b) plots the analogous Y_0-side profile, $\hat{p}_0(y) = \widehat{E}[\hat{q}_0^G(W, Y, C_R) \mid Y = y, D = 0]/\{1 + \widehat{E}[\hat{q}_0^G \mid Y = y, D = 0]\}$. The dotted horizontal line indicates the overall sample retirement rate $\widehat{P}(D{=}1) = 0.714$ (1,140/1,597). Both profiles are representative-specific: $\hat{q}^G$ is a selected element of the non-unique observed calibration set $\mathcal{Q}_1$, and the displayed curves are not in general invariant across different elements of $\mathcal{Q}_1$, even though μ_1 itself is invariant via the target-invariance certificate. The shaded bands reflect kernel-smoothing uncertainty only and *do not* incorporate first-stage calibration-estimation uncertainty, regularization choice, cross-fitting variability, or sampling uncertainty. The curves are representative-specific summaries of the selected Branch G calibration representative under the maintained Branch G assumptions; they should not be interpreted as nonparametrically identified structural treatment probabilities.

Mortality attrition.

A specific concern in retirement-cognition studies is mortality attrition: between wave 3 (1996) and wave 10 (2010), 156 individuals in the analytic sample died, and an additional 154 left the survey. Such mortality and attrition are correlated with both retirement status and cognitive decline, raising the possibility that the observed cognitive gap is partly an artifact of differential survival.

Survivor analysis. Restricting to wave-10 survivors yields essentially the same estimate ($\widehat{\tau}^G = -1.633$ vs. the full-sample -1.504; see R3 in Table 30). This indicates that the central estimate is not driven by attrition asymmetry.

Inverse-probability-of-survival adjustment. As an additional check, we apply an IPSW (inverse-probability-of-survival weighting) correction, modeling the wave-3-to-wave-10 survival probability as a logit in (W, V, S, Q, X, D). The IPSW-adjusted Branch G estimate is $\widehat{\tau}^{G,\text{IPSW}} = -1.621$ (SE 0.547), again close to the baseline estimate. The similarity of the survivor-only and IPSW-adjusted estimates suggests that the main Branch G estimate is not mechanically driven by the observed wave-3-to-wave-10 attrition pattern, although this check does not by itself identify a full survivor-average or principal-stratum causal effect.

Connection to censored ATE and principal causal interpretation. A fuller treatment of mortality could replace the wave-10 outcome with a restricted-mean cognition (e.g., cognition av-

eraged over years of subsequent life, with a censoring rule for deaths). The auxiliary-measurement framework extends to this restricted-mean parameter without modification: the orthogonal score and adjoint range condition apply to any contrast of the form $\mathbb{E}[\psi(Y_1)] - \mathbb{E}[\psi(Y_0)]$ with ψ a bounded function of the survival-and-cognition trajectory. We leave the formal RMST extension to future work but report it as a natural extension of the present framework.