CERTIFYPAR: ADEQUACY CERTIFICATION AND COEFFICIENT INFERENCE FOR PIECEWISE AFFINE REGRESSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Precise coefficient estimates in piecewise affine regression do not establish that a common affine description is adequate within each region. We introduce CertifyPAR, a reporting contract that certifies regional affine adequacy before releasing coefficients with valid confidence intervals. Existing learners supply regions fixed independently of the audit responses; coefficients are estimated anew from the audit data. The targets are least-squares projection coefficients on the observed audit inputs, defined without assuming an affine mean. In the two-axis construction, two equivalence tests assess non-affinity within prespecified sub-regions (audit cells) and disagreement among their affine projections. Both noise-normalized root-mean-square departures must be certified below their respective prespecified margins. A pooled alternative certifies total departure from a common affine fit. In either construction, a region's requested coefficients are released together only after certification. Under independent homoskedastic Gaussian errors, certification based on audit residual projections preserves the coefficient-error distribution while reusing audit observations for estimation. For unknown error variance, two pure-error samples, independent of each other and the audit data, provide separate certification and inference scales. Under these assumptions, we establish finite-sample control of the familywise false-certification probability and, separately, simultaneous confidence-interval coverage conditional on every complete reporting pattern of positive probability. Simulations using synthetic data and real-data covariates assess error control under the assumed noise model and quantify trade-offs among reporting availability, coefficient precision, and measurement cost. GPU-runtime and fluorescence-calibration examples illustrate reporting decisions under benchmark adequacy margins and highlight obstacles to satisfying the noise assumptions required for exact guarantees.

1 INTRODUCTION

Regional coefficients provide concise summaries of flexible predictors. Model trees, clusterwise linear regression, and piecewise affine regression and classification (PARC) fit affine predictors within learned regions (Wang & Witten, 1997; Loh, 2002; Bu et al., 2026; Bemporad, 2023). Even when predictions are accurate, however, a regional slope may summarize incompatible local response patterns. A narrow confidence interval quantifies how precisely this summary is estimated; it does not establish its adequacy as a common affine description. We ask when a regional slope can be reported as a common summary of the response within that region. This requires both evidence of adequacy and valid inference after the reporting decision. First, failure to detect lack of fit may reflect limited power; equivalence testing instead requires affirmative evidence that departures lie below prespecified tolerances (Berger & Hsu, 1996; Robinson & Froese, 2004). Second, when the same observations determine whether regional coefficients are reported and are also used to estimate them, the confidence intervals must account for that reporting decision (Bock et al., 1973; Giles & Giles, 1993).

To address these requirements, we propose CertifyPAR, an adequacy-certification and coefficient-inference procedure for piecewise affine regression. It reports a region's coefficients only when audit data support the prespecified adequacy requirement, while preserving confidence-interval coverage

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

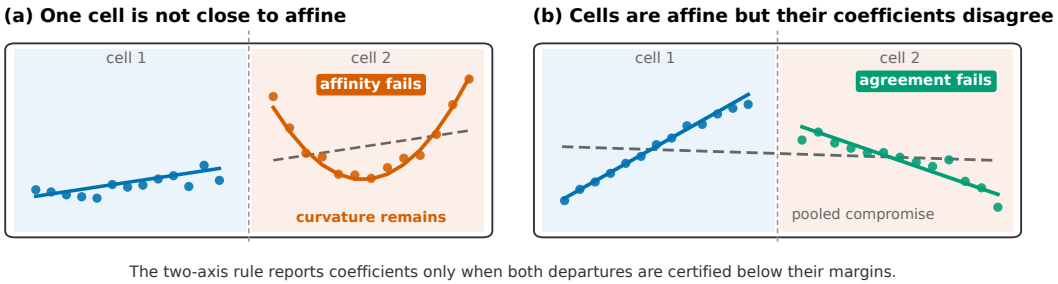

Figure 1: **Two failures of a pooled affine summary.** (a) One audit cell exhibits non-affinity even though its affine projection agrees with the other cell's projection. (b) Cellwise affine projections disagree, so the pooled projection is a compromise. The panels are schematic. Both certificates concern aggregate departures over a region's audit inputs, not worst-case departures in any individual cell or row.

after that decision. Existing learners supply regions fixed independently of the audit responses; the coefficients are then re-estimated from the audit data. The targets are the coefficients of the least-squares affine approximation to the mean responses at the observed audit inputs. These projection coefficients remain defined even when the mean is not affine. For each region, all coefficients requested in advance are reported together, or all are withheld. We study two certification rules within this reporting contract. A pooled rule certifies that total departure from a common affine fit lies below a margin. A two-axis rule separates non-affinity (A) within prespecified sub-regions, called *audit cells*, from disagreement (H) among their affine projections (Figure 1), and requires both departures to be certified below their respective declared margins. Both rules measure departures as noise-normalized root-mean-square deviations. Withholding does not by itself establish inadequacy.

Under independent homoskedastic Gaussian errors, coefficient estimation errors are independent of audit residual projections (Cochran, 1934; Battey et al., 2026). A reporting decision based only on these projections therefore leaves the coefficient-error distribution unchanged, allowing the same audit observations to be used for estimation. When the error variance is unknown, two pure-error samples–for example, suitably scaled differences between equal-mean replicate measurements–provide separate certification and inference scales. Independence of these samples from each other and the audit data preserves the Student-t distribution of the coefficient pivots conditional on the reporting pattern. This construction is sufficient for the stated conditional guarantee.

Our contributions are:

1. **A reporting contract with two error guarantees.** Calibrated adequacy tests control the familywise probability of false certification. Separately, residual-based reporting preserves simultaneous coefficient coverage conditional on every positive-probability pattern of reported and withheld regions (Corollaries 2 and 27).
2. **An exact inferential construction and its limits.** We specify the data separation and scale conditions needed for exact inference and provide counterexamples showing how scale reuse and heteroskedasticity can invalidate the gaurantees (Section 5; Appendix C).
3. **Reporting trade-offs and acquisition costs.** In growing known-noise designs with fixed margins and test levels, reporting probabilities at fixed nonzero departures strictly inside both margins tend to one for the two-axis rule and to zero for exact-affinity nonrejection at any fixed level (Proposition 34). In the matched-cost Gaussian repeat study, reuse yields narrower intervals than the implemented split procedure while retaining a similar fraction of audit rows in reported regions; the response budget includes scale measurements. The higher-dimensional and matched-cost studies show similar reporting availability for the pooled and two-axis rules; all unsuccessful settings are retained (Section 6).

The guarantees concern aggregate adequacy and projection coefficients at the observed audit inputs; extending them to future inputs requires additional assumptions. Applications also require scientifically justified margins and valid noise information. Under the benchmark margins, neither the GPU-runtime example nor the fluorescence-calibration example yields a coefficient report. In the calibration example, most observed within-cell residual variation occurs between wells at the same concentration. Both examples reveal obstacles to satisfying the required noise assumptions.

2 Related Work

Learning regional affine models. Model trees, PARC, and predictive clusterwise regression learn partitions of the input space together with region-specific affine predictors (Wang & Witten, 1997; Loh, 2002; Bemporad, 2023; Gitman et al., 2018; Bu et al., 2026). Hybrid system identification similarly estimates local dynamics and switching regions (Paoletti et al., 2007; Garulli et al., 2012). CertifyPAR addresses a different stage of the analysis: it takes the learned regions as given, fixes them independently of the audit responses, and asks whether each region supports a common affine description. The reported targets are least-squares projection coefficients on the realized audit design, following the approximation view of regression (Buja et al., 2019).

Affirmative evidence of adequacy. Testing whether a model is approximately valid, rather than exactly correct, is a classical statistical problem (Hodges & Lehmann, 1954; Dette & Munk, 1998). Building on equivalence testing (Berger & Hsu, 1996), direct precedents include model validation (Robinson & Froese, 2004) and noncentrality-based close-fit assessment in structural equation modeling (MacCallum et al., 1996; Yuan et al., 2016; Marcoulides & Yuan, 2017). The pooled rule applies this logic to regional regression residuals. The H component uses the reduction in residual sum of squares from pooled to cellwise fits familiar from Chow tests (Chow, 1960); the A component assesses remaining cellwise lack of fit (Draper & Smith, 1998). Each test certifies its departure below a nonzero margin, using an independent noise scale when variance is unknown. Learn then Test (Angelopoulos et al., 2025) calibrates predictive risks through multiple testing; our adequacy tests accompany a separate conditional coverage guarantee for reported coefficients. Our bounds for detectable departures build on classical Gaussian signal detection (Baraud, 2002).

Using the same data for adequacy assessment and inference. Classical Gaussian projection independence (Cochran, 1934) supplies our finite-sample argument. Sufficiency and ancillarity (Battey et al., 2026), estimator residualization (Sarfati & Vilfort, 2026), and inference under nonrejection (de Chaisemartin & D'Haultfœuille, 2026) address related questions, the latter two asymptotically. Under the stated Gaussian and scale assumptions, this orthogonal decomposition allows CertifyPAR to reuse the audit observations while preserving simultaneous coefficient coverage conditional on every positive-probability pattern of reported and withheld regions.

Partition selection and information splitting. Selective inference addresses fitted models (Lee et al., 2016), CART node means (Neufeld et al., 2022), cluster contrasts (Chen & Witten, 2023), and selected panel-group coefficients (Wan et al., 2026; Akgün & Okui, 2026); Berk et al. (2013) protect a prespecified family of models simultaneously. CertifyPAR removes partition selection from the audit by fixing regions and coefficient requests independently of audit responses; only the adequacy decision adapts. Row splitting separates observations (Rinaldo et al., 2019; Athey & Imbens, 2016), while carving, fission, and thinning divide information differently (Fithian et al., 2014; Leiner et al., 2025; Neufeld et al., 2024). Gaussian fission and thinning can preserve the full-design target. Their exact independent versions require correctly specified noise covariance, so a comparison under unknown variance must account for how that information is obtained and its cost. Our experiments compare reuse with row and repeat splitting; a matched comparison with fission and thinning remains open. Appendix H details these distinctions.

3 Problem Setting

Let (X_R, Y_R) be the finite audit sample. Conditional on the observed covariates, stable row IDs, and pre-response design information $\mathcal{F}_X$, assume

$$Y_R = \mu + \varepsilon, \qquad \varepsilon \sim N(0, \sigma^2 I), \tag{1}$$

with unrestricted μ. A frozen hard router $r_\psi : \mathbb{R}^{p_0} \to \mathcal{G}_{\text{slot}} \cup \{\perp\}$ assigns rows to finitely many potential region labels, called *slots*; $\perp$ denotes row-level abstention. A region's assigned rows form a *group*. A nonempty slot is a candidate; one passing the response-free design and power screens is eligible. An eligible group passing its prespecified certification rule is *released*: all its requested coefficients are reported. Discovery data may train the router under Appendix A.2; preprocessing, labels, abstention, and deterministic ties freeze independently of audit responses and scale measurements. Frozen transductive partitions may instead use a stable-ID map.

A response-free reference standardizes the $1 \le p \le p_0$ nonconstant coordinates to $z \in \mathbb{R}^p$; set $d = p + 1$. The fixed map T_{xz} returns coefficients to raw units (Appendix A.1). For candidate g, let N_g be its size, $A_g \in \mathbb{R}^{N_g \times d}$ have rows $(1, z_i^\mathsf{T})$, and Y_g, μ_g be its response and mean vectors. When $\mathrm{rank}(A_g) = d$, the least-squares projection target is

$$\theta_g^\circ = (A_g^\mathsf{T} A_g)^{-1} A_g^\mathsf{T} \mu_g, \qquad \theta_{x,g}^\circ = T_{xz} \theta_g^\circ. \tag{2}$$

This projection on the audit inputs requires no affine mean (Buja et al., 2019).

Let $\mathcal{F}_0$ collect $\mathcal{F}_X$, the completed discovery object, and all response-free frozen objects, excluding (Y_R, S_C, S_I). For unknown noise, each *pure-error bank* $b \in \{C, I\}$ supplies $m_b \ge 1$ independent $N(0, \sigma^2)$ errors. Conditional on $\mathcal{F}_0$, the banks are independent of each other and Y_R. Equal-mean replicate pairs provide such errors when all member errors are mutually independent $N(0, \sigma^2)$ across pairs, banks, and audit responses:

$$E_{b,r} = (Y_{b,r}^+ - Y_{b,r}^-)/\sqrt{2}, \quad S_b^2 = m_b^{-1} \sum_{r=1}^{m_b} E_{b,r}^2, \quad m_b S_b^2/\sigma^2 \sim \chi_{m_b}^2. \tag{3}$$

Certification uses S_C; independent S_I, opened after the reporting pattern and pooled fits freeze, preserves the pattern-conditional Student-t law. This sufficient design requires independence, not merely delayed opening (Appendix C). Replicate pairs cost $2(m_C + m_I)$ additional responses. An exact, externally justified $\sigma > 0$ frozen before responses removes both banks.

Predeclare coordinate requests, positive margins for the chosen rule, a familywise false-certification budget $\alpha_{\mathrm{cert}} \in (0, 1)$, and an interval noncoverage budget $\gamma_{\mathrm{FWER}} \in (0, 1)$ for every complete reporting pattern. Prospective power screens affect eligibility, not these budgets.

4 CertifyPAR

We first illustrate pooled certification, then define the separate two-axis requirements.

4.1 Core case: one frozen group with known noise

Let $A \in \mathbb{R}^{N \times d}$ be fixed and full-rank with $N > d$, $P = A(A^\mathsf{T} A)^{-1} A^\mathsf{T}$, and $Y = \mu + \varepsilon$ with $\varepsilon \sim N(0, \sigma^2 I)$ and known σ. The estimate depends on PY; gates using only $(I - P)Y$ are independent of its estimation error. For prespecified $\tau > 0$, the pooled rule tests $\|(I - P)\mu\|_2/(\sigma\sqrt{N}) \ge \tau$ and reports only after rejection, retaining ordinary Gaussian coefficient intervals. The two-axis rule uses the same independence for separate cellwise disagreement and non-affinity obligations. Both extend to multiple groups and coordinate requests, using independent pure-error scales for unknown noise. Algorithm 1 gives the two-axis response order; Appendix A.3 specifies its maps and data roles.

Algorithm 1 CertifyPAR two-axis implementation: freeze, certify, fit, and report

Inputs: Maps, margins, budgets, response roles, and coordinate requests predeclared before discovery responses; an independently trained or supplied frozen router.

1. **Before audit responses and scales:** Route X_R and evaluate the predeclared test-allocation map. Freeze cells, projectors, ranks, condition-number and variance limits, and prospective-power eligibility $\mathcal{G}_{\mathrm{aud}}$ under those levels. Retain the interval map for every reporting pattern.
2. **Certification:** Open Y_R and S_C. Apply the two lower-tail tests with static allocation, or group-wise IU/Holm (Section 4.2), using the frozen levels and eligibility screens.
3. **Freeze the decision:** Record the complete reporting pattern s. Evaluate its predeclared interval map without inspecting coefficients; then compute and freeze the reported groups' pooled fits.
4. **Inference:** Open S_I. Emit equation (8) for every requested coordinate of every $g \in s$, and nothing otherwise. No response-driven retries, retuning, or coefficient-based filtering are allowed.

Known noise omits both banks and uses Gaussian intervals.

4.2 Affirmative adequacy tests

For the two-axis rule, before opening (Y_R, S_C, S_I), partition each candidate g into $J_g \ge 2$ cells by a predeclared rule using audit covariates, tie-breaking IDs, and optionally the frozen discovery

object. Let A_{gj} be cell j's augmented design. Require $\mathrm{rank}(A_g) = \mathrm{rank}(A_{gj}) = d$ for every cell and $N_g > J_g d$. Ordering rows cellwise, define

$$\begin{aligned} P_g &= A_g(A_g^\mathsf{T} A_g)^{-1} A_g^\mathsf{T}, \quad P_{S,g} = \mathrm{diag}_j\{A_{gj}(A_{gj}^\mathsf{T} A_{gj})^{-1} A_{gj}^\mathsf{T}\}, \\ H_g &= P_{S,g} - P_g, \qquad\qquad W_g = I - P_{S,g}. \end{aligned} \tag{4}$$

The ranks are $r_{H,g} = (J_g - 1)d$ and $r_{A,g} = N_g - J_g d$. Predeclared condition-number, requested-coordinate unit-noise variance-factor, and power limits define the eligible set $\mathcal{G}_{\mathrm{aud}}$; all screens are response-free (Appendices A.4–A.5).

The pooled affine space is nested in the cellwise space, so P_g, H_g, W_g are mutually orthogonal and

$$\eta_{H,g} = \frac{\|H_g \mu_g\|_2}{\sigma\sqrt{N_g}}, \quad \eta_{A,g} = \frac{\|W_g \mu_g\|_2}{\sigma\sqrt{N_g}}, \quad \eta_{H,g}^2 + \eta_{A,g}^2 = \frac{\|(I - P_g)\mu_g\|_2^2}{\sigma^2 N_g}. \tag{5}$$

H measures disagreement among cellwise affine projections; A measures within-cell non-affinity. The observed H energy is the reduction in residual sum of squares from pooled to cellwise fits, as in a Chow comparison (Chow, 1960); our lower-tail test certifies departure below a nonzero margin. Predeclared margins $\tau_{H,g}, \tau_{A,g} > 0$ define adequacy as $\mathsf{Adeq}_g(\tau_{H,g}, \tau_{A,g}) = \{\eta_{H,g} < \tau_{H,g},\ \eta_{A,g} < \tau_{A,g}\}$. These are aggregate root-mean-square (RMS) departures, not worst-cell bounds.

The margins define a rectangle, whereas total deviation $\eta_{T,g} = (\eta_{H,g}^2 + \eta_{A,g}^2)^{1/2}$ gives

$$\{\eta_{T,g} < \min(\tau_{H,g}, \tau_{A,g})\} \subseteq \mathsf{Adeq}_g(\tau_{H,g}, \tau_{A,g}) \subseteq \{\eta_{T,g} < (\tau_{H,g}^2 + \tau_{A,g}^2)^{1/2}\}. \tag{6}$$

Neither reverse containment holds in general. Total deviation is cell-rule invariant; the H/A split and its separate adequacy obligations depend on the chosen resolution. For one group, a rule using only total residual energy and an independent scale (when needed), calibrated for the rectangular requirement, reports adequate departures with $\eta_T \geq \min(\tau_H, \tau_A)$ with probability at most its error level. On eligible growing known-noise designs with fixed margins and test levels, two-axis gates report each such fixed departure with probability tending to one (Remark 28).

Interpreting the margins. Margins permit structural RMS departures below $\tau_{u,g}\sigma$, excluding estimation error. They need scientific justification and freeze before discovery responses. Exact outcome-unit margins require predeclared exact σ (Appendix A.5); Appendix F.6 illustrates benchmark margins.

For $u \in \{H, A\}$, test $H_{0,u,g} : \eta_{u,g} \geq \tau_{u,g}$ against $H_{1,u,g} : \eta_{u,g} < \tau_{u,g}$. Predeclare a planning point $0 \leq \tau_{u,g}^- < \tau_{u,g}$ and power floor $1 - \beta_{u,g}$ to screen feasibility without changing the null. Put $M_{\mathrm{slot}} = \max\{1, |\mathcal{G}_{\mathrm{slot}}|\}$, $Q_{H,g} = H_g$, $Q_{A,g} = W_g$, and $r_{u,g} = \mathrm{rank}(Q_{u,g})$. Let $F_{r,m,\lambda}$ and $G_{r,\lambda}$ denote the noncentral-F and noncentral chi-square CDFs. Unknown-noise statistics and lower cutoffs are

$$F_{u,g}^{\mathrm{aud}} = \frac{\|Q_{u,g} Y_g\|_2^2 / r_{u,g}}{S_C^2}, \qquad c_{u,g} = F^{-1}_{r_{u,g}, m_C, N_g \tau_{u,g}^2}(\alpha_{u,g}). \tag{7}$$

Static allocation. Freeze positive axis levels totaling at most α_{cert} over all potential slots; the default is $\alpha_{u,g} = \alpha_{\mathrm{cert}}/(2M_{\mathrm{slot}})$. Require prospective power $F_{r_{u,g}, m_C, N_g(\tau_{u,g}^-)^2}(c_{u,g}) \geq 1 - \beta_{u,g}$, then release exactly when both $F_{u,g}^{\mathrm{aud}} < c_{u,g}$. Empty or ineligible slots retain their allocations.

Groupwise intersection–union/Holm (IU/Holm). Screen power at the conservative level $\alpha_{\mathrm{cert}}/M_{\mathrm{slot}}$. Form $p_{u,g} = F_{r_{u,g}, m_C, N_g \tau_{u,g}^2}(F_{u,g}^{\mathrm{aud}})$ and $p_g = \max_u p_{u,g}$ (Berger & Hsu, 1996); assign $p_g = 1$ to empty or ineligible slots. Release the groups rejected by strict Holm step-down at level α_{cert} (Holm, 1979). Known noise uses $T_{u,g}^{\mathrm{aud}} = \|Q_{u,g} Y_g\|_2^2/\sigma^2$ and the corresponding $G_{r,\lambda}$ quantiles, powers, and boundary p-values in either mode. Router abstention, design ineligibility, statistical non-release, and numerical failure remain separate outcomes (Appendix A.5).

Pooled equivalence. Alternatively, predeclare $\eta_{T,g} < \tau_{T,g}$ with $\tau_{T,g} > 0$, and response-free design and power screens requiring $\mathrm{rank}(A_g) = d$ and $N_g > d$; no cells are needed. Use equation (7) with $Q_{T,g} = I - P_g$, rank $r_{T,g} = N_g - d$, and boundary noncentrality $N_g \tau_{T,g}^2$. Allocate group levels totaling at most α_{cert} over all potential slots, or apply strict Holm at α_{cert} to the boundary p-values. Known noise uses the chi-square analogue. Appendix A.11 (Corollary 27) gives the calibration.

Table 1: Error guarantees for both rules under the frozen Gaussian model (Corollaries 2–3 and 27). Unknown noise uses the independent scales in equation (3). All rows condition on $\mathcal{F}_0$; only the second additionally conditions on a positive-probability complete reporting pattern.

Failure event	Additional conditioning	Upper bound
Any reported group violates a declared margin	None	α_{cert}
Any reported interval misses its target	$\widehat{\mathcal{S}} = s$	γ_{FWER}
Either failure occurs	None	$\alpha_{\mathrm{cert}} + \gamma_{\mathrm{FWER}}$

4.3 Atomic reports with audit-row reuse

For either rule, $\mathcal{G}_{\mathrm{aud}}$ is its frozen eligible set. A complete reporting pattern specifies reported groups $s \subseteq \mathcal{G}_{\mathrm{aud}}$ and all withheld eligible groups. Freeze a nonempty coordinate request $\mathcal{L}_g$ for each eligible g. With coordinate basis vector e_ℓ, define

$$\widehat{\theta}_g = (A_g^{\mathsf{T}} A_g)^{-1} A_g^{\mathsf{T}} Y_g, \quad \widehat{\theta}_{x,g} = T_{xz} \widehat{\theta}_g, \quad \widehat{\theta}_{x,g\ell} = e_\ell^{\mathsf{T}} \widehat{\theta}_{x,g}, \quad v_{g\ell}^x = e_\ell^{\mathsf{T}} T_{xz} (A_g^{\mathsf{T}} A_g)^{-1} T_{xz}^{\mathsf{T}} e_\ell.$$

Before responses, fix a map assigning $0 < \gamma_{g\ell}(s) < 1$ to every possible nonempty pattern, with $\sum_{g \in s, \ell \in \mathcal{L}_g} \gamma_{g\ell}(s) \le \gamma_{\mathrm{FWER}}$. Its evaluation may use the frozen design and pattern, never coefficient estimates. Using normal quantiles z_q or Student quantiles $t_{\nu,q}$, report

$$\mathrm{CI}_{g\ell}(s) = \widehat{\theta}_{x,g\ell} \pm \begin{cases} z_{1-\gamma_{g\ell}(s)/2} \sigma \sqrt{v_{g\ell}^x}, & \sigma \text{ known}, \\ t_{m_I, 1-\gamma_{g\ell}(s)/2} S_I \sqrt{v_{g\ell}^x}, & \sigma \text{ unknown}. \end{cases} \tag{8}$$

Report all requested coordinates for each $g \in s$; $s = \varnothing$ emits nothing. Values, signs, widths, and display choices cannot filter release. Only residual projections and S_C determine the pattern.

5 Theoretical Guarantees

Let $\widehat{\mathcal{S}} \subseteq \mathcal{G}_{\mathrm{aud}}$ be the reported groups and define

$$\mathscr{R} = \sigma(\{(I - P_g) Y_g : g \in \mathcal{G}_{\mathrm{aud}}\}, S_C), \tag{9}$$

omitting S_C when σ is known. Gaussian projection independence (Cochran, 1934) separates residual-based reporting from conditional coefficient inference. False certification requires separate test calibration (Table 1).

Lemma 1 (Residual-based reporting preserves coefficient inference)**.** *Under equation* (1) *and, when σ is unknown, the independent-scale contract in equation* (3)*, jointly over the disjoint audit groups,*

$$\left(\{\widehat{\theta}_g - \theta_g^\circ\}_{g \in \mathcal{G}_{\mathrm{aud}}}, S_I \right) \perp\!\!\!\perp \mathscr{R} \mid \mathcal{F}_0. \tag{10}$$

Known noise omits both scales and uses Gaussian pivots. Every predeclared $\mathscr{R} \vee \mathcal{F}_0$-measurable reporting rule preserves the joint coefficient-error and independent-scale laws conditional on every positive-probability complete reporting pattern. It may combine residual statistics across groups but cannot inspect $P_g Y_g$ or S_I.

Both rules satisfy this condition; their adequacy requirements and reporting probabilities differ.

Corollary 2 (Frozen-router certification and pattern-conditional inference)**.** *For the two-axis construction, every $g \in \mathcal{G}_{\mathrm{aud}}$ and $u \in \{H, A\}$ satisfy the unknown-noise law*

$$F_{u,g}^{\mathrm{aud}} \mid \mathcal{F}_0 \sim F_{r_{u,g}, m_C, N_g \eta_{u,g}^2}. \tag{11}$$

Under either the static axis allocation or the groupwise IU/Holm rule in Section 4.2,

$$\mathbb{P}\Big(\exists g \in \widehat{\mathcal{S}},\ u \in \{H, A\} : \eta_{u,g} \ge \tau_{u,g} \mid \mathcal{F}_0 \Big) \le \alpha_{\mathrm{cert}}. \tag{12}$$

Separately, for every complete reporting pattern s of positive conditional probability and every predeclared map satisfying $\sum_{g \in s, \ell \in \mathcal{L}_g} \gamma_{g\ell}(s) \le \gamma_{\mathrm{FWER}}$,

$$\mathbb{P}\Big(\exists g \in s,\ \ell \in \mathcal{L}_g : \theta_{x,g\ell}^\circ \notin \mathrm{CI}_{g\ell}(s) \mid \widehat{\mathcal{S}} = s, \mathcal{F}_0 \Big) \le \gamma_{\mathrm{FWER}}. \tag{13}$$

No latent-purity or routing-correctness premise is required. Known noise uses the noncentral chi-square analogue.

False certification is bounded before observing the pattern, not conditional on obtaining reports.

Corollary 3 (Joint reporting guarantee for the two-axis rule)**.** *Under the preceding two-axis procedure, suppose* $\alpha_{\mathrm{cert}} + \gamma_{\mathrm{FWER}} < 1$*. Conditional on* $\mathcal{F}_0$*, with probability at least* $1 - \alpha_{\mathrm{cert}} - \gamma_{\mathrm{FWER}}$*, every reported group is inside both declared margins, every reported interval covers its projection target, and*

$$\|(I - P_g)\mu_g\|_2/(\sigma\sqrt{N_g}) < \sqrt{\tau_{H,g}^2 + \tau_{A,g}^2} \tag{14}$$

for every reported g*.*

Audit-row reuse. Under exact affinity, full-audit OLS has no larger covariance in Loewner order than a full-rank subset fit. Appendix A.9 compares widths under matched allocations and information. Non-affine means generally yield different projection targets; unknown noise incurs scale costs.

Response-free planning. At fixed test levels, the known-noise Gaussian-sequence frontier is $\tau \asymp r^{1/4}/\sqrt{N}$ for axis rank r (Baraud, 2002): fixed cell count and dimension give $N^{-1/2}$ for H and $N^{-1/4}$ for A (Proposition 22, Appendix A.10).

Scale substitutions. Under exact affinity, reusing S_C in a Student-t interval is conservative for a single-group static all-pass gate: reporting favors a large denominator. This protection can fail when another group is withheld (Appendix C). Using the pooled residual scale instead causes single-group undercoverage because reporting selects small residual numerators.

Limits. The heteroskedastic Gaussian example in Proposition 29 has an exactly calibrated H-axis gate and correct unconditional sandwich variance, yet a nominal 95% interval covers only 86.91% after reporting; an independent A-axis pass does not repair it. Marginal calibration and a correct standard error do not replace reporting–coefficient independence. Heteroskedasticity can also invalidate the audit law. Withholding is inconclusive, and error budgets do not guarantee reports. Latent-class interpretation requires additional premises (Appendix B.4).

6 Experiments

We evaluate calibration, reporting availability, and measurement cost. Simulations provide known finite-design targets; measured responses support reporting decisions and assumption audits. We compare the two-axis rule (C), pooled equivalence, exact-affinity nonrejection (NR), sample splitting (S), and fresh full-design fitting responses (R). Coefficient- and regression-significance gates are negative controls for selection bias. The campaigns use benchmark margins $\tau_H = \tau_A = 0.75$ and budgets $\alpha_{\mathrm{cert}} = \gamma_{\mathrm{FWER}} = 0.05$; allocations are documented in Appendices E–G.

Affirmative adequacy and nonrejection answer different questions. On the same 41 eligible known-noise groups, both C and NR preserve conditional coefficient inference. At the inner departure $(0.2, 0.2)$, mean reporting probabilities are 0.93 for C and 0.81 for NR (Appendix E.9). NR's outside-margin reporting bound below 0.05 is specific to this screened roster; its usual exact-affinity level is not a general adequacy calibration. For growing eligible known-noise designs with fixed margins and levels, reporting at fixed nonzero departures strictly inside both margins tends to one for C and zero for NR (Proposition 34).

Calibration and audit-row reuse. E1 uses 200,000 Gaussian replications per arm and effect. Familywise noncoverage conditional on nonempty reporting is 0.0472–0.0475 against the 0.05 budget; this summary aggregates reporting patterns. The coefficient-inspecting controls reach about seven times that budget at zero effect (Appendix E.3). A fixed split retains only 11 of 41 eligible groups and has larger standard errors. Under non-affinity, scoring its fit-half intervals against the full-design target raises noncoverage from 0.0463 to 0.0556. Fresh full-design responses preserve the target at twice the audit-response cost (Appendix E.6).

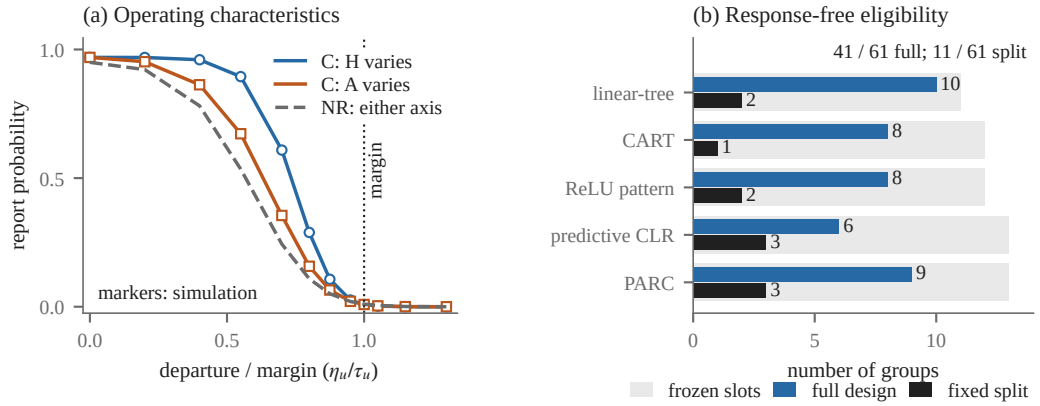

Figure 2: Reporting probability and response-free availability. (a) C and NR on the same 41 eligible known-noise groups, with the other departure zero. Lines average exact laws; C markers are saved simulation means. NR's level controls rejection at exact affinity; C's allocations control false certification at inadequacy boundaries. These levels are not interchangeable; Proposition 34 gives the fixed-level asymptotic contrast. (b) Full-audit and fixed-split eligibility among 61 frozen slots. Appendices E.4, E.6, and E.9 specify cohorts and allocations.

Independent noise information has a physical cost. Unknown-noise eligibility is zero with 50 or 100 certification pairs and reaches 31 of 61 slots with 400. Two 400-pair banks require 1,600 additional responses per arm (Appendix E.7). At matched total response cost in the balanced-repeat study, reuse and repeat splitting report essentially the same adequate mass, while reuse yields narrower intervals on their common distinct-input design. Pooled equivalence behaves similarly (Appendix G); these findings concern the implemented allocations.

Predictive fit and useful coefficient reporting. In the paired two-input curvature study, ungated reporting (U) emits an outside-margin group in every simulated family at amplitude 3. At that amplitude, C's reports cover about 55% and 27% of audit rows for linear trees and PARC, with observed familywise false-certification frequencies of 0 and 0.0006 (equal-design means over 20 designs; Appendix F.5). C and U use the same fits. The higher-dimensional study adds nuisance inputs to frozen functions. At $p = 5$, PARC achieves mean noisy-test $R^2 = 0.9682$ despite 46.47% inadequate audit mass (Figure 3a). Table 2 shows what each policy reports and retains the complete 22-cell decision roster.

The pooled comparator certifies the sufficient inner disk in equation (6), using C's cohort, audit OLS, Holm level, and interval allocation. Only one of 22 cells meets C's usefulness screens; pooled passes the same cell. All inadequate eligible groups in that cell violate A alone. No distinctive H-axis yield benefit appears in these designs; Remark 28 gives the complementary limitation of total-energy rules calibrated to the rectangle.

Measured audits explain why precise fits are withheld. In PLATERO's frozen gain50 calibration (González-Cebrián et al., 2023; 2022), an ungated analysis would report $R^2 = 0.9992$ and a slope of $14{,}415 \pm 10$ a.u./μM (rounded half-width of a working-reference 95% simultaneous interval). The H and A statistics exceed their certification cutoffs about 10- and 25-fold, so both tests fail. C and NR withhold both coefficients. Most observed within-cell residual variation lies between wells at the same concentration (Figure 3b; Appendix F.7). SGEMM repeats show scale incompatibility, higher PLATERO gains saturate, and BAM lacks a justified independent equal-mean repeat design (Appendix F.8). Earlier CCPP/Airfoil presets admit no slots through the geometry screen (Appendix F.9). These audits identify acquisition and modeling obstacles; none establishes an exact measured-data report or empirical true-coverage estimate.

Table 2: Reporting yield and its limits. Top: equal-design means for ten $p = 5$ PARC designs; masses use all audit rows as denominator. Bottom: counts out of ten within each frozen setting–learner cell. Screen A requires noisy-test $R^2 \geq 0.95$ and $\geq 10\%$ inadequate geometry-eligible mass. On A-positive designs with both mass types in the common power-eligible cohort, B requires a nonempty report, $\geq 80\%$ inadequate-mass removal, and $\geq 20\%$ adequate-mass retention. A cell needs ≥ 3 designs meeting A and B. The last row means zero in each of 19 cells. Appendix F.9 gives all definitions and rows.

Policy ($p = 5$, PARC)	Adequate mass (%)	Inadequate mass (%)
Ungated audit OLS (U)	18.4839	46.4673
Two-axis equivalence (C)	8.5469	0.0086
Pooled equivalence	7.5156	0.0069
Exact-affinity nonrejection (NR)	1.8367	0.0015

Frozen setting–learner cell	Cells	A	C-B	Pooled-B	NR-B
PARC demo, $p = 2$, PARC	1	3	2	2	2
PARC demo, $p = 5$, PARC	1	10	6	6	2
PARC demo, $p = 10$, PARC	1	2	2	2	0
Other frozen cells (each)	19	0	0	0	0

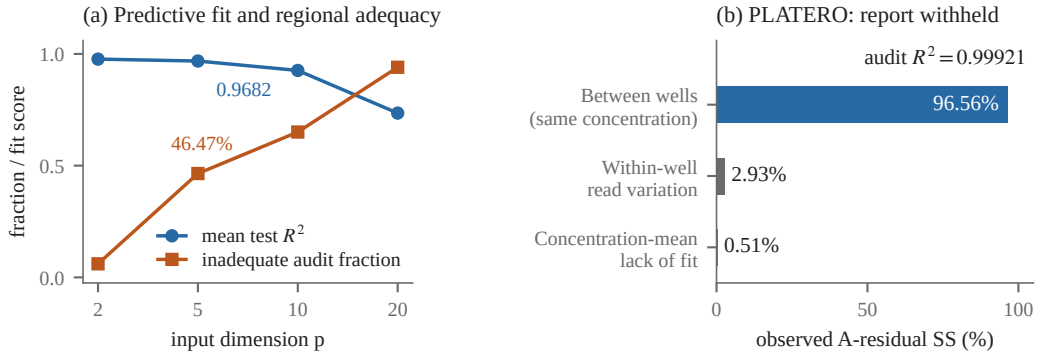

Figure 3: Two views of regional adequacy. (a) PARC predictive fit and inadequate audit mass as nuisance inputs are added; each point averages ten designs. (b) PLATERO gain50's observed A-axis residual sum of squares: 96.56% between wells at a common concentration, 2.93% within wells, and 0.51% concentration-mean lack of fit within cells. This post-freeze descriptive decomposition neither identifies variance components nor confirms the exact noise assumptions.

7 Conclusion

CertifyPAR is a reporting contract for regions supplied by existing learners: it reports least-squares projection coefficients on the observed audit inputs only after affirmative evidence of affine adequacy, not after a failure to detect lack of fit. Under the stated Gaussian and independent-scale assumptions, calibrated equivalence tests control familywise false certification, while residual-based reporting preserves simultaneous confidence-interval coverage conditional on every positive-probability complete reporting pattern. The two-axis and pooled rules certify different declared adequacy requirements. In simulations, accurate predictions coexist with inadequate affine summaries, and the experiments quantify trade-offs among reporting availability, coefficient precision, and measurement cost. The measured-data audits withhold even a precisely estimated calibration slope and expose obstacles to satisfying the required noise assumptions. Open questions include extensions to heteroskedastic errors, comparisons with fission and thinning at matched acquisition cost, and scientifically justified margin selection for specific applications.

REPRODUCIBILITY STATEMENT

The supplementary code provides the frozen protocols, generated-design seeds, pinned dependencies, complete numerical ledgers, and generators for Figure 2 and Appendices E and F. Separate source identities distinguish the E campaign, original F campaign, two-dimensional follow-up, and completion analysis. The package supports verification and rendering from saved full-run outputs, and separate commands for fresh experiments. A retained native Windows verification and render run passes all four components; its 45 display files reproduce either byte-for-byte or with exact decoded-image/rendered-page agreement under the recorded checks. This was a replay of saved results, not a new full campaign. A separate deterministic producer computes the residual-nonrejection comparison from the frozen E designs (Appendix E.9). Appendices E.8 and F.10 describe execution evidence and its limits. Protocol records document the intended response-opening order; they do not establish physical measurement independence. The artifacts are experimental reference implementations; the deployment requirements that remain outside them are specified in Appendix D.4.

AI-USE STATEMENT

The authors independently identified the research need and supplied the core theoretical ideas and substantial partial derivations. OpenAI ChatGPT assisted in formalizing and completing the theory, including formulating and refining mathematical claims, proposing additional results and proof steps, and drafting and checking proofs.

AI assistance covered the full experimental workflow, including experimental design, dataset sourcing, synthetic-data generation, code implementation and debugging, experiment execution, interpretation of results, and preparation of figures and reproducibility materials. The authors approved the experimental plans and the final inclusion of results. The authors determined the manuscript's overall structure; ChatGPT assisted with wording, sentence drafting, and editing.

AI-assisted validation included automated tests, numerical replays, and cross-checks between reported claims and recorded outputs. The supplementary materials document the scope and execution provenance of these checks. The authors made the final decisions about the included contributions and take full responsibility for the submission, including all AI-assisted text, mathematical claims and proofs, code, and experimental results.

REFERENCES

Oğuzhan Akgün and Ryo Okui. Robust inference methods for latent group panel models under possible group non-separation, 2026. URL https://arxiv.org/abs/2511.18550v2. Version 2, revised 27 July 2026; first posted in 2025.

Anastasios N. Angelopoulos, Stephen Bates, Emmanuel J. Candès, Michael I. Jordan, and Lihua Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. *The Annals of Applied Statistics*, 19(2):1641–1662, 2025. doi: 10.1214/24-AOAS1998.

Susan Athey and Guido W. Imbens. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360, 2016. doi: 10.1073/pnas.1510489113.

Yannick Baraud. Non-asymptotic minimax rates of testing in signal detection. *Bernoulli*, 8(5):577–606, 2002. URL https://projecteuclid.org/euclid.bj/1078435219.

Heather S. Battey, Daniel García Rasines, and Yanbo Tang. Post-reduction inference for confidence sets of models. *Biometrika*, art. asag045, 2026. doi: 10.1093/biomet/asag045. Accepted manuscript; published online 7 July 2026.

Alberto Bemporad. A piecewise linear regression and classification algorithm with application to learning and model predictive control of hybrid systems. *IEEE Transactions on Automatic Control*, 68(6):3194–3209, 2023. doi: 10.1109/TAC.2022.3183036.

Roger L. Berger and Jason C. Hsu. Bioequivalence trials, intersection–union tests and equivalence confidence sets. *Statistical Science*, 11(4):283–319, 1996. doi: 10.1214/ss/1032280304.

Richard Berk, Lawrence Brown, Andreas Buja, Kai Zhang, and Linda Zhao. Valid post-selection inference. *The Annals of Statistics*, 41(2):802–837, 2013. doi: 10.1214/12-AOS1077.

M. E. Bock, T. A. Yancey, and G. G. Judge. The statistical consequences of preliminary test estimators in regression. *Journal of the American Statistical Association*, 68(341):109–116, 1973. doi: 10.1080/01621459.1973.10481349.

Valentina Breschi, Dario Piga, and Alberto Bemporad. Piecewise affine regression via recursive multiple least squares and multicategory discrimination. *Automatica*, 73:155–162, 2016. doi: 10.1016/j.automatica.2016.07.016.

Thomas Brooks, D. Pope, and Michael Marcolini. Airfoil self-noise. UCI Machine Learning Repository, 1989. URL https://doi.org/10.24432/C5VW2C.

Dohyun Bu, Hyunho Kim, and Jong-Seok Lee. Covariate-guided clusterwise linear regression for generalization to unseen data. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/e878c8f38381d0964677fb9536c494ee-Paper-Conference.pdf.

Andreas Buja, Lawrence Brown, Richard Berk, Edward George, Emil Pitkin, Mikhail Traskin, Kai Zhang, and Linda Zhao. Models as approximations I: Consequences illustrated with linear regression. *Statistical Science*, 34(4):523–544, 2019. doi: 10.1214/18-STS693.

Harlan Campbell. Equivalence testing for linear regression. *Psychological Methods*, 29(1):88–98, 2024. doi: 10.1037/met0000596.

Marco Cerliani. linear-tree: Model trees with linear models at the leaves. Software, version 0.3.5, 2022. URL https://github.com/cerlymarco/linear-tree.

Yiqun T. Chen and Daniela M. Witten. Selective inference for k-means clustering. *Journal of Machine Learning Research*, 24(152):1–41, 2023. URL https://www.jmlr.org/papers/v24/22-0371.html.

Gregory C. Chow. Tests of equality between sets of coefficients in two linear regressions. *Econometrica*, 28(3):591–605, 1960. doi: 10.2307/1910133.

C. J. Clopper and E. S. Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413, 1934. doi: 10.1093/biomet/26.4.404.

William G. Cochran. The distribution of quadratic forms in a normal system, with applications to the analysis of covariance. *Mathematical Proceedings of the Cambridge Philosophical Society*, 30(2):178–191, 1934. doi: 10.1017/S0305004100016595.

Clément de Chaisemartin and Xavier D'Haultfœuille. Is inference conditional on not rejecting a pre-test less reliable than unconditional inference?, 2026. URL https://arxiv.org/abs/2407.03725v3. Version 3; first posted in 2024.

Holger Dette and Axel Munk. Validation of linear regression models. *The Annals of Statistics*, 26 (2):778–800, 1998. doi: 10.1214/aos/1028144860.

Holger Dette, Kathrin Möllenhoff, Stanislav Volgushev, and Frank Bretz. Equivalence of regression curves. *Journal of the American Statistical Association*, 113(522):711–729, 2018. doi: 10.1080/01621459.2017.1281813.

Norman R. Draper and Harry Smith. *Applied Regression Analysis*. John Wiley & Sons, New York, third edition, 1998. doi: 10.1002/9781118625590.

Giancarlo Ferrari-Trecate, Marco Muselli, Diego Liberati, and Manfred Morari. A clustering technique for the identification of piecewise affine systems. *Automatica*, 39(2):205–217, 2003. doi: 10.1016/S0005-1098(02)00224-8.

William Fithian, Dennis L. Sun, and Jonathan Taylor. Optimal inference after model selection, 2014. URL https://arxiv.org/abs/1410.2597. arXiv preprint arXiv:1410.2597; revised 2017.

Jerome H. Friedman. Multivariate adaptive regression splines. *The Annals of Statistics*, 19(1):1–67, 1991. doi: 10.1214/aos/1176347963.

Andrea Garulli, Simone Paoletti, and Antonio Vicino. A survey on switched and piecewise affine system identification. *IFAC Proceedings Volumes*, 45(16):344–355, 2012. doi: 10.3182/20120711-3-BE-2027.00332.

Judith A. Giles and David E. A. Giles. Pre-test estimation and testing in econometrics: Recent developments. *Journal of Economic Surveys*, 7(2):145–197, 1993. doi: 10.1111/j.1467-6419.1993.tb00163.x.

Igor Gitman, Jieshi Chen, Eric Lei, and Artur Dubrawski. Novel prediction techniques based on clusterwise linear regression, 2018. URL https://arxiv.org/abs/1804.10742. arXiv preprint arXiv:1804.10742.

Alba González-Cebrián, Joan Borràs-Ferrís, Yadira Boada, Alejandro Vignoni, Alberto Ferrer, and Jesús Picó. PLATERO: Calibration and analysis of GFP plate readers. experimental data sets and test software. Zenodo, version v1.0, 2022. URL https://doi.org/10.5281/zenodo.7224497.

Alba González-Cebrián, Joan Borràs-Ferrís, Yadira Boada, Alejandro Vignoni, Alberto Ferrer, and Jesús Picó. PLATERO: A calibration protocol for plate reader green fluorescence measurements. *Frontiers in Bioengineering and Biotechnology*, 11:1104445, 2023. doi: 10.3389/fbioe.2023.1104445.

J. L. Hodges, Jr. and E. L. Lehmann. Testing the approximate validity of statistical hypotheses. *Journal of the Royal Statistical Society: Series B (Methodological)*, 16(2):261–268, 1954. doi: 10.1111/j.2517-6161.1954.tb00169.x.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979. URL https://www.jstor.org/stable/4615733.

Jason D. Lee, Dennis L. Sun, Yuekai Sun, and Jonathan E. Taylor. Exact post-selection inference, with application to the lasso. *The Annals of Statistics*, 44(3):907–927, 2016. doi: 10.1214/15-AOS1371.

James Leiner, Boyan Duan, Larry Wasserman, and Aaditya Ramdas. Data fission: Splitting a single data point. *Journal of the American Statistical Association*, 120(549):135–146, 2025. doi: 10.1080/01621459.2023.2270748.

Wei-Yin Loh. Regression trees with unbiased variable selection and interaction detection. *Statistica Sinica*, 12(2):361–386, 2002. URL https://www3.stat.sinica.edu.tw/statistica/j12n2/j12n21/j12n21.htm.

Robert C. MacCallum, Michael W. Browne, and Hazuki M. Sugawara. Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods*, 1(2):130–149, 1996. doi: 10.1037/1082-989X.1.2.130.

Katerina M. Marcoulides and Ke-Hai Yuan. New ways to evaluate goodness of fit: A note on using equivalence testing to assess structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, 24(1):148–153, 2017. doi: 10.1080/10705511.2016.1225260.

Anna Neufeld, Ameer Dharamshi, Lucy L. Gao, and Daniela Witten. Data thinning for convolution-closed distributions. *Journal of Machine Learning Research*, 25(57):1–35, 2024. URL https://www.jmlr.org/papers/v25/23-0446.html.

Anna C. Neufeld, Lucy L. Gao, and Daniela M. Witten. Tree-values: Selective inference for regression trees. *Journal of Machine Learning Research*, 23(305):1–43, 2022. URL https://jmlr.org/papers/v23/21-0722.html.

Simone Paoletti, Aleksandar Lj. Juloski, Giancarlo Ferrari-Trecate, and René Vidal. Identification of hybrid systems: A tutorial. *European Journal of Control*, 13(2–3):242–260, 2007. doi: 10.3166/ejc.13.242-260.

Enrique Paredes and Rafael Ballester-Ripoll. SGEMM GPU kernel performance. UCI Machine Learning Repository, 2017. URL https://doi.org/10.24432/C5MK70.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12(85): 2825–2830, 2011. URL https://jmlr.org/papers/v12/pedregosa11a.html.

Alessandro Rinaldo, Larry Wasserman, and Max G'Sell. Bootstrapping and sample splitting for high-dimensional, assumption-lean inference. *The Annals of Statistics*, 47(6):3438–3469, 2019. doi: 10.1214/18-AOS1784.

Andrew P. Robinson and Robert E. Froese. Model validation using equivalence tests. *Ecological Modelling*, 176(3–4):349–358, 2004. doi: 10.1016/j.ecolmodel.2004.01.013.

Reca Sarfati and Vod Vilfort. Integrating diagnostic checks into estimation, 2026. URL https://arxiv.org/abs/2604.16690v1.

Markus Schilling, Steffen Glaubitz, Kathrin Matzak, Birgit Rehmer, and Birgit Skrotzki. Full dataset of several mechanical tests on an S355 steel sheet as reference data for digital representations. Zenodo, 2022. URL https://doi.org/10.5281/zenodo.6778336. Version 1.0.0, CC BY 4.0.

Nikolai Spuck, Matthias Schmid, Malte Monin, and Moritz Berger. Confidence intervals for tree-structured varying coefficients. *Computational Statistics & Data Analysis*, 207:108142, 2025. doi: 10.1016/j.csda.2025.108142.

Pinar Tüfekci and Heysem Kaya. Combined cycle power plant. UCI Machine Learning Repository, 2014. URL https://doi.org/10.24432/C5002N.

Chuang Wan, Jiajun Sun, and Xingbai Xu. Conditional selective inference for the selected groups in panel data, 2026. URL https://arxiv.org/abs/2511.04466v2. Version 2, revised 4 August 2026; first posted in 2025.

Yong Wang and Ian H. Witten. Induction of model trees for predicting continuous classes. In *Poster Papers of the 9th European Conference on Machine Learning*, pp. 128–137, Prague, Czech Republic, 1997. University of Economics, Faculty of Informatics and Statistics.

Stefan Wellek. *Testing Statistical Hypotheses of Equivalence and Noninferiority*. Chapman and Hall/CRC, Boca Raton, FL, second edition, 2010. ISBN 9781439808184.

Ke-Hai Yuan, Wai Chan, George A. Marcoulides, and Peter M. Bentler. Assessing structural equation models by equivalence testing with adjusted fit indexes. *Structural Equation Modeling: A Multidisciplinary Journal*, 23(3):319–330, 2016. doi: 10.1080/10705511.2015.1065414.

APPENDIX GUIDE

Question	Evidence and scope
Why does coefficient inference survive reporting?	Appendices A.7 and A.8.
What guarantees does pooled certification provide?	Appendix A.11, Corollary 27; rectangular-corner limitation in Remark 28.
What is gained by reusing audit rows?	Appendices A.9 and E.6.
When is spherical noise essential?	Appendix B, especially Proposition 29.
What does a reported projection mean?	Appendix B.4; aggregate adequacy and its limits in Appendix B.
Why separate certification and inference scales, and at what cost?	Appendices C and E.7.
How does equivalence compare with residual nonrejection?	Appendix E.9, Proposition 34.
What do measured responses reveal?	SGEMM in Appendix F.3 and calibration in Appendix F.7.
Where are reporting availability and comparison workflows documented?	Appendices E and F; Table E.11.
What remains at matched total response cost?	Appendix G, Table G.1.
How does the contract differ from prior methods?	Appendix H.
What do the supplied implementations establish?	Appendix D.4.

A ROUTER-AGNOSTIC AUDIT: SPECIFICATION AND HEADLINE PROOFS

This appendix first specifies the frozen objects and response order, then proves the projector decomposition, residual-release independence, false-certification control, pattern-conditional inference, audit-row reuse, and the adequacy frontier in that order. Appendix A.11 gives the pooled calibration (Corollary 27) and the scope of energy-only certification (Remark 28). Appendix B isolates target and distributional limits plus the conditional coefficient interpretation. Appendix C analyzes scale selection and the cost of unknown-noise inference. Appendix D describes the reference implementation. The statistical results are exact-real statements; software outputs are floating-point evaluations unless stated otherwise.

A.1 NOTATION, PROBABILITY SPACES, AND PARAMETER DOMAINS

All vectors are columns. For a symmetric matrix Q, $Q \succeq 0$ denotes positive semidefiniteness. For a full-column-rank matrix A, $P_A = A(A^\mathsf{T}A)^{-1}A^\mathsf{T}$. The symbol $F_{r,m,\lambda}(t)$ denotes the CDF at t of a noncentral F random variable with numerator degrees r, denominator degrees m, and noncentrality λ; $F^{-1}_{r,m,\lambda}(a)$ is its left-continuous quantile. Similarly, $G_{r,\lambda}(t)$ denotes the noncentral chi-square CDF and $G^{-1}_{r,\lambda}(a)$ its left-continuous quantile. Write Φ and ϕ for the standard-normal CDF and density, $z_q = \Phi^{-1}(q)$, and $t_{\nu,q}$ for the Student-t_ν quantile. All test distributions are continuous, so the quantile convention does not affect their probabilities. Every operational pass uses a strict inequality.

The base error model is specified conditional on $\mathcal{F}_X$, the sigma-field generated by all observed covariates, stable IDs, and pre-response design information. Formally, ordinary responses satisfy

$$Y = \mu(X) + \varepsilon, \qquad \varepsilon \mid \mathcal{F}_X \sim N(0, \sigma^2 I),$$

with mutually independent errors across the response roles below. Here $\mu(X)$ may be an arbitrary fixed vector after conditioning. The results here concern this conditional finite-design model.

Before audit responses or scales are opened, a declared response-free reference table X_{ref} fixes the coordinate system below. It may use independent discovery covariates or the complete audit covariate table. After deleting exactly zero-variance reference coordinates, the core audit requires retained

dimension $p \geq 1$ and sets $d = p + 1$; an intercept-only design is outside the stated domain. Its finite audit-cell-count set $\mathcal{J} \subset \{2, 3, \ldots\}$ is nonempty. The scale degrees m_C, m_I are positive integers in the unknown-noise audit. The condition-number bounds use the spectral condition number $\mathrm{cond}_2(M) = \lambda_{\max}(M)/\lambda_{\min}(M)$ for $M \succ 0$. All variance, condition, and precision bounds are finite and strictly positive. Each type-I and interval allocation is in $(0, 1)$; each power-failure allocation is in $[0, 1)$; and every displayed global error or failure budget is in $[0, 1)$ and is strictly positive whenever its opportunity family is nonempty. Audit margins obey the strict orderings stated below. Nonfinite configuration or design inputs lie outside the parameter space and cause pre-response ineligibility. A nonfinite value observed after a response or scale role opens instead causes the corresponding numerical or terminal failure.

For X_{ref}, let a_j and s_j be the empirical mean and standard deviation using divisor $|X_{\mathrm{ref}}|$. Delete coordinate j exactly when $s_j = 0$, put $D = \mathrm{diag}(s_j)$ on retained coordinates, and set $z = D^{-1}(x - a)$. With $\widetilde{z} = (1, z^{\mathsf{T}})^{\mathsf{T}}$,

$$T_{xz} = \begin{bmatrix} 1 & -a^{\mathsf{T}} D^{-1} \\ 0 & D^{-1} \end{bmatrix}, \qquad \theta_x = T_{xz}\theta_z. \tag{15}$$

The entire coefficient vector and covariance are transformed with T_{xz}; transforming marginal endpoints separately is invalid. Slopes on deleted coordinates are outside the estimand because, on the declared reference design, they cannot be distinguished from the intercept and retained slopes. This is an estimand convention tied to the frozen reference, not a rank necessity for X_R: a coordinate deleted there remains excluded even if it varies on the later audit rows.

A.2 Independent roles and exact scale premise

The response roles are independent discovery data (X_D, Y_D) when used, audit/report rows (X_R, Y_R), and, for unknown noise, certification and inference scale banks $D_{\sigma C}, D_{\sigma I}$. Write $\mathcal{B}_\sigma = \{C, I\}$. Rows have unique stable identifiers. Conditional on all covariates, the ordinary-response errors and all supplied scale errors are mutually independent. Covariate-only preprocessing may reuse covariate records from any role. No response-bearing measurement may be reused across discovery, audit/report, or scale-bank roles. Reuse within discovery is allowed; it does not permit reuse of an audit or scale-bank response. A physical unit may recur across roles only through distinct measurements satisfying this error-independence contract; any shared random unit effect inducing residual dependence violates it.

The balanced-repeat study in Appendix G uses a separate sufficient-statistic construction with shared physical repeats. It applies the statistical guarantees through the orthogonal transformation and rescaling specified there; it is outside this disjoint-measurement acquisition protocol.

For each supplied unknown-noise bank $b \in \mathcal{B}_\sigma$, the protocol supplies either independent direct pure errors $E_{b,r} \overset{\mathrm{iid}}{\sim} N(0, \sigma^2)$ or distinct replicate pairs. A pair must satisfy the scientific design premise

$$\mathbb{E}(Y^+_{b,r} \mid \mathcal{F}_X) = \mathbb{E}(Y^-_{b,r} \mid \mathcal{F}_X), \qquad E_{b,r} = (Y^+_{b,r} - Y^-_{b,r})/\sqrt{2}. \tag{16}$$

All members of all pairs are distinct and have independent $N(0, \sigma^2)$ errors. Identical covariates imply equal means only under an additional common regression-function model; covariate identity alone is not sufficient under the arbitrary fixed-mean model. Define

$$S_b^2 = m_b^{-1} \sum_{r=1}^{m_b} E_{b,r}^2, \qquad b \in \mathcal{B}_\sigma. \tag{17}$$

The two scale roles are not interchangeable under the present exact contract. S_C enters the release statistics and cannot also serve as the release-independent Student denominator. S_I remains unopened until the complete statistical reporting pattern and all released-group pooled fits are frozen. A residual scale from the possibly misspecified audit fits is not a uniform substitute: lack of fit can inflate the denominator and make the lower-tail equivalence gate anti-conservative. These are requirements of this procedure, not a claim that every conceivable method needs two scales. In the known-noise branch, an exact, externally justified $\sigma > 0$ is frozen before responses, and all scale banks and scale-reading operations are absent.

The theory permits either direct pure errors or replicate pairs, combined with either a frozen covariate router or a response-blind transductive partition. The reference implementation accepts replicate

pairs S_C, S_I for an external serialized stable-ID-to-label-or-$\perp$ map. The external adapter accepts no callable and makes no unseen-query claim; it hashes source, dataset, and payload identities and requires attested blindness to Y_R, S_C, S_I, a scientific premise. Direct pure-error input is unimplemented.

Lemma 4 (Pure-error laws)**.** *Under the cross-role independence premise above and either direct pure errors or replicate pairs satisfying equation* (16)*, $m_b S_b^2/\sigma^2 \sim \chi^2_{m_b}$. All supplied scales are mutually independent and independent of all ordinary-response roles.*

Proof. The equal means cancel. The difference of the two independent errors has variance $2\sigma^2$, so division by $\sqrt{2}$ yields $E_{b,r} \sim N(0, \sigma^2)$. Distinct rows and the stipulated independence give independent standardized normals. Summing their squares proves the chi-square law, and disjoint roles give the joint independence. The same conclusion is immediate for directly supplied pure errors. □

A.3 Response order and frozen configuration

The detailed two-axis procedure follows four response-ordered stages; Corollary 27 gives the pooled specialization.

A1. Before any optional discovery response, freeze all response roles, X_{ref}, finite slot caps and naming rules, admissible cell and tie rules, coordinate-request, margin, and budget maps, release mode, interval-allocation map, and every design and prospective-power rule. Also accept a compliant hard router (or transductive response-blind partition), or freeze a discovery-training procedure independently of audit responses.

A2. If discovery training is used, open its response role once and serialize exactly one router. From that router or the accepted frozen object, instantiate and freeze the realized slot family, coordinate requests, and allocations by the predeclared maps. Using only that object, X_R, and stable IDs, route rows, construct cells and projectors, and screen response-free eligibility. Empty or ineligible realized slots retain their multiplicity cost.

A3. Open (Y_R, S_C) once in the unknown-noise branch, or Y_R once when σ is known. Apply the declared two-axis rule, freeze the complete statistical release pattern s, evaluate the predeclared interval map at s, and compute and freeze the pooled fits for released groups.

A4. Only after both s and those pooled fits are frozen, open S_I in the unknown-noise branch and issue each released report atomically.

No response-driven fallback, retry, router portfolio, scale search, relabeling, coordinate request, or post-release filtering is permitted. Evaluating a map frozen for every possible s is not budget recycling.

Let $\mathcal{F}_{\mathrm{freezeA}}$ be generated by the router, X_R, stable IDs, and the frozen audit algorithm. Every response-free audit object is $\mathcal{F}_{\mathrm{freezeA}}$-measurable. Let $\mathcal{F}_0$ augment $\mathcal{F}_X$ with all discovery information and realized frozen audit objects, but exclude (Y_R, S_C, S_I). By the cross-role conditional-independence premise in Appendix A.2, conditioning further on this discovery information and these response-free objects does not alter the Gaussian response or pure-error laws. All core audit results condition on $\mathcal{F}_0$.

Audit cell count may be selected from the finite set $\mathcal{J}$ by a deterministic function of the frozen router and X_R. This preserves the conditional laws but does not make the resulting resolution scientifically informative; the candidate set and selection rule must be reported. Scientific margins and global budgets are protocol inputs. They may be common constants or a deterministic slot map fixed before discovery responses, but they may not be tuned from discovery outcomes. They are signal-to-noise tolerances: with unknown σ, an observed scale may evaluate the frozen test but may not retroactively convert an absolute outcome-unit tolerance into $\tau_{u,g}$. At A1 the protocol freezes the label cap and naming rule; at A2 the serialized router enumerates the realized finite namespace. Let $\mathcal{G}_{\mathrm{slot}}$ include every such label, including labels receiving no audit row, and put $M_{\mathrm{slot}} = \max\{1, |\mathcal{G}_{\mathrm{slot}}|\}$.

The reference implementation exposes two predeclared bundles. Its default static mode assigns $\alpha_{\mathrm{cert}}/(2M_{\mathrm{slot}})$ to every H/A opportunity and fixes interval allocations over all slot–coordinate opportunities. Its optional adaptive mode uses intersection–union group p-values with Holm over the

same complete slot family; after s is frozen and before S_I is opened, a predeclared rule divides γ_{FWER} over the requested coordinates in s. More general deterministic allocations are covered if their constraints are verified for every possible reporting pattern.

A.4 Routed groups, audit cells, and eligibility

Apply the serialized router to X_R. Rows routed to $\perp$ are never reported, and the hard-routed groups I_g are disjoint; set $N_g = |I_g|$. To construct cells, start with one leaf containing I_g. A leaf of size n is splittable exactly when $n \geq 2$. Repeatedly select the splittable leaf of largest size (tie: lexicographically smallest sorted ID tuple), select its retained covariate of largest range (tie: smallest coordinate index), sort by that coordinate and stable ID, and split into lower and upper halves of sizes $\lfloor n/2 \rfloor$ and $\lceil n/2 \rceil$. For each $J \in \mathcal{J}$, the current leaf set immediately after the tree first reaches J leaves defines the disjoint candidate cells $C_{g1}^{(J)}, \ldots, C_{gJ}^{(J)}$; their order is the deterministic leaf-list order inherited when a selected leaf is replaced by its lower child and then its upper child. Let $A_{gj}^{(J)}$ be the augmented design on candidate cell $C_{gj}^{(J)}$. A candidate is infeasible if the deterministic splitting tree terminates before reaching J leaves. A candidate is design-feasible when

$$\begin{aligned}
&\operatorname{rank}(A_g) = \operatorname{rank}(A_{gj}^{(J)}) = d, \quad r_{H,g}(J) = Jd - d > 0, \quad r_{A,g}(J) = N_g - Jd > 0, \\
&\operatorname{cond}_2(A_g^{\mathsf{T}} A_g) \leq \kappa_{\mathrm{pool}}, \quad \max_j \operatorname{cond}_2((A_{gj}^{(J)})^{\mathsf{T}} A_{gj}^{(J)}) \leq \kappa_{\mathrm{cell}}, \\
&v_{g\ell}^{x} = e_\ell^{\mathsf{T}} T_{xz} (A_g^{\mathsf{T}} A_g)^{-1} T_{xz}^{\mathsf{T}} e_\ell \leq v_{g\ell}^{\max} \quad \text{for every requested coordinate } \ell.
\end{aligned} \tag{18}$$

For each design-feasible J, form the projectors below and the exact prospective pass probabilities $\pi_{u,g}^{-}(J)$ from equation (24) in the unknown-noise branch or from the noncentral chi-square replacements in Corollary 16 when σ is known. Remove any candidate failing $\pi_{u,g}^{-}(J) \geq 1 - \beta_{u,g}$ on either axis, and among the remainder set

$$J_g \in \arg\min_J L_g(J), \qquad L_g(J) = \sum_{u \in \{H,A\}} \{1 - \pi_{u,g}^{-}(J)\}, \tag{19}$$

breaking an exact score tie in favor of the larger J. If none qualifies, delete the group before Y_R is opened. This complete construction and score are deterministic functions of frozen covariates, margins, allocations, and, in the unknown-noise branch only, scale degrees of freedom. Requested coordinates index the d-dimensional transformed vector (intercept plus retained slopes); discarded raw slopes cannot be requested. The largest-range split is a geometric heuristic and need not align with the location of coefficient changes; it is not asserted to be alternative-optimal. Response-free selection preserves the conditional audit laws, but J_g remains a scientific resolution choice: coarse cells may miss local changes, whereas fine cells spend rank and power. The score in equation (19) sums marginal pass-probability shortfalls; it is neither joint-release probability nor an alternative-detectability objective. A complete record must preserve the frozen cell construction, every candidate resolution and diagnostic, the selected J_g, and every deletion reason. Response-dependent resolution selection or multi-resolution certification requires a separately frozen opportunity family, selection rule, and error allocation.

Order rows cellwise and define

$$\begin{aligned}
P_g &= A_g (A_g^{\mathsf{T}} A_g)^{-1} A_g^{\mathsf{T}}, \\
P_{gj} &= A_{gj} (A_{gj}^{\mathsf{T}} A_{gj})^{-1} A_{gj}^{\mathsf{T}}, \\
P_{S,g} &= \operatorname{diag}\{P_{g1}, \ldots, P_{gJ_g}\}, \\
H_g &= P_{S,g} - P_g, \qquad W_g = I_{N_g} - P_{S,g}.
\end{aligned} \tag{20}$$

The finite-design pooled target is

$$\theta_g^{\circ} = (A_g^{\mathsf{T}} A_g)^{-1} A_g^{\mathsf{T}} \mu_g, \qquad \theta_{x,g}^{\circ} = T_{xz} \theta_g^{\circ}. \tag{21}$$

Define

$$\eta_{H,g} = \frac{\|H_g \mu_g\|_2}{\sigma \sqrt{N_g}}, \quad \eta_{A,g} = \frac{\|W_g \mu_g\|_2}{\sigma \sqrt{N_g}}, \tag{22}$$

and $\mathsf{Adeq}_g(\tau_{H,g}, \tau_{A,g}) = \{\eta_{H,g} < \tau_{H,g}, \eta_{A,g} < \tau_{A,g}\}$. The target exists whether or not this property holds; release licenses its interpretation as one adequate affine summary at the declared resolution.

A.5 AUDIT, RELEASE, AND ATOMIC REPORT

Let $Q_{H,g} = H_g$ denote the between-cell affine-projection-disagreement axis, $Q_{A,g} = W_g$ the within-cell non-affinity axis, and $r_{u,g} = \operatorname{rank}(Q_{u,g})$. Before Y_R, S_C, S_I, freeze $0 < \tau_{u,g}$, $0 \le \tau^-_{u,g} < \tau_{u,g}$, and allocations $\alpha_{u,g}, \beta_{u,g}$. Each $\tau_{u,g}$ is a noise-normalized aggregate RMS tolerance with absolute boundary $\sigma\tau_{u,g}$; an exact outcome-unit margin requires an exact σ frozen before responses. No discovery response or opened S_C, S_I may retune a margin. A failed gate may guide only a separately frozen future audit; no same-run repair is permitted. Also require a nonempty requested-coordinate set $\mathcal{L}_g$ for each auditable group, put $\mathcal{I}(s) = \{(g,\ell) : g \in s, \ell \in \mathcal{L}_g\}$, and freeze $\Gamma : s \mapsto \{\gamma_{g\ell}(s) : (g,\ell) \in \mathcal{I}(s)\}$ with $0 < \gamma_{g\ell}(s) < 1$ for nonempty s. For every possible complete statistical reporting pattern s, require

$$\sum_{(g,u)\in\mathcal{G}_{\text{slot}}\times\{H,A\}} \alpha_{u,g} \le \alpha_{\text{cert}}, \qquad \sum_{(g,\ell)\in\mathcal{I}(s)} \gamma_{g\ell}(s) \le \gamma_{\text{FWER}}. \tag{23}$$

$\tau^-_{u,g}$ is solely a planning and eligibility point; the tested null is always $\eta_{u,g} \ge \tau_{u,g}$. The default static allocation is $\alpha_{u,g} = \alpha_{\text{cert}}/(2M_{\text{slot}})$ for every slot–axis opportunity, including empty or later-ineligible slots. For $s = \varnothing$, Γ emits no interval allocation. Static preallocation is the constant-map special case and the implementation default. Another predeclared map may, for example, divide γ_{FWER} uniformly over the requested coordinates of groups in s; it is evaluated after release but before S_I and is never chosen from the coefficient estimates or interval widths. For unknown noise, freeze

$$c_{u,g} = F^{-1}_{r_{u,g},m_C,N_g\tau^2_{u,g}}(\alpha_{u,g}), \qquad \pi^-_{u,g} = F_{r_{u,g},m_C,N_g(\tau^-_{u,g})^2}(c_{u,g}). \tag{24}$$

The group is prospectively auditable only if $\pi^-_{u,g} \ge 1 - \beta_{u,g}$ on both axes. After opening (Y_R, S_C), compute

$$F^{\text{aud}}_{u,g} = \frac{\|Q_{u,g}Y_g\|_2^2/r_{u,g}}{S_C^2}. \tag{25}$$

Release g if and only if it is prospectively auditable and both strict comparisons $F^{\text{aud}}_{u,g} < c_{u,g}$ pass. A finite statistic above its cutoff is statistical non-release, and exact equality is recorded as a distinct boundary status. Nonfinite arithmetic or an unresolved numerical comparison is a numerical failure.

The following sharper intersection–union multiplicity mode is also admissible (Berger & Hsu, 1996). Define boundary p-values

$$p_{u,g} = F_{r_{u,g},m_C,N_g\tau^2_{u,g}}(F^{\text{aud}}_{u,g}), \qquad p_g = \max\{p_{H,g}, p_{A,g}\}. \tag{26}$$

Set $p_g = 1$ for every empty or ineligible slot. Under the group null that at least one axis lies at or beyond tolerance, p_g is super-uniform: the event $p_g \le t$ implies $p_{u,g} \le t$ for whichever axis null is true. Hence fixed group allocations summing to α_{cert} or Holm's step-down procedure over $\mathcal{G}_{\text{slot}}$ controls the probability of any bad group release under arbitrary dependence. Holm acts on the frozen candidate family and *defines* release; it is never applied to an already released set. For prospective screening, the uniform threshold $\alpha_{\text{cert}}/M_{\text{slot}}$ is a conservative sufficient floor because Holm contains every corresponding Bonferroni rejection. In that mode the prospective cutoff and inner pass probability use $\alpha_{\text{cert}}/M_{\text{slot}}$ in place of $\alpha_{u,g}$ in equation (24). The reference implementation provides both the default static bundle and this Holm bundle with realized-pattern interval allocation. Either release mode is residual-measurable and therefore shares the inference theorem below; the two bundles are not independently mixed in the reference implementation. Appendix E uses the pattern-uniform interval map in either mode; the high-dimensional study uses a fixed allocation.

For every eligible group define the potential estimator

$$\widehat{\theta}_g = (A_g^{\mathsf{T}}A_g)^{-1}A_g^{\mathsf{T}}Y_g, \qquad \widehat{\theta}_{x,g} = T_{xz}\widehat{\theta}_g. \tag{27}$$

It is computed and frozen only for a released group. In the unknown-noise branch, after the complete statistical reporting pattern and all released-group pooled fits are frozen, read S_I and report

$$\widehat{\theta}_{x,g\ell} \;\pm\; t_{m_I,1-\gamma_{g\ell}(s)/2}S_I\sqrt{v^x_{g\ell}} \tag{28}$$

for every pre-requested coordinate. Coefficient-bearing output is group-atomic: every requested coordinate is reported for a released group and no coefficient or interval is reported for an unreleased

group. Audit statistics and terminal status reasons remain logged for all attempted groups. No coefficient value, sign, interval width, or later display decision may filter release. The known-noise Gaussian interval is stated in Corollary 16.

A complete empirical record under the protocol must include a yield ledger: total, routed, and router-abstained audit rows; rows in prospectively auditable and released groups; and one exclusive status for every frozen slot (empty, response-free ineligible with reason, numerical failure, strict-decision equality, statistically non-released, or released). Pre-response statuses are fixed first; among attempted slots, numerical failure takes precedence over boundary and statistical classifications. Any threshold used to qualify a run on routed, router-abstained, or prospectively auditable audit rows must be frozen at A1 and checked before Y_R; released yield is post-response and descriptive. Such a threshold concerns only the frozen audit design. The core safety and conditional-coverage guarantees require no positive yield and imply neither population routing coverage nor unconditional availability.

Practitioner's recipe. First choose the adequacy requirement: one total-departure margin (pooled rule; no cells) or separate obligations for between-cell affine disagreement (H) and within-cell non-affinity (A). Then justify the margins. The prediction-risk identity in equation (F.3) interprets these choices; it does not set the scientific margins or remove coefficient-estimation error. With unknown noise, freeze noise-standardized margins or their deterministic slot map before discovery responses; an opened scale cannot convert an outcome-unit tolerance retrospectively. Define equal-mean replicate pairs at the scientific measurement unit whose noise variance governs the audit responses. Their Gaussian member errors must share that variance and be mutually independent across both banks and all response roles (Appendix A.2); repeated readings need not supply the scale for independent experimental units. Choose m_C by the exact response-free power calculation in equation (24) or Appendix A.11, using the planned group sizes, ranks, margins, inner points, and multiplicity levels. Choose m_I separately for planned interval precision under equation (28), accounting for design variance factors, Student critical values, and scale uncertainty; certification power does not size this bank. Freeze these choices and the design rules before responses. Neither bank has a universal 400-pair prescription. Without suitable pairs, an exact externally justified σ frozen before responses enables the known-noise branch; the theory also permits independent direct pure errors, whose input is unimplemented. If neither exact scale premise is available, the stated exact procedure is unavailable. Follow the staged freeze order in Appendix A.3 without response-driven retuning.

A.6 Gaussian tools for the core audit

Lemma 5 (Gaussian projection and whitening)**.** *If $Y = \mu + \varepsilon$, $\varepsilon \sim N(0, \sigma^2 I)$, and $Q_1, \dots, Q_L$ are pairwise orthogonal projectors, then the vectors $Q_\ell Y$ are mutually independent and*

$$\|Q_\ell Y\|_2^2/\sigma^2 \sim \chi^2_{\mathrm{rank}(Q_\ell)}(\|Q_\ell \mu\|_2^2/\sigma^2).$$

If $Z \sim N(\delta, \sigma^2 V)$ with $V \succ 0$, then $Z^\mathsf{T} V^{-1} Z/\sigma^2 \sim \chi^2_d(\delta^\mathsf{T} V^{-1}\delta/\sigma^2)$.

Proof. The cross-covariance of two projected vectors is $\sigma^2 Q_j Q_k = 0$; joint Gaussianity converts zero covariance to independence. Orthonormal coordinates for each projector range give the noncentral chi-square norm law. For the second claim, whiten with $V^{-1/2}$. □

Lemma 6 (Lower-tail monotonicity)**.** *For fixed degrees of freedom and cutoff, the noncentral-F and noncentral chi-square CDFs are nonincreasing in their noncentrality.*

Proof. A noncentral chi-square variable is a Poisson mixture of central chi-square variables with $r + 2K$ degrees of freedom, where $K \sim \mathrm{Poisson}(\lambda/2)$. Coupling Poisson variables shows stochastic increase in λ. For an F variable, condition on its independent positive central chi-square denominator and average the ordered numerator lower-tail probabilities. □

A.7 Orthogonal audit and conditional inference

Let $\mathcal{G}_{\mathrm{aud}} \subseteq \mathcal{G}_{\mathrm{slot}}$ be the $\mathcal{F}_0$-measurable set of nonempty groups that pass every response-free design and prospective-power gate.

Lemma 7 (Nested projector decomposition)**.** *For every eligible group, $P_g P_{S,g} = P_{S,g} P_g = P_g$. Thus P_g, H_g, W_g are pairwise orthogonal projectors summing to I_{N_g} with ranks $d, J_g d - d, N_g - J_g d$, respectively. Moreover,*

$$\frac{\|(I_{N_g} - P_g)\mu_g\|_2^2}{\sigma^2 N_g} = \eta_{H,g}^2 + \eta_{A,g}^2. \tag{29}$$

Proof. A pooled affine fitted vector restricts to an affine vector in every cell, so $\mathrm{col}(P_g) \subset \mathrm{col}(P_{S,g})$. Projectors onto nested subspaces satisfy the two product identities. Subtraction then shows that H_g and W_g are symmetric idempotents, every pairwise product vanishes, and the three sum to identity. Full cell and pooled ranks give the ranks. Finally $I_{N_g} - P_g = H_g + W_g$ and orthogonality give the squared-norm identity. □

Remark 8 (Rectangular versus total-residual adequacy). Let $\eta_T = (\eta_H^2 + \eta_A^2)^{1/2}$. Lemma 7 implies

$$\{\eta_T < \min(\tau_H, \tau_A)\} \subseteq \{\eta_H < \tau_H, \eta_A < \tau_A\} \subseteq \{\eta_T < (\tau_H^2 + \tau_A^2)^{1/2}\}.$$

Both reverse inclusions fail whenever both margins are positive: points near the rectangle's diagonal corner lie outside the inner disk, while points near either coordinate axis can lie inside the outer disk and violate that axis's margin. Hence no threshold depending only on the Euclidean total deviation η_T can encode both declared obligations without either extra non-release or false licensing relative to the two-axis property. This does not preclude a scalar encoding of the rectangle itself, such as a maximum of normalized axes. Because η_T is cell-rule invariant, common margins give a deterministic sensitivity bracket over any frozen rule family: the inner disk is adequate for every rule, the exterior of the outer disk is inadequate for every rule, and only the intervening annulus can change status. This is not a stability bound between cell rules.

Proof of Lemma 1. For every auditable group,

$$\widehat{\theta}_g - \theta_g^\circ = (A_g^\mathsf{T} A_g)^{-1} A_g^\mathsf{T} \varepsilon_g. \tag{30}$$

The coefficient error is a linear function of $P_g \varepsilon_g$ because $A_g^\mathsf{T}(I_{N_g} - P_g) = 0$. Orthogonality of P_g and $I_{N_g} - P_g$ makes it independent of $(I_{N_g} - P_g)Y_g$ under joint Gaussianity. For the two-axis construction, this residual generates the same sigma-field as $(H_g Y_g, W_g Y_g)$: the two components sum to the residual and each is a fixed projection of it. The general inference argument therefore needs full-rank group designs, not a cellwise full-rank decomposition. Hard-routed groups use disjoint rows, and the scale banks are independent, so the independence holds jointly across groups and includes S_C, S_I. Every allowed release rule is therefore independent of the displayed errors and S_I. Conditioning on one of its positive-probability reporting-pattern fibers leaves their joint law unchanged; dividing each Gaussian coordinate by the independent S_I/σ gives the Student statement. In the known-noise branch, omitting both scales gives the corresponding Gaussian pivots. □

Theorem 9 (Exact unknown-noise audit law)**.** *Conditional on $\mathcal{F}_0$,*

$$F_{u,g}^{\mathrm{aud}} \sim F_{r_{u,g}, m_C, N_g \eta_{u,g}^2}, \qquad u \in \{H, A\}. \tag{31}$$

Proof. Lemma 5 applied to H_g, W_g gives the noncentral chi-square numerators. Lemma 4 gives the independent common denominator, yielding each marginal noncentral-F law. No joint independence of the H/A statistics is asserted or needed. □

Theorem 10 (Static-rule false-certification control)**.** *For any realized discovery algorithm and router obeying the freeze contract, let $\widehat{\mathcal{S}}_{\mathrm{static}}$ be the release set defined by the fixed-axis cutoffs and static allocations in equation* (24)*. Then*

$$\mathbb{P}\{\exists g \in \widehat{\mathcal{S}}_{\mathrm{static}},\ u \in \{H, A\} : \eta_{u,g} \geq \tau_{u,g} \mid \mathcal{F}_0\} \leq \sum_{(g,u) \in \mathcal{G}_{\mathrm{slot}} \times \{H,A\}} \alpha_{u,g} \leq \alpha_{\mathrm{cert}}. \tag{32}$$

No graph-recovery, routing-correctness, or latent-purity premise is used.

Proof. At a true deviation at or beyond its outer margin, Lemma 6 and equation (24) give $\mathbb{P}(F_{u,g}^{\mathrm{aud}} < c_{u,g} \mid \mathcal{F}_0) \leq \alpha_{u,g}$. A statically released group violating a margin must contain such a false axis pass. Union over the complete preallocated opportunity set and use equation (23). □

Corollary 11 (Groupwise intersection–union and Holm control)**.** *For the p-values in equation* (26)*, either fixed group allocations summing to α_{cert} or Holm's step-down procedure over $\mathcal{G}_{\mathrm{slot}}$ controls*

$$\mathbb{P}\{\exists g \in \widehat{\mathcal{S}} :\ \eta_{H,g} \geq \tau_{H,g} \text{ or } \eta_{A,g} \geq \tau_{A,g} \mid \mathcal{F}_0\} \leq \alpha_{\mathrm{cert}}. \tag{33}$$

The release set remains measurable with respect to $\mathscr{R}$.

Proof. If group g is bad, at least one component null is true. Boundary monotonicity makes that component p-value super-uniform, and $\{p_g \leq t\}$ is a subset of the event that this true-null p-value is at most t. Thus p_g is super-uniform under the group null. Fixed-allocation control follows by a union bound; Holm gives strong FWER control for arbitrary dependence. Every p-value is a function only of H/A statistics and S_C, so the final statement follows. □

Remark 12 (Fixed-audit static–Holm comparison). Hold the slot family, selected cells, projectors, margins, planning points, power floors, and resulting component p-values fixed. Default-static release implies $p_g < \alpha_{\mathrm{cert}}/(2M_{\mathrm{slot}})$, hence groupwise Bonferroni at $\alpha_{\mathrm{cert}}/M_{\mathrm{slot}}$ and therefore Holm also releases g. Thus the default-static release set is pointwise contained in the IU/Holm set for this common audit object. This does not compare separately instantiated bundles whose mode-specific prospective selector chooses different cells, and no strictly positive additional Holm yield is distribution-free. The improvement from $\alpha_{\mathrm{cert}}/(2M_{\mathrm{slot}})$ to $\alpha_{\mathrm{cert}}/M_{\mathrm{slot}}$ comes already from forming one valid groupwise intersection–union p-value; later Holm thresholds may improve realized yield.

Lemma 13 (Independent-scale Student-t pivots)**.** *For each pre-requested raw coordinate,*

$$\frac{\widehat{\theta}_{x,g\ell} - \theta^{\circ}_{x,g\ell}}{S_I \sqrt{v^{x}_{g\ell}}} \sim t_{m_I}. \tag{34}$$

Conditional on $\mathcal{F}_0$, the fixed-index pivot collection is jointly independent of the complete statistical reporting pattern. The pivots need not be mutually independent: within-group Gaussian numerators can be correlated, and all pivots share S_I.

Proof. Equation (30) and the full transformation T_{xz} make the numerator, divided by $\sigma\sqrt{v^{x}_{g\ell}}$, standard normal. Lemma 4 supplies an independent $S_I/\sigma = \sqrt{\chi^2_{m_I}/m_I}$ denominator. Their ratio has a Student-t_{m_I} distribution. Lemma 1 gives joint independence from release. □

Theorem 14 (Release-pattern-conditional interval FWER)**.** *For every complete statistical reporting pattern s with positive conditional probability,*

$$\mathbb{P}\{\exists g \in s,\ \ell \in \mathcal{L}_g : \theta^{\circ}_{x,g\ell} \notin \mathrm{CI}_{g\ell}(s) \mid \widehat{\mathcal{S}} = s, \mathcal{F}_0\} \leq \sum_{(g,\ell)\in\mathcal{I}(s)} \gamma_{g\ell}(s) \leq \gamma_{\mathrm{FWER}}. \tag{35}$$

Proof. Conditioning on $\widehat{\mathcal{S}} = s$ leaves each marginal pivot in Lemma 13 with its t_{m_I} law. Each requested interval therefore misses with probability $\gamma_{g\ell}(s)$. A union bound proves the claim; dependence among pivots is irrelevant. □

Theorem 15 (Audit power)**.** *In static mode, for a prospectively auditable group with $\eta_{u,g} \leq \tau^{-}_{u,g}$, axis u passes its fixed cutoff with conditional probability at least $\pi^{-}_{u,g} \geq 1 - \beta_{u,g}$. In Holm mode, the same statement holds for the conservative $\alpha_{\mathrm{cert}}/M_{\mathrm{slot}}$ prospective cutoff; if both axes of a group pass that cutoff, Holm releases it. Hence, in either declared mode, if a nonempty $\mathcal{F}_0$-measurable set S_0 of prospectively auditable groups lies at both inner margins,*

$$\mathbb{P}(S_0 \subseteq \widehat{\mathcal{S}} \mid \mathcal{F}_0) \geq 1 - \sum_{\substack{g\in S_0 \\ u\in\{H,A\}}} \beta_{u,g}. \tag{36}$$

This lower bound is positive whenever the displayed sum is below one.

Proof. Lower-tail monotonicity gives each stated pass probability. In static mode, failure to release some group in S_0 requires failure of one fixed axis gate. In Holm mode, every group whose two component p-values are strictly below $\alpha_{\mathrm{cert}}/M_{\mathrm{slot}}$ has $p_g < \alpha_{\mathrm{cert}}/M_{\mathrm{slot}}$ and therefore belongs to the Holm rejection set. Failure of this sufficient event again requires an axis failure. A union bound over those failures proves the result; the axis-failure events may remain dependent through their shared S_C. □

A.8 KNOWN-NOISE SPECIALIZATION

Corollary 16 (Exact known-noise procedure and guarantees)**.** *Suppose the exact value $\sigma > 0$ is frozen at stage A1 before any response and omit $D_{\sigma C}, D_{\sigma I}$. In the generic audit, replace the unknown-noise objects by*

$$\begin{aligned} T_{u,g}^{\mathrm{aud}} &= \|Q_{u,g}Y_g\|_2^2/\sigma^2, \\ c_{u,g} &= G^{-1}_{r_{u,g},N_g\tau^2_{u,g}}(\alpha_{u,g}), & \pi^-_{u,g} &= G_{r_{u,g},N_g(\tau^-_{u,g})^2}(c_{u,g}), \\ p_{u,g} &= G_{r_{u,g},N_g\tau^2_{u,g}}(T^{\mathrm{aud}}_{u,g}), & p_g &= \max\{p_{H,g}, p_{A,g}\}, \\ \mathrm{CI}_{g\ell}(s) &= \widehat{\theta}_{x,g\ell} \pm z_{1-\gamma_{g\ell}(s)/2}\sigma\sqrt{v^x_{g\ell}}. \end{aligned}$$

Use the same strict decisions and prospective gates. Conditional on $\mathcal{F}_0$,

$$T^{\mathrm{aud}}_{u,g} \sim \chi^2_{r_{u,g}}(N_g\eta^2_{u,g}), \qquad \frac{\widehat{\theta}_{x,g\ell} - \theta^\circ_{x,g\ell}}{\sigma\sqrt{v^x_{g\ell}}} \sim N(0,1),$$

and the Gaussian pivot is independent of the entire residual-measurable complete statistical reporting pattern. Consequently the corresponding known-noise conclusions of Lemma 1, Theorems 9 and 10, Corollary 11, and Theorems 14 and 15 hold under the displayed substitutions and the same allocations.

Proof. Lemma 5 gives the displayed audit law, and Lemma 6 gives the same boundary and power arguments without a random denominator. Projector orthogonality is unchanged, while

$$\frac{\widehat{\theta}_{x,g\ell} - \theta^\circ_{x,g\ell}}{\sigma\sqrt{v^x_{g\ell}}} \sim N(0,1)$$

is independent of the residual-measurable reporting pattern. The same union-bound argument and reporting-pattern conditioning therefore give the stated Gaussian coverage and power conclusions. □

Proof of Corollary 2. Equation (31) is Theorem 9. The false-certification statement is Theorem 10 in static mode and Corollary 11 in Holm mode; conditional FWER is Theorem 14. The known-noise analogues follow from Corollary 16. Their premises depend only on the frozen router and audit design, not on discovery correctness. □

Proof of Corollary 3. Average equation (35) over reporting patterns, combine with equation (32) in static mode or equation (33) in Holm mode, and union bound. The final deviation claim uses equation (29). The known-noise conclusion follows by substituting the corresponding false-certification and coverage statements from Corollary 16; equation (29) is unchanged. □

A.9 AUDIT-ROW REUSE VERSUS A SECOND AUDIT SPLIT

Proposition 17 (Exact-affine covariance cost of a second split)**.** *Partition a fixed full-rank design into certification and inference rows, $A = (A_C^\mathsf{T}, A_I^\mathsf{T})^\mathsf{T}$, with $A_I^\mathsf{T}A_I \succ 0$, and suppose $Y = A\theta + \varepsilon$, $\varepsilon \sim N(0, \sigma^2 I)$. Any conventional release decision measurable with respect to Y_C and independent auxiliary randomness is independent of $\widehat{\theta}_I = (A_I^\mathsf{T}A_I)^{-1}A_I^\mathsf{T}Y_I$. If, in addition, the release decision uses Y_C only through Q_CY_C, where Q_C is frozen before responses and $Q_CA_C = 0$, then*

$$\widehat{\theta}_{\mathrm{full}} = (A^\mathsf{T}A)^{-1}A^\mathsf{T}Y \tag{37}$$

is also independent of release. Conditional on every release outcome of positive probability,

$$\mathrm{Cov}(\widehat{\theta}_{\mathrm{full}}) = \sigma^2(A^\mathsf{T}A)^{-1} \preceq \sigma^2(A_I^\mathsf{T}A_I)^{-1} = \mathrm{Cov}(\widehat{\theta}_I). \tag{38}$$

The inequality is strict in contrast direction a exactly when $A_C^\mathsf{T}A_C(A_I^\mathsf{T}A_I)^{-1}a \neq 0$. If $A_I^\mathsf{T}A_I = qA^\mathsf{T}A$, every contrast variance is inflated by q^{-1} and every known-noise standard error by $q^{-1/2}$.

Proof. The split estimator is independent of every Y_C-measurable release rule by row independence. Under the additional residual-only premise,

$$\mathrm{Cov}(\widehat{\theta}_{\mathrm{full}} - \theta, Q_C Y_C) = \sigma^2 (A^\mathsf{T} A)^{-1} A_C^\mathsf{T} Q_C^\mathsf{T} = 0,$$

so joint Gaussianity makes the full estimator independent of release. Because $A^\mathsf{T} A = A_I^\mathsf{T} A_I + A_C^\mathsf{T} A_C \succeq A_I^\mathsf{T} A_I$, inversion reverses the Loewner order. The strictness characterization follows from the null space of the covariance difference; the proportional formula is immediate. □

Remark 18 (Why exact affinity is required). Under arbitrary μ, inference-split OLS targets $(A_I^\mathsf{T} A_I)^{-1} A_I^\mathsf{T} \mu_I$, which need not equal the full-design target $(A^\mathsf{T} A)^{-1} A^\mathsf{T} \mu$. Proposition 17 is therefore a same-target comparison only under exact affinity, or another assumption equating the two projections. It is not a blanket conditional-UMVU claim; the familiar $\sqrt{2}$ factor applies only to a proportional half- information split.

Corollary 19 (Idealized joint cost of a second audit split)**.** *Consider a known-noise sequence with the same fixed multiplicity and power targets. Let a conventional split assign fractions $q_C \in (0,1)$ and $q_I = 1 - q_C$ of the routed rows to certification and inference, while reuse uses all N rows for both. Suppose the exact-affine target is common, $A_I^\mathsf{T} A_I = q_I A^\mathsf{T} A$, and the A-axis ranks satisfy $r_A(n)/n \to \kappa_A \in (0,1)$. For each compared coordinate, use the same interval allocation and hence the same known-noise critical value under both designs; if the allocation is release-pattern-adaptive, compare the same complete reporting pattern. Then*

$$\frac{\textit{split interval half-width}}{\textit{reuse interval half-width}} = q_I^{-1/2}.$$

For the A-axis gate, the sufficient growing-rank tolerance order is $n^{-1/4}$, evaluated at $n = q_C N$ under splitting and at $n = N$ under reuse. Thus an equal split inflates interval half-width exactly by $\sqrt{2}$ and incurs fourth-root certification-row scaling up to constants. The rate statement alone does not identify an exact $2^{1/4}$ tolerance multiplier.

Proof. The interval identity follows from $(A_I^\mathsf{T} A_I)^{-1} = q_I^{-1} (A^\mathsf{T} A)^{-1}$. At fixed multiplicity and power levels, $r_A(n)/n \to \kappa_A$ gives $r_A(n)^{1/4}/\sqrt{n} \asymp n^{-1/4}$. Evaluating this rate at $q_C N$ and N gives the stated fourth-root row-fraction scaling; an exact multiplier would require a shared canonical envelope and a ratio limit. □

The synthesis is deliberately idealized. It requires the same target, proportional Gram information, coordinatewise interval allocation and critical value, testing level, multiplicity, power target, and scale law. A fixed-rank axis instead has an $n^{-1/2}$ envelope and pays $q_C^{-1/2}$; unknown-scale degrees of freedom can add further cost. This is not a universal efficiency theorem. On an independent inference half, residual Student intervals conservatively cover that half's own projection under an arbitrary fixed mean. With residual degrees of freedom $\nu_I > 0$, the normalized coefficient error is standard normal and independent of $\nu_I s_I^2/\sigma^2 \sim \chi^2_{\nu_I}(\lambda_I)$, where $\lambda_I = \|(I - P_I)\mu_I\|_2^2/\sigma^2$. Noncentral chi-square stochastic dominance and monotonicity of two-sided normal coverage in the scale imply conservatism; the central Student law is exact at $\lambda_I = 0$. This also holds conditional on a complete release decision independent of the entire inference-half data. It need not cover the full-audit projection. Thus splitting need not acquire an inference bank. Its certification scale is a separate issue: a noncentral residual denominator does not give the gate's calibrated noncentral-F law with a central denominator. CertifyPAR reserves S_I because reused audit residuals participate in release. The displayed factors do not establish fixed-total-budget dominance.

A.10 Rank-dependent adequacy frontier

Theorem 20 (Known-noise equivalence frontier)**.** *Fix one frozen group-axis opportunity, suppress its indices, and let Q be its rank-r orthogonal projector. With known σ, choose an orthonormal basis $U \in \mathbb{R}^{N \times r}$ for $\mathrm{col}(Q)$ and write*

$$Z = U^\mathsf{T} Y/\sigma \sim N_r(\nu, I_r), \qquad \|\nu\|_2^2 = N\eta^2.$$

For a declared outer margin $\tau > 0$, set $b = N\tau^2$ and

$$c_{r,b}(\alpha) = G_{r,b}^{-1}(\alpha), \qquad \phi^*(Z) = \mathbf{1}\{\|Z\|_2^2 < c_{r,b}(\alpha)\},$$

where $\alpha \in (0,1)$. Statements 1–3 take α fixed; statement 4 allows it to vary. Then:

1. *ϕ^* is level α for the composite inadequacy null $\|\nu\|_2^2 \geq b$. If $b/\sqrt{r} \to \infty$, its power at exact adequacy satisfies $\mathbb{E}_0\phi^* \to 1$.*

2. *Every possibly randomized certificate ϕ satisfying $\sup_{\|\nu\|_2^2 \geq b} \mathbb{E}_\nu\phi \leq \alpha$ obeys the finite bound*

$$\mathbb{E}_0\phi \leq \alpha + \frac{1}{2}\{\cosh(b/r)^r - 1\}^{1/2} \leq \alpha + \frac{1}{2}\{\exp(b^2/(2r)) - 1\}^{1/2}. \tag{39}$$

   Hence if $b/\sqrt{r} \to 0$, no uniformly level-α certificate has exact-adequacy power exceeding $\alpha + o(1)$.

3. *For an inner margin $b^- = N(\tau^-)^2 < b$, if $(b - b^-)/\sqrt{r + b} \to \infty$, then $\inf_{\|\nu\|_2^2 \leq b^-} \mathbb{E}_\nu\phi^* \to 1$.*

4. *For every possibly shrinking α, if*

$$b^2 \geq 16(r + 2b)\log(1/\alpha), \tag{40}$$

   then $\mathbb{E}_0\phi^* \geq 1 - 8r/b^2$.

Proof. Lower-tail monotonicity proves level control. Let $X_b \sim \chi^2_r(b)$, whose mean and variance are $r + b$ and $2(r + 2b)$. When $b/\sqrt{r} \to \infty$, Chebyshev's inequality gives

$$\mathbb{P}\{X_b \leq r + b/2\} \leq 8(r + 2b)/b^2 = o(1),$$

so $c_{r,b}(\alpha) > r + b/2$ eventually. For $X_0 \sim \chi^2_r$, the probability above that cutoff is at most $8r/b^2 = o(1)$.

For the lower bound, let ϵ be uniform on $\{-1, 1\}^r$ and mix the boundary alternatives $\nu_\epsilon = \sqrt{b/r}\,\epsilon$. If $\bar{P}$ is this mixture and $P_0 = N_r(0, I_r)$, its likelihood ratio L satisfies

$$1 + \chi^2(\bar{P}, P_0) = \mathbb{E}_0 L^2 = \mathbb{E}_{\epsilon,\epsilon'} \exp(\nu_\epsilon^\mathsf{T}\nu_{\epsilon'}) = \cosh(b/r)^r.$$

Every mixture component is in the null, so $\mathbb{E}_{\bar{P}}\phi \leq \alpha$. Total variation and $\log\cosh x \leq x^2/2$ yield equation (39). Coefficient-space components, other groups, and independent scales may be given identical laws under the mixture and exact adequacy; tensoring with common distributions leaves total variation unchanged. Thus the lower bound also applies when the full protocol data are observed. For the inner margin, put $x = r + (b + b^-)/2$. Chebyshev bounds the boundary probability below x and the inner-boundary probability above x by $8(r+2b)/(b-b^-)^2$ and $8(r+2b^-)/(b-b^-)^2$, respectively. Monotonicity completes the uniform claim.

Finally, the noncentral chi-square moment-generating function and a Chernoff optimization give, for every $t > 0$,

$$\mathbb{P}\{X_b \leq r + b - 2\sqrt{(r + 2b)t}\} \leq e^{-t}.$$

With $t = \log(1/\alpha)$, equation (40) places the α-quantile at or above $r + b/2$. Chebyshev's inequality for X_0 then gives $\mathbb{P}(X_0 \geq r + b/2) \leq 8r/b^2$, proving part 4. □

Corollary 21 (Explicit finite multiplicity envelope)**.** *For one frozen known-noise axis, fix $r \geq 1$ and $a, \beta \in (0, 1)$, and put $L = \log(1/a)$ and*

$$\mathsf{B}(r, L) = 16L + 4\sqrt{L(r + 16L)}. \tag{41}$$

For $b = N\tau^2 \geq 0$, the sufficient condition $b^2 \geq 16(r + 2b)L$ is equivalent to $b \geq \mathsf{B}(r, L)$. Therefore $b \geq \max\{\mathsf{B}(r, L), \sqrt{8r/\beta}\}$ guarantees that the bound in Theorem 20, part 4, is at least $1 - \beta$. Moreover,

$$\begin{aligned} \mathsf{B}(r, L) &= 4\sqrt{rL} + 16L + O(L^{3/2}/\sqrt{r}), && L = o(r), \\ &= 32L + r/2 + O(r^2/L), && L \gg r. \end{aligned} \tag{42}$$

Thus the multiplicity-envelope contribution $\{\mathsf{B}(r, L)/N\}^{1/2}$ is $2(rL)^{1/4}/\sqrt{N}\{1 + o(1)\}$ in the first regime and $4\sqrt{2L/N}\{1 + o(1)\}$ in the second, while the full displayed $1 - \beta$ sufficient tolerance is

$$\frac{1}{\sqrt{N}} \max\Big\{\sqrt{\mathsf{B}(r, L)}, (8r/\beta)^{1/4}\Big\}.$$

Default static and conservative IU/Holm planning substitute respectively $L = \log(2M_{\mathrm{slot}}/\alpha_{\mathrm{cert}})$ and $L = \log(M_{\mathrm{slot}}/\alpha_{\mathrm{cert}})$. Since $\mathsf{B}(r, L)$ is strictly increasing in L, the latter multiplicity envelope is smaller; the full fixed-power envelope is no larger and may coincide when its shared β term dominates. This comparison does not characterize realized step-down yield.

Proof. The inequality is $b^2 - 32Lb - 16rL \geq 0$, whose roots are $16L \pm 4\sqrt{L(r+16L)}$; the smaller is negative. The power-floor condition also requires $8r/b^2 \leq \beta$. Expanding the square root in equation (41) under $L/r \to 0$ and $r/L \to 0$ gives equation (42). Differentiation, or monotonicity of the defining quadratic condition, gives strict increase in L. □

Proposition 22 (Known-noise frontier specialization and multiplicity cost)**.** *Consider a sequence of frozen known-noise axes. After orthonormalization, write $Z_N \sim N_{r_N}(\nu_N, I_{r_N})$ and $b_N = N\tau_N^2$, so the inadequacy boundary is $\|\nu_N\|_2^2 = b_N$. At fixed level α, the lower-tail norm gate has power at $\nu_N = 0$ tending to one if $b_N/\sqrt{r_N} \to \infty$. If $b_N/\sqrt{r_N} \to 0$, every possibly randomized certificate ϕ_N satisfying $0 \leq \phi_N \leq 1$ and $\sup_{\|\nu\|_2^2 \geq b_N} \mathbb{E}_\nu \phi_N \leq \alpha$ obeys $\mathbb{E}_0 \phi_N \leq \alpha + o(1)$. For finite planning, let a_* be the conservative prospective axis level: $a_* = \alpha_{\mathrm{cert}}/(2M_{\mathrm{slot}})$ for static Bonferroni and $a_* = \alpha_{\mathrm{cert}}/M_{\mathrm{slot}}$ for the first IU/Holm step. With $b = N\tau^2$ and axis rank r, the condition*

$$b^2 \geq 16(r+2b)\log(1/a_*) \tag{43}$$

guarantees exact-adequacy pass probability at least $1 - 8r/b^2$ at that level.

Proof. This is Theorem 20, parts 1, 2, and 4, with $\alpha = a_*$ in the finite sufficient statement, together with Corollary 21. □

Proposition 23 (Operational inner-margin power)**.** *For one known-noise axis in a frozen slot family, let a_* be its conservative prospective level: $a_* = \alpha_{\mathrm{cert}}/(2M_{\mathrm{slot}})$ in static mode and $a_* = \alpha_{\mathrm{cert}}/M_{\mathrm{slot}}$ in Holm mode. Put*

$$L = \log\frac{1}{a_*}, \quad K = \log\frac{1}{\beta},$$

where $\beta \in (0,1)$ is the desired inner-margin failure probability. Let $b = N\tau^2$, $b^- = N(\tau^-)^2 < b$, and $\Delta_b = b - b^-$. If

$$\Delta_b \geq 2\sqrt{(r+2b)L} + 2\sqrt{(r+2b^-)K} + 2K, \tag{44}$$

then the boundary-calibrated prospective gate at level a_* passes with probability at least $1-\beta$, uniformly over noncentralities at most b^-.

Proof. For $X_\lambda \sim \chi_r^2(\lambda)$, the noncentral chi-square Chernoff bounds give

$$\mathbb{P}\{X_\lambda \leq r + \lambda - 2\sqrt{(r+2\lambda)t}\} \leq e^{-t},$$
$$\mathbb{P}\{X_\lambda \geq r + \lambda + 2\sqrt{(r+2\lambda)t} + 2t\} \leq e^{-t}.$$

The first inequality lower-bounds the boundary quantile by $r + b - 2\sqrt{(r+2b)L}$. Condition (44) places this value above the upper $1-\beta$ bound at b^-. The second inequality and noncentrality monotonicity complete the proof. □

Proposition 24 (Sufficient unknown-scale power regime)**.** *Fix $\alpha \in (0,1)$ and consider a sequence with $N, r, m \in \{1, 2, \ldots\}$, allowing r and m to vary with N. Let $V \sim \chi_m^2/m$ be independent and use the boundary-calibrated noncentral-F gate with $b^- = N(\tau^-)^2 < b = N\tau^2$. If*

$$\frac{(b-b^-)^2}{r+b} \to \infty, \qquad \frac{m(b-b^-)^2}{(r+b)^2} \to \infty, \tag{45}$$

then pass probability tends to one uniformly over noncentralities at most b^-. The known-noise impossibility bound remains valid for every m.

Proof. Let $x = r + (b+b^-)/2$ and $h = (b-b^-)/\{4(r+b)\}$. At the outer boundary, on $V \leq 1+h$, the event $X_b/(rV) \leq x/r$ implies $X_b \leq r + b - (b-b^-)/4$. At the inner boundary, on $V \geq 1-h$, the reverse event implies $X_{b^-} \geq r + b^- + (b-b^-)/4$. Chebyshev's inequality and $\mathrm{Var}(V) = 2/m$ bound each relevant tail by

$$\frac{32(r+2b)}{(b-b^-)^2} + \frac{32(r+b)^2}{m(b-b^-)^2},$$

up to replacing b by b^- in the first term. Equation (45) makes these bounds vanish, and noncentrality monotonicity gives uniformity. The lower-bound mixture may be augmented by the independent V, whose law is identical under the boundary and exact adequacy, so it cannot increase total variation. □

Remark 25 (Scope of the unknown-scale result). Proposition 24 is a fixed-level sufficient result only. It is not a necessary condition, a matching lower bound, or a sharp unknown-scale minimax frontier. The known-noise Rademacher lower bound remains valid after adjoining an independent scale statistic with the same law under both hypotheses, but does not quantify the additional random-denominator difficulty. We therefore claim neither that equation (45) is necessary nor that its constants or dependence on m are optimal.

Remark 26 (Rate interpretation). Beyond the rate summary in Section 5, nonzero inner margins are governed by the squared-margin gap $N\{\tau^2 - (\tau^-)^2\}$. Proposition 23 shows explicitly how the power failure target contributes $\log(1/\beta)$; equation (40) is sufficient, not claimed sharp.

A.11 Pooled total-departure certification

Corollary 27 (Pooled certification and conditional coefficient inference)**.** *Under the Gaussian response and independent-scale premises, let $\mathcal{G}_{T,\mathrm{aud}} \subseteq \mathcal{G}_{\mathrm{slot}}$ be a response-free eligible roster of disjoint groups with $\mathrm{rank}(A_g) = d$ and $N_g > d$. Freeze its design and prospective-power screens, positive margins $\tau_{T,g}$, coordinate requests, and test and interval-allocation maps under the response-order contract. No audit-cell decomposition is required. Put*

$$R_g = I_{N_g} - P_g, \qquad \nu_g = N_g - d, \qquad \eta_{T,g} = \frac{\|R_g \mu_g\|_2}{\sigma\sqrt{N_g}}.$$

For $g \in \mathcal{G}_{T,\mathrm{aud}}$, define the boundary p-value

$$p_{T,g} = \begin{cases} G_{\nu_g, N_g \tau_{T,g}^2}\big(\|R_g Y_g\|_2^2/\sigma^2\big), & \sigma \text{ known}, \\ F_{\nu_g, m_C, N_g \tau_{T,g}^2}\big(\|R_g Y_g\|_2^2/(\nu_g S_C^2)\big), & \sigma \text{ unknown}. \end{cases}$$

Set $p_{T,g} = 1$ for empty or ineligible slots. With positive static group levels $\alpha_{T,g}$ summing to at most α_{cert} over all potential slots, report exactly when $p_{T,g} < \alpha_{T,g}$; alternatively use strict Holm step-down at level α_{cert}. Writing $\widehat{\mathcal{S}}_T$ for the resulting complete reporting pattern,

$$\mathbb{P}\{\exists g \in \widehat{\mathcal{S}}_T : \eta_{T,g} \geq \tau_{T,g} \mid \mathcal{F}_0\} \leq \alpha_{\mathrm{cert}}.$$

With the same Gaussian or independent-scale Student intervals and a predeclared map satisfying $\sum_{g \in s, \ell \in \mathcal{L}_g} \gamma_{g\ell}(s) \leq \gamma_{\mathrm{FWER}}$, equation (35) *holds with $\widehat{\mathcal{S}}_T = s$ for every positive-probability complete pattern. Consequently, conditional on $\mathcal{F}_0$, the probability of either a reported group violating its total-departure margin or any reported interval missing its target is at most $\alpha_{\mathrm{cert}} + \gamma_{\mathrm{FWER}}$. These statements concern total-departure adequacy. If a two-axis decomposition is also fixed and $\tau_{T,g} \leq \min(\tau_{H,g}, \tau_{A,g})$, the certificate is additionally sufficient for that rectangular adequacy requirement.*

Proof. Lemma 5 gives $\|R_g Y_g\|_2^2/\sigma^2 \sim \chi^2_{\nu_g}(N_g \eta_{T,g}^2)$. Lemma 4 supplies the independent central denominator when needed. By Lemma 6, each boundary p-value is super-uniform under $\eta_{T,g} \geq \tau_{T,g}$. A union bound or Holm's argument gives false-certification control without requiring independence of p-values. The complete decision depends only on the pooled residuals and S_C, so Lemma 1 and the interval union bound give conditional coverage. Averaging the conditional noncoverage bound over complete patterns and taking a union with false certification gives the joint bound. Finally, Lemma 7 yields $\eta_{H,g}, \eta_{A,g} \leq \eta_{T,g}$. □

Remark 28 (Total-energy rules for a rectangular adequacy requirement). Condition on the frozen information $\mathcal{F}_0$, fix $\sigma > 0$, and consider one group with fixed full-rank design and cell decomposition, suppressing its index. Assume $r_H, r_A \geq 1$ and $\tau_H, \tau_A > 0$. Let a certificate, with independent randomization if used, read the audit responses only through $T = \|(I - P)Y\|_2^2$, together with the independent certification scale S_C when noise is unknown. Suppose its reporting probability is uniformly at most $\alpha \in (0, 1)$ over the rectangular inadequacy null $\{\mu : \eta_H \geq \tau_H \text{ or } \eta_A \geq \tau_A\}$. Then its reporting probability is at most α at every mean satisfying

$$\eta_T = (\eta_H^2 + \eta_A^2)^{1/2} \geq \min(\tau_H, \tau_A).$$

In particular, this bound holds for the adequate corner departures

$$\mathcal{C} = \{(h, a) : 0 \leq h < \tau_H,\ 0 \leq a < \tau_A,\ h^2 + a^2 \geq \min(\tau_H, \tau_A)^2\}.$$

In the known-noise design sequences of Proposition 34, with fixed positive static axis levels and fixed margins, the two-axis rule instead reports with probability tending to one at each fixed point of $\mathcal{C}$. For equal margins, $\mathcal{C}$ occupies the fraction $1 - \pi/4$ of the nonnegative adequacy square by area; this is not a probability or an empirical prevalence claim. The comparison concerns total-energy-only rules required to certify the same rectangle. A pooled rule certifying a different total-departure margin addresses a different adequacy requirement.

Proof. For fixed σ and frozen design, the law $T/\sigma^2 \sim \chi^2_{N-d}(N\eta_T^2)$ depends on the unrestricted audit mean only through η_T; the independent S_C law is unchanged. Choose an axis u attaining $\min(\tau_H, \tau_A)$ and a unit vector v_u in its projector range, and set $\widetilde{\mu} = P\mu + \sigma\sqrt{N}\,\eta_T v_u$. This mean has the same total-residual law and the same projected coefficient target, but its u-axis departure is $\eta_T \geq \tau_u$ and its other departure is zero. Uniform validity at $\widetilde{\mu}$ therefore bounds the original mean's reporting probability by α. The known-noise two-axis limit follows from Proposition 34, since each fixed point of $\mathcal{C}$ is strictly within both margins and has nonzero total departure. The area calculation subtracts a quarter disk of radius τ from a square of side τ when $\tau_H = \tau_A = \tau$. □

B Scope limits and coefficient interpretation

B.1 Aggregate rather than worst-cell adequacy

The normalization in equation (22) defines an average scientific target. If $\mu_g = A_g\theta + \delta_g$, where δ_g is nonzero on at most k_g rows and $\|\delta_g\|_\infty \leq C$, then projector contraction and Lemma 7 give

$$\eta_{H,g}, \eta_{A,g} \leq \frac{\|\delta_g\|_2}{\sigma\sqrt{N_g}} \leq \frac{C}{\sigma}\sqrt{\frac{k_g}{N_g}}.$$

Thus a bounded anomaly on $k_g = o(N_g)$ rows is diluted. Conversely, placing either axis outside its RMS margin requires at least $k_g \geq \sigma^2 N_g \tau_{u,g}^2/C^2$ perturbed rows, and alignment with that projector is still necessary. At the fixed-level frontier this lower bound is of order $\sigma^2\sqrt{r_{u,g}}/C^2$. Writing $L_* = \log(1/a_*)$, the multiplicity-envelope contribution is of order $\sigma^2\sqrt{r_{u,g}L_*}/C^2$ when $L_* = o(r_{u,g})$ and $\sigma^2 L_*/C^2$ when $L_* \gg r_{u,g}$. For fixed desired failure probability β, choosing the sufficient planning boundary $N_g\tau_{u,g}^2 = \max\{\mathsf{B}(r_{u,g}, L_*), \sqrt{8r_{u,g}/\beta}\}$ makes the corresponding necessary affected-row lower bound $(\sigma^2/C^2)\max\{\mathsf{B}(r_{u,g}, L_*), \sqrt{8r_{u,g}/\beta}\}$. This count alone is not sufficient to violate a margin: perturbation magnitude and projector alignment also matter. The procedure does not control a maximum over rows or cells; such a target needs separately calibrated cellwise tests and a new power analysis.

B.2 Why spherical noise is load-bearing

For a general Gaussian covariance Σ_g, Euclidean projector orthogonality does not imply $P_g\Sigma_g H_g = P_g\Sigma_g W_g = 0$. The quadratic-form laws and independence from release can therefore fail under heteroskedasticity. If $H_g\mu_g = W_g\mu_g = 0$, the actual joint quadratic release region is centered and symmetric, so a centered jointly Gaussian coefficient/residual pair has zero conditional coefficient mean; its conditional law and covariance still change, however, and the homoskedastic OLS variance formula is generally wrong. At a nonzero residual margin, the shifted region can additionally induce selection bias, as follows.

Proposition 29 (A heteroskedastic conditional-coverage counterexample)**.** *Take $N = 6$, $d = 2$, two cells of three rows, and*

$$A = \begin{bmatrix} 1 & -1 \\ 1 & 0 \\ 1 & 1 \\ 1 & -1 \\ 1 & 0 \\ 1 & 1 \end{bmatrix}, \qquad \Sigma = \mathrm{diag}(8/5, 8/5, 8/5, 2/5, 2/5, 2/5).$$

Let A_1 and A_2 be the first and last three rows of A, respectively, put $P = P_A$, $P_S = \mathrm{diag}(P_{A_1}, P_{A_2})$, $H = P_S - P$, and $W = I - P_S$. For orthonormal directions

$$u_P = \frac{(1,1,1,1,1,1)^\mathsf{T}}{\sqrt{6}}, \quad h_1 = \frac{(1,1,1,-1,-1,-1)^\mathsf{T}}{\sqrt{6}}, \quad h_2 = \frac{(-1,0,1,1,0,-1)^\mathsf{T}}{2},$$

$u_P \in \mathrm{col}(P)$ and (h_1, h_2) spans $\mathrm{col}(H)$. Let $\varepsilon \sim N(0, \Sigma)$ have independent heteroskedastic coordinates and $\mu = 2h_1$. Then, for $X = u_P^\mathsf{T}\varepsilon$ and $Z_j = h_j^\mathsf{T}\varepsilon$,

$$(X, Z_1, Z_2)^\mathsf{T} \sim N\left(0, \begin{bmatrix} 1 & 0.6 & 0 \\ 0.6 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}\right), \qquad \|HY\|_2^2 = (2 + Z_1)^2 + Z_2^2.$$

Because the covariate column is centered, X is the pooled-intercept error divided by its exact sandwich standard deviation. At boundary noncentrality 4, let $c = G_{2,4}^{-1}(0.05) \approx 0.6455990496$ and release on $R = \{\|HY\|_2^2 < c\}$. Although $\mathbb{P}(R) = 0.05$ and $\mathbb{P}\{|X| \le z_{0.975}\} = 0.95$, one-dimensional quadrature gives

$$\mathbb{P}\{|X| \le z_{0.975} \mid R\} \approx 0.86908546. \tag{46}$$

For this design, $W\Sigma P = W\Sigma H = 0$ and $W\mu = 0$. The A-axis statistic is therefore independent of (X, Z_1, Z_2), so conditioning additionally on any positive-probability known-noise A-axis pass event leaves the displayed conditional coverage unchanged and embeds the construction in the full two-axis release rule.

Proof. Direct multiplication gives the displayed covariance and covariance orthogonality with the W component. Conditional on $Z_1 = z$, $X \sim N(0.6z, 0.8^2)$, while integrating out Z_2 weights $z \in [-2 - \sqrt{c}, -2 + \sqrt{c}]$ by

$$w_c(z) = \phi(z)\{2\Phi(\sqrt{c - (2+z)^2}) - 1\}.$$

Dividing the integral of the conditional normal coverage probability against w_c by 0.05 yields equation (46). The same quadrature gives $\mathbb{E}(X \mid R) \approx -1.03212548$. These decimals are reproducible floating-point evaluations, not interval-enclosed certificates. □

For a fixed outcome-unit margin, lower-endpoint variance calibration need not remain uniform over a known bracket. In a scalar Gaussian illustration, let $Y \sim N(\mu, \sigma^2)$, $\sigma \in [\sigma_-, 2\sigma_-]$, and test $H_0 : |\mu| \ge \rho$ with $\rho = 2\sigma_-$. Choose $a > 0$ from $\Phi(a - 2) - \Phi(-a - 2) = 0.05$, so $a \approx 0.425190$, and release on $R_- = \{|Y| < a\sigma_-\}$. At $\sigma = \sigma_-$, release probability decreases with $|\mu|$, so the rule is level 0.05 over this null. At $\mu = \rho, \sigma = 2\sigma_-$, however, it equals approximately 0.102880. This example concerns a fixed outcome-unit null, not the protocol's noise-normalized null. For the latter, substituting a valid lower bound for σ in the known-noise statistic while retaining its noise-normalized boundary quantile is conservative by event inclusion. Neither observation rules out a separately calibrated bracket-aware procedure.

Exact audit-row reuse can extend when covariance shape is known and $P_g\Sigma_g H_g = P_g\Sigma_g W_g = 0$, with the quadratic forms and interval variances recalibrated; that is a different GLS contract and is not claimed.

An asymptotic relaxation would require a uniform joint approximation for the coefficient and residual pivots strong enough to survive conditioning on the complete statistical reporting pattern (or explicit lower bounds on the relevant pattern probabilities), uniform calibration of the residual quadratic forms, and an independent consistent covariance or scale estimate. None follows from the present contract. Robustness beyond the spherical-Gaussian model or a separately calibrated GLS extension remains open.

B.3 Why latent purity needs additional design structure

Proposition 30 (Under-excited designs can leave latent purity unidentified)**.** *Let $p = 1$ and suppose every response-bearing covariate lies in $\{-1, 1\}$. A one-mechanism model with raw coefficient $(1, 0)^\mathsf{T}$ and a two-mechanism model with assignment map $c(1) = 1$, $c(-1) = 2$, coefficients $(0, 1)^\mathsf{T}$ and $(0, -1)^\mathsf{T}$, respectively, induce the same response distribution. For a row set containing both*

signs, one-mechanism purity is true in the first model and false in the second. Any procedure whose probability of falsely certifying purity is at most α uniformly over impure models therefore certifies purity with probability at most α in this observationally identical pure model.

Proof. At $x = 1$, the first coefficient in the two-mechanism model gives mean 1; at $x = -1$, its second coefficient also gives mean 1. The one-mechanism model has mean 1 at both designs. With the same Gaussian variance, every joint response distribution is identical. Hence every randomized or deterministic certificate event has the same probability in both models although its truth value differs. Uniform false-certification control in the impure model gives the claimed upper bound in the pure model. □

In the two-mechanism representation, each mechanism is observed at only one distinct covariate value, so its class-specific augmented design has rank one rather than $d = 2$. The example proves non-identifiability over the unrestricted design class; it does not establish impossibility under cell purity and uniform excitation. It explains why the assumption-free target is the observable finite-design projection in equation (21), while the positive result in Appendix B.4 licenses a stronger class interpretation under explicit additional structure.

B.4 Interpreting a released projection

For cell j, let $\theta^{\circ}_{z,gj} = (A_{gj}^{\mathsf{T}} A_{gj})^{-1} A_{gj}^{\mathsf{T}} \mu_{gj}$ and $w_{gj} = |C_{gj}|/N_g$. By a piecewise-affine class model we mean that there are fixed class coefficients $\{\theta_{z,k}\}$ and a row-level class map $c_g(i)$ such that $\mu_i = \widetilde{z}_i^{\mathsf{T}} \theta_{z,c_g(i)}$. A cell is *pure* when $c_g(i)$ is constant over its rows; writing that value as $c_g(j)$ gives $\mu_{gj} = A_{gj} \theta_{z,c_g(j)}$. Cell purity is a scientific premise and is not inferred by the audit.

Proposition 31 (Cell-projection identities)**.** *For every eligible group,*

$$\|H_g \mu_g\|_2^2 = \sum_j \|A_{gj}(\theta^{\circ}_{z,gj} - \theta^{\circ}_g)\|_2^2, \tag{47}$$

$$\|W_g \mu_g\|_2^2 = \sum_j \|(I - P_{gj})\mu_{gj}\|_2^2. \tag{48}$$

Proof. Put $\widetilde{\mu}_g = P_{S,g}\mu_g$. The residual $\mu_g - \widetilde{\mu}_g$ is orthogonal to $\mathrm{col}(P_{S,g})$, which contains $\mathrm{col}(P_g)$; hence $P_g \mu_g = P_g \widetilde{\mu}_g$. Within cell j, the cell-j blocks of $P_{S,g}\mu_g$ and $P_g \mu_g$ are respectively $A_{gj}\theta^{\circ}_{z,gj}$ and $A_{gj}\theta^{\circ}_g$. Summing blockwise norms proves the first identity. The block-diagonal definition of W_g proves the second. □

Proposition 32 (Excitation bound)**.** *If every cell satisfies $|C_{gj}|^{-1} A_{gj}^{\mathsf{T}} A_{gj} \succeq \kappa I_d$ for $\kappa > 0$, then*

$$\eta_{H,g}^2 \geq \frac{\kappa}{\sigma^2} \sum_{j<k} w_{gj} w_{gk} \|\theta^{\circ}_{z,gj} - \theta^{\circ}_{z,gk}\|_2^2. \tag{49}$$

Proof. Proposition 31 and excitation yield

$$\eta_{H,g}^2 \geq \frac{\kappa}{\sigma^2} \sum_j w_{gj} \|\theta^{\circ}_{z,gj} - \theta^{\circ}_g\|_2^2 \geq \frac{\kappa}{\sigma^2} \min_{\vartheta} \sum_j w_{gj} \|\theta^{\circ}_{z,gj} - \vartheta\|_2^2.$$

The minimum equals the pairwise sum in equation (49) by the weighted-variance identity. □

Corollary 33 (Mass-indexed coefficient-class exclusion)**.** *Fix an $\mathcal{F}_0$-measurable collection $\mathcal{G}_0 \subseteq \mathcal{G}_{\mathrm{aud}}$. Suppose its audit cells are pure under a piecewise-affine class model, and define the observable design excitation*

$$\kappa_g = \min_j \lambda_{\min}\big(|C_{gj}|^{-1} A_{gj}^{\mathsf{T}} A_{gj}\big) > 0. \tag{50}$$

For any analyst-declared mass floor $\pi_0 \in (0, 1/2]$, a group satisfying $\eta_{H,g} < \tau_{H,g}$ cannot contain two coefficient classes that each occupy at least π_0 of its rows and satisfy

$$\frac{\|\theta_{z,a} - \theta_{z,b}\|_2}{\sigma} \geq \frac{\tau_{H,g}}{\sqrt{\kappa_g}\,\pi_0}. \tag{51}$$

Conditional on $\mathcal{F}_0$, the probability that any released group in $\mathcal{G}_0$ violates this exclusion is at most $\sum_{g \in \mathcal{G}_0} \alpha_{H,g}$ under the fixed-axis rule and at most α_{cert} under the groupwise rule.

Proof. Cell purity and full rank make each cell target equal its class coefficient. If classes a and b each have aggregate mass at least π_0, the sum of $w_{gj}w_{gk}$ over cell pairs drawn from these two classes is at least π_0^2. Proposition 32 therefore gives $\eta_{H,g} \geq \sqrt{\kappa_g}\,\pi_0\|\theta_{z,a} - \theta_{z,b}\|_2/\sigma$. Equation (51) contradicts the true H margin. Under the fixed-axis rule, any released violation requires a false H pass, and a union bound gives the stated sum. Under the groupwise rule, a released violation is a falsely certified group, so Corollary 11 gives the α_{cert} bound. □

The mass floor is a declared interpretation scale, not an estimated class mass, and cell purity remains an unaudited scientific premise. If the analyst also posits that every represented class has mass at least π_0 and that distinct classes exceed the displayed separation, the corollary reduces to a single-class interpretation and θ_g° equals that class coefficient. In raw coordinates the separation in equation (51) is $\|T_{xz}^{-1}(\theta_{x,a} - \theta_{x,b})\|_2$; using the untransformed raw Euclidean norm would change the excitation constant. Excitation, class mass, and cell purity are scientific assumptions, not quantities certified by the H/A gate. Without these premises, the licensed target remains the pooled finite-design projection.

C Unknown-noise scale design and cost

C.1 An inference scale selected by the certification gate

Coefficient-error independence alone does not preserve a Student pivot whose scale also participates in selection. To see the issue within the static unknown-noise gate, condition on $\mathcal{F}_0$ and consider one eligible group, suppressing g. Let $U_u = \|Q_u Y\|_2^2/\sigma^2 \sim \chi^2_{r_u, N\eta_u^2}$ for $u \in \{H, A\}$, and $V = S_C^2/\sigma^2 \sim \chi^2_{m_C}/m_C$. These three variables are independent. The standardized error Z of any requested coefficient is $N(0,1)$ and independent of all three. With the positive static cutoffs c_u in equation (7), the group's release event is $R = \{U_H < r_H c_H V,\ U_A < r_A c_A V\}$. Its conditional scale density satisfies

$$f_{V|R}(v) = \frac{f_V(v)\prod_{u\in\{H,A\}} G_{r_u, N\eta_u^2}(r_u c_u v)}{\mathbb{P}(R)}, \qquad v > 0. \tag{52}$$

This follows by conditioning on $V = v$ and multiplying the independent numerator pass probabilities. The multiplying weight is strictly increasing, so the selected scale is no longer distributed as $\chi^2_{m_C}/m_C$. Although Z remains independent of V conditional on R, the ratio $Z/\sqrt{V}$ consequently loses its ordinary Student law. In this single-group all-pass example, the selection favors larger scales: fixed-level two-sided Student intervals are conservative. The example establishes loss of the stated conditional pivot law, not inevitable undercoverage for every scale-reuse procedure or complete pattern.

A complete pattern can reverse the scale-selection direction. Take two disjoint exactly affine groups with $N = 8$, $d = 2$, two four-row cells, and an independent shared certification bank $m_C = 10$. Each group has inputs $(-4, -3, -2, -1, 1, 2, 3, 4)$, cells split by sign, and full-rank designs. Freeze admitting geometry screens, both margins at 2, and all four static axis levels at 0.0125. Ranks are $r_H = 2$, $r_A = 4$; boundary noncentrality is 32. Cutoffs are $c_H = F^{-1}_{2,10,32}(0.0125) = 4.781012817684$ and $c_A = F^{-1}_{4,10,32}(0.0125) = 2.642560027063$. Zero-departure planning powers are 0.96509 and 0.90309, exceeding a frozen 0.8 floor. This is a valid lower-tail equivalence gate. For $V = S_C^2/\sigma^2$, a group's conditional pass probability is

$$p(v) = (1 - e^{-c_H v})\{1 - e^{-2c_A v}(1 + 2c_A v)\}.$$

On the complete pattern $s = \{1\}$, its scale density is proportional to $f_V(v)p(v)\{1 - p(v)\}$. The nonrelease factor defeats the increasing-weight argument. Freeze both groups' intercept requests: allocate 0.05 to a singleton report, 0.025 per intercept if both release, and emit nothing on the empty pattern. This is the static certification rule with a predeclared pattern-dependent interval map, not the implementation's fixed-allocation default. Using S_C in the ordinary Student interval gives

$$\mathbb{P}(\text{coverage} \mid s = \{1\}) = \frac{\int_0^\infty [2\Phi\{t_{10,0.975}\sqrt{v}\} - 1] f_V(v)p(v)\{1 - p(v)\}\,dv}{\int_0^\infty f_V(v)p(v)\{1 - p(v)\}\,dv} = 0.9036617477$$

by numerical evaluation. The pattern probability is 0.08039325285. Conditioning only on group 1 passing gives 0.9600105647 for the fixed-level interval, consistent with single-group conservatism. Independent quadrature and a finite Gamma/Laplace-sum calculation agree; numerical tolerances are not interval-arithmetic certificates. This illustration does not assess the empirical bank size 400 or prove two-bank necessity. An independent S_I is a sufficient construction for the full-pattern law.

C.2 A POOLED RESIDUAL SCALE CAN UNDER-COVER AFTER THE SAME GATE

The preceding conservatism is specific to reusing S_C. Keeping the same independent certification bank and replacing only the inference scale by $S_p^2 = \|(I-P)Y\|_2^2/(N-d)$ can have the opposite effect, even at exact affinity. Let $a = r_H$, $b = r_A$, and $\nu = a + b = N - d$. At $\eta_H = \eta_A = 0$, $Q = U_H + U_A \sim \chi^2_\nu$ and $B = U_H/Q \sim \mathrm{Beta}(a/2, b/2)$ are independent. This factorization follows by changing variables in the product of their central chi-square densities. Since $S_p^2/\sigma^2 = Q/\nu$, write

$$k(B) = \max\left\{\frac{B}{ac_H}, \frac{1-B}{bc_A}\right\}, \qquad w(q) = \mathbb{P}(R \mid Q = q) = \mathbb{E}_B[\mathbb{P}\{V > qk(B)\}]\,.$$

The weight w is strictly decreasing: release selects smaller pooled residual sums of squares. The standardized coefficient error Z remains independent of (Q, B, V). For a fixed two-sided noncoverage allocation $0 < \gamma < 1$, put $h(q) = 2\Phi(t_{\nu,1-\gamma/2}\sqrt{q/\nu}) - 1$. This is strictly increasing, so

$$\mathbb{P}\Big\{|Z| \le t_{\nu,1-\gamma/2}\sqrt{Q/\nu} \mid R\Big\} = \frac{\mathbb{E}\{h(Q)w(Q)\}}{\mathbb{E}\{w(Q)\}} < \mathbb{E}\{h(Q)\} = 1 - \gamma. \tag{53}$$

Indeed, for independent copies Q, Q', twice the covariance of $h(Q)$ and $w(Q)$ equals $\mathbb{E}[(h(Q) - h(Q'))(w(Q) - w(Q'))] < 0$. Thus this replacement produces conditional undercoverage despite exact unconditional Student inference. This is an analytical single-group result; it does not quantify the size or direction of errors for every mean or complete multi-group pattern.

C.3 RESIDUAL-MEAN CONTAMINATION AND MEASUREMENT COST

Under an arbitrary fixed mean, $\nu S_p^2/\sigma^2$ instead has a noncentral chi-square law with $\nu = N - d$ degrees of freedom and noncentrality $\lambda = N(\eta_H^2 + \eta_A^2)$. Hence

$$\frac{\mathbb{E}(S_p^2 \mid \mathcal{F}_0)}{\sigma^2} = 1 + \frac{N(\eta_H^2 + \eta_A^2)}{N - d}, \qquad R_{\mathrm{RMS}} = \frac{\sqrt{\mathbb{E}(S_p^2 \mid \mathcal{F}_0)}}{\sigma}. \tag{54}$$

Here "unconditional" means before the release decision, still conditional on the frozen design. Table C.1 substitutes the 41 known-noise eligible designs in Appendix E into this exact expectation, weighting groups equally. They have $N = 100$–397 and $d = 7$. The outer corner $(0.75, 0.75)$ has total squared deviation 1.125; other points on the rectangular boundary need not have that value. This is a hypothetical scale-contamination calculation, not a new response simulation.

Table C.1: Pooled residual RMS scale relative to σ, before release. These deterministic calculations use the 41 known-noise eligible groups, not the 31 groups retained by the unknown-noise screen.

Fixed (η_H, η_A)	R_{RMS}		
	Minimum	Maximum	Mean
$(0, 0)$	1.000000	1.000000	1.000000
$(0.20, 0.20)$	1.039921	1.042124	1.040951
$(0.75, 0.75)$	1.464648	1.486498	1.474887

R_{RMS} is not $\mathbb{E}(S_p \mid \mathcal{F}_0)/\sigma$, and neither quantity is a release-conditional half-width ratio. Actual widths also depend on selection and critical values: a pooled-scale Student interval uses $N - d$ degrees of freedom, whereas the independent-bank interval uses m_I. The outer-corner factor is numerically similar to the 1.45 splitting SE factor in Section 6, but the quantities and eligible populations differ. This comparison does not establish equal-width or equal-cost performance.

An independent S_I has no residual-mean contribution and keeps its declared central chi-square law after every permitted complete reporting pattern. Together with Gaussian projection independence, this preserves the stated Student law. Scientific independence must be justified; the response order excludes S_I from selection. At $m_C = m_I = 400$, replicate-pair banks cost 1,600 additional responses per arm. Relative to an alternative retaining S_C for the same audit, the additional inference bank costs 800 responses. Two banks are a sufficient construction, not a necessity or a minimum-cost result. The application must justify whether this guarantee warrants that measurement cost.

D IMPLEMENTATION AND REPRODUCIBILITY

D.1 EXECUTABLE ALGORITHM AND FAILURE SEMANTICS

The implementation follows Appendix A.3: the generic audit starts from a frozen router or partition serialized before audit roles open. Its API enforces logical stage gates on value use; physical isolation of unopened files, credentials, or caller processes is a deployment premise, not something the in-process API proves. Frozen role identities, source hashes, and external-router blindness fields are verified only as serialized contracts. The implementation does not itself establish equal-mean sampling, independence, source fidelity, or physical outcome blindness.

A pre-response design ineligibility is scoped to the predeclared group or run gate; the statistical core does not silently reassign that scope. Before a future role is opened, invalid material such as an S_I bank may be replaced only by an A1-frozen, opened-value-independent correction rule that preserves the frozen role identities and sampling law. The in-process API validates the finally supplied bank against the frozen manifest but cannot establish that correction rule's independence.

A frozen-router failure before audit opening invalidates that router. After audit opening, a failure invalidates the complete replicate. Report construction is one-use: a call outside the exact pre-final stage fails without mutating an already finalized or invalid ledger. Response-driven replacement, fallback, and retry are forbidden.

D.2 NUMERICAL BOUNDARY

The results are exact-real statements; the implementation uses pinned NumPy/SciPy floating-point arithmetic, QR/SVD factorizations, noncentral laws, strict comparisons, and deterministic total-order tie rules. It fails closed on nonfinite values, but it provides neither outward-rounded quantiles nor verified linear algebra. Thus outputs evaluate, rather than formally certify, the exact procedure. Independence of the H_g and W_g residual components is an exact-real consequence of orthogonal projectors under spherical Gaussian noise; the floating-point backend only approximates those projectors.

All operative decisions are strict. In static mode, equality with the frozen axis cutoff is non-release. In intersection–union/Holm mode, a statistic exactly equal to the conservative prospective cutoff is assigned the exact boundary p-value $\alpha_{\mathrm{cert}}/M_{\mathrm{slot}}$ rather than a recomputed CDF value, avoiding CDF/quantile round-trip drift. The corresponding group fails at the first Holm threshold but may be released at a later rank after all preceding hypotheses reject and the realized rank-specific threshold becomes larger; equality with the realized Holm p-value threshold remains non-release.

Before report construction, the implementation rejects nonfinite critical values, half-widths, coefficient estimates, and interval endpoints.

D.3 COMPLEXITY AND SPARSE CONSTRUCTION

Let $N_{\mathrm{routed}} = \sum_g N_g \le |X_R|$. For the generic audit, the current implementation rebuilds the deterministic split tree separately for every candidate $J \in \mathcal{J}$, so its repeated-sort routine has conservative cost $O\{N_{\mathrm{routed}}(p + \log(1 + |X_R|)) \sum_{J\in\mathcal{J}} J\}$; for fixed p and $\mathcal{J}$ this is linear in N_{routed} up to the logarithm. The current code does not cache leaf keys or coordinate orders. QR/SVD diagnostic and factorization costs additionally sum over routed groups, candidate cell counts, and their cells. Because the selector retains every design-feasible candidate, per-group peak audit storage is $O(|\mathcal{J}|N_g^2)$ for dense projectors at $N_g \le 512$ and $O(|\mathcal{J}|N_g d)$ for reduced-QR bases above that threshold.

Tests compare dense and QR operators at $N_g = 80$ and separately exercise linear-size QR replay at $N_g = 2048$; both are numerical checks, not certified arithmetic. The latter applies P_g, H_g, and W_g without dense $N_g \times N_g$ matrices.

Experiments must report resource and numerical failures and excluded scopes rather than silently remove them.

D.4 Reproducibility and artifact boundary

Statistical core and experiment component. The statistical core and experiment implementations have different purposes. The reference statistical core implements the frozen response order, release-set and pooled-fit ordering, and group-atomic reporting. Its preserved historical acceptance target contains 107 tests across 13 modules. Three later overflow regressions in one added module bring the current target to 110 tests across 14 modules, all passing in the assistant Linux audit. These checks separate nonfinite numerical evaluation from statistical nonrelease. Later numerical hardening evaluates cancellation-prone campaign residual energies by squared residual norms and rejects nonfinite intermediates; historical producer sources and saved ledgers remain separately preserved. Software tests and numerical witnesses establish mechanics or feasibility, not empirical performance. Pooled-equivalence gates are supplied by the higher-dimensional and matched-cost experiment producers; the reference core implements two-axis gating and pooled coefficient fitting. The higher-dimensional implementation is documented at `highdim/highdim_experiments_v1/producer/audit_core.py`. The separate experiment component supplies the data-proposed Gaussian campaign, frozen designs and seeds, numerical ledger, and generators for Figure 2 and Appendix E. Its entry point supports integrity verification, numerical checks, rendering, and fresh campaign execution. Appendix F extends the evidence with separately identified producers and ledgers. Each component's run guide specifies its own acceptance target and dependencies; neither component is a complete offline operating-system and TeX image.

Implementation boundary. The supplied artifacts are experimental reference implementations of the statistical procedures. The mathematical joint reporting guarantee concerns the specified statistical events; it is not a certification that a deployed data-acquisition and publishing system meets every protocol requirement. The reference core records routes, selected cells, decisions, reports, abstentions, response hashes, and transition events. It does not serialize all rejected-cell-count diagnostics, every exclusive per-slot status required by Appendix A.5, or every terminal numerical/runtime failure. The preserved campaign bookkeeping, publisher, and evaluator components are not connected into a complete source-worker-to-protocol-record path. Operational source archives, exact adapter environments and source-fidelity fixtures, and application-specific data and scale justifications are not all bundled. The reference runners accordingly reject confirmatory status; their outputs are not designated full-protocol executions. The experiment component has its own retained ledger and reporting diagnostics, but does not remedy these deployment omissions or establish upstream-source fidelity for every preserved adapter.

Meaning of reproduction. Rebuilding figures or replaying the controlled campaign checks the supplied computational and statistical experiment. It does not establish physical response isolation, end-to-end application provenance, certified arithmetic, complete aggregation of deployment failures, or durable operational retention. A full-protocol designation requires those conditions and all required records to be bound and validated separately. The evidence now includes the finite-design experiment in Appendix E, the learned-design experiments and measured working-model cases in Appendix F, and the mathematical results under their stated assumptions. The additional experiment ledgers and response-role records do not remove the reference core's deployment boundary above. Each campaign retains its own producer identity and reporting scope.

D.5 EXECUTION PROVENANCE

Table D.1: Execution provenance. Historical author-supplied returns and later assistant-environment checks refer to different executions. A replay checks computation on existing results; it supplies no new statistical confirmation.

Result or component	Execution evidence	Scope of verification
Original E campaign	Linux baseline; recorded Windows replay.	Same-seed full reproduction is attributed to a retained independent report; its raw replay ledger and logs were not supplied.
Original F, F2D, and completion	Author-supplied full-run returns record Windows 11/Python 3.12.10.	Separate protocols, producer identities, and full result ledgers are retained.
Integrated E/F/F2D/completion replay	Later author-supplied Windows verification/rendering return.	Saved-result replay: all four components and 45 historical displays checked by byte or decoded-image/rendered-page agreement.
E.9 residual comparison	AI-assisted exact-law calculation and deterministic replay.	Independent Poisson/gamma-mixture check; assistant-environment replay matches all eight reference output files byte-for-byte.
G original and amendment	Linux simulation; original defective ledger retained; assistant-environment correction.	Frozen exposed-outcome correction and separate structural audit; corrected analysis replay reproduces JSON/CSV/PNG byte-for-byte.
F.9 higher-dimensional study	Linux execution and relocation verification.	Saved fits, ledgers, and numerical checks; no Windows rerun.
F.8 measured-response audits	Retrospective SGEMM, PLATERO, and BAM studies; assistant-environment Linux replays.	Raw sources, figures, saved-ledger checks, and bounded replays; no exact measured-data report or true-coverage estimate.
Later numerical hardening	Assistant Linux audit; 110 core tests and 13 campaign arithmetic regressions pass.	Saved-design, same-seed F/F2D response replays retain reporting events. Floating-point diagnostic differences are recorded separately; no router refits or new statistical confirmation.

Appendices E.8, F.10, and G give execution details. Protocol records document the implemented response-opening order; physical measurement independence remains an application assumption. The artifacts are experimental reference implementations. Appendix D.4 specifies the acquisition, failure-recording, and publishing requirements outside their deployment boundary.

E CONTROLLED EXPERIMENTS ON DATA-PROPOSED PARTITIONS

This appendix evaluates the statistical mechanisms of CertifyPAR on fixed audit designs induced by discovery-trained routers. The main questions are whether residual-based reporting preserves coefficient coverage, whether reporting follows the exact finite-design laws, and how eligibility and information costs limit reporting. Here C denotes the two-axis rule; NR denotes exact-affinity nonrejection in Appendix E.9. The experiment component supplements the reference statistical core described in Appendix D; it is a controlled simulation study, not a full-protocol deployment or a confirmatory application study.

E.1 DATA, ROUTERS, AND FROZEN AUDIT DESIGNS

Data roles and targets. We use the four datasets distributed with scikit-learn (Pedregosa et al., 2011). Table E.1 gives the source sizes and discovery responses. Breast-cancer mean radius and wine proline are removed from their respective covariates before any further processing. Digit labels are treated as numerical regression responses only to train a partition proposer; this is not a digit-classification benchmark. A source-specific seeded permutation assigns 40% of rows, rounded to the nearest integer, to discovery and the remainder to audit. All five proposers share these row roles within a source. Only discovery responses train the routers. Original audit responses are never used to evaluate coverage or select audit objects.

Table E.1: Real-data inputs to the partition and design construction. Every source retains six covariates. The response column describes discovery training only; all inferential audit responses are simulated.

Source	Rows	Discovery	Audit	Discovery response
Diabetes	442	177	265	Disease progression
Breast cancer	569	228	341	Mean radius
Digits	1797	719	1078	Numerical digit label
Wine	178	71	107	Proline

Covariates and reporting coordinates. We inspect columns in their source order using covariates alone, stopping after six are retained. A column is rejected if its variance is zero, if its most frequent value occupies more than 20% of rows, or if its absolute correlation with an already retained column exceeds 0.90. This transductive screen can use all observed covariates, but no audit response. All four sources retain $p = 6$ columns, giving affine dimension $d = 7$. Centering and scaling for the audit are fixed from audit covariates. We report the intercept and first two slopes *in these standardized coordinates*, with unit-noise covariance $(A_g^\mathsf{T} A_g)^{-1}$. The targets are the finite-design projections θ_g°, not the proposer's fitted coefficients or coefficients in original measurement units.

Proposers. The complete four-source by five-proposer crossing contains 20 arms. The requested capacity is $K = \min\{8, \max\{2, \lfloor n_R/(1.6 \times 90)\rfloor\}\}$, where n_R is the source's audit-row count; realized slot counts can differ from K. Each fitted proposer exports only a frozen hard covariate router.

- **Linear-tree:** `linear-tree` 0.3.5 with OLS leaf models (Cerliani, 2022). Maximum depth is $\max\{1, \lceil \log_2 K \rceil\}$, minimum leaf size is 81, and the split grid uses 25 bins.
- **CART:** scikit-learn regression tree with at most K leaves and minimum leaf size 81. Its leaves predict constants; CertifyPAR subsequently fits a new affine model on each eligible audit group.
- **ReLU pattern:** a one-hidden-layer `MLPRegressor` with $\max\{1, \lceil \log_2 K \rceil\}$ ReLU units, penalty 10^{-3}, and at most 2000 iterations. Discovery covariates and responses are standardized. The hidden activation sign pattern supplies the router. Patterns represented by fewer than 30 discovery rows are excluded, unless this would retain fewer than two patterns, in which case all observed discovery patterns are kept. Unseen or excluded patterns yield router abstention.
- **Predictive CLR:** a local reimplementation using up to 60 hard-assignment OLS iterations, followed by a 200-tree random-forest classifier (maximum depth 6, minimum leaf size 10) for

Table E.2: Statistical freeze card. Every value was fixed before any audit response was generated; the two digests bind the configuration and the code that produced the records.

Quantity	Value
familywise false-certification budget α_{cert}	0.05
simultaneous-interval budget γ_{FWER}	0.05
prospective power floor $1-\beta$	0.80
margins $\tau_H = \tau_A$	0.75
planning points $\tau_H^- = \tau_A^-$	0.2
candidate cell counts	2, 3, 4
requested coordinates per group	3
unit-noise variance-factor limit	0.5
minimum rows per frozen slot	90
discovery fraction	0.4
retained covariates p (cap)	6
pure-error banks $m_C = m_I$ (unknown noise)	400
frozen slots	61
eligible audit groups	41
arms (covariate source $\times$ proposer)	20
protocol hash	`a6c8e1976b0546e3...`
producer source digest	`db87b73a0dde517a...`

covariate-only routing. Labels with fewer than 10 discovery rows are removed unless fewer than two would remain. This implements a predictive-CLR construction; it is not the published CG-CLR implementation and is not presented as a fidelity benchmark.

- **PARC:** `pyparc` 2.0.4 (Bemporad, 2023), with requested K, Softmax separation, regularization 10^{-5}, at most 100 iterations, cost tolerance 10^{-4}, and ten k-means initializations. We use its retained piecewise-linear separation rule, not its fitted affine predictions. Discovery covariates are standardized; PARC's default region-pruning threshold is retained.

The source package records library defaults and compatibility adaptations, including input-validation shims, the NumPy 2.4 scalar-assignment adaptation for binary PARC separation, and a batch-shape wrapper for PARC prediction. There is no artificial three-region floor for PARC. These implementation details matter for reproducing the realized partitions; the study does not compare predictive accuracy or rank the five learners.

Eligibility and resolution. Table E.2 records the final campaign's fixed settings. Every slot receives a response-free status in the following precedence: empty, fewer than 90 rows, design/cutoff ineligible, prospective-power ineligible, or eligible. Candidate resolutions $J \in \{2, 3, 4\}$ use the deterministic balanced-cell rule of Appendix A.4: recursively bisect the largest splittable cell along its largest-range coordinate, resolving ties by the frozen rule and stable row IDs. Full pooled and cellwise rank, positive axis ranks, condition numbers at most 10^{12}, and requested-coordinate unit-noise variance factors at most 0.50 are required. Each axis must have prospective pass probability at least 0.80 at its inner point. Among qualifying resolutions, the core minimizes the sum of marginal power shortfalls and breaks exact ties toward larger J.

Static certification allocates $a_g = \alpha_{\mathrm{cert}}/(2M_{\mathrm{slot}})$ to each axis within an arm, retaining empty and ineligible slots in M_{slot}. Budgets are per arm, not a simultaneous budget over all 20 experimental arms. The outer margins 0.75 and inner points 0.20 are noise-normalized RMS values chosen for a feasible calibration study. They are not application-validated scientific tolerances. The settings were fixed for the reported campaign following protocol development; we make no preregistration claim.

E.2 KNOWN TARGETS, REPLICATION UNITS, AND ESTIMANDS

For each eligible group, let $u_{H,g}$ and $u_{A,g}$ be fixed unit vectors in $\mathrm{range}(H_g)$ and $\mathrm{range}(W_g)$, respectively. The simulation mean and responses are

$$\mu_g = A_g b_g + \sigma\sqrt{N_g}(\eta_{H,g}u_{H,g} + \eta_{A,g}u_{A,g}), \qquad Y_g = \mu_g + \sigma\varepsilon_g, \quad \varepsilon_g \sim N(0, I_{N_g}). \tag{E.1}$$

Here b_g is a generating coefficient vector, and $H_g A_g = W_g A_g = 0$ gives $\theta_g^\circ = b_g$ exactly in the mathematical construction. The two residual directions realize the declared deviations. For the three requested coordinates, $b_{g\ell} = \kappa\sigma\sqrt{[(A_g^\mathsf{T}A_g)^{-1}]_{\ell\ell}}$; all other coordinates are zero. The effect multiplier κ is distinct from the residual-deviation positions. The primary branch uses $\sigma = 1$. At the adequate truth used for E1, E3, E6, and E7, $\eta_H = \eta_A = 0.20 < 0.75$. E5 uses only design matrices, with no response draws.

Each arm has one frozen router, audit design, and set of residual directions. We vary independent Gaussian response draws, not dataset draws or retrained routers. The direct engine draws row-level errors and computes the actual pooled/cellwise projections through their orthonormal bases. Its analytical expansion of Y_g avoids storing dense $N_g \times N_g$ projectors. Headline estimates use this engine; the known-noise noncentral laws provide separate analytic reference probabilities. Table E.3 identifies the actual computational units; unused legacy replication constants are not treated as executed experiments.

Table E.3: Experiment index and actual repetition counts. A family is one dataset–proposer arm. Response replications do not increase the number of independent designs.

ID	Question	Repetitions or unit
E1	Coverage under three release rules; five κ values	200,000 per arm and κ
E2	Joint release law; two axes and 12 positions	100,000 per group/axis/position
E3	Slot eligibility and adequate-truth release	Design ledger and E1 at $\kappa = 1$
E4	Familywise false certification at the boundary	200,000 per arm and axis
E5	Full-design versus half-design coefficient SE	64 fixed splits per eligible group
E6a	Unknown-noise eligibility versus m_C	Five response-free settings
E6	Unknown-noise coverage and release	50,000 per arm
E7	Row-split and independent-replicate alternatives	50,000 per arm

Coverage and interval allocation. For a nonempty complete reporting pattern s within an arm, each released group reports its three requested coordinates, with $\gamma_{g\ell}(s) = \gamma_{\mathrm{FWER}}/(3|s|)$. Known-noise intervals have half-width $z_{1-\gamma_{g\ell}(s)/2}\sigma\sqrt{v_{g\ell}}$, where $v_{g\ell} = [(A_g^\mathsf{T}A_g)^{-1}]_{\ell\ell}$. Unknown-noise intervals replace σ by S_I and z by a Student-t_{m_I} quantile. One replication is a noncoverage event if any emitted interval misses its declared target. Empty reporting patterns contribute to release probabilities, but not to the conditional-noncoverage denominator.

For arm a and effect κ, write $n_{a\kappa}$ for the number of nonempty replications and $m_{a\kappa}$ for their familywise misses. The headline is

$$\widehat{q}_\kappa = \frac{\sum_a m_{a\kappa}}{\sum_a n_{a\kappa}}. \tag{E.2}$$

This pools complete patterns by observed frequency and arms by their nonempty counts. It is not an unweighted arm mean and does not itself test every conditional inequality in Corollary 2. Per-pattern estimates use only replications with the corresponding complete pattern. We retain all observed patterns, including small denominators.

Monte Carlo uncertainty. At each fixed binomial endpoint we report pointwise 99% Clopper–Pearson intervals (Clopper & Pearson, 1934). Pooled intervals apply the same count formula to heterogeneous arms or designs and are reported as pooled Monte Carlo summaries, not exact common-probability binomial intervals or uncertainty over a population of datasets. No displayed interval is simultaneous over the operating grid or over selected worst patterns. Arm-range shading reports heterogeneity, not a confidence band. The primary familywise coverage budget remains 0.05; the 99% intervals quantify simulation error in its estimated operating characteristics.

E.3 Coverage and response-dependent reporting: E1

E1 sweeps $\kappa \in \{0, 0.5, 1, 2, 3\}$ at the adequate truth. CertifyPAR releases only when both residual gates pass. Two diagnostic controls use the same eligible designs, requested coordinates, coefficient estimates, and pattern-dependent interval map, but replace the release rule. The *coefficient-gated* control releases if any requested coordinate has absolute known-noise z statistic above $z_{0.975}$. The *regression-gated* control uses an overall regression F statistic above its 95th percentile, with $(d-1, N_g - d)$ degrees of freedom. The latter is a conventional significance heuristic under these deliberately non-affine means, not a calibrated adequacy test. Both inspect pooled-fit information and fall outside Lemma 1. Neither is a published selective inference procedure or a competitive benchmark with selection-adjusted intervals.

Table E.4 reports the full three-rule sweep; Table E.5 details CertifyPAR's pooled and selected-pattern results. CertifyPAR's pooled noncoverage ranges from 0.0471838 to 0.0474636. At $\kappa = 0$, the two controls yield 0.3455863 and 0.3516994, approximately seven times the nominal 0.05 budget; at $\kappa = 3$ they yield 0.0471 and 0.0454 after rounding. The effect-size dependence is consistent with distortion induced by selecting low-signal fits. It is not evidence that any significance filter fails by a fixed factor in every regime. The largest CertifyPAR arm-level estimate across this sweep is approximately 0.0503; finite Monte Carlo estimates need not lie below the nominal bound.

Table E.4: Pooled familywise noncoverage for all three E1 reporting rules, conditional on a nonempty reporting pattern. Each entry divides pooled misses by pooled nonempty replications; it is not an unweighted arm average. CertifyPAR's intervals and pattern ledger appear in Table E.5; full control counts and interval endpoints are retained in the numerical evidence.

κ	CertifyPAR	Coefficient-gated	Regression-gated
0.0	0.0473	0.3456	0.3517
0.5	0.0474	0.2084	0.2730
1.0	0.0472	0.0876	0.1404
2.0	0.0475	0.0483	0.0520
3.0	0.0474	0.0471	0.0454

Why these controls, and what they do not compare. The preliminary-test literature motivates examining inference after significance screening (Bock et al., 1973; Giles & Giles, 1993); it does not establish that these exact rules describe all practitioner behavior. Here E1 isolates the effect of the reporting rule while keeping the audit designs and interval map fixed. This is a different question from comparing valid selective-inference procedures. For example, Chen & Witten (2023) test differences between cluster means conditional on intermediate k-means assignments, whereas Lee et al. (2016) construct confidence intervals for coefficients selected by the lasso. Adapting such procedures to this study would require specifying a common target and conditioning event and deciding how to treat the two adequacy gates. No such matched comparison was performed. The present evidence therefore supports the stated mechanism without establishing superiority over those methods; the five proposers are sources of partitions, not inferential baselines.

Table E.11 compares the targets, release information, and coverage statements of the workflows evaluated here.

Complete reporting patterns. There are 125, 126, 124, 124, and 125 recorded arm-pattern combinations across the five effect sizes (624 records in total). Their pointwise 99% lower endpoints never exceed 0.05; 115 intervals lie wholly below 0.05, consistent with conservative allocation. This is a diagnostic of the implemented conditional coverage, not simultaneous empirical verification of all possible patterns. At $\kappa = 0$, the selected largest estimate among patterns observed at least 1000 times is 0.0628 with interval $[0.0448, 0.0850]$ and denominator 1019; the selection makes its pointwise interval inappropriate as a confidence bound on the worst conditional risk. Figure E.1 shows 98 of 124 records at $\kappa = 1$, using the display threshold of 25 occurrences. The other 26 records remain in the numerical ledger.

Table E.5: Complete-pattern ledger. The pattern-averaged column pools release events; the worst-pattern column is the selected maximum over patterns with $n \geq 1000$ and carries a pointwise interval, so it is not a simultaneous validation.

κ	Patterns	Pattern-avg.	99% interval	Worst pattern	99% interval	n
0.0	125	0.0473	[0.0470, 0.0476]	0.0628	[0.0448, 0.0850]	1019
0.5	126	0.0474	[0.0471, 0.0477]	0.0526	[0.0361, 0.0734]	1008
1.0	124	0.0472	[0.0469, 0.0475]	0.0527	[0.0374, 0.0718]	1176
2.0	124	0.0475	[0.0472, 0.0478]	0.0540	[0.0387, 0.0729]	1222
3.0	125	0.0474	[0.0471, 0.0477]	0.0563	[0.0412, 0.0747]	1315

Budget $\gamma_{\mathrm{FWER}} = 0.05$. Every worst-pattern interval covers the budget.

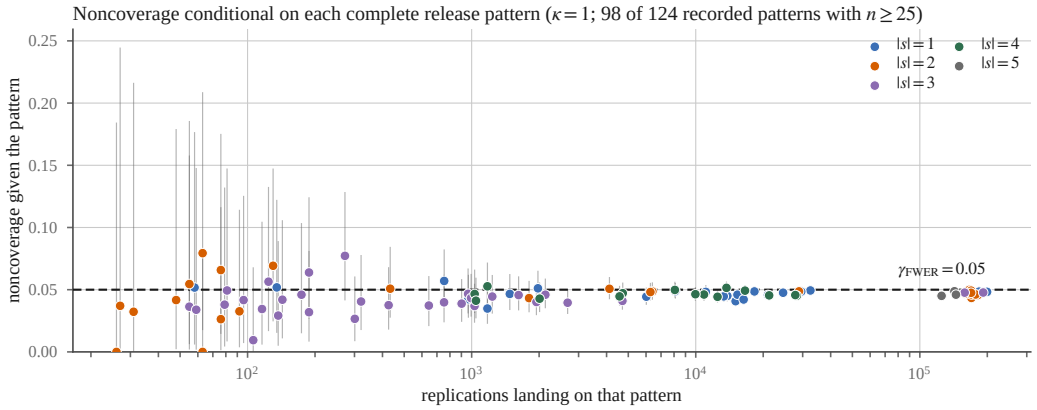

Figure E.1: Conditional noncoverage by complete reporting pattern at $\kappa = 1$. The plot retains the 98 arm-pattern records with at least 25 occurrences, including their full pointwise 99% interval ranges. It does not replace the full ledger or provide a simultaneous bound over patterns.

E.4 Joint operating characteristic and boundary error: E2 and E4

Exact joint reference. For known noise let $c_{H,g}, c_{A,g}$ be the frozen lower-tail cutoffs for $\|H_g Y_g\|_2^2/\sigma^2$ and $\|W_g Y_g\|_2^2/\sigma^2$. Orthogonality and Gaussian noise give

$$p_g(\eta_H, \eta_A) = G_{r_{H,g}, N_g \eta_H^2}(c_{H,g}) G_{r_{A,g}, N_g \eta_A^2}(c_{A,g}). \tag{E.3}$$

E2 varies η_u/τ_u over $\{0, 0.20, 0.40, 0.55, 0.70, 0.80, 0.875, 0.95, 1, 1.05, 1.15, 1.30\}$ for one axis u, while setting the other deviation to zero. Both gates must still pass. Thus the H and A curves in Figure 2a are joint release curves, not marginal axis-pass curves. The 41 eligible designs give $41 \times 2 \times 12 = 984$ operating points and 98.4 million group-level trials. These are reused fixed designs, not 98.4 million independent data designs.

Eight reference probabilities are outside their pointwise 99% intervals (three H and five A), compared with 8.43285 expected under the corresponding binomial laws. For point i, let $K_i \sim \mathrm{Binomial}(n_i, p_i)$, where $n_i = 100{,}000$ and p_i is its exact joint probability, and let $C_i(k)$ be its closed 99% Clopper–Pearson interval. The expected count is

$$\sum_{i=1}^{984} \mathbb{P}\{p_i \notin C_i(K_i)\} = \sum_{i=1}^{984} \sum_{k=0}^{n_i} \mathbf{1}\{p_i \notin C_i(k)\} \mathbb{P}(K_i = k) \approx 8.43285. \tag{E.4}$$

We evaluate these probabilities with binomial tails. Linearity of expectation needs no independence between points. The nominal $984 \times 0.01 = 9.84$ is only an upper bound because the intervals are conservative. Comparing eight to its expectation is descriptive, not an omnibus test or a simultaneous coverage claim. The maximum absolute probability difference is 0.00534. Figure E.2 retains all 984 raw probability pairs, including 71 zero Monte Carlo estimates, on a symmetric-log scale near zero.

Table E.6: Exact-law concordance. Monte Carlo against the exact joint noncentral chi-square law at every design–axis–grid point, and the realised release probability at the margin against the declared per-axis level (≤ 0.025).

Axis	Points	Outside 99%	$\max\lvert\Delta\rvert$	mean $\lvert\Delta\rvert$	mean at $\eta = \tau$	max at $\eta = \tau$
H	492	3	0.00347	0.00049	0.00836	0.02426
A	492	5	0.00534	0.00054	0.00873	0.02603

Intervals are pointwise Clopper–Pearson, which is conservative, so the expected exclusion count is below the nominal 1%.

This agreement checks numerical implementation against the analytic law; it neither proves the theorem nor tests robustness to another noise model. We do not interpret a normal-reference KS statistic for standardized binomial errors as an exact calibration test.

At the A-axis margin, the largest estimated joint release probability is 0.02603 (wine/CART, slot 0), while its exact probability is 0.02499347 and the allocated axis level is 0.025. Its pointwise 99% interval, $[0.02475062, 0.02735471]$, contains the reference value. Hence an estimate above 0.025 is not evidence that the exact bound has been exceeded.

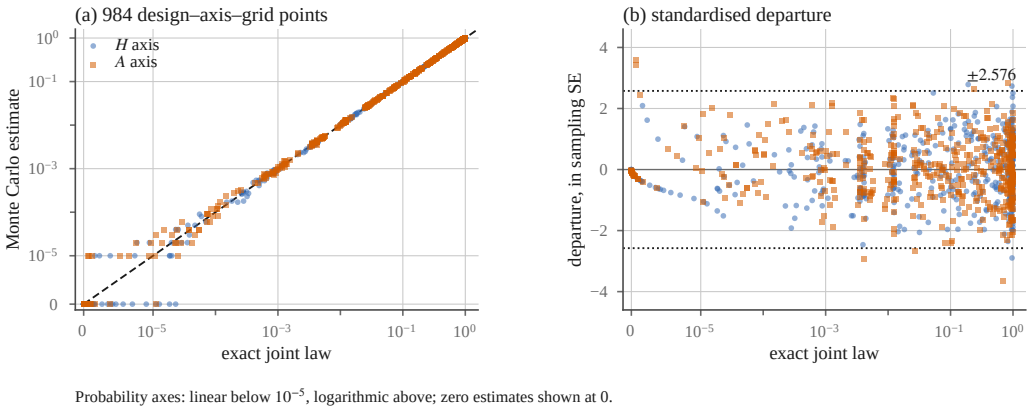

Figure E.2: Direct simulation versus the known-noise joint law at every E2 point. (a) All raw probability pairs, including zero estimates. (b) Departures divided by their binomial standard errors; the ± 2.576 lines are normal-reference diagnostics, not exact or simultaneous calibration bands.

Familywise false certification. E4 places every eligible group in an arm simultaneously at the boundary on one axis, with the other deviation zero, and records whether any group releases. Each such release is false certification because adequacy uses strict margins. Across the 17 arms with eligible groups, each axis has $17 \times 200{,}000 = 3{,}400{,}000$ family-level trials. Pooled rates are 0.02015 for H (99% interval $[0.01996, 0.02035]$) and 0.02104 for A ($[0.02084, 0.02124]$); the largest arm estimates are $4899/200000 = 0.024495$ and 0.02558. The three zero-eligible arms have structurally zero familywise events and are retained in the availability ledger but excluded from this pooled trial count. These checks concern the stated boundary configurations, not an exhaustive empirical supremum over all inadequate mean vectors. The uniform guarantee comes from Corollary 2.

E.5 Response-free availability and audit resolution: E3

Of the 61 frozen slots, 41 are eligible (67.2%), 15 have fewer than 90 rows, four fail design/cutoff gates, and one fails prospective power. None is empty. These 20 exclusions are deliberate response-free abstentions from an unsupported audit, not observed failures of the statistical adequacy tests. They reduce availability; the count alone does not demonstrate safety or show that the excluded means are inadequate. By proposer, the eligible/frozen counts are 10/11 for linear-tree, 8/12 for CART, 8/12 for ReLU pattern, 6/13 for predictive CLR, and 9/13 for PARC. Wine with ReLU

pattern, predictive CLR, or PARC contributes no eligible group; all its slots fail the row floor. These outcomes remain in the complete crossing. Separately, the digits/ReLU router abstains on 18 audit rows; row-level router abstention is distinct from the 61-slot eligibility denominator.

Figure E.3 shows how each proposer's frozen slots are removed by the row-count, design, and prospective-power gates. These are response-free counts, not realized statistical release frequencies. Group sizes, selected cell counts, axis ranks, and prospective powers are retained in the numerical ledger. In particular, the A-axis rank $r_{A,g} = N_g - J_g d$ can make small-margin certification demanding (Proposition 22); this finite collection does not estimate an asymptotic scaling exponent.

E3's realized release summaries reuse E1 at $\kappa = 1$ rather than drawing another response sample. The mean probability of at least one release is 0.9882 over the 17 eligible arm families. It answers a different question from 41/61 slot eligibility: conditional on reaching an eligible family, adequate simulated means often pass. High conditional yield does not remove the response-free exclusions.

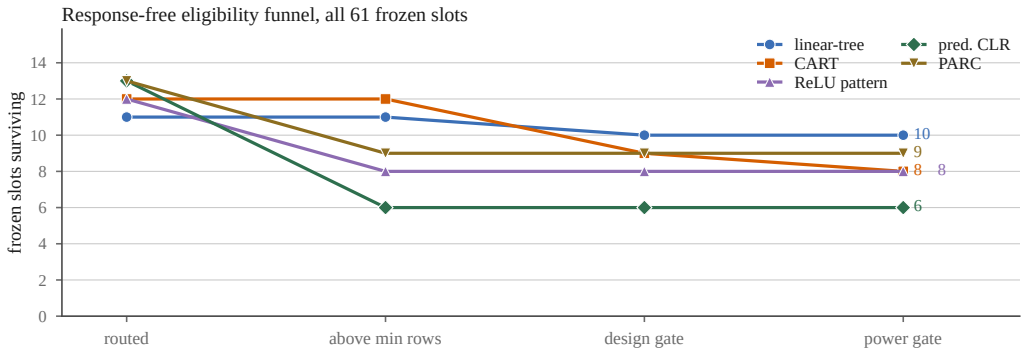

Figure E.3: Response-free eligibility funnel for all 61 frozen slots, grouped by proposer. The successive row-count, design, and power gates retain 41 eligible groups. Statistical release after observing audit responses is a separate stage.

E.6 Information and target costs of separation: E5 and E7

Design-only SE comparison (E5). For each of the 41 eligible groups, E5 averages 64 fixed random row splits. For a fit-half design $A_{g,F}$, its requested-coordinate ratio is

$$R_{g\ell} = \sqrt{\frac{[(A_{g,F}^{\mathsf{T}} A_{g,F})^{-1}]_{\ell\ell}}{[(A_g^{\mathsf{T}} A_g)^{-1}]_{\ell\ell}}}. \tag{E.5}$$

Averaging over the 64 splits and three requested coordinates first gives one ratio per group. These group ratios have mean 1.4504 and range 1.4234–1.5057. This is design arithmetic, with no coverage experiment or release simulation. The idealized proportional-information value is $\sqrt{2}$ under the matched-target, matched-critical-value conditions in Appendix A.9. Realized Gram matrices need not be proportional.

Operational alternatives (E7). At $\kappa = 1$ and $\eta_H = \eta_A = 0.20$, E7 compares reuse with two alternatives. The row-split alternative draws a separate fixed partition of each full group into gate and fit halves. It keeps the original selected cell count on the gate half, rebuilds its response-free audit objects, and applies its prospective power gate; it does not rerun the full resolution search on the half design. The fit half must have full column rank. Only 11 of 41 groups remain eligible: 29 fail gate-half prospective power and one fails gate-half rank. This is the measured protocol's availability, not the best availability achievable by every split procedure.

Under the non-affine mean in equation (E.1), the fit-half target is

$$\theta^{\circ}_{g,F} = (A_{g,F}^{\mathsf{T}} A_{g,F})^{-1} A_{g,F}^{\mathsf{T}} \mu_{g,F}, \tag{E.6}$$

Table E.7: Audit-row reuse against its alternatives. Under the non-affine worlds the finite-design projection is row-set dependent, so an estimator fitted on a subset of the rows targets $\theta_{g,F}$ rather than the full-group θ_g that CertifyPAR reports; the row split is therefore scored against both.

Procedure	Targets	Noncoverage	Available	SE ratio	Responses
CertifyPAR (audit-row reuse)	θ_g	0.0472	41/41	1.00	N
Row split, own target	$\theta_{g,F}$	0.0463	11/41	1.45	N
Row split, scored on θ_g	θ_g	**0.0556**	11/41	1.45	N
Independent replicate responses	θ_g	0.0471	41/41	1.00	$2N$

Budget 0.05. Largest target shift across the split-eligible groups: 1.039 full-design standard errors. SE ratios average over available group–coordinates and equal width ratios at a common critical value; realized simultaneous widths also depend on release-pattern size. Noncoverage pools the eligible families of each procedure, which need not be the same families.

which generally differs from θ_g° even though gate and fit errors are independent. We score each split interval against both targets. The largest requested-coordinate shift over split-eligible groups is 1.039 full-design standard errors. The independent-replicate alternative instead gates on one full-design response vector and fits on an independent vector with the same design and mean. It retains the full target and all 41 eligible groups, with twice as many audit responses.

In the preserved display labels, θ_g and $\theta_{g,F}$ abbreviate θ_g° and $\theta_{g,F}^\circ$, respectively.

The row split has 18,549 familywise misses among 400,897 nonempty family replications when scored against its own target (0.0462687), and 22,306 against the full target (0.0556402). The independent-replicate alternative has 39,605/840,014 misses (0.0471480). The corresponding reuse estimate is 0.0471838 from E1 at $\kappa = 1$. Split results pool 11 groups in nine arm families; reuse and replicates pool 41 groups in 17 families. Therefore these aggregate rates are not a matched-population causal comparison of procedures. The full-target rate exceeds 0.05 because these intervals are being used for a different target; it does not contradict valid inference for their fit-half projection target. The full-target excess is also not universal: four of nine split-family estimates are below 0.05.

E7's mean coefficient SE ratio is 1.44965024 over eligible group-coordinate pairs. It equals a half-width ratio only at a common critical value; realized simultaneous widths use different pattern-dependent critical values. Legacy `width_ratio` fields retain this SE meaning. Mean per-group release probabilities are 0.8639 for splitting and 0.9343 for independent replicates; the arm-family any-release averages are 0.8909 and 0.9883. These denominators and row-set-dependent targets govern the interpretation of Figure E.4.

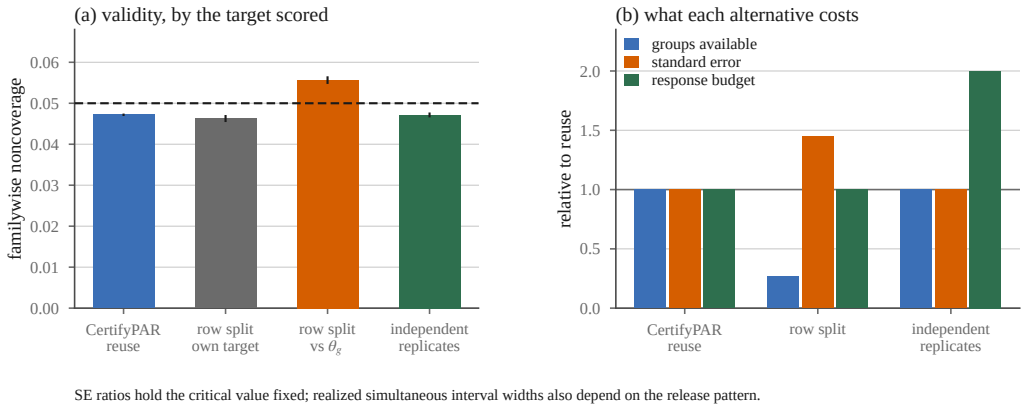

Figure E.4: Audit-row reuse, fixed row splitting, and independent full-design replicates. Split intervals are evaluated against both the fit-half and full-design projection targets. Availability and noncoverage refer to the different eligible populations described in the text; plotted SE ratios exclude pattern-dependent interval critical values.

E.7 Unknown-noise scales and their cost: E6 and E6a

E6 replaces known noise by independent scale statistics satisfying $m_C S_C^2/\sigma^2 \sim \chi^2_{m_C}$ and $m_I S_I^2/\sigma^2 \sim \chi^2_{m_I}$. Both are independent of audit errors; the certification scale is shared across groups and axes within an arm. The simulator draws these scale laws directly. It tests the independent-scale statistical contract, not the practical acquisition of equal-mean replicate pairs. Because H and A share S_C, their unknown-noise joint release probability cannot be calculated by multiplying marginal noncentral-F probabilities as in the known-noise equation (E.3).

Eligibility is recomputed for the unknown-noise branch. E6a varies certification degrees of freedom $m_C \in \{50, 100, 200, 400, 800\}$ on the same slot roster; eligible counts are respectively 0, 0, 9, 31, and 37 out of 61. Table E.8 includes illustrative equal-mean replicate costs $2(m_C + m_I) = 4m_C$ when $m_I = m_C$. Only m_C changes the displayed eligibility sweep; it is not a coverage sweep over all five matched-bank choices. At $m_C = m_I = 400$, the 1600 extra responses refer to the two shared scale banks for an arm-level audit, in addition to its audit rows. The zero-availability statement refers to the tested values 50 and 100; no continuum sweep of degrees of freedom was performed. These are response-free evaluations of the noncentral-F power formulas for the frozen designs, margins, budgets, and power floor. Thus 400 is neither a universal requirement nor the minimum for any availability: nine slots qualify at 200.

E6 evaluates 31 eligible groups in 17 arm families at $m_C = m_I = 400$, using 50,000 replications per family. Static Bonferroni gives pooled conditional noncoverage 0.0470 (99% interval $[0.0464, 0.0476]$). The groupwise intersection–union/Holm rule gives 0.0471 ($[0.0465, 0.0477]$). Mean arm-family any-release probabilities are 0.9457 and 0.9706, respectively. The two multiplicity modes are applied to the same static-screened designs and draws; this comparison does not rerun the mode-specific prospective-design selection for Holm. It therefore isolates a multiplicity-mode change on these objects, rather than the full possible mode-dependent design comparison. Shared scales can create dependence across reports; the declared Bonferroni allocation does not require those reports to be independent. These results support calibration for the simulated scale contract and expose the cost of adequate prospective power, but do not show that suitable pure-error sources exist in the original datasets.

Availability including response-free exclusions. Table E.9 combines eligibility and release, using only the existing E6 ledger at the inner planning points. Each of 17 eligible arms contributes 50,000 simulated families; the other three arms contribute structural zeros. The all-arm mean gives each of the 20 frozen arms equal weight. Released-group occurrences divided by $61 \times 50{,}000$ give the all-slot fraction, including ineligible slots; division by $31 \times 50{,}000$ gives the fraction among eligible slots. These are different denominators, and the all-arm mean does not imply that simulations

Table E.8: Unknown-noise measurement cost. The prospective power gate is re-evaluated in the unknown-noise branch, where the cutoff is a noncentral-F quantile with m_C denominator degrees of freedom. Extra responses are an illustrative matched-bank cost of $2(m_C + m_I) = 4m_C$ with $m_I = m_C$ at each swept point; the eligibility sweep varies m_C only. At the tested values $m_C = 50$ and 100, no group is eligible.

Branch	m_C	Extra responses	Eligible	Availability
known	–	–	41/61	67.2%
unknown	50	200	0/61	0.0%
unknown	100	400	0/61	0.0%
unknown	200	800	9/61	14.8%
unknown	400	1600	31/61	50.8%
unknown	800	3200	37/61	60.7%

At the frozen $m_C = m_I = 400$: noncoverage 0.0470 under static Bonferroni and 0.0471 under groupwise IU/Holm, against a 0.05 budget.

were run for the three ineligible arms. Groups share scales within each arm, so we do not attach a binomial interval that treats all slot events as independent.

Table E.9: Simulated unknown-noise availability on the same static-screened objects. Both modes assume the independent scale supply; physical scale acquisition is not measured.

Quantity	Static	IU/Holm
Eligible slots / all slots	$31/61$	$31/61$
Nonempty simulated families	803,808	825,052
Any-report mean, 17 eligible arms	0.945656	0.970649
Any-report mean, all 20 arms	0.803808	0.825052
Released-group occurrences	1,352,063	1,455,452
Released fraction, 31 eligible slots	0.872299	0.939001
Released fraction, all 61 slots	0.443299	0.477197

Every simulated eligible family receives an independent inference scale. The experiment measures neither the probability of obtaining valid replicate pairs nor the frequency of physical measurement or source-validation failures. Full practical availability is therefore not identified by these results.

E.8 Reproducibility and scope of the evidence

Reproduction. The experiment archive contains the frozen protocol, statistical core, proposer adapters, numerical ledger, and display generators. Its `submission.py` entry point verifies the package, audits numerical quantities, renders existing results, or runs the full campaign into a fresh external output directory. Installation and exact commands are given in its run guide; rendering does not rerun the experiments.

Design and response random streams use separate hash-derived namespaces. The baseline uses `CERTIFYPAR/FIG2-ICLR/design/v1` and `CERTIFYPAR/FIG2-ICLR/response/v1`; the first eight SHA-256 bytes of slash-joined namespace/labels define each integer seed. Seeds for row roles, router fitting, residual directions, and splits are fixed independently of response streams. The evidence here uses one response namespace. Same-seed replays check computational reproduction, not new statistical confirmation.

The baseline was generated in a Linux container with Python 3.11.15, NumPy 2.4.4, SciPy 1.17.1, scikit-learn 1.8.0, Matplotlib 3.10.9, linear-tree 0.3.5, and pyparc 2.0.4. The recorded Windows 11/Python 3.12.10 execution was a replay. A later independent report records full same-seed reproduction on the baseline platform and regeneration of all 15 display files. Its raw replay ledger and logs were not supplied; this evidence remains attributed to the retained report. The reported numerical results and displays retain the baseline values and the verified reporting corrections.

What is established and what remains open. The experiments check Gaussian calibration and its availability costs on one frozen design per arm. They do not assess real-response coverage, heteroskedastic or non-Gaussian robustness, repeated-training variability, prediction accuracy, large-scale performance, or the scientific usefulness of the margins. Small Monte Carlo error does not replace design diversity. The controls are selection diagnostics, not published selection-adjusted inference competitors. Splitting changes information and potentially the finite-design target; full-design replicates preserve the target at extra cost. Unknown-noise inference additionally requires justified independent pure-error information. These boundaries remain necessary for any application interpretation.

E.9 Residual nonrejection on the same frozen designs

A comparator with valid coefficient inference. The significance controls in E1 inspect coefficient-bearing information. A conventional residual lack-of-fit test addresses a closer question: should reporting continue when the exact affine model is not rejected (Draper & Smith, 1998)? We compare this nonrejection rule (NR) with CertifyPAR's static rule (C) on the same 41 previously eligible known-noise groups, retaining their original targets, audit rows, cell ranks, and C's multiplicity allocations. Both rules are restricted to this C-screened population; this is not a whole-pipeline availability comparison over all 61 slots.

With $\nu_g = N_g - d = r_{H,g} + r_{A,g}$, let $q_g = G^{-1}_{\nu_g,0}(0.95)$. NR reports the pooled coefficients when $\|(I - P_g)Y_g\|_2^2/\sigma^2 < q_g$. Both C and NR use only residuals for selection, so Lemma 1 preserves the same known-noise projection intervals under complete-pattern conditioning. NR tests exact affinity, while C tests the declared adequacy margins. No additional responses, model fits, or Monte Carlo draws are used here.

What the nominal levels control. NR's 0.05 level controls rejection at exact affinity; C's allocated axis levels control false certification at the corresponding inadequacy boundaries. The exact-affinity reporting probabilities, 0.950 for NR and mean 0.969 for C, therefore reflect different calibrations. Equating an NR level to a C axis allocation would match neither these reporting probabilities nor the outside-contract error. Matching exact-affinity retention alone would require the design-specific NR level $1 - p_g^{\mathrm{C}}(0,0)$, still a different criterion from matching false certification. We compare the prespecified rules at common audit resources, rather than treating their nominal levels as interchangeable. The asymptotic separation in Proposition 34 holds at every fixed NR level.

Exact reporting probabilities. For any prescribed departures (h, a), the noncentral laws give

$$p_g^{\mathrm{NR}}(h,a) = G_{\nu_g, N_g(h^2+a^2)}(q_g), \qquad p_g^{\mathrm{C}}(h,a) = G_{r_{H,g}, N_g h^2}(c_{H,g}) G_{r_{A,g}, N_g a^2}(c_{A,g}),$$

where $c_{u,g}$ is the retained known-noise static cutoff. The product for C uses the independent H/A numerators; it would not apply to F ratios sharing an estimated scale. Assigning the same (h, a) to all eligible groups in an arm gives $p_{\mathrm{any}} = 1 - \prod_g (1 - p_g)$ by disjoint-group Gaussian independence. Empty families contribute zero. If (h, a) lies outside the strict rectangle, any report in that homogeneous scenario is an outside-contract report. NR does not attach an affirmative adequacy claim to its report.

The calculation retains every original E2 axis point (12 levels on each of two axes), plus the joint inner point $(0.2, 0.2)$: 1,025 group records, 500 arm records, and summaries over all 20 arms and separately the 17 with eligible groups. Table E.10 selects the exact-affine, inner, and two boundary points; all curves and full-precision records are included in the separate analysis package.

Retention and boundary behavior. At the inner planning point, C's mean group-report probability is 0.934132, versus 0.807008 for NR; the corresponding any-report means are 0.988211 and 0.938044. C therefore retains more reporting at this adequate departure while giving a margin-based interpretation. This is not a universal yield or optimality comparison.

The comparator also controls outside-contract reporting at 0.05 on this particular screened roster. Indeed, any outside group has total noncentrality at least $N_g \min(\tau_H^2, \tau_A^2)$. Monotonicity of the noncentral chi-square lower CDF bounds its NR reporting probability by the single-axis boundary value b_g. Independent-group aggregation bounds each arm's outside reporting by $1 - \prod_g (1 - b_g)$;

Table E.10: Exact-law comparison at matched known-noise audit resources. Group probabilities have equal weights over 41 eligible groups; any-report probabilities have equal weights over the 17 arms containing those groups. The first two rows are adequate; the last two are outside the strict contract by boundary equality. Values are analytical probabilities, not Monte Carlo estimates or interval-noncoverage rates. C denotes CertifyPAR and NR denotes residual lack-of-fit nonrejection.

	Mean group-report probability		Mean any-report probability	
(η_H, η_A)	C	NR	C	NR
$(0, 0)$	0.969296	0.950000	0.995029	0.987044
$(0.2, 0.2)$	0.934132	0.807008	0.988211	0.938044
$(0.75, 0)$	0.008404	0.010165	0.020169	0.024416
$(0, 0.75)$	0.008680	0.010165	0.020829	0.024416

the largest is 0.049815. For C, the largest homogeneous H- and A-boundary any-report probabilities are 0.024338 and 0.024993. Thus these designs show higher inner-point retention with C, not an NR error violation. NR's roster-specific bound is established by this additional calculation; its usual exact-affinity test level is not a general adequacy calibration for other margins or designs.

Why more data separate the reporting questions.

Proposition 34 (Asymptotic separation of equivalence and nonrejection)**.** *Consider a sequence of eligible frozen group designs under the Gaussian known-noise model, with $N \to \infty$, positive residual ranks $r_H, r_A \geq 1$ satisfying $r_H + r_A = N - d \leq N$, fixed margins $\tau_H, \tau_A > 0$, and fixed static axis levels $\alpha_H, \alpha_A \in (0, 1)$. Let C release exactly when both known-noise equivalence gates pass at these levels, and let NR retain the exact-affinity test at any fixed level $\alpha_{\mathrm{NR}} \in (0, 1)$ (0.05 above). At fixed departures $(\eta_H, \eta_A) = (h, a)$ satisfying $h, a \geq 0$, $0 < h^2 + a^2$, $h < \tau_H$, and $a < \tau_A$, the group-report probabilities obey*

$$p^{\mathrm{NR}}(h, a) \longrightarrow 0, \qquad p^{\mathrm{C}}(h, a) \longrightarrow 1.$$

Proof. Put $\nu = N - d$ and $\delta^2 = h^2 + a^2 > 0$. For fixed α_{NR}, the central NR cutoff is $\nu + O(\sqrt{\nu} + 1)$, while its statistic has mean $\nu + N\delta^2$ and variance $2(\nu + 2N\delta^2) = O(N)$. Chebyshev's inequality gives nonrejection probability tending to zero. For C, each fixed squared-margin gap $N(\tau_u^2 - \eta_u^2)$ dominates $\sqrt{r_u + N\tau_u^2}$; Theorem 20, part 3, gives both axis pass probabilities tending to one. A union bound on the two axis-failure events then gives the joint C limit. □

Thus a sufficiently precise exact-affinity test withholds a fixed nonzero departure even when it satisfies the declared adequacy margins; the equivalence gates eventually retain it.

At exact affinity NR instead reports with probability $1 - \alpha_{\mathrm{NR}}$ at every N. The contrast above concerns a strictly inner point, not margin equality, and is a theoretical design sequence, not an additional experiment or a monotonic finite-sample claim. The static axis allocations remain fixed as N grows; the proposition makes no claim for C or NR levels that shrink with N, including a growing slot family or an NR level matched to C's exact-affinity retention. It also does not apply automatically to unknown noise with a fixed certification-bank size: the independent-scale uncertainty remains. Proposition 24 states a sufficient unknown-scale regime when the bank size also grows.

Timing and reproducibility. This exact-law comparison was added during manuscript review after the E and F outcomes were available. Its separate protocol was recorded before computing the group curves; a subsequent recorded amendment added family aggregation after the initial group summaries were inspected. Both stages and every specified result are retained. The producer has its own identity and reads the original unmodified E ledger; it reproduces all 984 stored E2 exact probabilities. An independent Poisson/gamma-mixture evaluation checks the noncentral CDF calculations. This is a deterministic operating-law extension, not another response campaign, a new independent dataset, or a modification of the original E results.

Published selective-inference procedures provide conceptual context and are not empirically evaluated here. For a CART-region mean, Tree-values (Neufeld et al., 2022) conditions its Gaussian

pivot on selection of the region and the orthogonal response projection; its computationally efficient variant also conditions on the region's ancestors. The k-means procedure of Chen & Witten (2023) tests equality of two selected clusters' mean vectors, conditioning its pivot on all intermediate assignments, the orthogonal data projection, and the contrast direction. These are different targets and selection events from fixed-router affine-coefficient reporting with adequacy gates. Table E.11 is consequently a map of this paper's comparisons, not a benchmark of those published procedures.

Table E.11: Operational choice among reporting workflows. "Extra" counts physical responses beyond one response on each audit row; a pair bank with m contrasts uses $2m$ responses. Conditional coverage means familywise coefficient coverage given the frozen design and every positive-probability complete reporting pattern, under the stated Gaussian and scale contracts. Adequacy error control is before release-pattern conditioning.

Workflow	Release claim	Coefficient target	Scale and extra responses	Conditional coverage
C, known σ	Separate H/A equivalence margins	Full-audit θ_g°	Frozen known σ; no extra responses	Yes, by residual independence
C, unknown σ	Separate H/A equivalence margins	Full-audit θ_g°	Independent S_C, S_I; $2(m_C + m_I)$ pair responses	Yes, by residual/scale independence
Pooled equivalence	Total-departure equivalence; the benchmark uses the sufficient inner disk	Full-audit θ_g°	Same known- or unknown-scale contract as C	Yes, by the same independence
NR: residual nonrejection	Exact-affinity nonrejection; no affirmative margin claim	Full-audit θ_g°	Known σ in this paper; no extra responses	Yes, by residual independence
S: row split	H/A margins only on the gate half	Fit-half $\theta_{g,F}^\circ$, not the full-audit target	Fit residuals can give a conservative own-target interval; the gate still needs justified scale	Yes for its own target, by independent fit errors
R: replicate fit	C's release rule on the original audit responses	Full-audit θ_g°	One additional response per audit row, plus C's scale contract	Yes, by independent replicate errors

Table E.11 separates the reporting questions. The known-noise C–NR comparison holds audit rows, projection targets, and interval allocations fixed, and both retain conditional coverage. It therefore isolates margin-based release versus exact-affinity nonrejection, including their different retention probabilities. NR's outside-contract bound is checked separately for the screened roster in Section E.9. NR is evaluated only with known noise. The original unknown-noise F comparison gave C, S, and R the same two banks; R additionally used a full-design response replicate (Table F.2). This isolates audit-row reuse but is not the least-cost split. The matched-total-response study therefore also gives S its own response-free resolution search and residual-based, own-target interval scale while charging discovery and every repeat role.

Predeployment checklist. These are scientific inputs, not outcome-adaptive tuning choices.

1. **Target and comparison.** Decide whether all audit rows must define the estimand. If a fit-half projection is acceptable, S can avoid a separate inference bank by using a conservative fit-residual interval under the stated Gaussian model. Use pooled equivalence when total departure is the desired adequacy requirement.
2. **Margins and resolution.** State tolerable departure in scientific units, justify normalization by σ, and freeze cell counts and the response-free selection rule. Values 0.75 and 0.20 are not defaults.

3. **Geometry.** Before responses, verify the chosen rule's rank, condition-number limits, prospective power, requested-coordinate precision, and any required cell sizes. Failure means ineligibility, not poor local fit.
4. **Noise acquisition.** For unknown noise, document how equal-mean, independent, homoskedastic repeats produce S_C and, when needed, S_I. Duplicate covariates alone do not establish equal means; audit-fit residuals do not validate a misspecified lower-tail adequacy gate.
5. **Interpretation.** Freeze the full slot family, multiplicity, and report map. Release licenses only the stated finite-design projection; non-release is inconclusive, and an empty eligible set remains a limitation.

F LEARNED-ROUTER EXPERIMENTS AND MEASURED CALIBRATION

This appendix evaluates the complete discovery-to-reporting workflow, its response cost, and its behavior when the noise model is only a working reference. It adds three kinds of evidence to Appendix E: repeated training and auditing on generated designs, a two-dimensional amplitude sweep, and measured-response reporting examples. The synthetic studies have known finite-design targets and adequacy truth. The SGEMM runtime and PLATERO fluorescence studies have real responses, but unknown adequacy truth; their intervals and gates remain working-model references. All original settings, follow-up amplitudes, and measured outcomes are retained.

F.1 PROTOCOL, COMPARISON WORKFLOWS, AND SUMMARIES

Designs and learned partitions. For each setting in Table F.1, we generate 20 independent discovery/audit designs. Coordinates are independent uniform variables on $[-1, 1]$; discovery and audit rows are independent. Discovery responses add independent $N(0, 0.2^2)$ errors to the specified mean. A linear tree and PARC share each setting/design's data, producing $6 \times 20 \times 2 = 240$ learner/design objects from 120 setting-specific data designs. The learners are paired conditions, not independent datasets. Conditional on each learned router and audit design, the core campaign generates 2,000 response repetitions with independent certification and inference pair banks. The same response and bank streams are used across the two learners within a setting/design. Every synthetic coefficient family requests the intercept and all raw-coordinate slopes.

The linear-tree proposer uses linear regression leaves, maximum depth 2, minimum leaf size 81, and 25 candidate bins (Cerliani, 2022). PARC requests four regions, with regularization 10^{-5}, at most 100 iterations, softmax separation, stopping tolerance 10^{-4}, and ten k-means initializations (Bemporad, 2023). Its input standardization is fitted only on discovery covariates. Only the learned hard routing rule is retained: audit coefficients are newly fitted on the frozen audit inputs. Both proposers use a four-slot namespace, including empty or unavailable slots. The statistical comparisons below hold the proposer fixed; they are not prediction comparisons between CertifyPAR and these partition learners.

Table F.1: Original synthetic settings. $n_D = n_R$ is the number of rows in each independent discovery and audit role, not their combined size. Every row uses noise SD 0.2, 20 generated designs, two paired learners, and 2,000 audit-response repetitions.

Setting	Mean $\mu(x)$	p	$n_D = n_R$
S1	$0.5 + x_1$	1	2,048
S2	$\lvert x_1 \rvert$	1	2,048
S3	$x_1 - \mathbf{1}\{x_1 \le 0\}$	1	2,048
S4	$(\lvert x_1 + x_2 \rvert + 0.5(x_1 - x_2))/\sqrt{2}$	2	2,048
S5	$0.75(x_1^2 + x_2^2)$	2	2,048
S6	$\lvert x_1 \rvert + 0.5 \sum_{j=2}^{8} x_j/\sqrt{7}$	8	4,096

Audit contract. The common benchmark sets $\alpha_{\text{cert}} = \gamma_{\text{FWER}} = 0.05$, $\tau_H = \tau_A = 0.75$, inner planning points $\tau_H^- = \tau_A^- = 0.20$, and marginal prospective-power floor 0.80. Each candidate has two balanced cells, constructed without responses by splitting the widest-range coordinate at its median with stable-ID tie breaking. This is the minimum cell count supporting an H axis; we do not evaluate a cell-count sweep. Full and cellwise augmented designs must have full column rank, positive residual ranks, condition number at most 10^{12}, and requested-coordinate diagonals of the unit-noise pooled covariance at most 0.50. A candidate needs at least 90 rows. Failure of these requirements is recorded before opening audit responses.

The default banks each contain 400 disjoint pairs. Synthetic pair members have equal means and independent errors, making the required pure-error construction exact. Prospective screening deliberately uses the more conservative axis level $0.05/(2 \times 4) = 0.00625$. Actual release uses groupwise intersection–union p-values $p_g = \max(p_{H,g}, p_{A,g})$ and strict Holm step-down over all four slots; unavailable or unpowered slots receive p-value one. Thus the stored planning cutoffs are not necessarily the thresholds used by the observed Holm decision. For each nonempty reporting pattern, a

Bonferroni interval family divides the 0.05 budget uniformly across its requested coordinates, using only S_I and t_{400} quantiles.

What the four workflows compare. Table F.2 specifies the response-use alternatives. C (the two-axis rule) and U share the same full-group OLS estimator and predictions. U omits prospective-power and response-dependent adequacy screening while retaining the full audit geometry requirements for this controlled comparison; ordinary ungated OLS need not impose these extra cell requirements. R uses exactly C's reporting pattern and an independent response replicate at every full-design input for its fit. The fixed sample-splitting baseline S follows the established principle of separating selection and estimation (Rinaldo et al., 2019; Athey & Imbens, 2016). Within each frozen full-design cell, alternate rows are assigned to gate and fit halves using the deterministic row order. Each half must pass its own applicable geometry requirements; the gate half must also pass the prospective-power screen. Failed splits are not retried.

S audits the gate-half mean and estimates the fit-half projection. Its own-target coverage is the primary interval assessment; scoring those same intervals against θ_g° on the full group is a separate target-mismatch diagnostic. No adequacy transfer to the fit half or full group is claimed. These four workflows isolate response reuse, target changes, and report selection; they are not matched implementations of all selective-inference procedures in Section 2.

Table F.2: Response use of the implemented workflows, not minimum possible costs. n_R counts all audit rows in an object, including rows in unavailable groups. Costs include audit and used bank measurements, but exclude discovery and secondary prediction-assessment responses. At the default bank sizes, $2(m_C + m_I) = 1{,}600$ and $2m_I = 800$.

Workflow	Release and target	Fit responses	Measurements
C: two-axis rule	Full-group adequacy; full-group projection	Original audit rows	$n_R + 2(m_C + m_I)$
U: ungated	No adequacy claim; full-group projection	Same fit as C	$n_R + 2m_I$
R: fresh replicate	Same release and target as C	Independent full-design replicate	$2n_R + 2(m_C + m_I)$
S: sample splitting	Gate-half adequacy; fit-half projection	Fixed fit half	$n_R + 2(m_C + m_I)$

The S row retains C's two-bank convention to isolate row separation; it does not establish that splitting requires both banks. An independent fit half's residual variance gives conservative intervals for its own projection under an arbitrary fixed Gaussian mean (exact under affinity). Keeping the certification bank but using this interval scale would cost $n_R + 800$; inference without an adequacy certificate can use n_R. Neither alternative was run here. A residual scale contaminated by lack of fit does not by itself calibrate the lower-tail adequacy gate. The observed 1.45 design factor does not determine either alternative's random interval width.

Estimands and uncertainty. A false-certification event is a repetition with at least one C report whose true finite-group H or A departure lies outside the declared rectangle. For U, the analogous outside-contract reporting frequency is descriptive: U makes no adequacy claim. Conditional familywise noncoverage counts repetitions with at least one missed requested coefficient, divided by repetitions with a nonempty report. This averages over observed nonempty patterns and does not empirically verify every individual pattern. Reported-slot fractions use all four slots, and reported-row fractions use all audit rows. In the S comparison the latter refers to the original groups whose labels are reported. A zero denominator is NA, not zero.

Plots retain individual-design points and their pointwise 95% Clopper–Pearson intervals (Clopper & Pearson, 1934); diamonds are equal-design means over defined rates. Those intervals quantify response Monte Carlo uncertainty within a fixed learned design, not uncertainty across designs or simultaneous uncertainty over the panels. General mean half-widths average the procedure's issued coordinates and can involve different reporting populations. R and C have identical widths by construction because their pattern, full design, and inference bank coincide. The additional S1 width

comparison uses only paired C/S reports of the same group and coordinate in a repetition, retaining their actual pattern-dependent critical values.

F.2 Original synthetic campaign: reporting, targets, and resource costs

The one-dimensional atlas in Figure F.1 shows how released affine summaries relate to smooth, kinked, and discontinuous mean functions. Figure F.2 retains the original oblique and quadratic settings. Both atlases use design 0 and response draw 0, fixed throughout the display recipe. The status maps extend a frozen group label to a plotting grid; their colored areas do not assert pointwise validity on the continuum. In particular, a globally quadratic response does not require within-cell failure in every learned region: the audited quantities are the actual noise-relative finite-group departures at the chosen resolution. H/A failure labels identify the first failed Holm step; subsequent unreleased eligible slots are marked "Holm blocked." Detailed response-free ineligibility reasons remain in the ledger.

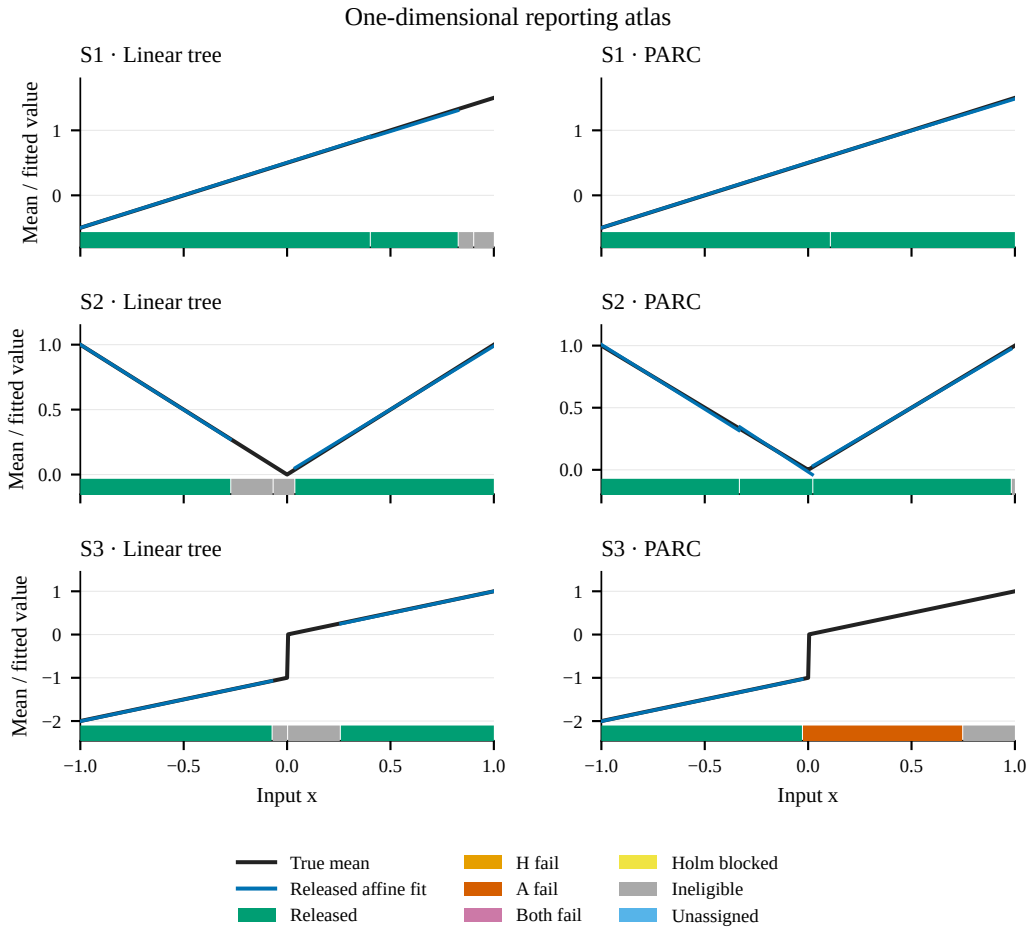

Figure F.1: Original one-dimensional reporting atlas for S1–S3, with linear-tree and PARC routers. Black curves show the declared means; colored fitted segments and status strips show the saved first-draw outcome on design 0. Status strips distinguish release, statistical withholding, and response-free ineligibility. No whole-function confidence band or continuum guarantee is shown.

Coverage and useful reporting are distinct outcomes. Across the twelve setting/learner combinations, C's equal-design conditional-noncoverage means range from 0.03298 to 0.04818. Its mean false-certification frequency is zero in eleven combinations and 0.00030 in S3/PARC. Of 734 geometry-qualified full-design groups, 18 are outside the contract, all in S3/PARC on 18 designs.

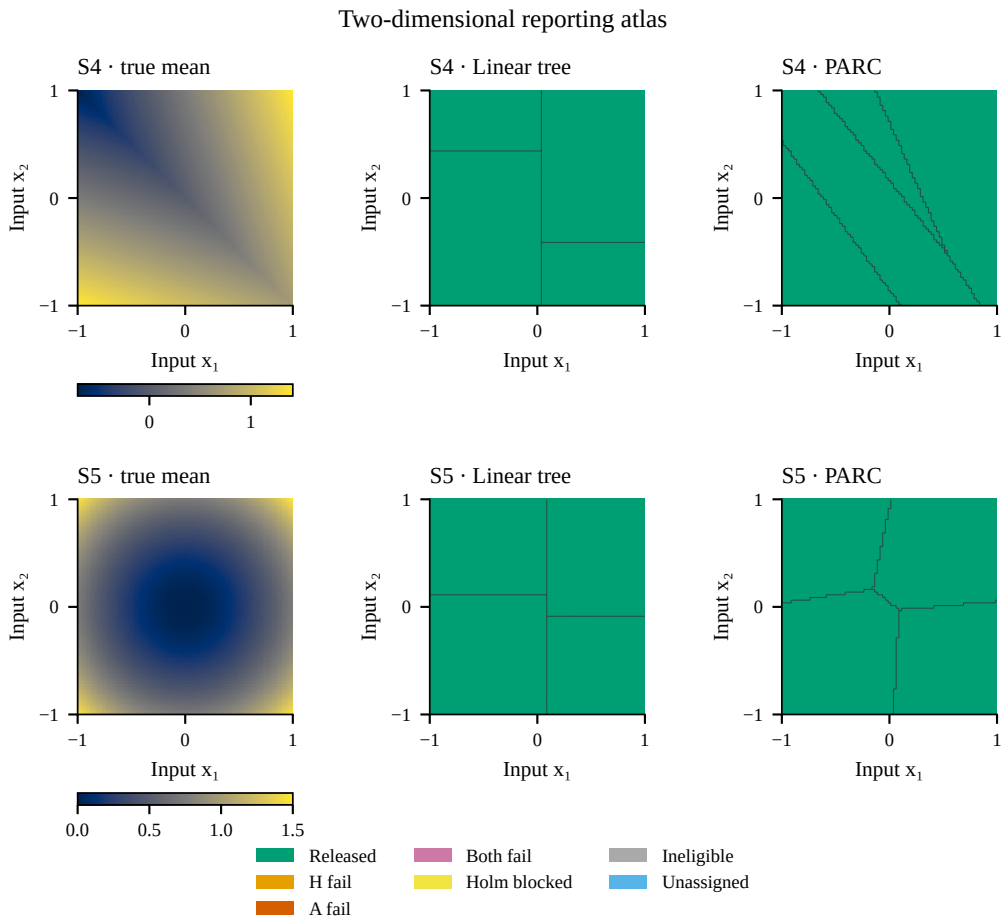

Figure F.2: Original two-dimensional atlas for S4 and S5. The first column shows the true mean; the remaining columns show the saved reporting status for each learner on design 0, draw 0. Local H/A departures, geometry, and prospective power determine the result, not the global description "piecewise affine" or "quadratic." All original outcomes are retained alongside the additive amplitude sweep in Section F.5. Boundaries follow adjacent changes in categorical router labels; slot numbers are not interpolated.

Thus the other eleven combinations contain no outside-contract groups and cannot produce false certification; Appendix F.5 supplies the learned-router outside-margin stress test. In S3/PARC, U reports at least one outside-contract group on 90% of the designs' repetitions on average. C reports 47.00% of audit rows there, compared with U's 92.35%: the gate changes which projection coefficients are presented with an adequacy qualification. This is a reporting distinction, not a predictive advantage, since their common fits are identical.

Splitting changes both availability and the target. For S4 with a linear tree, S's equal-design noncoverage is 0.04265 for its fit-half target and 0.06031 for the full-group target. The corresponding S5 values are 0.04431 versus 0.05966 for the tree and 0.04426 versus 0.05552 for PARC. These comparisons directly exhibit the target change; they do not show invalidity of sample splitting for its own target. In exactly affine S1 the targets coincide, and the mean across designs of the paired common-report S/C half-width ratio is 1.3899 for the tree and 1.4143 for PARC. General width averages in Figure F.4 should not be interpreted as matched efficiency ratios. R preserves C's full-design target and width while costing another n_R responses: 5,696 versus 3,648 measurements in S1–S5.

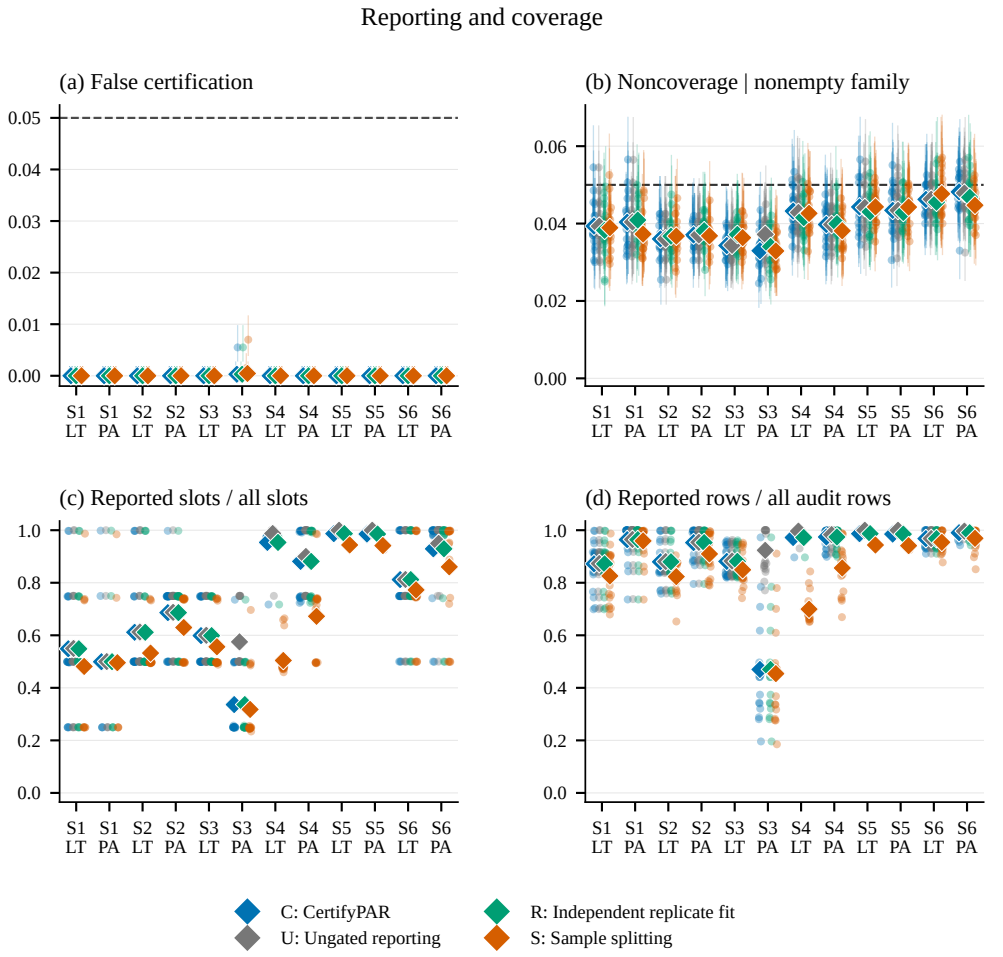

Figure F.3: Original synthetic comparison: certification, coverage, and reporting. Panels (a)–(d) show false certification, conditional interval noncoverage, issued slots, and issued audit rows. Points are individual designs; diamonds are equal-design means; light intervals in (a,b) are pointwise per-design 95% Clopper–Pearson intervals. Dashed lines mark the 0.05 error budget. C/U share coefficients; C/R share release and half-widths. Panel (a) omits U, which makes no adequacy claim; its outside-contract reporting rate is given separately in the text. LT denotes linear tree; PA denotes PARC. Missing conditional rates remain NA.

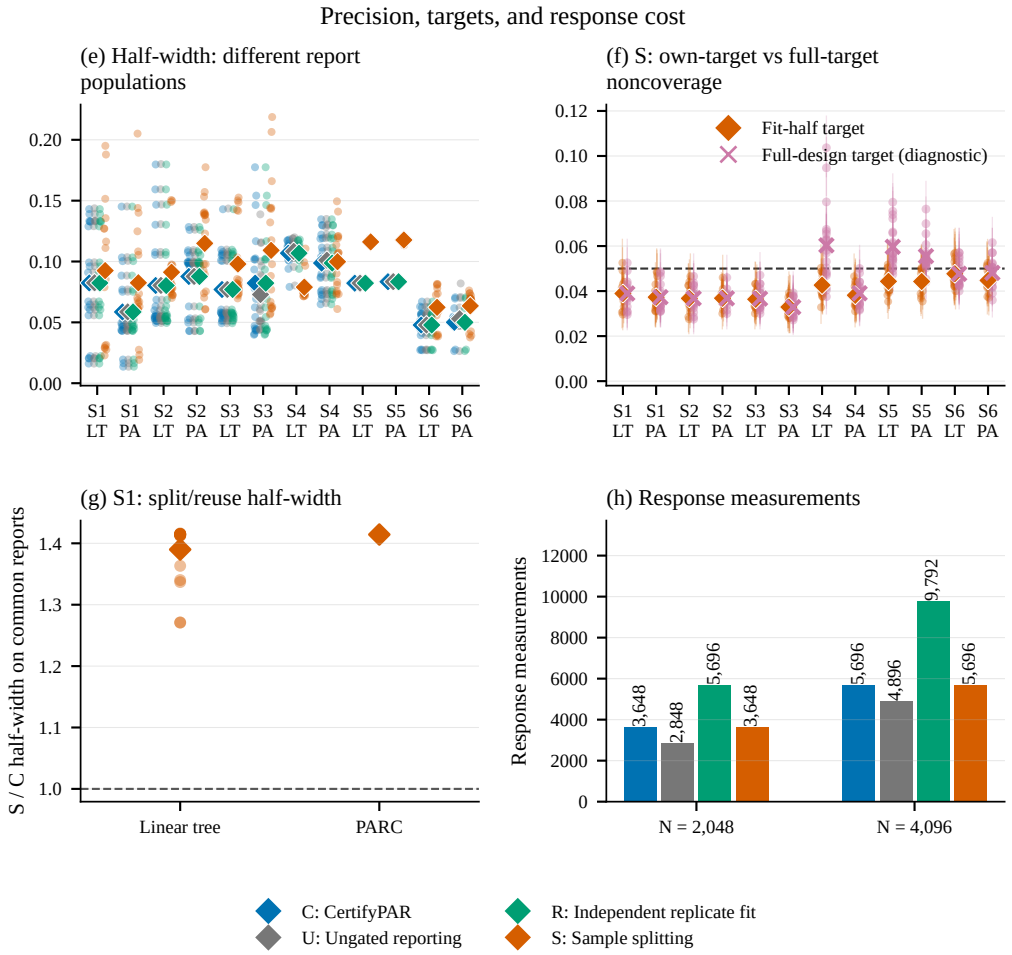

Figure F.4: Original synthetic comparison continued: precision, targets, and response cost. Panels (e)–(h) retain the descriptive widths, split-target comparison, paired S1 widths, and response measurements. Points are per-design summaries; diamonds are equal-design means over defined values. Width populations differ. S is scored against its fit-half target except where the full-design target is explicitly labeled; its interval bars are pointwise per-design 95% Clopper–Pearson intervals. S1 ratios use paired C/S reports of the same group and coordinate. Response counts include audit and bank measurements and R's extra audit responses; discovery and prediction assessment are excluded. The dashed line in (g) marks equal half-widths. LT denotes linear tree; PA denotes PARC.

F.3 SGEMM: MEASURED LACK OF FIT RELATIVE TO REPEAT PRECISION

The SGEMM GPU-kernel dataset contains 241,600 input configurations, fourteen configuration parameters, and four measured runtimes per configuration (Paredes & Ballester-Ripoll, 2017). The recorded role allocation uses 10,000 development configurations, 4,096 discovery configurations, 8,192 audit configurations, 8,192 secondary test configurations, and 400 configurations per pure-error bank. Roles are disjoint at the configuration level. All four runtimes are used in development, the first runtime is used in discovery, audit, and test, and the first two runtimes form each bank pair. This accesses 62,080 measurements in total, including 40,000 development measurements.

Responses are natural-log runtimes. Ten positive configuration parameters are encoded on a $\log_2$ scale; four binary parameters retain their original coding. The fourteen features are standardized using discovery covariates, and the two discovery-trained routers use the same four-slot protocol. Requested coefficients concern MWG, NWG, VWM, and VWN; displayed slopes are transformed back to natural-log runtime per unit of the corresponding $\log_2$ input. They are finite-design associations, not causal tuning effects or recommendations for the fastest configuration. Both routers' complete patterns and pooled fits are frozen before the inference bank is parsed.

The certification-bank estimate gives $S_C = 0.01448$ log units, about 1.45% relative timing variation. On the five eligible groups, observed H/A residual magnitudes relative to the 0.75 margins range from 22.12 to 50.30 (Table F.3). The larger contribution is A, the within-cell residual component. These are the raw plug-in quantities $\widehat{\eta}_{u,g} = \|Q_{u,g}Y_g\|/(\sqrt{N_g}S_C)$, including response noise; they are not bias-corrected estimates or confidence bounds on the unknown mean departures. All five groups are withheld by C, whereas U reports their requested slopes. The remaining three slots fail response-free requirements. Saturated lower-tail p-values of one are retained, together with the observed statistics, so withholding is not reported as an unexplained zero yield.

Table F.3: Measured SGEMM groups with eligible C designs. Both axis p-values are one in every row. The final column is a descriptive magnitude-to-margin ratio, not a lower confidence bound on real-data misspecification. Slot numbers preserve the frozen roster.

Router	Slot	N_g	F_H	F_A	$\widehat{\eta}_H, \widehat{\eta}_A$	$\max_u \widehat{\eta}_u/0.75$
Linear tree	0	1,338	6,083.40	281.56	$8.26, 16.59$	22.12
Linear tree	2	1,618	4,412.88	1,049.59	$6.40, 32.10$	42.79
Linear tree	3	4,354	13,803.85	1,432.78	$6.90, 37.72$	50.30
PARC	2	1,035	2,526.74	911.11	$6.05, 29.74$	39.66
PARC	3	2,445	1,442.10	721.23	$2.97, 26.69$	35.59

After release freezes, the bank comparison gives $S_C^2 = 2.0971292 \times 10^{-4}$, $S_I^2 = 1.3894439 \times 10^{-4}$, and ratio 1.50933. Its two-sided $F_{400,400}$ reference p-value is 4.14×10^{-5}. This is evidence against the joint common-Gaussian-pair reference, not a robust test identifying a specific failure mechanism, a conditional-on-release p-value, or an extra gate. The development strata also show substantially different repeat variances. Consequently, Figure F.5 demonstrates the recorded reporting decision and its scale-relative discrepancies; it does not establish calibrated real-data coverage. The practical importance of these particular runtime slopes remains unestablished.

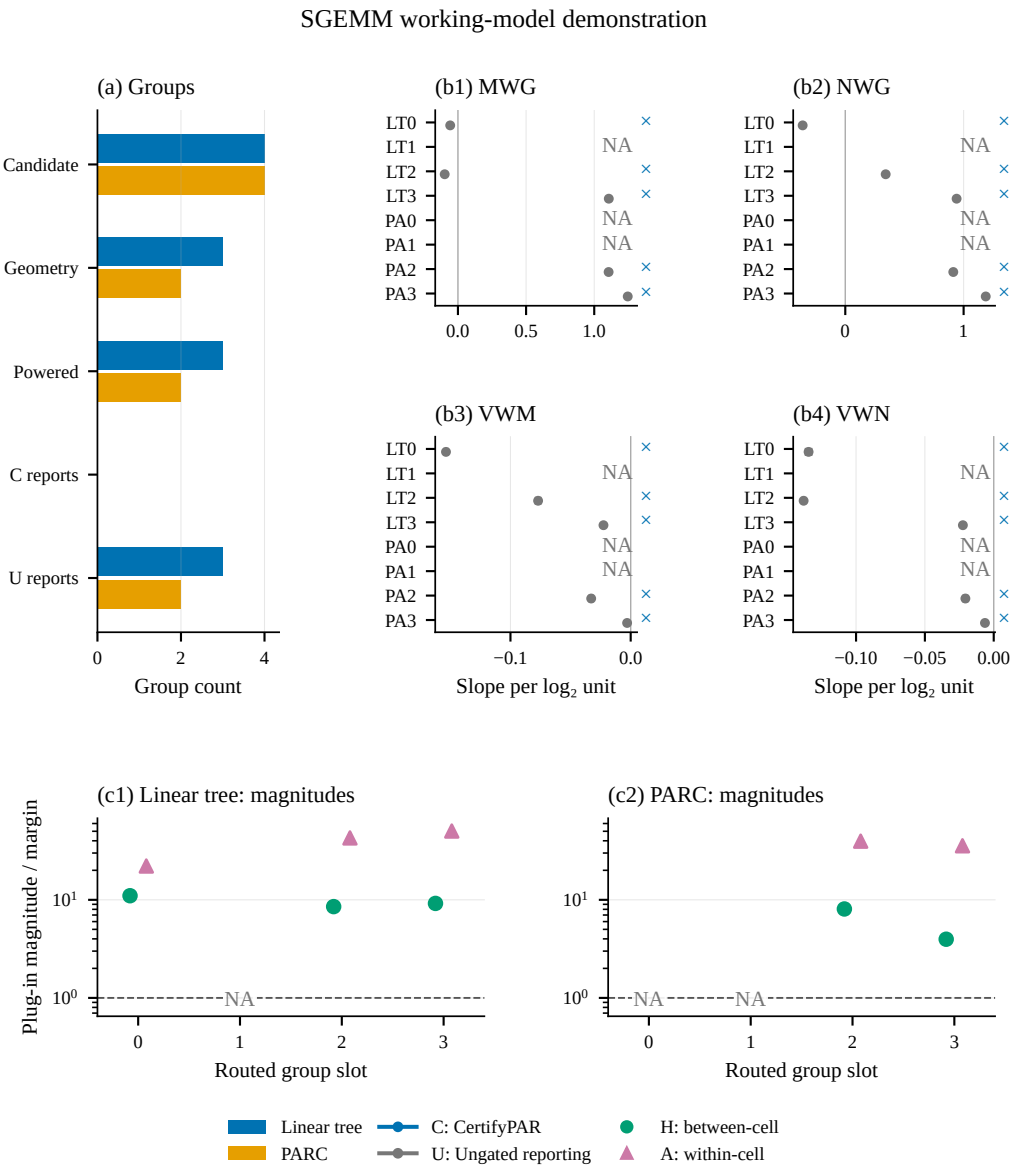

Figure F.5: Measured SGEMM working-model analysis. Candidate, geometry-qualified, powered, and reported group counts retain the complete roster. Ungated coefficient intervals are finite-design association summaries in encoded input units. The magnitude panels show the observed H/A components relative to the fixed benchmark margins. C withholds all five eligible groups; this is not an empirical coverage result or evidence of an improved kernel-tuning decision. LT denotes linear tree; PA denotes PARC; digits denote group slots. A cross marks C withholding; NA marks an unavailable interval or undefined geometry. Plug-in magnitudes are noisy descriptive estimates, not adequacy bounds.

F.4 NOISE-BANK BUDGETS AND DEPARTURES FROM THE REFERENCE MODEL

The sensitivity block reuses the first five designs of S1 and S4 for both learners. It evaluates $(m_C, m_I) \in \{(100, 400), (400, 100), (400, 400), (800, 400), (400, 800)\}$ with 2,000 response repetitions per condition. Eligibility is recomputed when m_C changes. At $m_C = 100$, every object in this block withholds all slots because the prospective-power requirement is not met; conditional noncoverage is consequently undefined. Increasing the inference-bank size changes interval construction, whereas increasing the certification-bank size can change eligibility and reporting. For example, in S4/PARC the mean reported-slot fraction rises from 0.7901 at $m_C = 400$ to 0.84575 at $m_C = 800$, holding $m_I = 400$.

The same sensitivity designs also receive two prespecified stress variants. The heavy-tailed variant replaces Gaussian errors, including pair-member errors, by $\sqrt{3/5}\, t_5$ errors scaled to SD 0.2. The heteroskedastic variant uses the SGEMM development data described above: four repeated log runtimes estimate each configuration's sample variance, averaged within the four fixed MWG strata. The stratum weights partition the synthetic x_1 quantiles; their variances are then normalized to audit-average variance 0.2^2. Each synthetic pair draws one shared scale from that profile, with independent errors for its two members. Average bank and audit variance match by construction, but the common-variance premise does not. The estimated profile is held fixed during this stress simulation.

Figure F.6 separates these synthetic outcomes from their measured variance-profile input. For C, the four setting/learner noncoverage means are 0.0097–0.0316 under this heteroskedastic profile and 0.03733–0.04089 under the heavy-tailed variant. The heteroskedastic example is conservative on this reported metric and also reduces reporting; it does not reproduce the direction of every counterexample in the theory. Neither favorable error frequencies nor matching average variances extend the exact theorem to these error laws. This block is a conditional stress study, not an estimate of coverage on SGEMM responses.

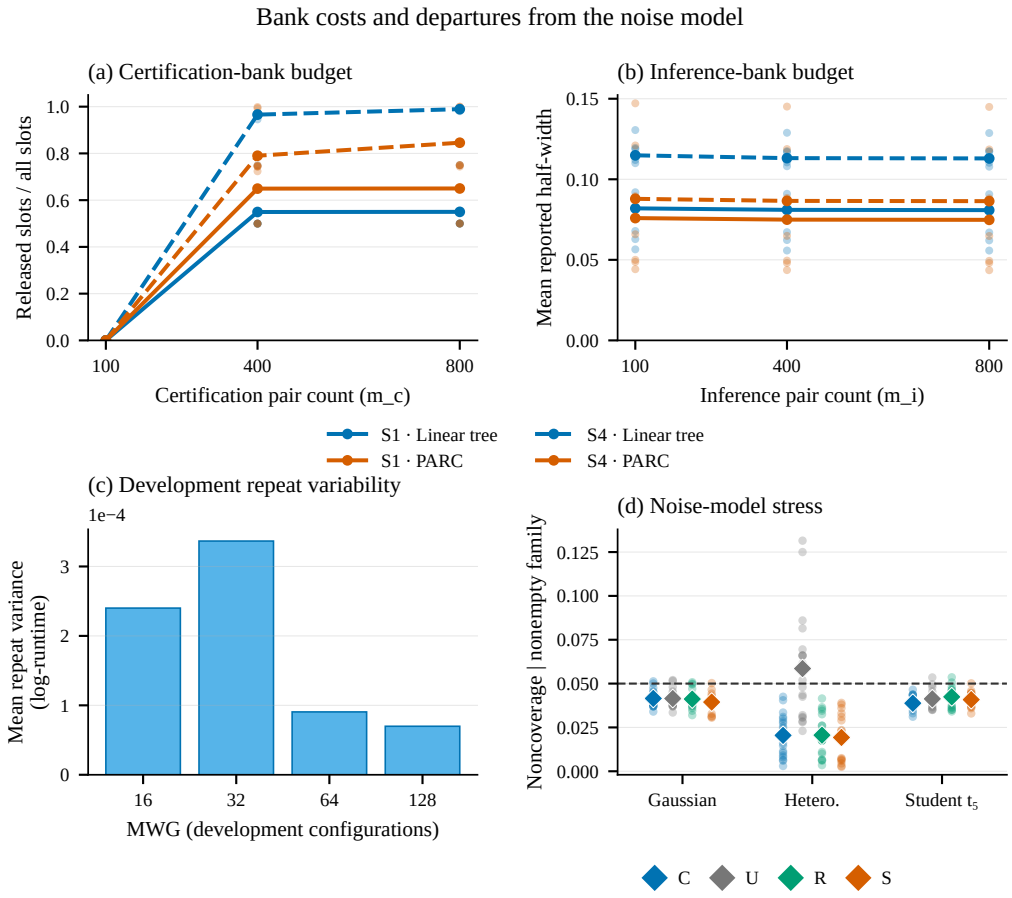

Figure F.6: Bank resources and model stress. Certification-bank sizes change prospective eligibility and yield; inference-bank sizes change interval uncertainty. The measured SGEMM development profile supplies the four variance strata used in the synthetic heteroskedastic stress. Stress non-coverage is computed on simulated responses, with five designs per setting/learner; no-report cases are NA. Displayed reference levels are nominal Gaussian-model budgets, not guarantees under the stress laws.

F.5 TWO-DIMENSIONAL FOLLOW-UP WITH ALL AMPLITUDE LEVELS RETAINED

The original S5 example left a useful question: how does reporting change when *actual local* departure, rather than a global curvature label, becomes large? An additive follow-up uses

$$\mu_a(x) = 0.5x_1 - 0.25x_2 + a\max(x_1, 0)^2(1 + x_2^2), \qquad a \in \{0, 0.75, 1.5, 3, 6, 12\}. \tag{F.1}$$

It retains the same contract, learners, two cells, $\sigma = 0.2$, and 400-pair banks, with 2,048 discovery and 2,048 audit rows. There are 20 independent base designs; inputs, discovery errors, audit errors, banks, and learner initialization streams are paired across amplitudes, with inputs/errors/banks also paired across learners. The $20 \times 6 \times 2 = 240$ objects are therefore repeated conditions on 20 base designs. Each receives 2,000 response repetitions. Oracle η_H and η_A are used only to evaluate the already frozen groups.

This follow-up was designed after inspecting the original two-dimensional results. It is not an untouched confirmatory test or a replacement chosen to conceal the original outcomes. Both atlas files retain all six amplitudes using base design 0 and response draw 0 (Figures F.7 and F.8). Within each row, the mean map and status maps describe the same saved design; the true-mean color scale varies across amplitude rows.

The contract separates reporting decisions. At amplitudes 0, 0.75, and 1.5, no geometry-qualified group lies outside the oracle adequacy rectangle. Nevertheless, release can decline as adequate groups approach its edges: at $a = 1.5$, the equal-design release fraction among powered adequate groups is 0.75055 for the tree and 0.87993 for PARC. Adequacy does not promise high release probability near its boundary.

At $a \geq 3$, every design under both learners has an outside-contract U report. C's equal-design false-certification frequencies remain zero for the tree and are 0.000600, 0.000075, and zero for PARC at $a = 3, 6, 12$, respectively. The tree continues to report about 54% of audit rows; PARC reports 27.24%, 14.61%, and 12.12% at the same levels. This illustrates a consequential partition effect: stronger global curvature need not eliminate every adequate finite group, and the two learners need not preserve the same reporting domain. It is not a uniform dominance comparison between learners.

Table F.4 and Figure F.9 retain both error and yield denominators. Across the twelve amplitude/learner cells, the equal-design conditional-noncoverage means range from 0.03690 to 0.04638. At PARC amplitudes 6 and 12, only five of twenty designs produce a nonempty report, so those conditional means describe five defined rates; the other fifteen remain NA. Replacing their missing rates by zero would spuriously improve the displayed coverage.

Table F.4: All two-dimensional conditions. Entries are equal-design means of per-design rates. C/U outside denotes any outside-contract report in a repetition; U's column is descriptive. Row fraction uses all audit rows. NC is conditional familywise noncoverage when at least one group is reported. Parentheses give the number of designs with a defined NC rate out of 20; all other displayed means use all 20 designs.

a	Router	C outside	U outside	C row fraction	C NC (defined)
0	Linear tree	0	0	0.91131	0.04331 (20)
0	PARC	0	0	0.95868	0.04638 (20)
0.75	Linear tree	0	0	0.92442	0.04124 (20)
0.75	PARC	0	0	0.97107	0.03946 (20)
1.5	Linear tree	0	0	0.77946	0.03995 (20)
1.5	PARC	0	0	0.91864	0.04139 (20)
3	Linear tree	0	1	0.54769	0.04300 (20)
3	PARC	0.000600	1	0.27244	0.03690 (20)
6	Linear tree	0	1	0.54533	0.04293 (20)
6	PARC	0.000075	1	0.14613	0.03956 (5)
12	Linear tree	0	1	0.54324	0.04270 (20)
12	PARC	0	1	0.12119	0.04319 (5)

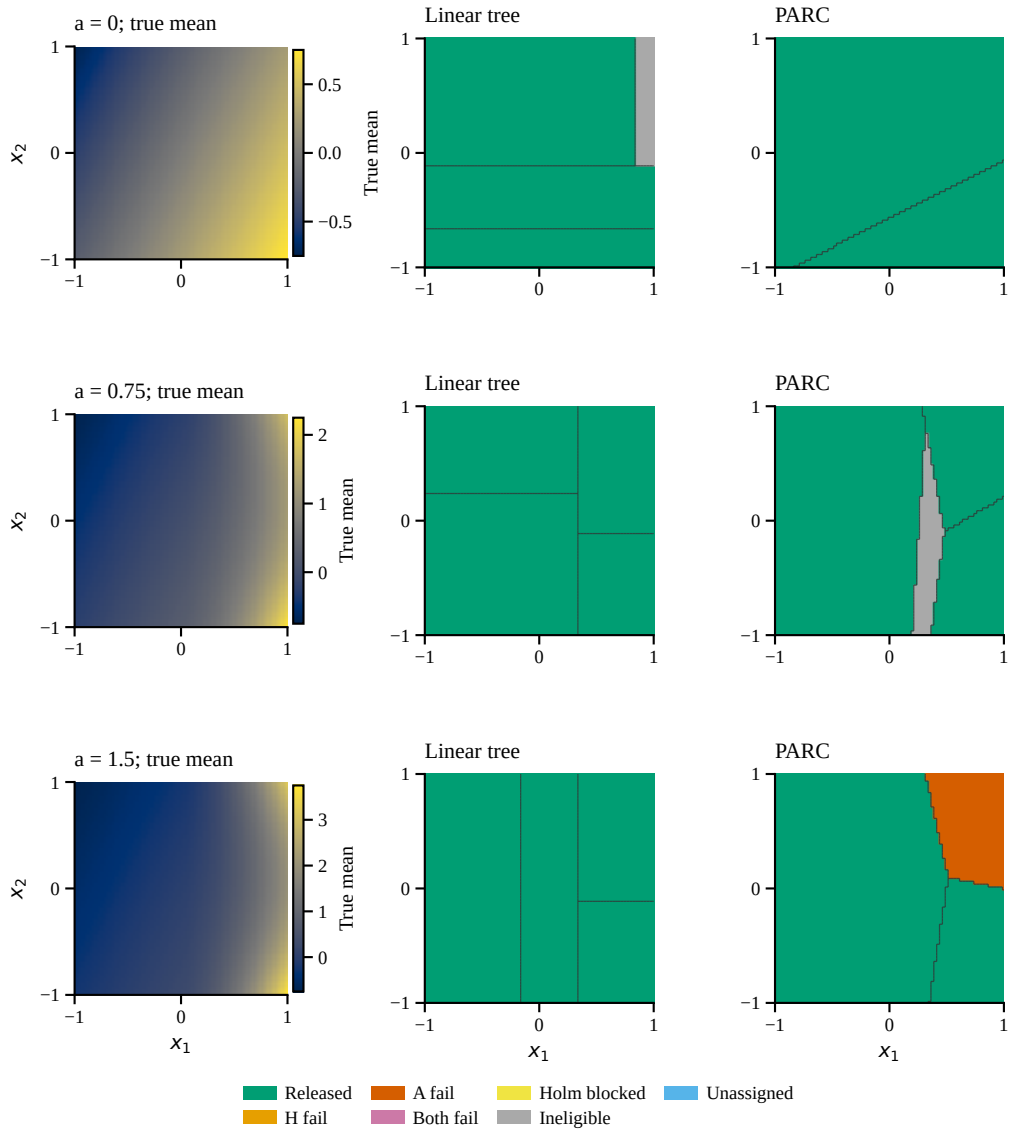

Figure F.7: Two-dimensional follow-up at amplitudes 0, 0.75, and 1.5. All panels use the fixed base design 0 and response draw 0. Statuses refer to finite-audit groups, with colors extended to a plotting grid by the learned router. The true-mean scales vary by row. Mild global curvature can coexist with local departures inside the unchanged H/A margins.

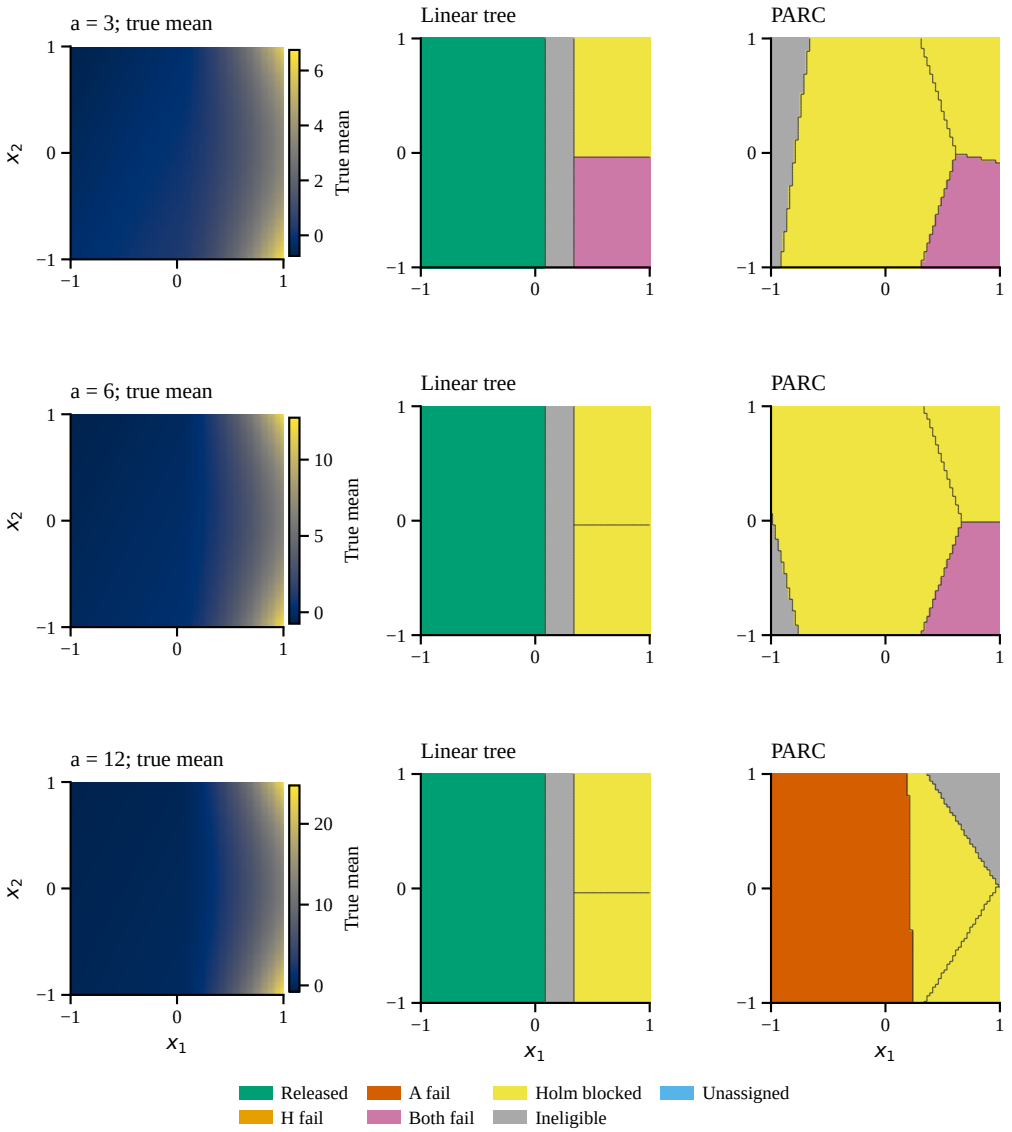

Figure F.8: The same saved-design atlas at amplitudes 3, 6, and 12, retained together with all lower levels and the original campaign. Changes in reporting reflect each learner's realized finite-group geometry, departures, and statistical decisions. These maps do not assert that every point in a colored region satisfies the contract.

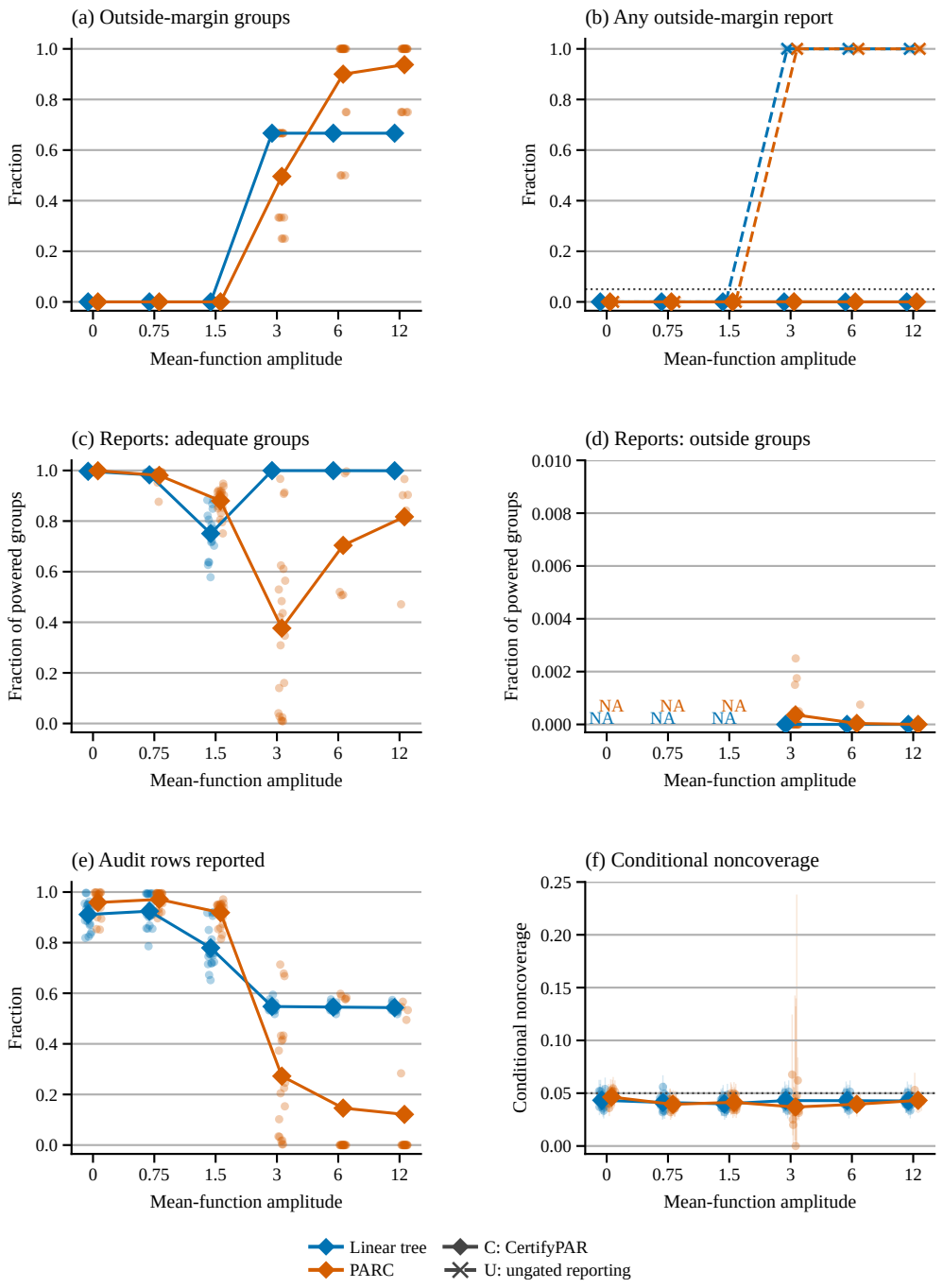

Figure F.9: Full two-dimensional amplitude sweep. Panels distinguish oracle geometry, outside-contract family reports, C reporting among powered adequate and outside groups, reported audit rows, and conditional noncoverage. Small points are base designs; diamonds average defined per-design rates. Light bars in (b, C) and (f) are pointwise per-design 95% Clopper–Pearson intervals, not across-design intervals. In (c,d), each design's denominator is powered groups times response draws. The certification budget applies to C. U asserts no adequacy guarantee. Amplitude and learner conditions are paired, and empty denominators remain NA rather than receiving zero error.

F.6 PREDICTION RISK AND ADEQUACY ANSWER DIFFERENT QUESTIONS

For every available saved draw-zero fit from the full amplitude sweep, we compute

$$R_g = \left(\sigma^2 + N_g^{-1}\|\mu_g - A_g\widehat{\theta}_g\|_2^2\right)^{1/2}. \tag{F.2}$$

This is the root expected squared prediction error for a fresh response at those same fixed audit inputs, conditional on the saved fit and known synthetic noise law. It is not an empirical held-out RMSE, the expectation of a sample RMSE, or a new response campaign. Orthogonal projection gives

$$\frac{R_g^2}{\sigma^2} = 1 + \eta_{H,g}^2 + \eta_{A,g}^2 + \frac{\|A_g(\widehat{\theta}_g - \theta_g^\circ)\|_2^2}{N_g\sigma^2}. \tag{F.3}$$

Thus prediction risk combines irreducible noise, both adequacy departures, and coefficient-estimation error. A single predictive score does not encode the two-axis rectangle or prospective eligibility. The axes in Figure F.10 are algebraically related, so the figure explains this distinction rather than providing independent validation. C and U have identical risk on common inputs.

Interpreting and choosing the margins. Write $e_g = \|A_g(\widehat{\theta}_g - \theta_g^\circ)\|_2^2/(N_g\sigma^2)$. When the true group satisfies both margins, equation (F.3) gives $R_g/\sigma < \sqrt{1 + \tau_{H,g}^2 + \tau_{A,g}^2 + e_g}$. For the oracle projection $e_g = 0$; the benchmark pair $(0.75, 0.75)$ then allows total structural RMS departure just below $\sqrt{1.125}\sigma \approx 1.06\sigma$, a structural squared-error contribution below $1.125\sigma^2$, and an oracle root expected prediction error below $\sqrt{2.125}\sigma \approx 1.458\sigma$. The fitted predictor retains the estimation term inside the square root. This interpretation concerns fresh responses at the same fixed inputs; the certification error budget still allows false releases.

An application should first specify which aggregate departures matter and the measurement precision relative to which they are judged. Separate margins allow different limits on disagreement between cellwise descriptions and departure within cells; their squared sum sets the corresponding total structural-error ceiling. The risk identity explains that ceiling but does not determine its scientific acceptability or how to allocate it between axes. Our 0.75 choices are research benchmarks. Under unknown noise the protocol requires a standardized margin fixed in advance; an observed bank cannot be used to retune it into a desired absolute-unit tolerance. A larger true σ permits a larger absolute departure at the same standardized margin, so a change of instrument or replication unit changes the meaning of the qualification as well as its acquisition design.

The record contains all 960 slots: 734 have available fits, including 13 that fail prospective power; 226 are unavailable (70 empty, 33 below the row minimum, and 123 failing the unit-noise variance bound). The first draw releases 402 available slots and withholds 332. Of the latter, 270 lie outside the oracle margins and 62 lie inside. No outside-margin slot is released in that saved draw, a descriptive outcome that is not an independent estimate of false-certification probability. The 2,000-draw release frequencies remain separately available in the complete table. Slot, learner, and amplitude records are dependent observations on 20 base designs.

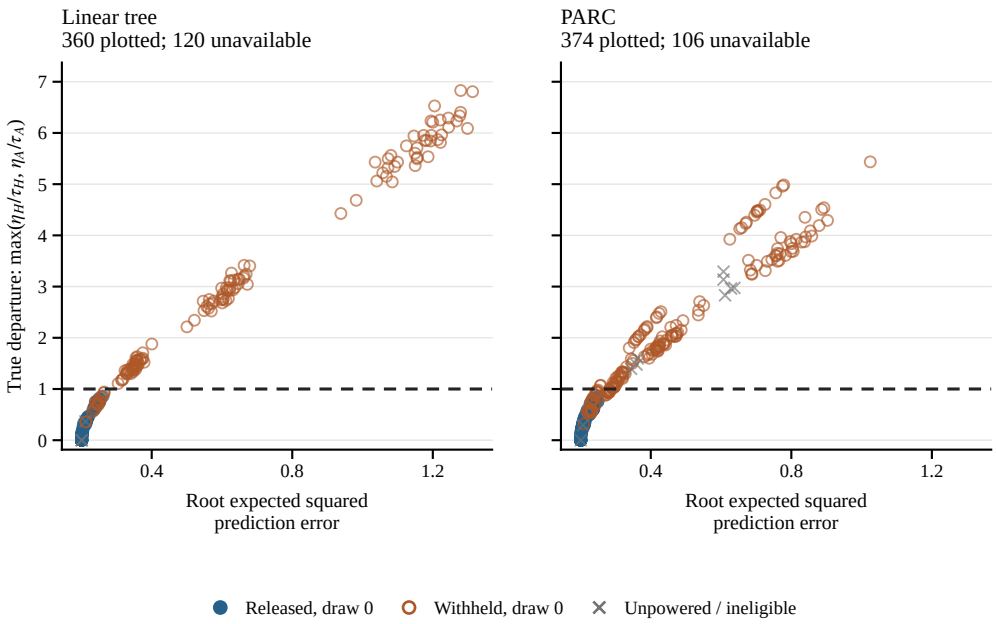

Figure F.10: Analytic fixed-input prediction risk versus oracle departure for saved first-draw fits. The dashed horizontal boundary is the declared two-axis adequacy threshold; markers record first-draw C decisions. There are 734 plotted fits among 960 retained slots, including available unpowered fits. The risk incorporates known fresh-response noise and is not empirical test RMSE. Equation (F.3) links the axes; C/U predictions coincide. The plotted fits represent 463,724 audit-row occurrences across 20 independent base designs; groups and amplitudes are dependent. No risk is assigned to an absent fit.

F.7 PLATERO: A FORWARD-CALIBRATION REPORT WITH EXCELLENT IN-SAMPLE FIT

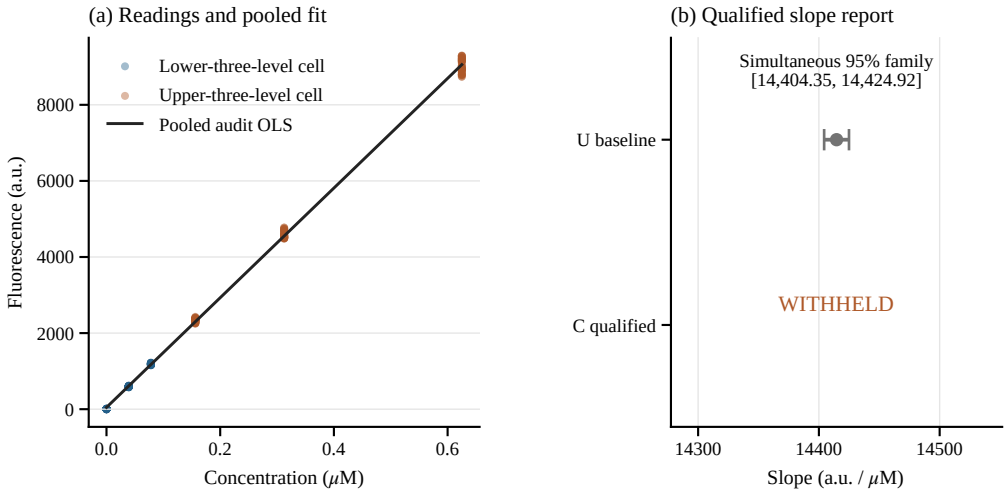

Figure F.11: A measured calibration report changes under the adequacy contract. The same pooled fit has in-sample $R^2 = 0.99921$. U reports a slope interval; C withholds the qualified report after both gates fail. The slope interval is from a simultaneous 95% intercept/slope family; no whole-curve band is shown. True adequacy is unknown and the exact-noise premises are unverified. Most A residual variation is between wells at the same concentration.

Reporting question and fixed case. Fluorescence calibration gives the coefficient an explicit measurement meaning: a forward affine slope expresses fluorescence units per nominal concentration unit. We use the PLATERO experimental data and accompanying calibration study (González-Cebrián et al., 2023; 2022), fixing the first documented reader workbook, `PlateReader1.xlsx`, at gain 50. The planned question is whether one pooled raw-fluorescence line can receive the declared adequacy qualification over the six sampled concentrations. This case evaluates the reporting contract on a fixed group, not a newly learned partition. Its lower-three-level and upper-three-level cells provide the H/A resolution.

The levels are 0, 0.039063, 0.078125, 0.15625, 0.3125, and 0.625 μM, with 16 physical wells per level and eight readings per well. The 96 wells supply 768 readings. Within each level, wells are ordered by plate row and column; the eight-position role sequence (D, A, A, A, C, C, I, I) is cyclically shifted by the well index modulo eight, using each shift twice. Each well supplies three audit readings and one same-well pair to each bank. D supplies one reserved reading per well and is opened only for post-freeze diagnostics; it does not select the model. This counterbalanced allocation yields 288 audit readings, 192 C readings, 192 I readings, and 96 D readings. Audit repetitions do not create 288 independent wells, and all roles are used, leaving no untouched response-validation set.

Scale, cells, and opening order. Responses remain on their raw scale, with a free intercept and no estimated blank subtraction, log transform, outlier removal, gain search, or post-outcome cell search. The single group has pooled rank 2, separate-cell rank 4, $r_H = 2$, and $r_A = 284$. Each cell has 144 readings from 48 wells. Both banks contain 96 same-well pairs, with normalized differences as in equation (3). The margins remain 0.75, the inner planning points 0.20, and the marginal power floor 0.80. Condition-number and raw-coordinate covariance-diagonal caps are 10^8 and 0.50, respectively. With one group, both lower-tail axes use level 0.05; the group p-value is their maximum. The completion implementation uses $p_g \leq 0.05$ at the endpoint, rather than the strict convention in the general algorithm; both recorded p-values equal one, so this endpoint distinction does not affect the result. The interval family requests the intercept and slope, using

a simultaneous 95% Bonferroni construction with $t_{96,0.9875}$. The margins and precision caps are research benchmarks, not externally validated laboratory acceptance limits.

The program reads metadata, then audit and certification values, freezes the concrete pooled fit and release decision, and only then reads inference values. D values and all-response diagnostics are accessed after this freeze. Guarded data access enforces this ordering in the supplied program; it does not establish independence of the physical measurements. Table F.5 records the resulting decision.

Table F.5: The single PLATERO calibration case. Statistics and interval endpoints are working-model references on the measured design. Axis powers are marginal powers at the declared inner alternatives, not the joint release probability. U issues the candidate two-coordinate family; C issues neither coordinate.

Quantity	Recorded value
Physical wells / total readings	96 / 768
Audit readings / C pairs / I pairs	288 / 96 / 96
Response-free eligibility	All declared checks pass
Prospective power, H / A	0.9999999965 / 0.807897
Observed F_H / H cutoff	563.98117 / 57.38578
Observed F_A / A cutoff	29.96470 / 1.21141
Lower-tail p-values, H / A	1 / 1
Pooled intercept / slope	46.74973 a.u. / 14,414.63623 a.u./μM
U slope interval	[14,404.34877, 14,424.92369] a.u./μM
C issued coefficient intervals	None: eligible group withheld
Audit in-sample R^2 / residual RMSE	0.99921136 / 86.84377 a.u.
S_C^2 / S_I^2	225.36458 / 270.328125 a.u.^{2}

What changes in the report. Figure F.11 displays the common audit fit and the issued-report comparison. Its in-sample $R^2 = 0.99921$ coexists with both H and A statistics exceeding their release cutoffs, by factors 9.83 and 24.74. U reports a narrow forward-slope interval from the same fit and inference bank; C withholds an adequacy-qualified coefficient report. The R^2 and residual RMSE are post-return descriptive checks on the same audit responses, not predictive validation. Withholding does not make the finite-design OLS projection false, validate an inverse-calibration procedure, or show that a downstream laboratory decision was improved.

The measured case also does not isolate an advantage over ordinary residual nonrejection. From the same recorded sums of squares and certification scale, the pooled lack-of-fit statistic is

$$F_T = \frac{\|(I-P)Y\|^2/(N-d)}{S_C^2} = \frac{2F_H + 284F_A}{286} = 33.69908.$$

Its central $F_{286,96}$ working-reference 95th percentile is 1.33161, so conventional residual nonrejection would also withhold. This calculation uses the existing record; it neither adds a response experiment nor validates the noise premises. The example's contribution is the measured contrast between aggregate fit, the declared qualification, and the source of the residual discrepancy.

The dominant discrepancy is between wells, not curvature. To interpret the A-axis result, let $\bar{y}_w$ be a well's three-audit-reading mean and $\bar{y}_c$ the mean over all audit readings at its concentration. The observed A residual sum of squares splits into within-well residuals, differences of well means within concentration, and departures of concentration means from their cell lines. These orthogonal, weighted components sum to 1,917,847.1151 a.u.^{2}. Between-well differences at the same nominal concentration account for 1,851,817.3542 a.u.^{2} (96.56%); within-well audit variation contributes 56,222 (2.93%); departures of concentration means from their separate cell lines contribute 9,807.7609 (0.51%). This is an algebraic description of observed variation, not a causal variance decomposition or an identified random-effects model. The principal contrast is a common affine description across wells against the much narrower within-well repeat-reading scale. Large curvature is not the main explanation supported by these data.

Why the exact-model claim remains conditional. The banks estimate repeat-reading variability, not total preparation uncertainty. Pooling normalized C/I pair differences within each concentration gives root mean square values of approximately 3.20–31.73 a.u.; the individual-bank values plotted by concentration range from 2.78 to 34.49 a.u. The read-order diagnostics show additional structure (Figure F.12). The aggregate ratio $S_C^2/S_I^2 = 0.83367$ has a two-sided $F_{96,96}$ reference p-value of 0.37439. This aggregate agreement does not establish common variance across concentration, independent readings, stable well means, Gaussian errors, or accurate nominal concentrations. Repeated-reading roles are disjoint, but their statistical independence is a scientific assumption. No diagnostic is used to revise the frozen gate or intervals.

The source describes measurement-procedure differences, including mixing between readings. The operational workbook header is `F:G50:485,528`; the accompanying paper reports different wavelength numbers. We retain the exact workbook/column identity without inventing a reconciliation. These limitations bound the empirical conclusion: an excellent aggregate fit statistic can accompany a withheld report under an explicitly tighter repeat-scale contract, and diagnosing the source of that mismatch is necessary before treating it as laboratory certification.

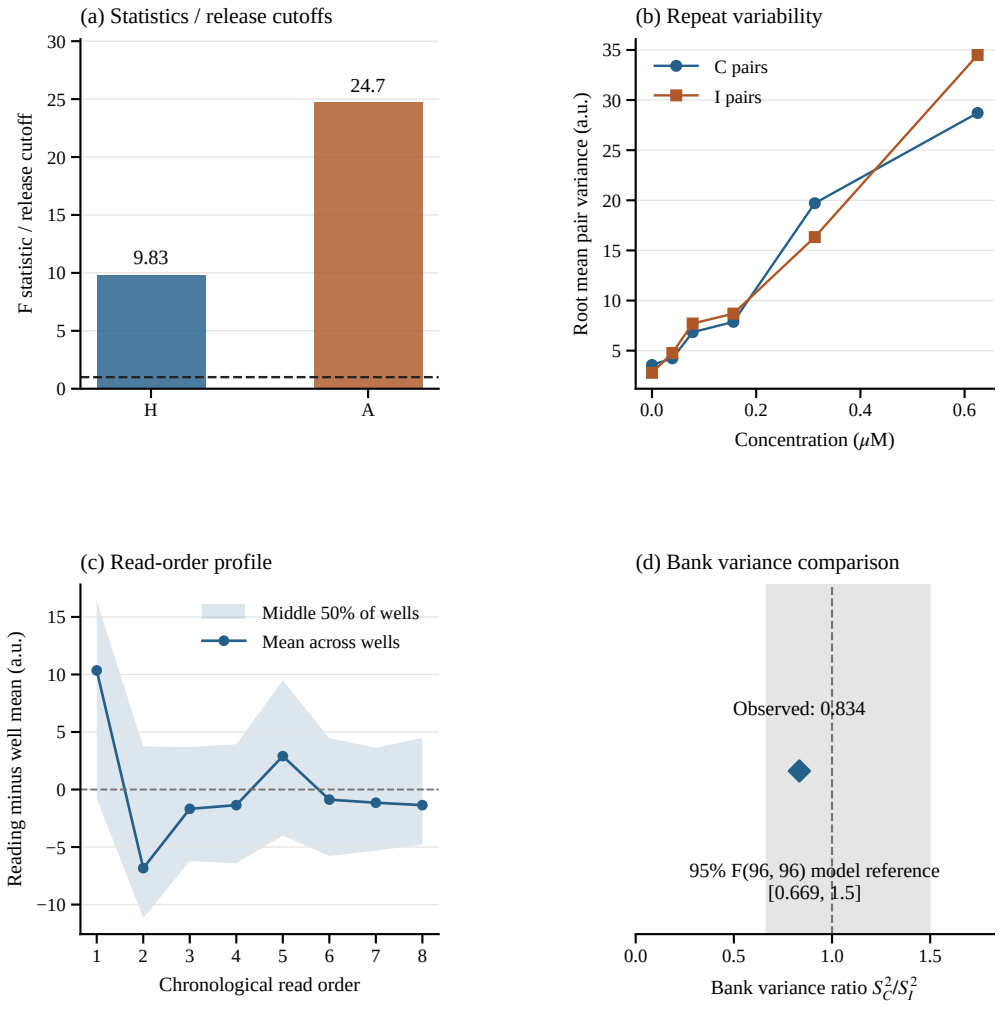

Figure F.12: Post-freeze PLATERO diagnostics. (a) Observed H/A statistics divided by their lower-tail cutoffs. (b) Concentration-specific root mean squared normalized pair differences within each bank. (c) Within-well-centered read-order means and the middle 50% of wells; shading is descriptive, not a confidence band. (d) Aggregate bank ratio against a central F reference. These checks do not alter reporting or validate the noise assumptions. The A-axis discrepancy is dominated by between-well differences at a fixed nominal concentration, as quantified in the text.

F.8 POST-FREEZE MEASURED-RESPONSE AUDITS

After the original measured examples were complete, we froze three bounded audits to ask whether the H component and regional predictive structure recur across measured units. These are retrospective replications on public data, not untouched external confirmation. They do not estimate true adequacy or coverage because the mean responses and required independent homoskedastic-Gaussian noise references are unknown. Dataset rosters, transformations, cells, role rules, metrics, and stopping criteria were fixed before each complete response analysis; every declared gain and specimen is retained.

SGEMM repeat stability. From the public SGEMM data (Paredes & Ballester-Ripoll, 2017), we excluded the 31,280 configurations used by earlier roles and fixed 100,000 remaining configurations by a covariate hash. Twenty disjoint blocks of 5,000 configurations were each split into 2,500 fit and 2,500 test configurations without using responses. Responses are log runtimes, positive configuration parameters are encoded on a $\log_2$ scale, and the four MWG levels define the cell rule fixed from the earlier analysis. The global affine model has mean test $R^2 = 0.5260$, versus 0.6802 for the MWG-cell model. The paired block difference is 0.1542 with a 95% t interval $[0.1506, 0.1579]$. These blocks are disjoint summaries from one dataset, not twenty independently acquired datasets.

The mean plug-in H and A magnitudes are 38.23 and 52.99, with paired H/A ratio 0.7217 (95% t interval $[0.7090, 0.7345]$). Requested coefficient estimates are stable over the four recorded executions, but estimated regional departures are large; repeat stability alone does not license the pooled interpretation. The MWG-specific repeat-variance max/min ratio averages 5.41, and the two disjoint diagnostic-bank variance ratio is inside the predeclared range $[2/3, 1.5]$ in only $13/20$ blocks. Run-order shifts are also detectable. The common-scale reference therefore fails, and no coefficient is certified.

PLATERO gains. The fixed gain-50 PLATERO analysis in Appendix F.7 was extended without gain selection to all four documented gains, using the same 96 wells, eight readings, role rotation, and lower-three/upper-three concentration cells (González-Cebrián et al., 2023; 2022). Gains 50 and 60 are numerically available: their H/A RMS ratios are 0.3641 and 0.4181, respectively, and both axes have lower-tail p-value one. Concentration-specific repeat-variance max/min ratios are 64.66 and 153.53 at gain 50 and 880.18 and 1004.25 at gain 60 for the two banks. Gains 70 and 80 contain $128/768$ and $256/768$ recorded `OVRFLW` cells. We mark each such gain unavailable rather than delete or impute rows. Thus the recurring H component is measured, but every declared gain is either withheld or unavailable.

BAM tensile specimens. The BAM S355 archive contains ten tensile-test traces (Schilling et al., 2022). We carried the original specimen's elongation window $[0.25, 0.65]$ percent and cut 0.45 unchanged to the other nine specimens. Within each trace, alternating positions defined fit and test roles while preserving trace order; the specimen, rather than the serial row, is the uncertainty unit. A fixed two-cell affine model improves held-out R^2 over one pooled affine model on all $10/10$ specimens. The mean difference is 0.2588 with specimen-level 95% t interval $[0.1404, 0.3772]$, and the median descriptive H/A RMS ratio is 1.9616. A quadratic model is better than the two-cell model by 0.0192 R^2 on average (equivalently, cell-minus-quadratic 95% t interval $[-0.0277, -0.0108]$). The result supports a repeatable transition and the need to audit a requested affine coefficient; it does not establish piecewise-affine predictive superiority. Serial dependence and the absence of equal-mean independent repeats preclude certification.

Together these audits show that regional coefficient change recurs in measured systems, while the assumptions needed for exact qualification can fail in observable ways (Table F.6). They do not show an H/A reporting advantage over pooled equivalence in these designs, validate the benchmark margins in application units, or turn withholding into proof that a coefficient is scientifically false.

Table F.6: Measured-response evidence added after the original campaigns. Intervals summarize blocks or specimens as indicated; they are not empirical coverage intervals. Exact reporting means CertifyPAR coefficient reporting under the stated measurement contract.

Data	Fixed positive result	Limitation for certification	Exact reporting
SGEMM	Cell-minus-global test $R^2 = 0.1542$; H/A $= 0.7217$	Repeat variance and bank compatibility fail	No
PLATERO	H/A $= 0.3641, 0.4181$ at gains 50, 60	Both available gains withheld; gains 70, 80 saturated	No
BAM	Cell-minus-pooled test $R^2 = 0.2588$ on $10/10$ specimens	Quadratic better; no independent equal-mean bank	No

F.9 HIGH-DIMENSIONAL LEARNED-REGION STRESS TEST

This additional frozen known-noise stress test uses PARC and LinearTree on a public PARC regression example (Bemporad, 2023), Friedman 1 (Friedman, 1991), and an affine positive control. Inputs are independent uniform variates. The PARC example has mean $\mu(x) = \sin\{4x_1 - 5(x_2 - 0.5)^2\} + 2x_2$ and noise standard deviation 0.1; Friedman 1 uses its standard five-active-variable mean and noise standard deviation 1. Variables appended beyond the active coordinates are irrelevant. Consequently, the experiment probes a low-intrinsic-dimensional function with nuisance inputs, not a fully interacting high-dimensional regression surface.

Each function uses ten master designs. The base discovery, audit, and test sizes are 2,048, 4,096, and 4,096. The $p = 20$ doubled-budget cells double discovery and audit sizes. Four capacity candidates (2, 4, 8, and 16 regions) are selected by discovery-only validation MSE and refit on discovery data. Audit and test responses do not enter this selection. Across 11 setting cells and two learners, all 220 final models and 110,000 audit-family evaluations are retained.

The margins are $\tau_H = \tau_A = 0.75$, with inner planning point 0.2, marginal power floor 0.8, and $\alpha = \gamma = 0.05$. Criterion A requires a design to have noisy-test $R^2 \geq 0.95$ and at least 10% of all audit rows in truly inadequate, geometry-eligible groups; a setting passes A at $3/10$ designs. Criterion B is evaluated only in A-positive designs with nonzero adequate and inadequate mass on the common power-eligible cohort. It requires removal of at least 80% of inadequate mass, retention of at least 20% of adequate mass, and a nonempty report; a setting passes at $3/10$ designs. The same B rule is applied to C (the two-axis rule), pooled equivalence, and exact-affinity nonrejection (NR). These criteria are author-fixed evidence screens, not hypothesis tests.

At $p = 5$, PARC has mean noisy-test $R^2 = 0.9682$ while 46.4673% of all audit rows fall in inadequate eligible groups. Ungated audit OLS reports 18.4839% adequate and 46.4673% inadequate mass. C reports 8.5469% and 0.0086%; pooled equivalence reports 7.5156% and 0.0069%; NR reports 1.8367% and 0.0015%. C and pooled both meet B in $6/10$ designs. All 43 inadequate eligible groups in this cell are A-only, so the result motivates affirmative adequacy gating but does not establish an empirical necessity for the two-axis split in this design.

Operational pooled specification. The producer `highdim/highdim_experiments_v1/producer/audit_core.py` uses total residual energy divided by known σ^2, degrees of freedom $N_g - d$, and boundary noncentrality $N_g \min(\tau_{H,g}^2, \tau_{A,g}^2)$. Holm at 0.05 runs over discovery-selected capacity $K \in \{2, 4, 8, 16\}$, including empty slots; ineligible p-values are one. C uses Holm on the maximum of its two axis p-values. Both use C's same response-free power cohort, screened at axis level $0.05/(2K)$, the same draws, and the same OLS. Their known-scale interval allocation is fixed over all geometry-eligible group–coordinate pairs. The sufficient pooled disk does not equal the rectangular alternative; realized selections need not be nested. This is not a strongest-baseline claim.

Retrospective design-level uncertainty. The ten frozen $p = 5$ summaries give mean C-minus-pooled adequate mass 1.0313 percentage points, design standard deviation 1.9533, and standard error 0.6177. An unadjusted 95% t interval with nine degrees of freedom is $[-0.3660, 2.4285]$. All

ten differences are positive; one is 6.5099 points and the other nine range from 0.0933 to 1.2923. This is an approximate retrospective summary on a small, skewed design sample, without selection or multiplicity correction, not new confirmatory evidence. Response repetitions are averaged within design, not counted as independent designs (Figure F.13).

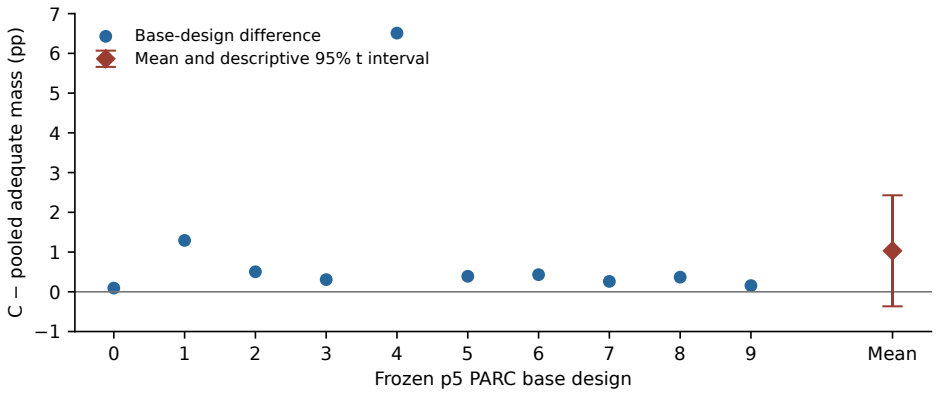

Figure F.13: All ten paired $p = 5$ design differences in adequate reported mass. The mean and descriptive interval quantify design variation, not response-level Monte Carlo error. The frozen utility criterion remains tied at $6/10$ designs.

The wider saved H/A decomposition contains five H-only and twenty joint violations among 936 geometry-eligible group-condition records. Repeated dimensions and learners share base designs. These show a distinct H truth violation, not incremental H reporting utility, and do not change the closed $p = 5$ A-only result. The absence of an observed H reporting advantage in these designs does not conflict with the analytical distinction for rectangular adequacy in Remark 28; the experiment does not test its limiting regime.

The complete decision roster is in Table F.7. Of the 22 setting–learner cells, only PARC on the PARC example at $p = 2$ and $p = 5$ passes A, and only $p = 5$ passes C's B criterion; pooled passes the same cell. The $p = 10$ PARC cell reaches A and B in $2/10$ designs, below the frozen $3/10$ threshold. All null and negative cells remain in the roster.

The measured-data extension uses grouped splits of CCPP (9,568 rows, four inputs) (Tüfekci & Kaya, 2014) and Airfoil Self-Noise (1,503 rows, five inputs) (Brooks et al., 1989). Neither dataset supplies a known mean or independently validated homoskedastic Gaussian scale, so no true adequacy, exact certification, or coefficient-coverage claim is made. None of 452 declared region slots passes the preset raw-coordinate geometry: 230 fail minimum rows, 152 raw-unit variance, 56 are empty, and 14 fail the condition cap. Descriptive pooled OLS is computable for 354 slots but is not certification evidence.

The high-dimensional archive retains the protocol, data provenance, grouped splits, seeds, candidate and final fits, masks, confidence-interval events, full tables, figures, and independent numerical checks. Three missing candidate logs were repaired by 30 isolated refits and 3,000 identical audit replays; these restore records and are not additional scientific evidence. Its actual execution and relocation verification were on Linux. Windows commands are supplied, but the high-dimensional package was not rerun on Windows.

Table F.7: All 22 frozen synthetic condition–learner cells. Entries are numbers of designs (out of ten) meeting empirical criterion A and the common criterion-B rule under the two-axis rule (C), pooled equivalence (P), and exact-affinity nonrejection (NR). B is base response budget and 2B doubles discovery and audit sizes. These are evidence-screening counts, not statistical tests.

Function	p	Budget	Learner	A	C-B	P-B	NR-B
PARC demo	2	B	PARC	3	2	2	2
PARC demo	2	B	LinearTree	0	0	0	0
PARC demo	5	B	PARC	10	6	6	2
PARC demo	5	B	LinearTree	0	0	0	0
PARC demo	10	B	PARC	2	2	2	0
PARC demo	10	B	LinearTree	0	0	0	0
PARC demo	20	B	PARC	0	0	0	0
PARC demo	20	B	LinearTree	0	0	0	0
PARC demo	20	2B	PARC	0	0	0	0
PARC demo	20	2B	LinearTree	0	0	0	0
Friedman 1	5	B	PARC	0	0	0	0
Friedman 1	5	B	LinearTree	0	0	0	0
Friedman 1	10	B	PARC	0	0	0	0
Friedman 1	10	B	LinearTree	0	0	0	0
Friedman 1	20	B	PARC	0	0	0	0
Friedman 1	20	B	LinearTree	0	0	0	0
Friedman 1	20	2B	PARC	0	0	0	0
Friedman 1	20	2B	LinearTree	0	0	0	0
Affine control	2	B	PARC	0	0	0	0
Affine control	2	B	LinearTree	0	0	0	0
Affine control	20	B	PARC	0	0	0	0
Affine control	20	B	LinearTree	0	0	0	0

F.10 Exposure, reproducibility, and the limits of the evidence

The original F, two-dimensional follow-up, and completion analyses have separately recorded protocols, source identities, and result ledgers. The follow-up was motivated by the original outcomes; the prediction-risk plot was designed after the saved follow-up fits existed. For PLATERO, source documentation, workbook identity, and response-free concentration/well metadata informed the fixed case before the preparation-stage program accessed numerical response cells. Synthetic fixtures were used for implementation checks. These records document the implemented order and exposure; they are not external preregistration, independently attested timestamps, or claims that all published source findings were unknown. The post-release SGEMM bank comparison and post-return PLATERO R^2 and residual decomposition are explicitly descriptive additions. No measured outcome was replaced to obtain a desired mix of release and withholding.

The supplement includes full protocols, producer sources, dependency pins, raw-data identities and notices, saved full-run ledgers, all figures, and complete CSV records. All declared setting/design/learner objects and unavailable slots remain accessible. PLATERO and SGEMM retain their source attributions and CC BY 4.0 notices; the PLATERO workbook is redistributed unchanged, without its MATLAB software. The numerical stack pins NumPy 2.4.4, SciPy 1.17.1, scikit-learn 1.8.0, and Matplotlib 3.10.9, with linear-tree 0.3.5 and PyPARC 2.0.4. Full author-run returns supply the experimental ledgers; subsequent verification and rendering reproduce their summaries without claiming a new scientific campaign. A native Windows verification/render return also passes all four integrated experiment components. Display comparisons account explicitly for byte differences that preserve decoded pixels or rendered PDF content.

The synthetic evidence supports model-based operating characteristics across learned designs, a concrete target/cost comparison with sample splitting, and reporting changes as oracle departures cross fixed margins. The measured cases show how the reporting contract changes what is issued and why the scale interpretation matters. They do not establish universal margins, calibrated guarantees for either measured noise process, better predictions, improved downstream decisions, or dominance

over all published selective-inference methods. Those distinctions are part of the experimental finding, not exceptions omitted from the reporting contract.

G MATCHED TOTAL PHYSICAL-RESPONSE COMPARISON

This appendix asks what remains of audit-row reuse after splitting is allowed its own response-free resolution search and every physical response is charged. The study was frozen before its valid response simulation. It is a balanced-repeat Gaussian benchmark, not a measured-data validation or an application-derived cost model. After response opening, an independent code review found that the original SciPy 1.17.1 noncentral-F CDF evaluation returned nonfinite values at large boundary noncentralities and that the implementation silently treated them as ordinary non-reporting. We retain that original ledger and report a single frozen numerical-correction replay below. The correction preserves the designs, response-free choices, margins, budgets, seeds, and 5,000 draws; it changes only the qualified outcome-side noncentral-F decision. It is therefore an exposed-outcome correction, not a second untouched confirmation. Its amended protocol hash is `713d4d6f64de93bd47bb583d3231536ea86c143f299901e716982bd2fc4c6d66`.

Fixed designs and acquisition. The roster crosses four synthetic settings, linear-tree and PARC routers, and ten design IDs, giving 80 fixed design units. Each has 800 common discovery responses and 400 distinct audit inputs. Audit and scale-response budgets are 2,400, 3,200, 4,000, and 5,600, so total physical-response costs are 3,200, 4,000, 4,800, and 6,400. Repeat measurements at one input provide pure-error information but not additional covariate excitation. The construction requires independent, homoskedastic Gaussian repeats with equal means at each repeated input. The benchmark margins retain the paper's individual-noise-normalized scale; they are not calibrated outcome-unit tolerances.

For every group, each policy that uses cells chooses among $J \in \{2, 3, 4\}$ using only the frozen design and prospective power. Stable ties and all response-free exclusions are fixed. CertifyPAR uses predetermined independent within-input contrast blocks for certification and inference scale. The split procedures use disjoint repeats for their gate and fit, and select their own J rather than inheriting CertifyPAR's resolution. We include a pooled rule on the same cohort. External-scale-only rows are retained as restricted ablations; the primary comparisons use internal or hybrid scale information.

The frozen planning rule for both split-internal policies retains a conservative scale convention: with r_{gate} repeats, candidate power screening and response-free choice of J use $\lfloor 400(r_{\mathrm{gate}} - 1)/2 \rfloor$ certification degrees of freedom, whereas their executed gate scale uses all $400(r_{\mathrm{gate}} - 1)$ gate-side pure-error degrees of freedom. This difference between planning and execution is part of the preserved selection rule; no frozen choices or reported outcomes are changed by this clarification.

Independent statistics from shared repeats. For C and the pooled rule, put $k = r_{\mathrm{gate}} = r_{\mathrm{fit}}$. At each input, an orthogonal transformation sends the k repeats to $\sqrt{k}\,\overline{Y}_i$ and $k-1$ zero-sum, unit-norm contrast coordinates. Under the declared Gaussian repeat model, their errors are mutually independent $N(0, \sigma^2)$. Fixed disjoint blocks of contrasts supply the two pure-error banks, independently of all input means; hybrid banks add independent external pure errors to the corresponding blocks. This is the sufficient-statistic extension distinguished from the physical-role protocol in Appendix A.2. For the mean-response audit, substitute

$$\sigma_{\mathrm{mean}} = \sigma/\sqrt{k}, \qquad S_{b,\mathrm{mean}} = S_b/\sqrt{k}, \qquad \tau_{u,\mathrm{mean}} = \sqrt{k}\,\tau_u.$$

The scale degrees of freedom are unchanged, and both the departure and its margin acquire the same factor $\sqrt{k}$, preserving the individual-noise adequacy requirement. Inner planning margins transform identically. The independent-scale Gaussian argument therefore gives the following implemented gate and interval formulas without requiring disjoint physical measurements.

With r_{gate} gate repeats, residual energy Q, rank r, and scale S_C^2, the implemented statistic and boundary noncentrality are

$$F = \frac{r_{\mathrm{gate}} Q}{r S_C^2}, \qquad \lambda_0 = r_{\mathrm{gate}} N_g \tau^2.$$

An interval based on r_{fit} fit repeats has variance $S_I^2 V / r_{\mathrm{fit}}$, where V is the unit-noise coefficient covariance on the distinct-input design. A split residual scale gives conservative intervals for its own finite projection under an arbitrary fixed Gaussian mean; exact central Student-t calibration requires an affine mean. This does not by itself validate a lower-tail adequacy gate with a contaminated scale.

Table G.1: Primary internal-repeat comparison at matched total physical-response cost. Mass is averaged over all audit rows and equally over the 80 fixed design units. C denotes the two-axis rule. The split owns its response-free resolution search.

Audit/scale	Total	Adequate reported mass		Mean half-width		Δ mass	Width red.
cost	cost	C	Split	C	Split	(pp)	(%)
2,400	3,200	.775844	.775486	.057384	.081183	+.036	29.32
3,200	4,000	.775844	.775787	.049667	.070262	+.006	29.31
4,000	4,800	.775844	.775841	.044415	.062820	+.0003	29.30
5,600	6,400	.775844	.775844	.037526	.053071	.000	29.29

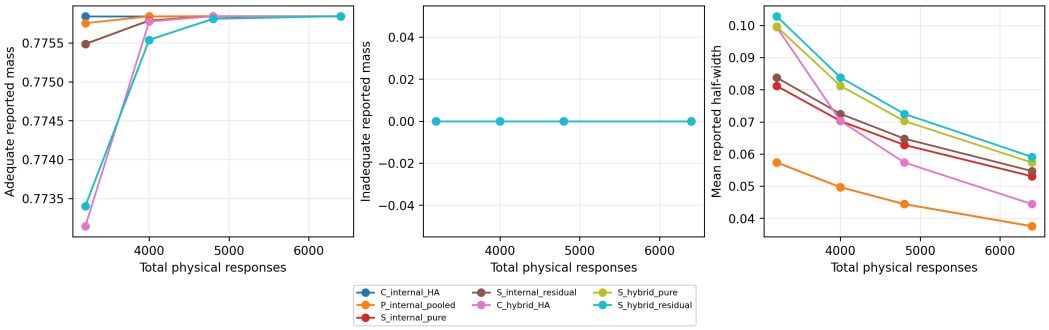

Figure G.1: Matched-budget operating characteristics. C, P, and S denote the two-axis rule, pooled rule, and split procedures, respectively. Curves are equal-design means over the 80 fixed units. All selected truths are adequate on this support, so the middle panel is identically zero and cannot establish H/A-specific filtering. The full ledger reports paired design-cluster uncertainty; connecting lines are visual guides, not a claim of monotonicity.

Frozen endpoints and uncertainty. Each design–policy–budget row receives 5,000 paired conditional response draws. These draws estimate operating characteristics for one design and are not counted as independent datasets. For a paired policy contrast, we first average over the eight setting–router combinations within design ID and then form a t_9 interval across the ten design-ID clusters. The frozen dominance rule requires a positive lower 95% bound for adequate reported mass and non-positive upper bounds for both inadequate reported mass and mean reported half-width. H/A distinctiveness applies the same rule against pooled equivalence and additionally requires binding-axis truth. Failure of this rule is reported as a trade-off, not rescued by another endpoint.

CertifyPAR's intervals are 29.29–29.32% narrower than the own-resolution pure-error split at every cost, with essentially the same adequate reported mass (Table G.1). The width reduction is the finite-design realization of the prespecified proportional-information identity $1 - 1/\sqrt{2} = 29.29\%$, not an empirical discovery. The paired C-minus-split mass differences are $+0.0358$ percentage points (t_9 interval $[0.0067, 0.0650]$) at audit/scale cost 2,400 and $+0.0056$ ($[0.0005, 0.0108]$) at 3,200; the differences are statistically unresolved at 4,000 and exactly zero at 5,600. Thus the frozen three-endpoint dominance rule passes at the two smallest budgets, but only through negligible mass differences. Against internal pooled equivalence, the mass differences are $+0.0091$ percentage points ($[-0.0001, 0.0183]$), $+0.0003$ ($[-0.0003, 0.0009]$), and then exactly zero; interval widths are also effectively identical. No H/A-distinctive dominance is established in these designs. One hybrid comparison at total cost 4,000 retains its 0.0236-point mass difference ($[0.0057, 0.0416]$), but it does not recur at neighboring budgets.

Coverage-pattern diagnostics and interpretation. The reporting family has four slots, so every policy–budget retains all 80 designs times 16 predeclared patterns. Of 51,200 cells, 47,069 are unobserved and receive upper bound one. No Holm-adjusted binomial violation is detected, but the maximum simultaneous one-sided Clopper–Pearson upper bound is one. Nonrejection therefore does not empirically verify the paper's full reporting-pattern conditional statement. Familywise miss

rates are about 4.0% for pure-error intervals and 3.3–3.4% for the more conservative split residual scale.

The result supports a narrower claim than H/A-specific cost dominance: under valid balanced repeats, residual reuse can buy coefficient precision without a material loss of reportable mass relative to splitting. Because no selected truth binds either axis, the experiment gives no empirical advantage to H/A separation over pooled equivalence in these designs (Figure G.1). Remark 28 concerns a different comparison: total-energy rules calibrated for the same rectangular requirement at adequate departures with larger total energy. All negative cells, restricted ablations, input identities, response-free choices, and complete pattern counts are retained in the supplement.

The amended replay completed all 3,200 design–policy–budget rows. It changed 56,306 group decisions in 103 rows that had been exposed to the legacy nonfinite calculation, and it recorded zero unresolved guard-band decisions. These counts are numerical diagnostics, not additional independent designs.

H DETAILED RELATION TO PRIOR WORK

Learning regions and qualifying coefficients. Model trees (Wang & Witten, 1997; Loh, 2002) and PARC (Bemporad, 2023) learn regional affine predictors. Hybrid and piecewise-affine identification estimates local dynamics and switching regions; surveys cover algebraic, Bayesian, clustering, and bounded-error methods (Paoletti et al., 2007; Garulli et al., 2012). Related constructions learn regressors and partitions by clustering (Ferrari-Trecate et al., 2003), recursive regression plus discrimination (Breschi et al., 2016), or joint regression–classification (Bemporad, 2023). Predictive and covariate-guided clusterwise regression address assignment of unseen covariates to local regressors (Gitman et al., 2018; Bu et al., 2026). CertifyPAR audits these groups for a common affine description at declared tolerances. The least-squares projection target on the realized audit design remains defined without exact mean specification (Buja et al., 2019). It starts after a router is frozen, identifies neither dynamics nor mode boundaries, and remains finite-audit even with a predictive router. Time-series or PWARX use needs dependence-aware theory beyond the independent-row setting.

Affirmative evidence of adequacy. The logic of testing approximate rather than exact validity goes back to Hodges & Lehmann (1954). In regression, Dette & Munk (1998) validate a prescribed linear form through an empirical L^2 discrepancy, while Baraud (2002) studies non-asymptotic minimax separation in Gaussian signal detection. These works establish that affirmative adequacy and its detectable separation are classical statistical questions; our contribution is not a new generic equivalence principle. Lack-of-fit and coefficient-stability tests detect departures from an exact model (Draper & Smith, 1998; Chow, 1960); equivalence tests reverse the burden and require evidence of departures below margins (Berger & Hsu, 1996; Wellek, 2010; Campbell, 2024; Dette et al., 2018). The RMSEA tradition's "not-close fit" tests are a noncentral-distribution precedent (MacCallum et al., 1996); CertifyPAR instead tests realized-design regression residual components and couples reporting to coefficient intervals. Model validation by equivalence testing (Robinson & Froese, 2004) and RMSEA-based assessment (Yuan et al., 2016; Marcoulides & Yuan, 2017) are direct precedents, including lower-tail tests at a noncentrality boundary. Our two-axis rule assigns separate margins to non-affinity within audit cells and disagreement among their affine projections; both must pass for reporting. Learn then Test (Angelopoulos et al., 2025) similarly uses multiple testing to calibrate predictive risks; our structural adequacy claims are paired with a separate conditional coefficient-coverage guarantee. A pooled residual-equivalence rule is therefore a required comparator. It certifies a different requirement, total departure, without cells. The rectangular requirement need not improve power or yield over every valid total-departure certificate. However, for one group, a rule using only total residual energy and an independent scale, when needed, and controlling false certification uniformly over the rectangular null at level α reports rectangle-adequate departures with $\eta_{T,g} \geq \min(\tau_{H,g}, \tau_{A,g})$ with probability at most α (Remark 28). Our experiments report both rules rather than treating marginal tests at uncalibrated levels as the only baseline.

Using the same data for checks and inference. Battey et al. (2026) use sufficiency and ancillary separation for post-reduction model assessment and parameter inference. Sarfati & Vilfort (2026) residualize estimators against diagnostic checks to obtain asymptotic independence and relate check orthogonality to efficiency. de Chaisemartin & D'Haultfœuille (2026) establish asymptotically conservative inference under specified nonrejection conditions without requiring independence, characterize asymptotic exactness, and analyze local alternatives. Building on these foundations, CertifyPAR couples affirmative adequacy requirements to finite-sample coefficient inference conditional on every complete reporting pattern of positive probability; Gaussian projection orthogonality (Cochran, 1934) justifies audit-row reuse. Under unknown noise, independent certification and inference measurements keep the Student denominator outside the reporting decision.

Partition selection and information splitting. Selective inference accounts for fitted models (Lee et al., 2016), k-means clusters (Chen & Witten, 2023), and CART node means (Neufeld et al., 2022). Wan et al. (2026) test slope contrasts after panel-group selection; Akgün & Okui (2026) develop selective tests and confidence sets for panel-group coefficients, including projection targets without requiring true grouping. Spuck et al. (2025) use a parametric bootstrap for coefficients in selected varying-coefficient trees. PoSI instead constructs intervals valid simultaneously over a prespecified family of model selections (Berk et al., 2013). CertifyPAR uses a narrower setup: the router and

requested coordinates are frozen, and the only response-driven decision reads a residual subspace independent of coefficient error. Its exact conditional guarantee does not cover unrestricted post-audit adaptation of the requests. CertifyPAR does not solve this response-driven partition-selection problem. We freeze a covariate-only router independently of audit responses and handle the subsequent response-dependent adequacy reporting. Tree-values and clustering-selective inference target selected node means or cluster contrasts under different conditioning events; a direct comparison would change our coefficient-reporting question (Appendix E). Sample splitting and honest trees separate observations (Rinaldo et al., 2019; Athey & Imbens, 2016), while data carving, fission, and thinning separate information in other ways (Fithian et al., 2014; Leiner et al., 2025; Neufeld et al., 2024). Our fixed row-splitting baseline makes splitting's effects on the coefficient target and measurement cost explicit alongside audit-row reuse. Gaussian fission and thinning can preserve full-design targets using noise-covariance information; their exact independent versions require that information to be correctly specified. They do not automatically remove the unknown-scale acquisition problem. The row-splitting baseline isolates row separation; Appendix G additionally compares own-resolution repeat splitting at matched total response cost. Neither comparison optimizes over all splitting methods.