Sortition Upstream of NTQR

How Panel Formation and Size Shape Ground-Truth-Free Evaluation

Daniel Ari Friedman

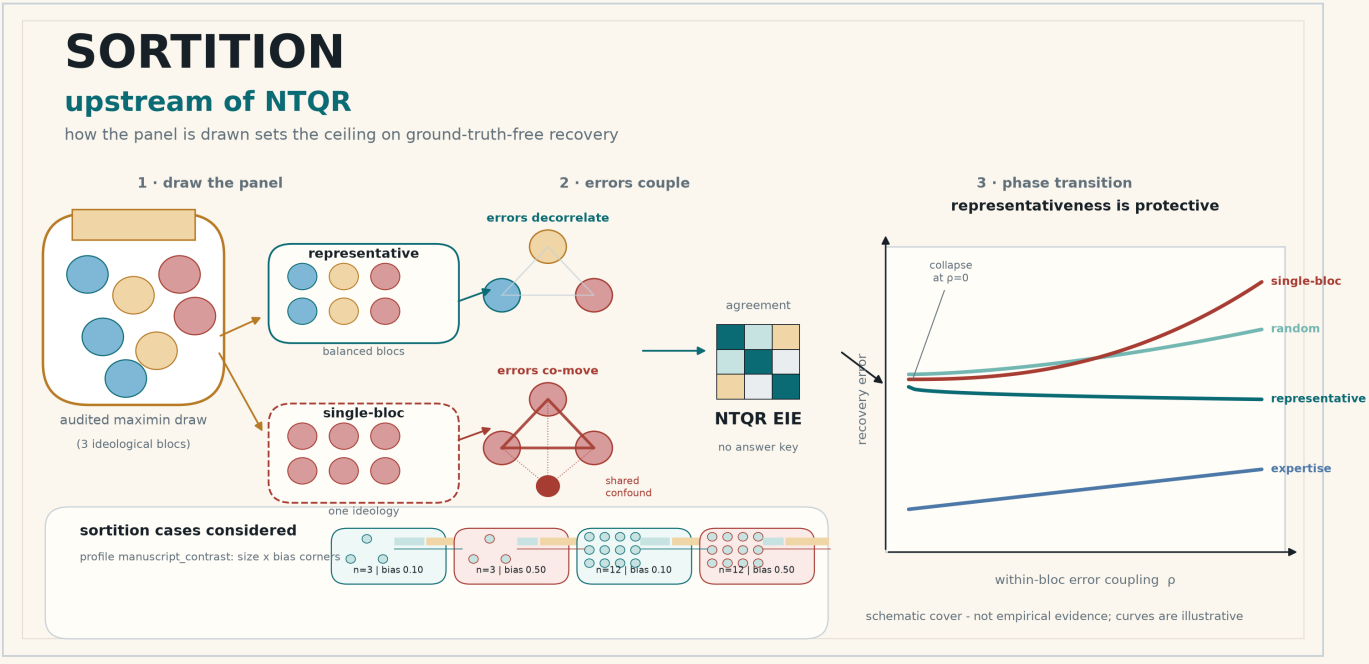
Active Inference Institute

danielarifriedman@gmail.com

ORCID: 0000-0001-6232-9096

DOI: 10.5281/zenodo.21083779

2026-06-25



Contents

1 Abstract

How should you choose the judges, jurors, or reviewers who form a panel — and does that upstream choice change how well you can evaluate them *without an answer key*? A panel can be selected many ways — by competence, by a representative lottery (**sortition**), by ideological bloc, or at random — and, separately, its noisy judgments can be evaluated blind: given the agreement/disagreement pattern among three binary judges, the `ntqr` package's error-independent (EIE) evaluator returns logically consistent estimates of item prevalence and per-judge accuracy with no labels at all. But that evaluator takes the panel as given. We join the two questions and ask whether the *rule that forms the panel* changes the oracle-referenced error of the no-answer-key evaluation — how far the blind estimate lands from the answer-key result, lower being better.

On a fully deterministic instrument (96 seeds, 96 experts, 300 items), the dominant lever is *which* rule forms the panel, not its size: competence-first selection recovers best (0.037), while representative, single-bloc, and random selection collapse together — *by construction*, because with independent judge errors composition cannot move an estimator that only sees agreement. Supplying the missing channel — same-group judges sharing a latent, marginal-accuracy-preserving error confound — makes the strategies fan out monotonically as within-bloc coupling rises: representative sortition stays flat while single-bloc selection degrades, the gap widening from 0.000 to 0.112. Within this instrument the relationship is closed-form: recovery error tracks the panel's **Herfindahl concentration index** over the axis a shared error rides on — minimized exactly by a balanced (representative) draw, maximized by a single bloc — and a continuous representativeness dial confirms error rises monotonically with it.

The protection is **conditional**: re-keying the confound to an axis the lottery does not balance erases the protection (0.147→0.229). The lesson for selecting and evaluating panels is thus a falsifiable, simulation-bounded prediction, not a preference for any one rule — representativeness protects blind recovery precisely when the panel balances the attribute a shared error rides on. Evidence is synthetic and oracle-scored; in a single small live model (gemma3:4b) the synthetically-best competence-first rule was the worst, illustrating that a selection rule validated on parameterized judges need not carry over to prompted ones — a hypothesis to test, not an established caution. All methods and documentation are openly available at the public repository **docxology/ntqr_allotment**.

2 Introduction: panel formation before blind evaluation

2.1 Blind evaluation begins before the estimator

Evaluating decision-makers without an answer key is a recurring problem: in crowdsourcing, in ensembles of classifiers, in deliberative bodies, and in any setting where the truth is expensive, contested, or unavailable at evaluation time — the problem Dawid and Skene first formalized for observer error-rate estimation without ground truth Dawid and Skene (1979). Later learning-from-crowds work made the same inferential problem operational for noisy human labels and missing gold standards Raykar et al. (2010), and budget-optimal crowdsourcing work shows why worker reliability, redundancy, and task assignment cannot be separated when answers are inferred from noisy repeated judgments rather than gold labels Karger et al. (2014). A parallel line estimates classifier accuracy *without any labels* by exploiting the structure of agreements among multiple predictors — logic- and constraint-based [Platanios et al., 2014] and spectral [Parisi et al., 2014] — and, like the exact NTQR solver, these methods lean on an **error-independence** (or low-rank residual-dependence) assumption whose violation is exactly the failure mode we make tunable here. The `ntqr` package (v0.8) frames the problem as algebraic logic for unsupervised evaluation from unlabeled decision data Corrada-Emmanuel (2026). Our generative model for *violating* that assumption — same-group judges sharing a latent shock through a Gaussian copula on a probit-thresholded competence variable — is the standard construction for correlated binary outcomes [Emrich and Piedmonte, 1991, Nelsen, 2006], which lets us preserve each judge's marginal accuracy exactly while dialing cross-judge error correlation. Given the joint pattern of agreements and disagreements among three binary judges, its error-independent (EIE) evaluator returns the possible logically consistent combinations of item prevalence and per-judge accuracy that could have produced those votes — *with no labels at all*.

The `ntqr` evaluator, however, takes the panel as a fixed input. Its claims are about which evaluations are logically consistent for a given set of judges. The judges themselves arrive from *somewhere*: a hiring process, a volunteer pool, a citizens' assembly lottery, a top-k leaderboard. That upstream step — **panel formation** — is the part this project studies. Our central question is deliberately one level above NTQR:

> Across matched synthetic population and corpus settings, how do the *strategy that forms the panel* and the *size of the panel* change the oracle-referenced error of no-answer-key NTQR-style evaluations?

Two scope notes frame everything below. First, *oracle-referenced recovery error* is the distance between the blind, no-answer-key estimate and the estimate you would have obtained *with* the answer key — lower is better. Second, NTQR's exact error-independent evaluator solves for *exactly three* binary judges, so a panel of any larger size is evaluated as an ensemble of its constituent trios; "panel size," throughout, means how many such trios the panel contributes, not a larger joint solve.

2.2 Application review as the empirical stress test

Application review is the manuscript's concrete empirical stress test because it combines expert judgment, scarce labels, panel formation, and plausible nuisance bias in one setting. Studies of academic and grant peer review emphasize that expert panels are not transparent measurement devices: judgment is field-situated Lamont (2009), replication of review decisions can be fragile Peters and Ceci (1982), reviewer agreement can be low even on the same proposals Cole et al. (1981); Pier et al. (2018), and statistical analysis of NIH review scores shows that panel-scoring uncertainty can materially change which proposals would be funded Johnson (2008). Productivity follow-up work adds a separate caution: NIH peer-review percentile scores can be weak predictors of grant productivity Fang et al. (2016), and peer-review bias is a documented design concern rather than an exotic failure mode Lee et al. (2013); Tomkins et al. (2017); Helmer et al. (2017). The postdoctoral fellowship version is also historically apt: Wennerås and Wold's analysis of postdoctoral fellowship review found nepotism and sexism in peer review Wennerås and Wold (1997), while broader funding studies report racial disparities in award outcomes Ginther et al. (2011).

Those literatures motivate the *mechanism* we stress-test, not the conclusion. This project uses fictitious postdoctoral applications and synthetic age metadata so the hidden quality label is generated independently of age. Age is included as a protected-attribute nuisance axis because age bias and age-discrimination effects are documented in employment-relevant settings North and Fiske (2013); Neumark et al. (2019). The manuscript therefore asks a controlled design question, not a policy question: if reviewer expertise and irrelevant age bias are known in the generator or prompt profile, which sampling rule best limits oracle-referenced NTQR error and age-conditioned recommendations?

2.3 Sortition from civic lotteries to evaluator sampling

Sortition — selection by lottery, often with quotas that make the drawn body mirror the population — is the canonical *fair*, non-comparative panel-formation rule Stone (2011). Modern implementations use auditable maximin algorithms Flanigan et al. (2021) that maximize the minimum selection probability subject to representativeness constraints. Sortition is attractive precisely because it does **not** select on competence; it selects on representativeness. That makes it a sharp test case for NTQR: a representative panel is heterogeneous and, by design, *not* curated for accuracy. Does representativeness help or hurt an estimator that depends on the statistical independence of judges' errors?

That procedural tension has a long pre-1800 lineage. Aristotle's political theory treats lot and election as constitutional signals — lot as democratic, election as oligarchic — while the Athenian institutional account describes juries and offices allocated by lot Aristotle (Politics); Aristotle (Athenian Constitution). Medieval and early-modern writers kept the same problem visible under different vocabularies: Aquinas (Summa Theologiae II-II q.95 a.8) distinguished practical uses of lots from divinatory misuse, and Contarini (1599) described Venetian mixed selection machinery as an anti-factional civic design. Enlightenment writers then made the lot-versus-choice contrast explicit again: Montesquieu (1748) and Rousseau (1762) both distinguish election by lot from election by choice. We cite these sources as procedural history, not as direct empirical precedent: this manuscript does not claim that Athenian juries, Venetian offices, or scholastic accounts of chance anticipate NTQR. They do show that the upstream choice between lot, choice, status, and asserted competence is an old institutional design problem.

Randomness is already a serious proposal in adjacent research-funding design, but usually at a different point in the pipeline: collective-allocation and modified-lottery proposals address how funds might be allocated after review or thresholding Bollen et al. (2014); Fang and Casadevall (2016). Lottery arguments also arise from the maverick-science problem: if high-variance projects are hard to rank reliably, randomized allocation can be an epistemic risk-control device rather than only an administrative convenience Avin (2019). This manuscript moves the lottery upstream. It asks how reviewer *sampling* changes an unlabeled evaluator before any funding decision is made.

The tension is not artificial. Classical jury-theorem results make group accuracy depend on competence, independence, and aggregation rule Grofman et al. (1983), while diversity results show that heterogeneous problem-solving groups can outperform ability-selected groups under specific search conditions Hong and Page (2004). Deliberative public-consultation work likewise treats representative participation as a normative and epistemic design choice, not just a sampling convenience Fishkin (2009). The formal voting-theory lineage is also historical rather than merely modern: Borda (1781) proposed a scored ballot for elections, and Condorcet (1785) analyzed the probability of correct plurality decisions. Those works mostly take the voters or jurors as given. The present instrument asks the upstream question they leave exogenous: how does the rule that forms the evaluator panel change the blind recovery problem before aggregation begins? This manuscript does not assume which rationale wins under NTQR; it measures the trade-off in a controlled binary-evaluation instrument.

We compare four panel-formation strategies against each other and against the supervised oracle: auditable representative sortition, uniform random selection (the honest baseline), single-bloc ideological selection (a deliberately correlated, non-representative comparator), and competence-first expertise thresholding. The oracle and the strong baselines are first-class comparators, not strawmen — the point of the instrument is to *measure*, including measuring our own preferred narrative against an honest null.

2.4 Falsifiable claims and negative controls

We state the study as five falsifiable hypotheses (H1–H5) and let the regenerated artifacts adjudicate each. The synthetic deterministic track tests H1–H4 against a known oracle; the live single-model companion tests H5. Methods (Table tbl. 2) states the load-bearing assumption and the negative-control check behind each, and the Discussion returns an explicit verdict hypothesis by hypothesis.

1. **H1 — Formation strategy is the dominant lever.** Different panel-formation rules yield materially different oracle-referenced EIE recovery error, even at a fixed population and panel size. *Tested by* the weighted-mean strategy ranking with a bootstrap separation gate and a power budget (fig. 2).
2. **H2 — Concentrating correlated bias degrades recovery.** Single-bloc selection, which seats judges whose errors are correlated, recovers with higher EIE error than a representative draw, because the error-independence assumption NTQR rests on is most stressed by a correlated panel. *Tested by* the representative-minus-single-bloc contrast across the full regime grid against analytical sign predictions (fig. 3), and resolved by the composition-coupled error confound that fans the strategies apart as within-bloc coupling rises (fig. 9).
3. **H3 — Size is a sampling knob, not guaranteed improvement.** Forming a larger ensemble gives more trios to average over, but whether that helps or hurts recovery — and whether any effect comes from the larger ensemble violating the solver's error-independence assumption more — is an empirical question; the power question is how

many observations at the analyzed grain would be needed to resolve a contrast of the observed magnitude. *Tested by* the per-strategy size sweep, a paired regime-controlled size contrast, a per-trio conditioning diagnostic, and the power/MDE budgets (fig. 6, fig. 12, fig. 7).

4. **H4 — Error-correlation is measurable, and recovery degrades with it.** A controlled correlation injection produces a realized error-correlation that NTQR itself reports, and oracle-referenced recovery error should rise as that realized correlation grows. *Tested by* the tolerance sweep, reporting the realized-correlation trend and an OLS recovery-vs-correlation slope with a bootstrap interval (fig. 8); the recovery slope, unresolved by the global-injection sweep, is resolved positive once the correlation is composition-coupled and marginal-preserving (fig. 9).
5. **H5 — The synthetic ranking transfers to a live single-model panel.** The strategy ordering measured against the synthetic oracle reproduces when one local `gemma3:4b` model is prompted as different reviewers. *Tested by* a matched-grain cross-track ranking comparison and cell-level directional alignment (fig. 20, fig. 19).

These hypotheses are written to be refutable, and the data refute several: H3 and H5 are rejected; H2 and H4 are unresolved on the baseline grid and resolve only once correlation is coupled to panel composition; and H1's ranking collapses to competence-first versus a bunched remainder. The instrument, its resolved cells, its explicit design-limited cells, and the axis-conditional negative control that keeps the sortition result honest — not a manufactured sortition win — are the contribution.

3 Methods: instrument, assumptions, and companion track

3.1 Synthetic deterministic track: seeded panels, blind estimates, oracle scoring

The first methodological track is a seeded synthetic instrument. It generates known populations and corpora, hides the answer key from the `ntqr` estimator, and then scores the returned logically consistent evaluations against the supervised oracle that is available only because the data are synthetic.

3.1.1 Pipeline: panel formation precedes no-answer-key estimation

The instrument runs strictly upstream-to-downstream and is deterministic end to end (every stochastic step is seeded with `numpy.random.default_rng`; figures use `MPLBACKEND=Agg`). One **trial** is:

1. **Generate** a synthetic expert population of known properties.
2. **Form** a panel from that population with one of four strategies.
3. **Judge** a corpus of items whose ground-truth labels are *known to us but hidden from the estimator*.
4. **Evaluate** the panel with the `ntqr` package *without the answer key*.
5. **Score** the unsupervised estimate against the supervised **oracle** computed from the known labels.

fig. 1 shows this upstream-to-downstream flow at a glance.

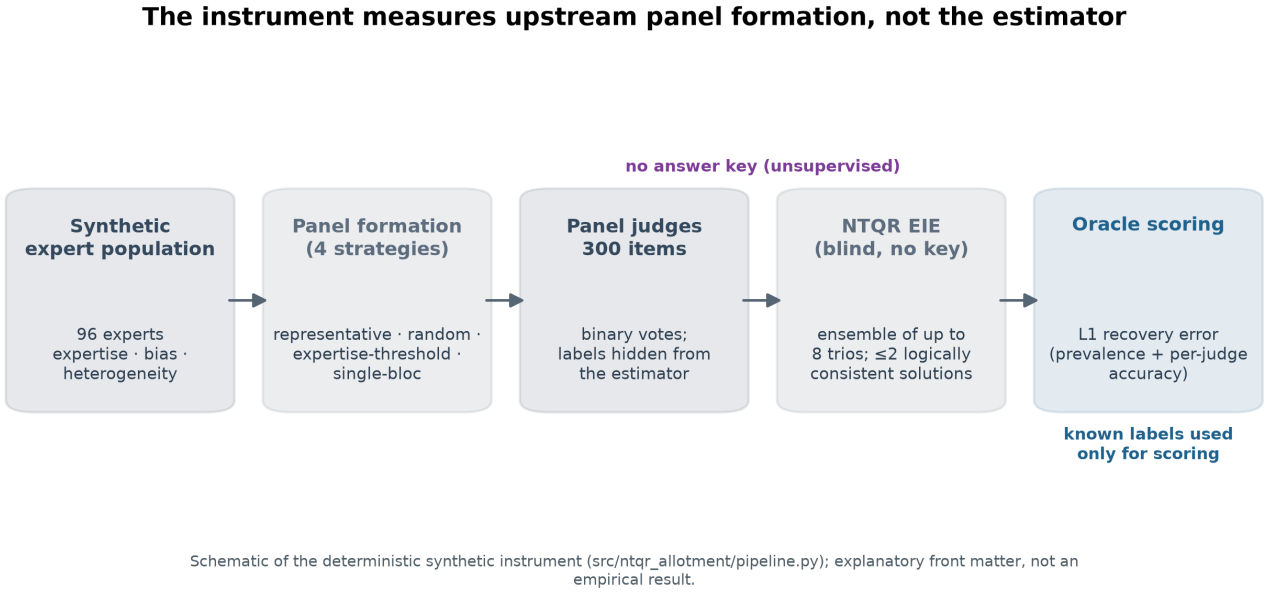


Figure 1: Left-to-right pipeline of the deterministic synthetic instrument (steps 1–5 in text; count tokens annotated from `96`, `300`, `8` so the figure cannot drift from the reported configuration): a known population is sampled → a panel is formed by one of four strategies → the panel judges a key-hidden corpus → the EIE evaluator runs blind over trios → only scoring reads the labels. The bracket marks the unsupervised region; the manipulated variable is the upstream formation rule. Explanatory schematic only; quantitative results are in the Results section.

The key methodological move is step 5: because the ground truth is synthetic and therefore known, the unlabeled `ntqr` evaluation and the supervised oracle can be compared on equal footing. Recovery error is the L1-style distance between the two evaluations — the absolute prevalence error plus the mean absolute per-judge accuracy error. Throughout, "ground-truth-free" is project shorthand for this no-answer-key estimator path, not an expansion or alternative name for `ntqr`.

Synthetic expert population. Each expert is a noisy binary judge with label-conditional accuracies `accuracy_a = P(vote a | true a)` and `accuracy_b = P(vote b | true b)`, derived from a continuous `expertise` (mean precision) and a signed `bias` that skews errors toward one label. The population sampler draws expertise from a normal centered at `mean_expertise` with standard deviation `expertise_heterogeneity`, and draws bias with a sign **correlated with each expert's ideology** (left → negative, right → positive). Because this only shifts each judge's *marginal* accuracy by ideology — every judge still errs from an independent stream — it does not by itself make single-bloc panels more error-correlated than a representative draw; supplying that missing cross-judge error-correlation channel is the job of the

composition-coupled confound introduced below. Each population in the sweep has 96 experts; each corpus has 300 items sampled at the configured prevalence.

3.1.2 Panel-formation strategies: four upstream rules

All four strategies are deterministic given their seed.

- **representative_sortition** — an auditable maximin lottery implemented on the open-source `allotment` engine (Citizen-Infra (2024), the AGPL-3.0 sortition library this project imports and uses directly), which realizes the fair stratified-selection algorithm of Flanigan et al. (2021). Ideology quotas are set by largest-remainder (Hamilton) apportionment so the panel mirrors the population's ideology composition as closely as integer seats allow; the draw carries the engine's SHA-256 audit hash for reproducibility.
- **random_selection** — a uniform draw without replacement. This is the simplest honest baseline and is treated as a first-class comparator throughout.
- **ideological_selection** — fill the panel from a single ideology bloc first, spilling over only if the bloc is too small. This deliberately concentrates correlated biases and is the *non-representative* comparator.
- **expertise_threshold** — select the top-k experts by expertise. This is the competence-first comparator and ignores representativeness entirely.

3.1.3 NTQR evaluation: trio EIE, oracle scoring, and majority voting

The `ntqr` package's exact error-independent evaluator is **trio-only**: it solves the error-independent algebraic system for *exactly three* binary judges, returning logically consistent (prevalence, per-judge accuracy) evaluations. The system admits up to two real solutions; complex or non-finite roots are dropped honestly rather than coerced. To resolve the two-fold ambiguity in the synthetic track we select the consistent solution **closest to the oracle** — this is the most charitable reading of the unlabeled estimate, so any residual error is a real failure to match the supervised oracle, not a sign ambiguity.

We compute two `ntqr` evaluations per trio: the error-independent evaluation (EIE, our headline) and the majority-voting evaluation (a comparator that partitions solutions into crowd-right and crowd-wrong). The **supervised oracle** is read directly from the label-conditioned vote counts; it is always real and finite, and a degenerate oracle is treated as a contract violation that fails loudly.

Ensemble-of-trios for panels larger than three. Because the exact solver is trio-only, a panel of size greater than three is evaluated by **ensemble-of-trios**. We scan trios (combinations of panel members) in deterministic order, collecting up to 8 *usable* trios — a trio is skipped if its vote pattern admits no real error-independent solution — and average their oracle-referenced errors. A single bad expert therefore does not starve the ensemble. If *every* trio in a panel is degenerate, the trial records an honest NaN (zero usable trios) so a sweep surfaces "no recovery possible here" rather than crashing or inventing a number. A panel of exactly three reduces to a single trio, so the ensemble result coincides with the single-trio result there.

Historically, Borda- and Condorcet-style work asks how votes should be aggregated once a voting body exists; here, the formation rule is part of the experimental treatment. The Methods therefore keep two objects separate throughout: the upstream panel draw and the downstream no-answer-key estimator.

3.1.4 Companion diagnostics: alarm cost, ternary feasibility, and maximin fairness

Three companion tracks measure structural properties of the instrument rather than oracle-referenced recovery error.

Alarm scaling is a small-Q constraint. The `ntqr` package also ships an **alarm**: it tests whether all judges can be simultaneously consistent with some answer key at a stated safety specification, a constraint system that gains more panel-size-indexed checks as judges are added. Unlike the trio evaluator, this project's alarm path enumerates the answer-key simplex, and our local benchmark shows roughly **cubic scaling in the corpus size** Q. A shipped benchmark (`scripts/bench_alarm.py`) reproduces this scaling on demand; indicative single-machine timings rise from about 0.7 s at $Q = 20$ to 8.9 s at $Q = 50$ to 97.9 s at $Q = 100$ (the exact constants vary with machine load — the cubic *scaling*, not the constants, is the robust local finding). This $O(Q^3)$ cost is a scaling limit on the statistical-power / alarm track as implemented here, so we report it as a finding and cap any alarm use at $Q \leq 30$ (opt-in only); larger corpora must raise the cap deliberately.

The ternary $R = 3$ axiom-consistency track (consistency only). A companion track (`src/ntqr_allotment/ternary.py`) extends the axiomatic surface from binary ($R = 2$) to ternary ($R = 3$) responses, but **only at the level of axiom-consistency and feasibility — never $R = 3$ recovery**. It checks whether an observed three-way vote profile is *consistent* with the NTQR algebraic axioms (the response counts sum correctly and lie in the feasible simplex), not whether

the unsupervised (prevalence, accuracy) state can be *solved*. Exact $R = 3$ recovery is unsolved upstream and is explicitly **out of scope / anti-vision** for this work: we make no claim to recover ternary evaluations. The track exists so the consistency/feasibility axioms can be exercised and tested at $R = 3$ without overstating what NTQR can do there.

N-judge alarm power is consistency-only. A second companion track (`src/ntqr_allotment/ensemble.py`) generalizes the single-trio consistency check to an **N-judge** observed-vote-count alarm and measures how the consistency signal scales with panel size (`alarm_power_curve`). In the current small-Q diagnostic, the tight safety setting is already saturated across the plotted panel sizes, so the figure demonstrates that the N-judge alarm is executable and panel-size-indexed rather than establishing a monotone growth law. The underlying answer-key enumeration is the same $O(Q^3)$ cost described above, so the N-judge track is exercised only at **small** Q. It is a panel-size-indexed *consistency* signal, not a recovery method.

Maximin fairness is a selection metric. The representative-sortition strategy is an auditable maximin lottery, and a fairness track (`src/ntqr_allotment/fairness.py`) characterizes the allotment's **selection-probability distribution** over the population — the probability each expert is seated across the lottery. The maximin objective is the **minimum selection probability**: a fairer lottery raises the floor on who can be seated. This track measures the representation properties of the draw itself and is independent of the downstream NTQR recovery numbers.

3.1.5 Notation: cells, trios, and inferential units

The manuscript keeps each statistic tied to the unit that generated it; this is the guardrail that prevents synthetic, power, and live empirical claims from borrowing strength from one another. Table tbl. 1 is the compact ledger for symbols, estimators, inferential units, and artifact ownership.

Table 1: Notation and inferential units for the manuscript's reported statistics.

Symbol	Surface / estimator	Unit	Aggregation and uncertainty	Source artifact
E, Q, π	Experts, items, and label prevalence hidden from NTQR	one seeded population/corpus	profile metadata, config hash, seed list	`output/data/sweep_results.json`
V_{ij}	Binary vote matrix by panel member i and item j	one panel trial	supervised oracle retained only for scoring	`src/ntqr_allotment/pipeline.py`
$\hat{\theta}_{\text{EIE}}$	NTQR error-independent evaluation	one usable trio	oracle-referenced recovery error	`src/ntqr_allotment/ntqr_eval.py`
$\bar{e}_{\text{trio}}$	Ensemble aggregation over usable trios	up to 8 trios per panel	NaN/sentinel if every trio is degenerate	`output/data/sweep_aggregated.csv`
$\bar{e}_s$	Strategy ranking by weighted mean EIE error	strategy over active profile cells	pooled 95% CI, seed count, profile/hash	`output/data/sweep_aggregated.csv`
$\Delta_{\text{ideo}-\text{rep}}$	Ideological-minus-representative contrast	active-profile regime cells	observed-vs-predicted alignment, descriptive intervals	`output/data/analytical_predictions.json`
ρ_{NTQR}	Realized pairwise error correlation	non-degenerate (ρ, strategy) cell	OLS slope with bootstrap CI over unique cells	`output/data/independence_sweep.csv`
d, n, MDE	Two-sample power design quantities	per-strategy EIE observations at fixed panel size	Cohen's d, permutation p, Holm correction, MDE, per-group observation budget	`output/data/power_analysis.csv`
Δ_{age}	Live postdoctoral-review age-disparity stress test	strategy x panel-size under one Gemma model	older-minus-younger recommendation-rate difference, descriptive intervals	`output/data/postdoc_panel_results.json`

Symbol	Surface / estimator	Unit	Aggregation and uncertainty	Source artifact
A_{align}	Analytical-vs-Gemma postdoc alignment	strategy x panel-size cells	directional sign agreement and unresolved-cell count	`output/data/postdoc_panel_alignment.json`

3.1.6 Assumption ledger: how each claim can fail

The analysis is organized as falsifiable claims rather than a single success story. Table tbl. 2 states what would count against each claim and which artifact carries the check. The rows map onto the Introduction's hypotheses: the representative-vs-ideological, panel-size, tolerance-sweep, and real-Ollama rows are the negative-control checks for H2, H3, H4, and H5 respectively; the NTQR-EIE-recovery row guards the estimator the whole study depends on (and so underpins H1, the strategy ranking tested directly in Results); and the null-and-significance row fixes the design-limited-vs-resolved interpretation discipline applied throughout.

Table 2: Assumption and falsification ledger for the manuscript's main claim families.

Claim family	Load-bearing assumption	Negative-control or falsification check	Current interpretation
NTQR EIE recovery	The three-judge error-independent algebra is the right estimator for a usable trio.	Complex/non-finite roots and every-degenerate panels are retained as failures, not coerced into numbers.	Residual recovery error is scored only after a real logically consistent solution exists.
Representative-vs-ideological contrast	Bias concentration should affect oracle-referenced EIE error through error dependence.	Cellwise ideological-minus-representative heatmap plus analytical directional checks can disagree with the predicted sign.	Design-limited on the baseline grid (independent errors); resolved once the composition-coupled confound supplies the error channel (fig. 9), not a universal sortition win.
Panel size	Enlarging the panel averages more trios; whether that helps or hurts, and by what mechanism, is measured rather than assumed.	Paired per-strategy size contrast can show error rising with panel size; the per-trio diagnostic locates the cause.	The active profile falsifies a uniform "larger is better" rule and refutes the error-correlation explanation for it.
Tolerance sweep	Injected ρ should be visible in NTQR-measured realized error correlation.	The measured ρ_{NTQR} must rise with injected ρ before any recovery-slope story is considered.	The diagnostic works; the recovery slope is unresolved under global injection but resolves positive under the marginal-preserving composition-coupled instrument (fig. 9).
Real-Ollama postdoc companion	The same sampling mechanism should shape age-bias expression and ranking under one prompted local LLM.	Gemma-only reviewer-panel rows can disagree with the analytical sign or remain unresolved by cell.	Reported as n-limited empirical companion evidence, not human-review validation.
Null and significance language	A non-significant contrast is not evidence of no effect unless the design could detect the relevant effect size.	Permutation p-values, Holm correction, MDE, and sample-size budgets are all reported together.	Nulls are split into resolved, underpowered, and well-powered design statements.

3.1.7 Sweep profiles: profiles, seeds, and aggregation units

The synthetic track is a deterministic grid sweep over the four strategies, panel sizes, expert stringency, bias spread, and the population/corpus parameters in `manuscript/config.yaml`. That file now defines named profiles: the reported sweep uses `manuscript_contrast` (config hash `fda4da941cf0`), while `smoke`, `manuscript_main`, `tolerance`, `power`, `panel_ladder`, and `research_broad` keep CI, legacy manuscript, assumption-tolerance, design-budget, finer panel-size, and broader sensitivity settings explicit. `live_postdoc_panel` separately stores the required-live Gemma model settings, reviewer/application counts, decode controls, and vote-cache path. Each reported grid cell is repeated over 96 seeds; per cell we report the mean EIE error, its sample standard deviation, and a 95% confidence interval. Degenerate cells (no usable trio) are excluded from aggregation via the same sentinel the emitter respects. A single seed is treated as an illustration, never a finding — all reported effects are seed-aggregated with confidence intervals. The aggregated table (`output/data/sweep_aggregated.csv`) and per-seed JSON (`output/data/sweep_results.json`) carry the profile name, config hash, seed list, and degenerate-row count; manuscript numbers are emitted from those artifacts by `src/ntqr_allotment/manuscript_variables.py`, so no result is hand-transcribed.

3.1.8 Controlled-correlation sweep: injected dependence as a diagnostic

The error-independence assumption is probed directly. `dependence.py`'s `sample_votes_correlated(experts, items, *, rho, seed)` injects a controllable shared-error latent of strength ρ, and `measure_error_correlations` reports the *realized* pairwise and three-way correlation NTQR itself computes from the votes — so the knob (ρ) and the measured quantity are independent. `independence_sweep.py` sweeps $\rho \times$ strategy at the trio over multiple seeds and aggregates recovery error against realized correlation (`output/data/independence_sweep.csv`), yielding the error-correlation **tolerance curve** reported in Results. This correlation sweep uses its own smaller grid — 24 experts and 120 items, four injected ρ levels by two strategies over up to six seeds per cell (eight non-degenerate (ρ, strategy) cells) — deliberately fixing the panel at the trio (the exact solver's unit) so panel size cannot confound a trio-level correlation study.

3.1.9 Composition-coupled confound: when group membership carries shared error

The tolerance sweep above injects correlation *globally*, identically for every panel, so it cannot test whether *how the panel is formed* changes the correlation the estimator sees. `bloc_confound.py` supplies that missing channel. `sample_votes_bloc_correlated(panel_experts, items, *, bloc_correlation, seed, axis)` drives each judge's correctness through a Gaussian copula whose shared component is keyed on a grouping attribute: $z_j = \sqrt{\rho}\, g_{\text{group}(j)} + \sqrt{1-\rho}\,\varepsilon_j$, correct iff $z_j < \Phi^{-1}(\text{acc}_j)$. Judges in the same group share the standard-normal stream g (keyed by a stable hash of the group value, so it is identical across panels and worker processes); judges in different groups stay independent. Because z_j is marginally standard normal, $P(z_j < \Phi^{-1}(\text{acc})) = \text{acc}$ exactly per label: the construction is **marginal-accuracy preserving**, so any recovery change is attributable to error correlation rather than a confounded accuracy shift, and $\rho = 0$ recovers the independent baseline. The inverse-normal Φ^{-1} uses a dependency-free Acklam rational approximation. `run_bloc_phase` sweeps strategy $\times$ $\rho \times$ bias-spread $\times$ stringency $\times$ panel-size $\times$ seed (`scripts/run_bloc_phase.py`, `output/data/bloc_phase.csv`). The default `axis="ideology"` keys the confound on the axis representative sortition balances; a negative-control grid keys it on `axis="expertise_tier"`, an axis the lottery does not balance, to test whether the representative robustness is innate or conditional. Recovery is scored against the same supervised oracle as the main sweep, and the realized correlation is read back with the same `measure_error_correlations` diagnostic.

3.1.10 Herfindahl exposure: concentration predicts shared-error risk

The fan-out is not arbitrary: keying the shared shock on the grouping axis makes a trio's confound exposure equal to its same-group pair count — the Herfindahl index — so the composition-to-exposure relationship has a closed-form backbone. That backbone is a designed, internally-consistent property of this instrument, not an independent empirical law: the closed form follows *by construction* from how the confound is keyed. What the simulation then genuinely tests — the falsifiable link — is whether NTQR's exact recovery actually degrades as that exposure rises. Let a panel have seat fractions p_b across the confound's grouping axis (here ideology, with B groups). The probability that two seats drawn with replacement fall in the same group is the **Herfindahl–Hirschman index** $H = \sum_b p_b^2$ (`theory.herfindahl_index`); for distinct seats it is the finite-panel correction $\sum_b c_b(c_b-1)/[N(N-1)]$ (`theory.same_group_pair_probability`), and the expected number of same-group pairs among the three pairs of a trio is three times that. Because the shared error shock is keyed on the group, a trio's exposure to it is exactly its same-group pair count. Holding competence fixed, the realized error-correlation NTQR measures is therefore monotone increasing in H, and — since the exact error-independent solver is the one whose assumption that exposure violates — so is recovery error. H is minimized at $1/B$ by a perfectly balanced panel and maximized at 1 by a single-group panel, which is precisely the representative-versus-single-bloc axis. The maximin sortition quota makes this exact: a representative draw attains $H = 1/B$ (here 1/3), single-bloc selection attains $H = 1$, and random selection

sits between — an ordering that matches their measured error-correlation ordering cell for cell (`tests/test_theory.py::test_herfindahl_predicts_strategy_correlation_ordering`).

This also licenses a *continuous* reading of representativeness rather than four discrete strategies. `bloc_confound.concentration_panel` forms a panel with a dial $c \in [0, 1]$: a fraction c of seats massed in one group and the rest balanced. In the large-panel limit its Herfindahl index is $H(c) = (c + \frac{1-c}{B})^2 + (B-1)(\frac{1-c}{B})^2$ (`theory.concentration_herfindahl`), monotone increasing from $1/B$ at $c = 0$ to 1 at $c = 1$. Sweeping c at fixed coupling (`run_concentration_sweep`, `output/data/bloc_concentration.csv`) traces recovery error against the dial and tests the predicted monotonicity directly (fig. 10). The contribution is thus a closed-form chain — composition → Herfindahl exposure → realized error-correlation → no-answer-key recovery error — verified end to end in simulation, with the conditional caveat (the law is stated over *the confound's* axis) built in. Operationally this suggests a panel diagnostic that needs no votes and no answer key: compute a *proposed* panel's concentration index over the attribute a shared error might ride on. *If* a shared error exists *and* its axis is known, a lower index over that axis implies lower modeled shared-error exposure in this instrument. Whether a real shared error exists, and on which axis, is outside what this simulation establishes — so this is a modeling diagnostic, not a validated trust signal for real panels.

3.1.11 Statistical power: separating rankings from resolved contrasts

Because the recurring nulls are computed on bounded per-strategy observation groups, a power layer (`power_analysis.py`, `power_study.py`) makes design adequacy explicit, following the standardized-effect and power/sample-size framework Cohen (1988). This design-budget framing also matches the review-panel literature's concern that reviewer counts and score precision are design parameters, not afterthoughts Kaplan et al. (2008). The pure-numpy toolkit provides a normal CDF/PPF checked against published constants, analytic two-sample power, Monte-Carlo `simulate_power` using the *actual* Welch-t / permutation test, `sample_size_for_power`, and a minimum-detectable-effect (MDE) solver; each primitive is bound to an independent reference (analytic vs simulation, Type-I rate vs alpha) and **no retrospective observed power is ever reported**. `power_study.py` applies this to every pairwise strategy contrast from the real per-seed sweep (`output/data/power_analysis.csv`), turning each soft null into an experiment budget (per-group observations at the analyzed trial/cell grain for 80% power). Separately, `statistics_analysis.strategy_separation` compares two strategies by their separately bootstrapped mean intervals and emits an explicit CI-overlap verdict (`separated` / `overlapping`) that must read `separated` before any "beats" wording is justified. Bootstrap intervals are used as descriptive uncertainty summaries, following the nonparametric bootstrap framing of Efron and Tibshirani (1993). Because the sweep compares every pair of strategies, the family of pairwise permutation p-values is corrected with the Holm-Bonferroni step-down procedure Holm (1979) (`statistics_analysis.holm_bonferroni`) before any significance count is reported, controlling the family-wise error rate without plain Bonferroni's conservatism. The Gemma postdoctoral panel artifact is reported with descriptive intervals and cell-level directional alignment only; it is not folded into the synthetic power family and is not reported as retrospective observed power.

3.2 Real-Ollama reviewer-panel track: single-model live companion

The second methodological track uses one live local language model through Ollama: `gemma3:4b`. It is deliberately **single-model** and deliberately not a model-family comparison. The empirical question is whether the same sampling mechanisms studied analytically — representative sortition, random selection, same-bias bloc selection, and expertise-threshold selection — remain visible when one real local LLM is prompted as different postdoctoral-review panelists.

The live track is therefore closer to an instrumented LLM-judge stress test than to an ethnography of human review. LLM-as-judge work has shown that prompted model judgments can be useful but also vulnerable to evaluator-specific artifacts such as position, verbosity, and self-enhancement biases Zheng et al. (2023); systematic position-bias tests likewise show that judge outputs can change with answer order rather than only answer quality Shi et al. (2025). Broader language-model risk work likewise warns against treating fluent model output as an unmediated measurement of social reality Bender et al. (2021). We use one model, bounded decoding, serialized provenance, and synthetic applicant/reviewer metadata precisely so the empirical surface remains auditable and does not masquerade as human-review validation. The two tracks are kept strictly separate and their evidence is never pooled: the synthetic deterministic track remains the controlled spine, while the live track is a companion stress test with separate artifacts, a separate config hash (`5161ffe474b3`), and separate caveats — neither supplies evidence for the other's claims.

3.2.1 Postdoctoral corpus: protected-attribute stress test

The empirical setting is a fictitious postdoctoral fellowship review panel. Each application has generated dossier text, a hidden latent quality label used only for oracle scoring, and synthetic age metadata in the range configured by `live_postd`

`oc_panel`. True latent quality is generated independently of age by default. Age is therefore a nuisance/protected-attribute stress test: any age-conditioned recommendation shift is reviewer bias expression, not signal. We report the observable older-minus-younger recommendation-rate disparity as a diagnostic in the spirit of protected-attribute error-rate auditing Hardt et al. (2016) (age is a probe, not an endorsement; see Ethics).

3.2.2 Reviewer profiles: expertise and age-bias prompts

Each synthetic reviewer has an expertise level and an irrelevant age-bias factor. Positive age bias means the reviewer erroneously favors older applicants; negative age bias means the reviewer erroneously favors younger applicants. Expertise controls sensitivity to merit evidence. The `ideological_selection` strategy key is kept internally for compatibility with the synthetic pipeline, but in the postdoctoral panel it is displayed as **same-bias selection**: a deliberately non-representative bloc that concentrates one bias direction.

`scripts/run_postdoc_panel.py` first runs the analytical postdoc vote model and then, unless explicitly asked for an offline smoke run, runs the live Gemma panel against Ollama with `--require-live`. The manuscript-facing configuration uses 12 seeds, 48 synthetic reviewers, 72 fictitious applications per seed, panel sizes 3, 6, 4 sampling strategies, temperature 0.2, `num_predict=1`, and timeout 20.0 s. Vote-cache keys include the config hash, seed, reviewer id, application id, model digest, and decode parameters, so interrupted live runs can resume without mixing incompatible votes.

3.2.3 Postdoc aggregation: analytical-vs-Gemma alignment

For each sampled panel, selected reviewers vote on the same fictitious applications. Panels of size greater than three are evaluated with the same ensemble-of-trios rule as the synthetic track: usable trios are passed through the exact three-classifier EIE and majority-vote evaluators, scored against the hidden oracle, and averaged. The live artifact records per-seed/per-strategy/per-size EIE error, majority-vote error, usable-trio counts, degeneracy counts, older-minus-younger recommendation-rate disparity, panel composition, model digest, decode parameters, and vote-cache provenance in `output/data/postdoc_panel_results.json`. The companion `output/data/postdoc_panel_alignment.json` compares analytical and Gemma directional signs cell by cell and marks unresolved cells explicitly. A generated local web explorer (`output/web/ntqr_explorer.html`) is a non-publishing reader/QA aid over these same source artifacts; it exposes filters and source tables but does not change the PDF claim boundary, and a statistic is eligible for manuscript prose only after it is regenerated into the static artifacts and the token/caption contract.

4 Results: what resolved and what stayed bounded

Numbers below are token-injected by `src/ntqr_allotment/manuscript_variables.py` from the artifact named in each section: `sweep_aggregated.csv` for strategy and size summaries, `independence_sweep.csv` for tolerance, `postdoc_panel_results.json` and `postdoc_panel_alignment.json` for the live Gemma reviewer-panel companion, `power_analysis.csv` / `sweep_results.json` for design budgets, and `alarm_timings.csv` for alarm timing. None are hand-transcribed. Errors are mean EIE recovery error against the supervised oracle, aggregated over 96 seeds in the active sweep profile.

This section adjudicates the five hypotheses from the Introduction. H1, H3, and H5 each resolve in one subsection. H2 and H4 are addressed in two stages: first against the baseline grid (where, by construction, both are design-limited because the baseline judges err independently), then resolved together by the composition-coupled confound. Remaining subsections report supporting diagnostics. The Discussion returns the per-hypothesis verdicts.

4.1 Synthetic deterministic results: controlled spine for H1-H4

The synthetic results are generated from seeded populations and corpora with a known oracle. They support strategy, regime, tolerance, power-budget, alarm, and diagnostic claims for the active deterministic profile only. Throughout, a *regime* is one (expert-stringency × bias-spread × panel-size) cell of the sweep grid (sixteen cells in the active profile); contrasts are evaluated cell by cell and averaged only where the text says so.

4.1.1 Formation strategy sets the recovery floor

H1 (formation strategy is the dominant lever). The strategies do not fan out into a graded four-way ranking; instead one rule stands far apart and the other three collapse together (fig. 2). Competence-first selection recovers at mean EIE error 0.037 — roughly a quarter of the error of any other rule — while representative sortition, random selection, and single-bloc ideological selection are **statistically indistinguishable from one another**, clustered at 0.147–0.148:

Strategy	Mean EIE error	95% CI
expertise threshold (far best)	0.037	±0.003
representative sortition	0.147	±0.014
random selection	0.148	±0.014
single-bloc ideological selection	0.148	±0.015

The dominant lever is therefore *whether the panel is curated for competence at all*, not a graded property of the formation rule. The competence-first-vs-sortition contrast is a **resolved** separation on this instrument: the trio-level bootstrap separation gate reads **separated** (means 0.037 vs 0.122), the two strategies' pooled confidence intervals in the table above are disjoint (0.037 ± 0.003 versus 0.147 ± 0.014), and the power analysis resolves it as **significant** (well-powered, not design-limited). Every contrast *among* the other three is, by contrast, design-indistinguishable — including the representative-vs-single-bloc pair, whose effect is inconclusive (its 95% CI crosses zero, inconclusive (95% CI crosses zero)). So competence-first selection beats the representative draw by a resolved interval, but representativeness, randomness, and single-bloc concentration are interchangeable for recovery on this instrument — a sharper and more honest result than a four-way ranking would suggest.

4.1.2 Sortition only separates when the confound rides on the balanced axis

H2 (concentrating correlated error degrades recovery). fig. 3 reports the single-bloc-minus-representative EIE error contrast across the full active regime grid. Positive cells mean the representative draw has lower post-NTQR recovery error. The axes expose the design levers directly: expert stringency (`mean_expertise`), ideological bias spread (`bias_std`), and sortition size (`panel_size`).

On this baseline grid the contrast is design-limited and must be read cell by cell: the analytical prediction is directional (single-bloc ideological selection should not beat representative sortition when bias is the manipulated dependence source), and the heatmap shows where the regenerated synthetic data aligns, where it remains uncertain, and where the active design is too small to resolve a sign. This grid cannot fan the strategies apart because it never realizes H2's premise — its judges err independently; the composition-coupled confound that supplies the missing channel, and resolves H2, is reported in §Composition-coupled correlation fans the strategies apart.

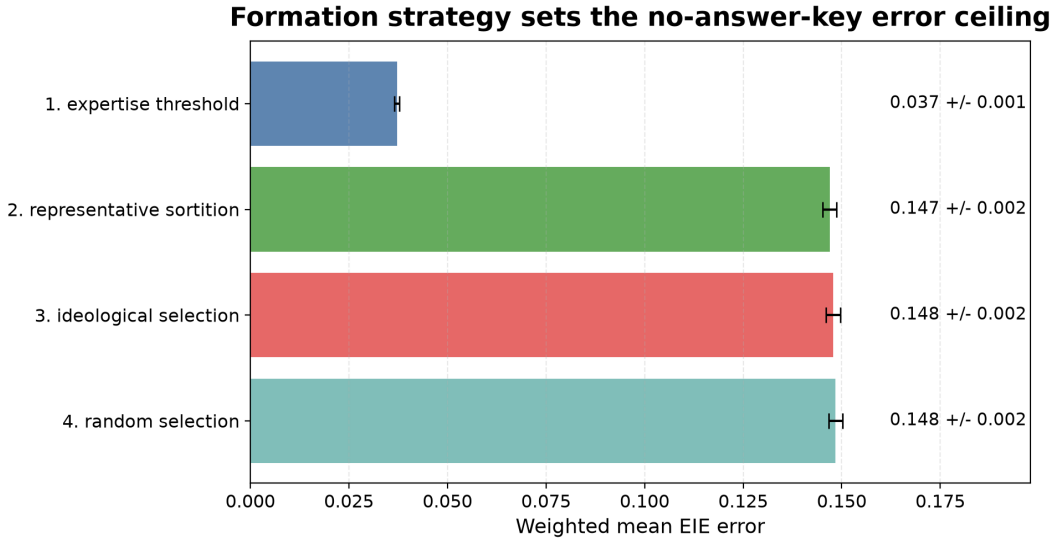


Figure 2: Horizontal bars give the mean oracle-referenced EIE recovery error for each of the four panel-formation strategies, ordered best (lowest, top) to worst (highest, bottom); the whiskers are 95% confidence intervals over 96 seeds and the value beside each bar is the mean +/- half-interval, all from source `output/data/sweep_aggregated.csv` (profile manuscript_contrast, hash fda4da941cf0). Read it as the ceiling each upstream choice imposes on the downstream blind estimator: competence-first selection sits far left (near-zero error) while the other three strategies cluster together to its right. Claim: the upstream formation rule, not the estimator, sets the no-answer-key error ceiling, and the competence-first-vs-rest gap dwarfs anything the estimator does on a fixed panel; caveat: only the competence-first separation is resolved — representative, random, and single-bloc selection are statistically indistinguishable from one another, as the power/separation layer certifies.

4.1.3 NTQR beats majority voting only in selected regimes

For the pre/post comparison, "pre-NTQR" means the supervised majority-vote baseline already stored as `mv_error`; "post-NTQR" means the ground-truth-free EIE recovery error stored as `eie_error`. fig. 4 plots `EIE - MV`, so negative cells mean the NTQR recovery estimate is closer to the supervised oracle than the majority-vote baseline for that regime.

The companion alignment map (fig. 5) makes the analytical layer auditable rather than rhetorical. Each cell reports how many expertise levels in that size-bias slice match the directional prediction that ideological-minus-representative EIE should be positive. The same JSON artifact also records the monotone checks for bias and expertise.

4.1.4 Larger panels are a neutral sampling knob here

H3 (size is a sampling knob, not a uniform improvement). fig. 6 plots EIE error against panel/ensemble size for each strategy. If size were a clean power knob, every curve would fall from size 3 to size 6. It does not:

Strategy	Size 3	Size 6	Pooled direction
expertise threshold	0.037	0.038	roughly flat
representative sortition	0.122	0.154	error rises
random selection	0.118	0.155	error rises
ideological selection	0.124	0.151	error rises

Those Size-3/Size-6 cells and the figure's end-labels are **pooled point estimates** over sixteen regimes. The powered test is a **paired** contrast that matches each regime-and-seed cell across the two sizes (`paired_size_contrast` in `src/ntqr_allotment/power_study.py`), removing the between-regime variance. Under that test 3 of the four strategies show a *resolved* trio-to-six-seat change, and every resolved change is a small **increase** in error: random selection (+0.015, 95% CI [+0.009, +0.021]), representative sortition (+0.007, CI [+0.001, +0.012]), and competence-first selection (+0.004, CI [+0.003, +0.005]); only single-bloc selection (-0.003, CI [-0.007, +0.002]) is within noise. These effects are **resolved but negligible** — the largest, +0.015, is about a tenth of the bottom-tier baseline near 0.148. So more experts do not help, and at most very slightly hurt: we reject the simple hypothesis that more experts always help, but the honest reading is that **size is essentially neutral** at this grid, and the dominant lever is **strategy**, not size.

What the size diagnostic rules out. The error-independent solver assumes the three judges' errors are uncorrelated, so the natural guess is that larger panels feed the ensemble more error-correlated trios. A per-trio diagnostic refutes that guess (fig. 7, `src/ntqr_allotment/trio_conditioning.py`, over 17,126 usable trios). The realized mean absolute

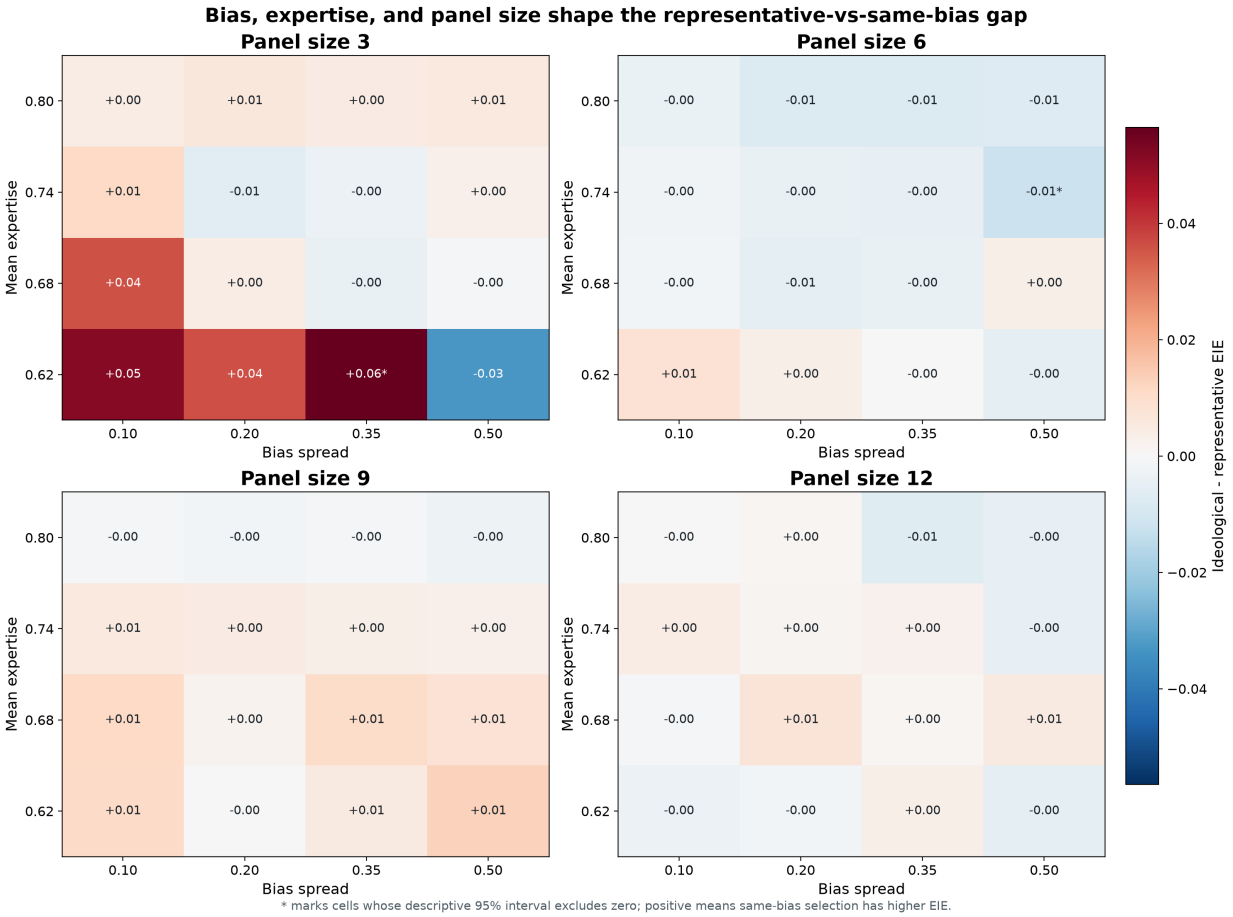


Figure 3: Faceted heatmaps of the single-bloc-minus-representative EIE error contrast across expert stringency (mean expertise, rows), ideological bias spread (columns), and panel size (one facet per size), from source `output/data/sweep_aggregated.csv` (profile manuscript_contrast, hash fda4da941cf0), with analytical sign predictions overlaid from `output/data/analytical_predictions.json`. Colour encodes the signed contrast on a diverging scale centred at zero: positive (red) cells mean the representative draw recovers with lower error than the single-bloc draw in that regime, negative (blue) the reverse. Statistic: cell-level mean contrast over 96 seeds; stars mark descriptive 95% intervals excluding zero. Read across a row to see how stringency modulates the gap and down a column to see the effect of bias concentration. Claim: the representative-vs-single-bloc distinction is not a single number but is regime-dependent, becoming visible only as bias, stringency, and size are jointly varied; caveat: the intervals are descriptive and synthetic-profile bounded, and they are deliberately not pooled with the real Ollama evidence, which lives in the separate Gemma postdoc companion in fig. 17, fig. 18, and fig. 19.

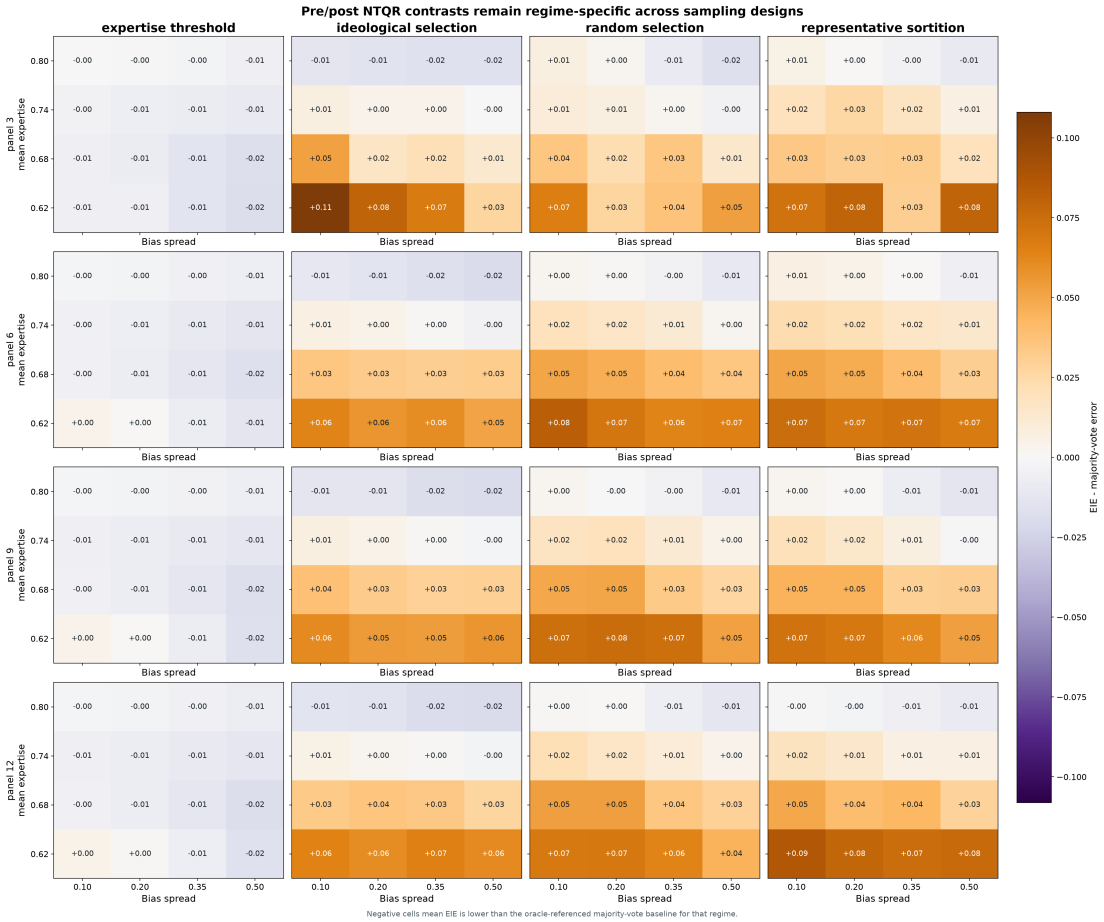


Figure 4: Faceted heatmaps contrasting the post-NTQR error-independent estimate against the pre-NTQR supervised majority-vote baseline, broken out by strategy, panel size, expert stringency, and bias spread, from source `output/data/analytical_predictions.json`, derived in turn from `output/data/sweep_aggregated.csv` (profile manuscript_contrast, hash fda4da941cf0). Colour encodes the signed difference on a diverging scale: negative (blue) cells are where the blind NTQR recovery lands closer to the oracle than the majority-vote baseline did, positive (red) where it does not. Statistic: cell-level mean difference over 96 seeds; metric is `eie_mean - mv_mean`, so the figure isolates what the estimator adds (or costs) on top of naive voting. Claim: panel size and bias act on each formation strategy differently before and after NTQR recovery, so there is no uniform pre/post improvement; caveat: this is oracle-referenced simulation on synthetic labels, not a live-judge validation claim.

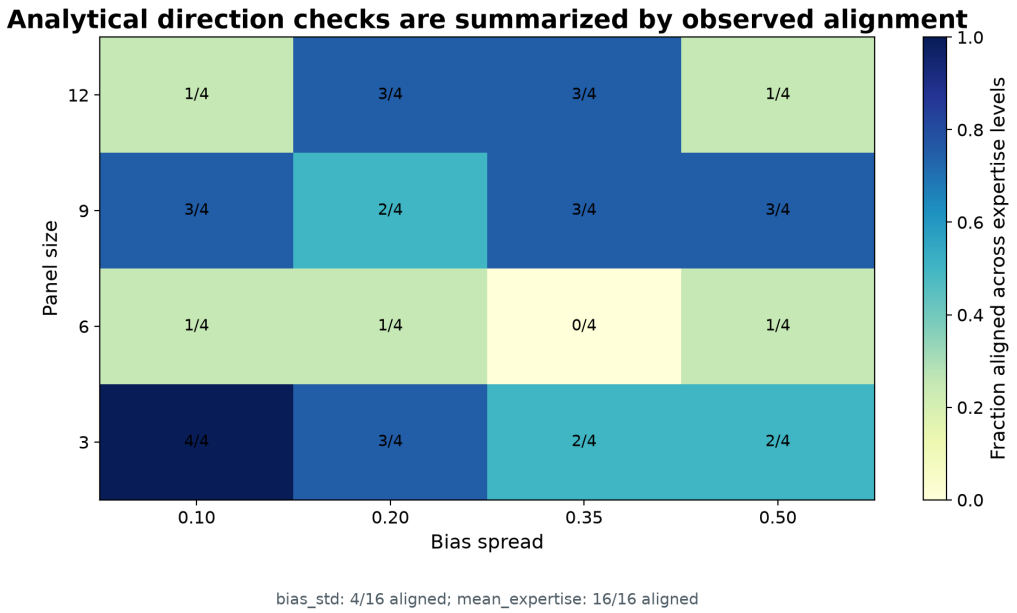


Figure 5: Audit heatmap of how often the analytical directional predictions match the regenerated synthetic cells, from source `output/data/analytical_predictions.json`. Each cell of a panel-size x bias-spread slice reports the count of expertise levels whose observed contrast sign agrees with the predicted sign; darker cells mean more of the expertise levels in that slice align with the prediction. Statistic: aligned expertise-level cells per panel-size x bias slice, over source sweep profile manuscript_contrast, 96 seeds. The figure exists to make the analytical layer falsifiable rather than rhetorical: a prediction that systematically disagreed with the data would show as pale cells. Claim: analytical expectations are checked against regenerated artifacts rather than asserted in prose; caveat: the predictions are directional and order constraints only, not closed-form numerical EIE laws, so partial agreement is expected and is reported honestly.

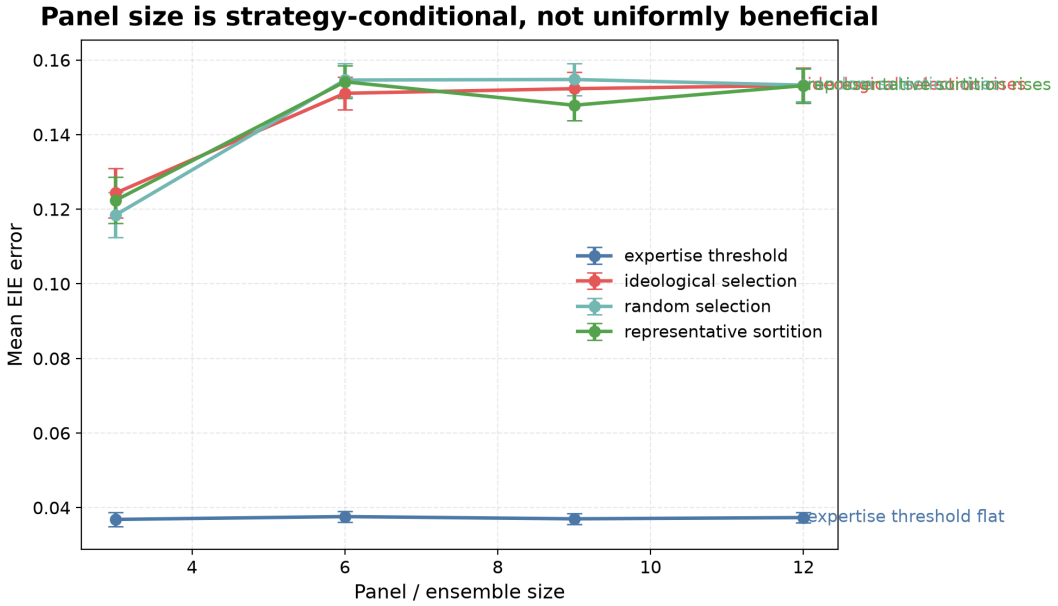


Figure 6: One line per panel-formation strategy tracing mean EIE recovery error against panel/ensemble size (3, 6, 9, 12 members), from source `output/data/sweep_results.json`, aggregated over 96 seeds with per-point 95% confidence intervals and a colour-matched end-label summarising each curve's trio-to-six-seat direction. If size were a clean power knob every curve would fall monotonically left to right; instead the curves cross. Read the vertical spread at any size as the strategy gap and each curve's slope as the strategy-specific effect of adding experts. Claim: a paired regime-controlled test (`paired_size_contrast`) resolves a trio-to-six-seat size effect for three of the four strategies, but every resolved increase is tiny and single-bloc is within noise, so size is essentially neutral at this grid and the dominant lever is which strategy forms the panel; caveat: this pooled curve marginalizes over sixteen regimes and is bounded to the active sweep profile.

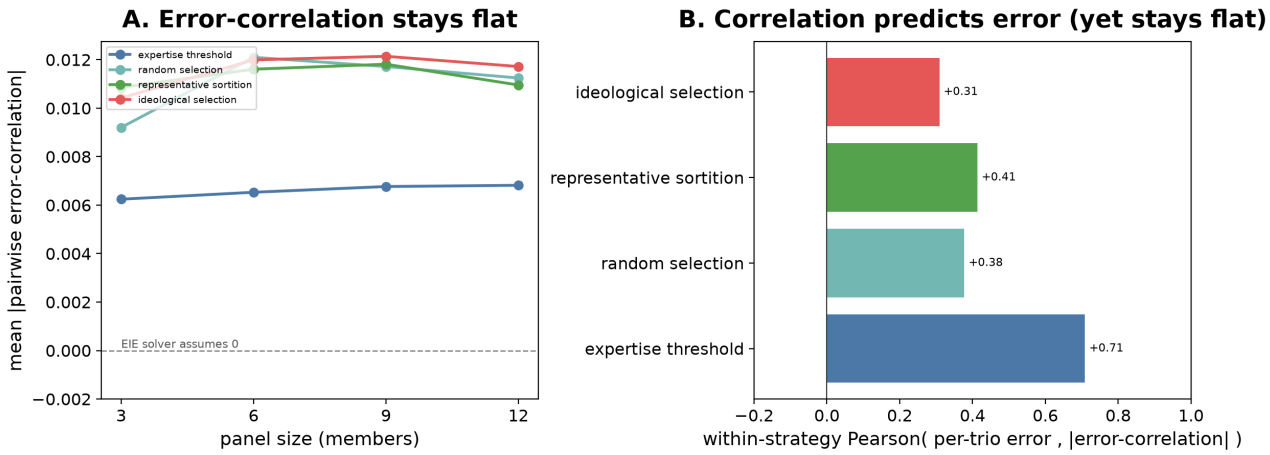


Figure 7: Two-panel per-trio mechanism diagnostic for the panel-size contrast, from source `output/data/trio_conditioning.json` (17,126 usable trios over 12 seeds of the reported regime grid). Panel A plots the mean absolute pairwise error-correlation of the usable trios against panel size, one line per formation strategy, with a dashed reference at zero (the value the error-independent solver assumes); every line holds the small 0.0087-to-0.0105 baseline rather than rising, so enlarging the panel does not pull in more error-correlated trios. Panel B plots the within-strategy Pearson correlation between a trio's recovery error and its absolute error-correlation; the bars are positive (up to +0.70 for competence-first), so correlation genuinely predicts per-trio error — it simply does not grow with size. Statistic: mean absolute error-correlation by size (A) and within-strategy Pearson of per-trio error against absolute error-correlation (B), measured over the same usable trios the ensemble-of-trios averages. Claim: the diagnostic rules out a size-growing error-correlation mechanism; caveat: it does not identify a positive mechanism and remains a 12-seed structural diagnostic on synthetic labels, not a headline confidence interval.

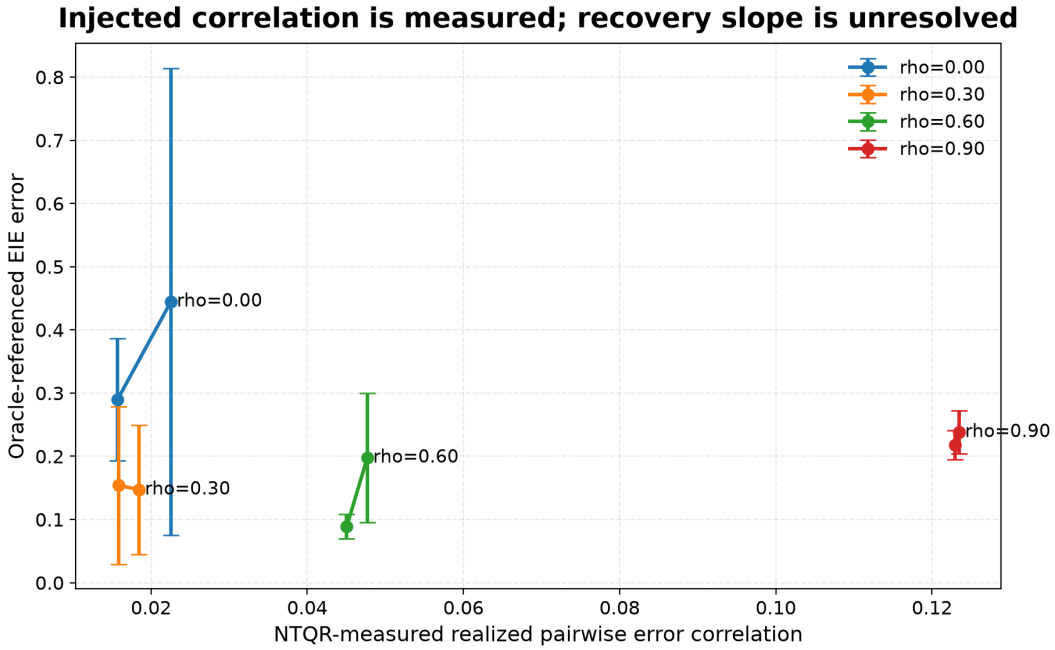


Figure 8: Oracle-referenced EIE error (y-axis) versus the realized pairwise error correlation ρ_{NTQR} that the NTQR estimator itself measures from the votes (x-axis), shown as a scatter with one point per non-degenerate (ρ, strategy) cell of the tolerance sweep, from source `output/data/independence_sweep.csv`. The x-axis is the quantity the exact solver assumes is zero; the controlled injection knob moves points rightward, which verifies the diagnostic, while the y-axis tests whether more correlation actually costs recovery accuracy. Claim: injected ρ creates measurable NTQR error correlation (points do move right), but the fitted recovery-error trend is unresolved at this grid; statistic: OLS error-vs-correlation slope -0.145 with 95% bootstrap CI [-4.335, 0.911], which crosses zero; caveat: this figure supports the correlation diagnostic only, not a resolved recovery-effect law.

none, and this grid is simply too small to choose between them (mean EIE error 0.367 at the lowest injected `rho` and 0.228 at the highest, with a non-monotone dip in between). We report the interval rather than a bare point estimate precisely because a single noisy slope on eight cells would overclaim a precision the data do not support.

What the instrument *does* establish here is narrower and robust: the realized correlation tracks the injected correlation (the monotone rise above). The *analytical* expectation encoded in `theory.py` (`predicted_error_vs_correlation`) is that recovery error should **not decrease** as positive error-correlation grows, because correlation violates the error-independence the exact solver assumes. That expectation is **not confirmed on this global-injection grid**: `independence_sweep.csv` shows a slight, non-monotone *decrease* (0.367 at the lowest injected `rho` to 0.228 at the highest, slope 95% CI crossing zero). We now attribute that non-monotonicity to a **disclosed limitation of this particular injection model**, not to small-grid noise alone: `dependence.sample_votes_correlated` mixes a shared and an independent *uniform* latent, and a convex combination of uniforms is not uniform, so the model's realized per-judge accuracy is **not preserved as `rho` varies** — it inflates and then deflates, peaking near `rho=0.5`. That accuracy confound moves recovery error in its own right and contaminates the recovery-vs-correlation slope. Rather than re-engineer this diagnostic (which would perturb a shipped result), we draw the H4 recovery conclusion from the **marginal-accuracy-preserving composition-coupled instrument** of the following subsection (§Composition-coupled correlation fans the strategies apart), which holds each judge's accuracy fixed by construction and resolves the recovery-vs-correlation relationship in the affirmative (fig. 9). The global-injection sweep remains in the paper as the correlation *diagnostic* (the realized correlation does rise with `rho`) with its accuracy artifact disclosed. This is a **measured behaviour under controlled correlation**, validated in simulation only; it does *not* show that sortition restores low oracle-referenced error on real prompted judges. The live Gemma postdoctoral-review panel that probes the same sampling mechanism on a local LLM is reported in the Real-Ollama results subsection.

4.1.6 Composition-coupled correlation exposes the sortition mechanism

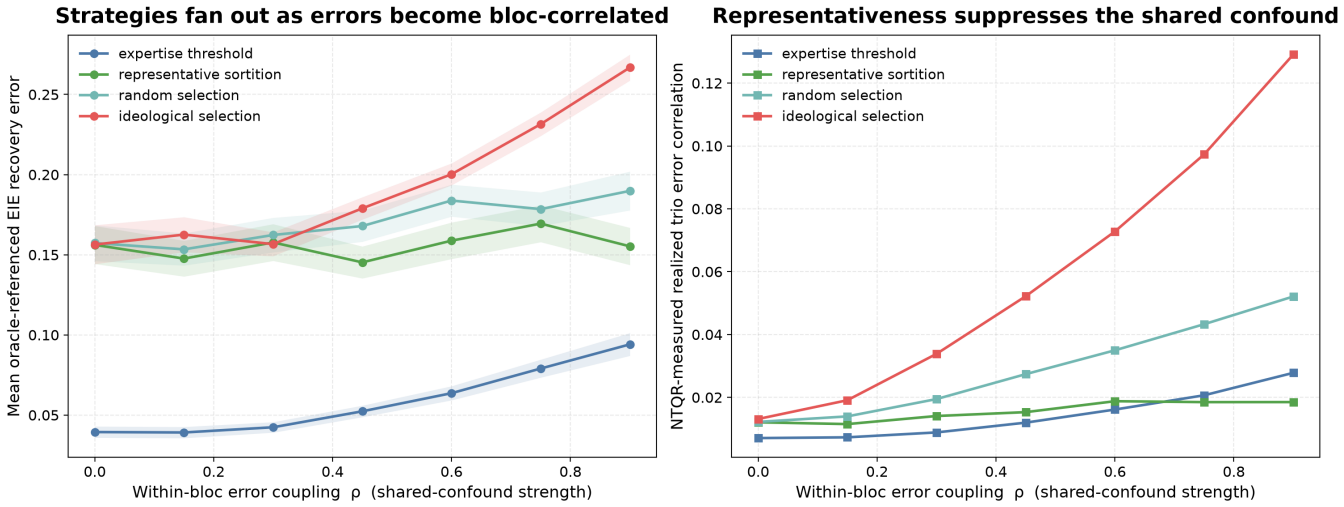
The H4 slope above is unresolved for a specific, fixable reason. The tolerance sweep injects a *global* correlation onto a *fixed* trio, identically for every strategy, so it measures sensitivity to correlation **decoupled from how the panel was formed**. That is the wrong instrument for H2, whose premise is that single-bloc selection *seats judges whose errors are correlated* — a premise the baseline generator never realizes, because there ideology shifts only each judge's *marginal* accuracy and every judge errs from an independent stream. Under that generator representative, random, and single-bloc panels are indistinguishable not by coincidence but **by construction**: the channel that would separate them does not exist, so no parameter sweep over the baseline can fan them out.

We close that gap with a composition-coupled confound (`src/ntqr_allotment/bloc_confound.py`). Judges who share an ideological bloc draw a shared latent *error shock* through a Gaussian copula of within-group strength ρ; the construction preserves each judge's per-label accuracy exactly, so any change in recovery is attributable to error *correlation*, not to an accuracy shift, and $\rho = 0$ reproduces the independent baseline. (The shared channel is a symmetric competence shock, not directional bias. ρ is the latent within-group correlation; what we plot as "realized correlation" is NTQR's own label-conditional error-correlation statistic, which is much smaller in magnitude than ρ — we report the quantity the solver assumes is zero, never the latent ρ.) We sweep ρ across 7 levels, aggregating on average 239 non-degenerate trials per (strategy, ρ) point over bias-spread, stringency, panel-size, and seed regimes (fig. 9).

The result is a clean, graded fan-out. At $\rho = 0$ the three composition strategies collapse exactly as the baseline reported — the ideological-minus-representative gap is 0.000. As coupling rises they fan out: representative sortition stays essentially flat (0.156 to 0.155 EIE error), random selection degrades (0.157 to 0.190), and single-bloc ideological selection degrades most (0.156 to 0.267), widening the gap to 0.112 at $\rho = 0.90$. NTQR's own correlation diagnostic makes the mechanism legible: at high coupling a representative trio carries measured error-correlation 0.018 while a single-bloc trio carries 0.129 — bloc-balancing *decorrelates* the shared shock that bloc-concentration *concentrates*. This resolves **H2** (concentrating correlated error does degrade recovery, once the correlation is composition-coupled) and answers the open **H4** recovery slope in the affirmative under a correctly specified, marginal-preserving instrument. It is not a pooling artifact: in a paired per-regime test at $\rho = 0.90$, single-bloc error exceeds representative error in 180/205 matched regimes (paired mean 0.102, 95% CI ± 0.012). Nor is the averaged subsample cherry-picked by the degenerate-trio skip: at high coupling the single-bloc panels have the *lowest* degenerate-trio rate of the four strategies (tracked per strategy in `bloc_phase_summary.json`), so dropping ill-posed trios cannot be what manufactures their degradation — if anything it makes the reported gap conservative.

This robustness is **conditional, not magical**, and a negative control says so. The protection appears because representative sortition balances the very axis — ideology — that the confound rides on. When the shared shock is re-keyed to an orthogonal axis the lottery does *not* balance (expertise tier), representative sortition loses its immunity: its error climbs from 0.147 to 0.229, and the large ideological-minus-representative gap of the matched axis (0.112 at $\rho = 0.90$) nearly closes under the orthogonal one (0.026). The point is not that some *other* strategy inherits the protection — competence-first selection draws the top experts, who span expertise tiers, so it does not maximally concentrate the tier axis either — but

that representativeness on ideology stops mattering once the confound no longer rides on ideology. The defensible claim is therefore precise: **balancing a panel on the axis a shared error rides on preserves no-answer-key recovery; balancing the wrong axis does not.** That is a statement about sortition design in simulation, not a blanket endorsement of representative panels.



Source: output/data/bloc_phase_summary.json — mean over bias-spread x stringency x panel-size x seed regimes. Left: recovery error; right: the mechanism (measured correlation).

Figure 9: Two-panel bloc-confound phase diagram from source `output/data/bloc_phase_summary.json`, aggregated over bias-spread, stringency, panel-size, and seed regimes. Left: mean oracle-referenced EIE recovery error (y) versus within-bloc error coupling ρ (x), one line per panel-formation strategy with 95% CI bands; the lines coincide at $\rho = 0$ (the reproduced baseline collapse) and fan out as ρ rises, representative sortition staying flat while single-bloc ideological selection climbs. Right: the NTQR-measured realized trio error correlation (y) versus ρ — the mechanism — showing representative sortition suppressing the shared confound (flat, low) while single-bloc selection concentrates it (steeply rising). Claim: composition-coupled error correlation makes panel-formation rule the dominant lever on recovery, with representativeness protective specifically when the panel balances the axis the confound rides on; caveat: synthetic, marginal-preserving simulation against a known oracle, and the protection is axis-conditional as the expertise-tier negative control demonstrates.

Representativeness is not only a four-way contrast but a **continuous dial**. Fixing the coupling at $\rho = 0.90$ and forming panels with a tunable single-bloc concentration c — from balanced ($c = 0$, Herfindahl index $1/B$) to single-bloc ($c = 1$, Herfindahl index 1) — recovery error climbs monotonically from 0.175 to 0.263 across the 6 dial levels (fraction of steps that increase error: 1.000; fig. 10). This is the closed-form Herfindahl account of §Methods made visible: the panel's concentration index over the confound axis, a closed-form combinatorial statistic, sets its shared-confound exposure and hence its no-answer-key recovery error.

4.1.7 Power budgets distinguish ranking from resolved contrasts

Before any "beats" wording, competence-first selection and representative sortition are compared by separately bootstrapped mean intervals at the trio (`strategy_separation`, `src/ntqr_allotment/statistics_analysis.py`): mean recovery error 0.037 (competence-first) versus 0.122 (representative), a signed difference of -0.086 with a CI-overlap verdict of **separated** — that is, the two strategies' separately bootstrapped mean intervals do not overlap, so the difference is not an artifact of within-strategy spread. The power study then separates resolved contrasts from design-limited nulls: of 28 pairwise strategy contrasts, 12 are well-powered and 16 are underpowered at the current observation count, while 13 of 28 reach raw permutation-test significance (12 after Holm-Bonferroni across the 28-test family). The strategy *ranking* is therefore a point estimate ordering plus a set of explicitly tested contrasts, not a blanket claim that every neighboring rank is a significant win. fig. 12 shows the analytic power-vs-sample-size curves that set these budgets. For a two-sample contrast, the design quantities are standardized effect d, per-group observation count n (seeded trials across the active profile cells), Type-I error α, target power $1 - \beta$, and the minimum detectable effect (MDE) at the chosen n. The budget is reported with a minimum detectable effect of 0.101 and between 5 and 543625 per-group observations required to reach 80% power **for effects of the magnitudes observed in these contrasts** — a prospective design target keyed to the observed effect sizes, never retrospective observed power (fig. 11, `output/data/power_analysis.csv`).

The remaining "inconclusive" verdicts are therefore statements about *design size*, made explicit through the minimum detectable effect — not evidence of no effect, and never reported as retrospective observed power.

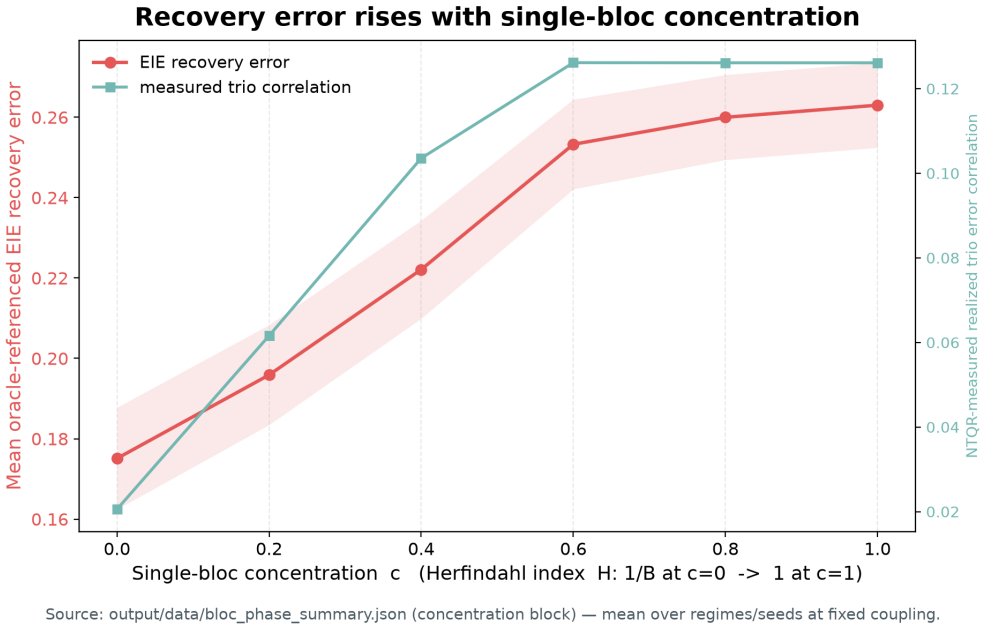


Figure 10: Single-panel concentration-dial figure from source `output/data/bloc_phase_summary.json` (concentration block), aggregated over bias-spread, stringency, and seed regimes at fixed within-bloc coupling. The x-axis is the single-bloc concentration dial c (the panel's Herfindahl index runs $1/B$ at $c = 0$ to 1 at $c = 1$). Left y-axis (circles, with 95% CI band): mean oracle-referenced EIE recovery error; right y-axis (squares): the NTQR-measured realized trio error correlation. Recovery error rises monotonically with concentration, while the realized correlation rises and then saturates once the scored trio becomes single-bloc. Claim: recovery error is a graded function of panel concentration over the confound axis, tracing the closed-form Herfindahl prediction rather than a binary representative-vs-single-bloc split; caveat: synthetic, marginal-accuracy-preserving simulation at one coupling level against a known oracle, and the protection implied by low concentration is axis-conditional — it holds only for the axis the confound rides on, as the expertise-tier negative control in fig. 9 shows.

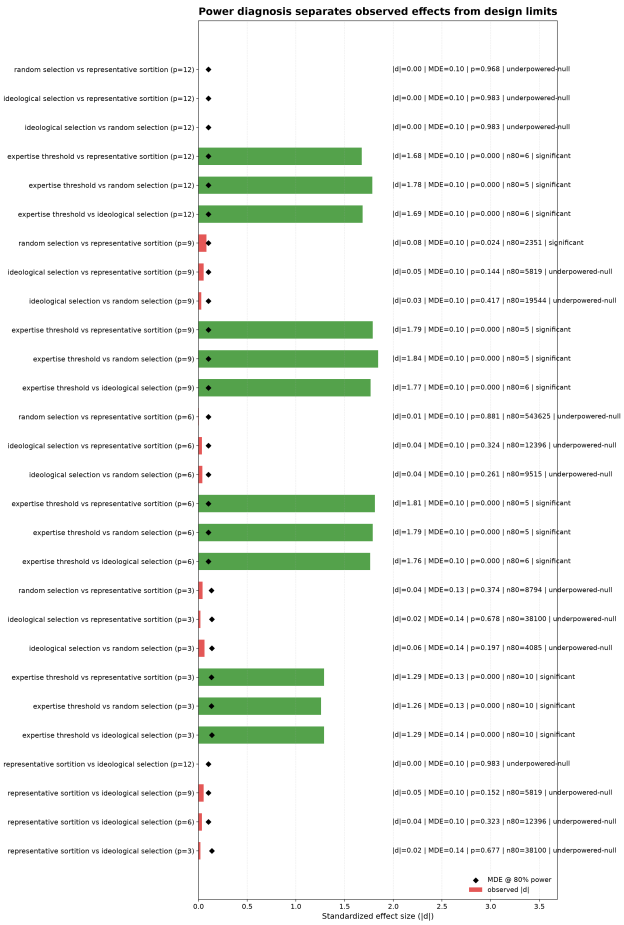


Figure 11: Design-adequacy diagnosis for every pairwise strategy contrast, from source `output/data/power_analysis.csv`: each row plots the observed standardized effect size (Cohen's d) against the minimum detectable effect (MDE) at 80% power, annotated with the permutation p-value and the per-group observation budgets needed to resolve an effect of that size. A contrast whose observed $|d|$ sits below its MDE marker is design-limited: the study could not have detected it even if it were real, so its non-significance is a statement about sample size, not about the absence of an effect. Claim: 16 of 28 contrasts are design-limited at the current observation count, which is why several neighboring-rank comparisons remain inconclusive; caveat: the budgets are keyed to the observed effect magnitudes and are prospective design targets, not retrospective observed power evidence.

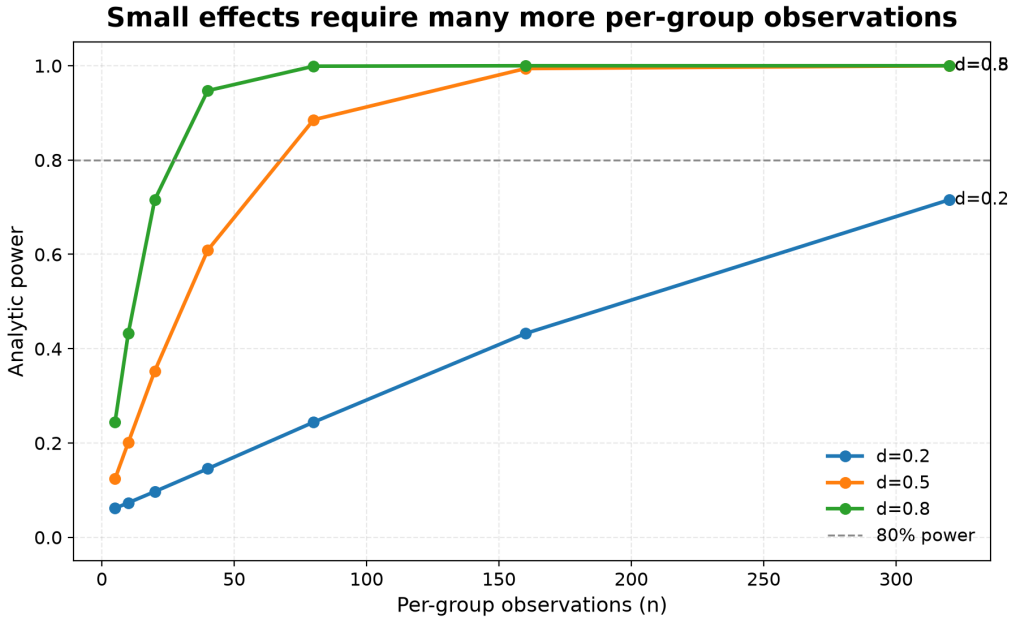


Figure 12: Analytic two-sample power $1 - \beta$ (y-axis) versus samples-per-group n (x-axis, log scale), shown as a family of curves with one curve per representative Cohen's d value, computed from source `output/data/power_analysis.csv` / `src/ntqr_allotment/power_analysis.py`; the horizontal dashed line marks the 80% power target at the chosen α, and where each curve crosses it gives the n that effect size requires. Read it as a budgeting tool: the smaller the standardized effect, the further right its curve crosses the dashed line, i.e. the more per-group observations are needed. Claim: small standardized effects require many more seeds per group than the current design provides, which is the mechanism behind the design-limited nulls; caveat: this is a prospective design-budget curve and an MDE visual, not retrospective observed power.

4.1.8 Companion diagnostics bound cost, correlation, fairness, and consistency

The companion alarm's answer-key enumeration is roughly cubic in corpus size (fig. 13): about 0.7 s at $Q = 20$, 8.9 s at $Q = 50$, and 97.9 s at $Q = 100$ (measured by `scripts/bench_alarm.py`, written to `output/data/alarm_timings.csv`). This is a real ceiling on the alarm track, so it is opt-in and capped at $Q \leq 30$. We report it as a finding: the alarm is usable as a small-corpus consistency check, not as a sweep-scale primitive.

Three further companion tracks measure structural properties of the pipeline rather than recovery error, and we report them as diagnostics. The error-correlation track records the mean realized correlation each formation strategy induces (fig. 14): single-bloc selection sits highest, consistent with its status as the deliberately correlated comparator. The maximin fairness track characterizes the representative lottery's selection-probability distribution over the population (fig. 15) — the maximin objective is the floor on who can be seated, independent of any downstream recovery number. The N-judge alarm-power track records a saturated small-Q alarm-firing rate across the plotted panel sizes (fig. 16). The ternary ($R = 3$) track is consistency/feasibility only — it confirms three-way vote profiles satisfy the NTQR axioms and is never an $R = 3$ recovery claim (out of scope) — so it yields a pass/fail check rather than a plotted number.

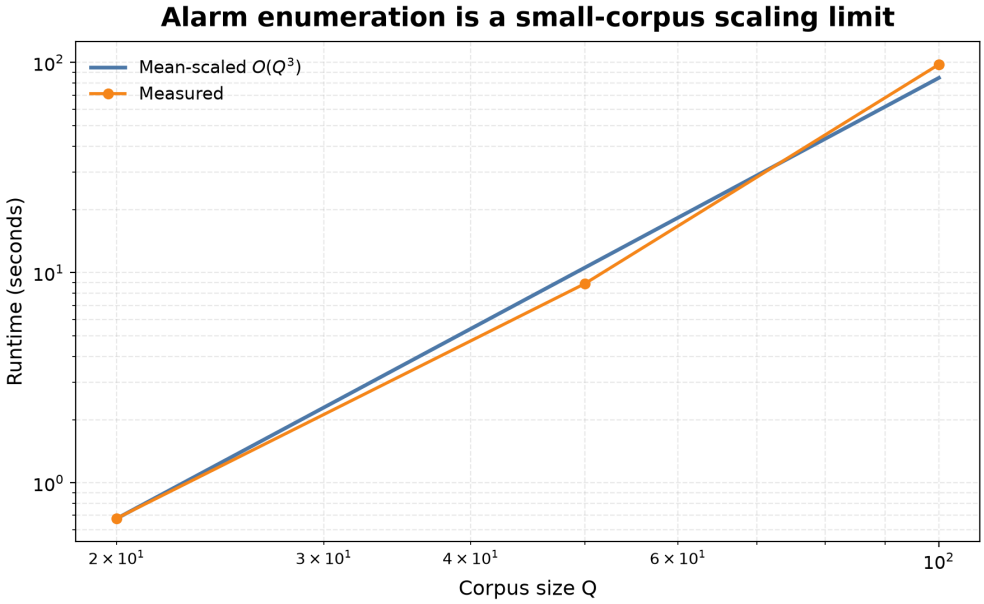


Figure 13: Measured alarm wall-clock time (seconds) versus corpus size Q on log-log axes, from source `output/data/alarm_timings.csv`, with a cubic $O(Q^3)$ reference line overlaid. On log-log axes a power law is a straight line whose slope is its exponent, so the measured points tracking the reference slope is the visual evidence that the answer-key-enumeration alarm scales cubically in Q. The practical consequence is a hard ceiling: the wall-clock cost rises steeply enough that the alarm is usable only as a small-corpus consistency check. Claim: at the current implementation the alarm is small-corpus only and is therefore opt-in and capped; caveat: the absolute wall-clock constants are machine-local and load-dependent, so it is the cubic scaling, not the individual timings, that is the robust finding.

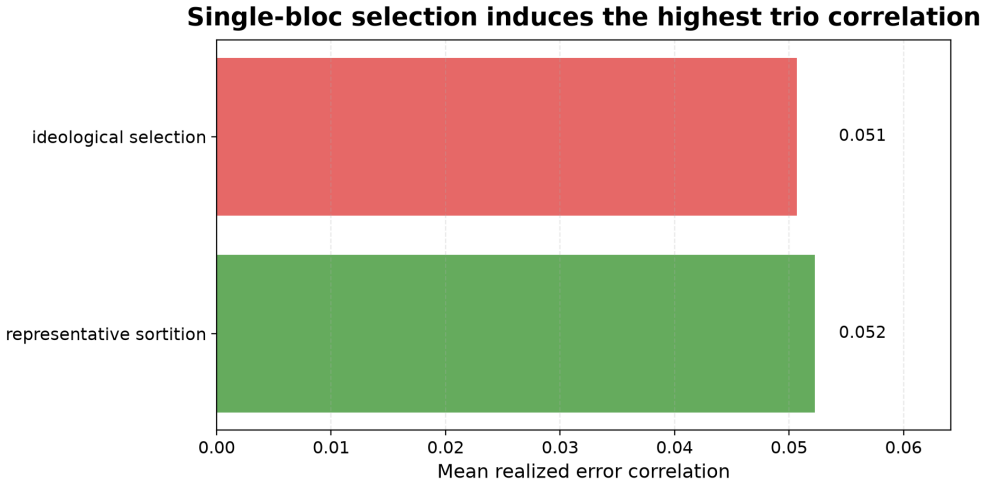


Figure 14: Bar chart of the mean realized pairwise error correlation that each panel-formation strategy induces among its judges, measured by NTQR's own supervised estimator over the tolerance sweep, from source `output/data/independence_sweep.csv`. Higher bars mean the strategy seats judges whose mistakes are more correlated, which is precisely the error-independence assumption the exact trio solver leans on. Single-bloc selection is the tallest bar, consistent with its design as the deliberately correlated comparator, while representative and random draws sit lower. Read this as a structural property of the draw itself, upstream of any recovery number. Caveat: this is a structural diagnostic of the formed panel, not a recovery-effect claim about downstream EIE error.

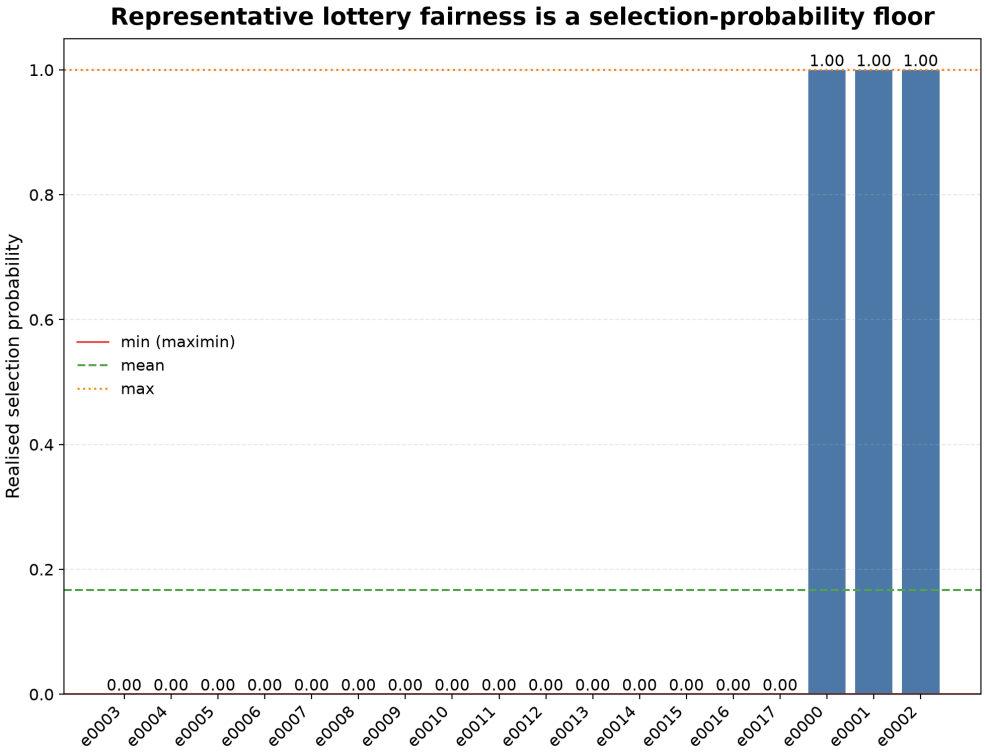


Figure 15: Per-candidate selection probabilities under the representative maximin sortition lottery, computed from `src/ntqr_allotment/fairness.py` over the feasible panel draws. Each bar is one expert's probability of being seated across the lottery; the maximin objective explicitly maximises the smallest of these probabilities, so the figure should be read by its floor (the shortest bar) rather than its average — a fairer lottery lifts the worst-off candidate's chance of selection. This characterises the representation properties of the draw and is fully independent of any downstream evaluation number. Caveat: this describes panel-formation fairness only, not NTQR recovery error.

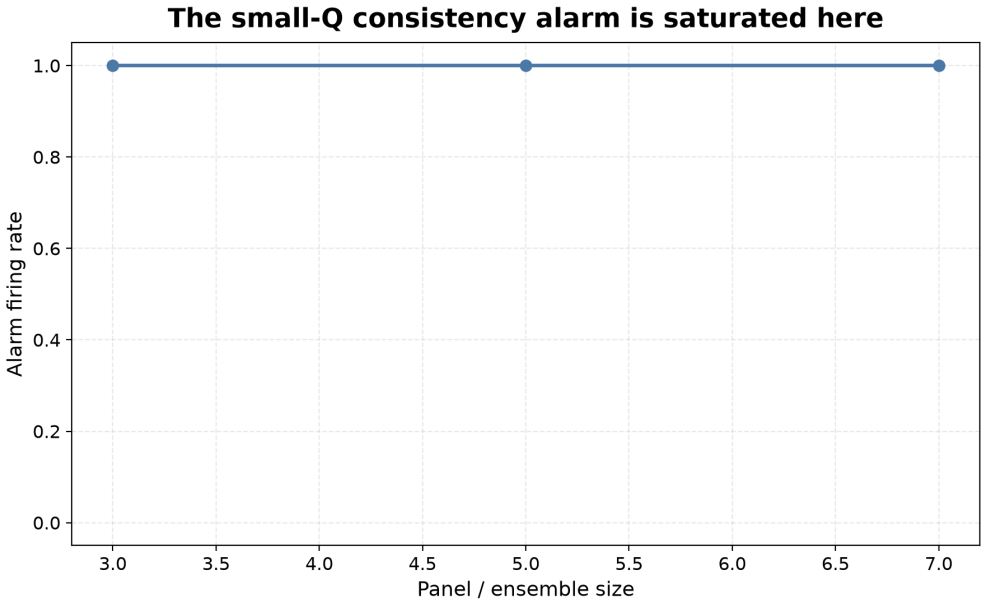


Figure 16: N-judge alarm firing rate as a function of panel size at small corpus size Q, computed live from `src/ntqr_allotment/ensemble.py`. The alarm fires when no single answer key can make all seated judges simultaneously axiom-consistent at the stated safety specification; the curve shows how often that happens as the panel grows. At the tight safety setting plotted here the rate is already saturated across the panel sizes shown, so the figure demonstrates that the N-judge alarm is executable and panel-size-indexed rather than establishing a monotone growth law (a looser setting would be needed to see a rising curve). Caveat: the alarm track is a consistency signal only, never a recovery method, and is bounded by the same $O(Q^3)$ answer-key enumeration, which confines it to small Q.

4.2 Real-Ollama postdoctoral panel results: live H5 companion

The live-Ollama results are separate empirical companion artifacts. They use one local `gemma3:4b` model prompted as synthetic postdoctoral reviewers over fictitious applications with synthetic age metadata. The result is not a model-family comparison, not a human-review validation, and not evidence that age belongs in real admissions or hiring review.

4.2.1 Gemma ranking asks the same sampling question under prompt labels

The live artifact uses 12 seeds, 48 reviewers, 72 applications per seed, panel sizes 3, 6, and live Ollama provenance (`gemma3:4b` digest a2af6cc3eb7f, config hash 5161ffe474b3, vote-cache entries 23688). The best live postdoc EIE point estimate is same-bias selection at 0.216; the worst is expertise threshold at 0.347. For the three-seat panels, representative sortition has EIE 0.225, same-bias selection has 0.228, expertise-threshold selection has 0.347, and random selection has 0.262.

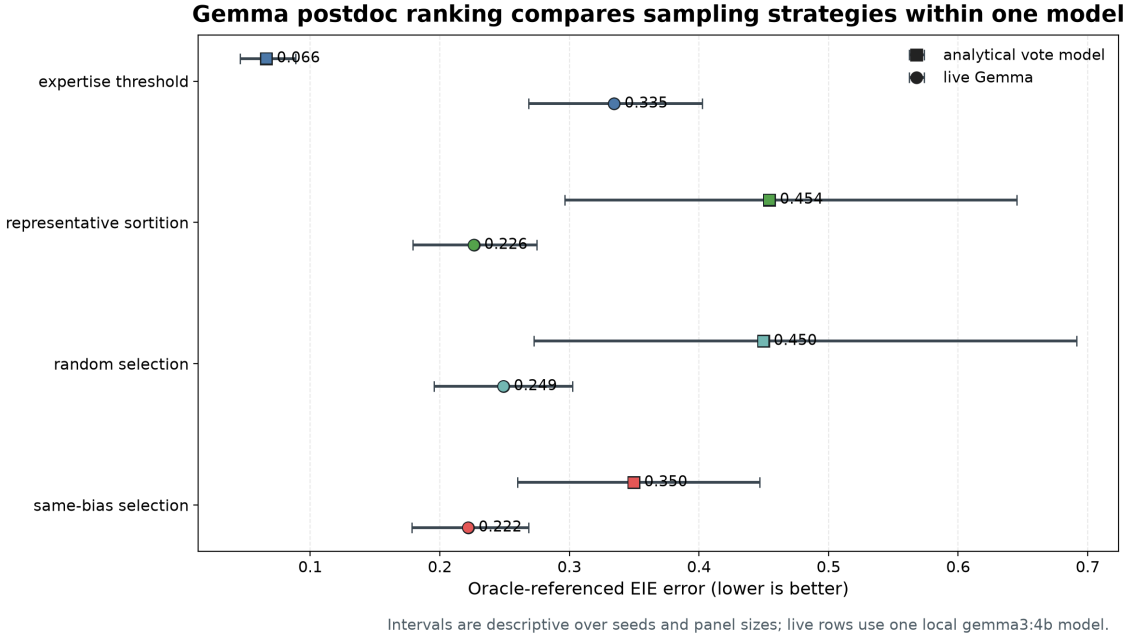


Figure 17: Side-by-side strategy ranking for the analytical vote model (square markers) and the live Gemma reviewer panel (circle markers), one row per sampling strategy, from source `output/data/postdoc_panel_results.json`. Metric: oracle-referenced EIE error (lower is better), aggregated by track, sampling strategy, and panel size, with descriptive intervals over 12 seeds. The two marker shapes are juxtaposed, never pooled, so the reader can see where the live model echoes the analytical ordering and where it departs; the horizontal gap between a strategy's square and circle is exactly that analytical-vs-live divergence. Claim: the live single-model panel is analysed as a within-model sampling-strategy stress test, not as an LLM-family comparison; caveat: it uses synthetic applications and age metadata only, one local Gemma model, and carries no human-review validation.

4.2.2 Same-bias panels expose age-conditioned recommendations

The age-bias outcome is older-minus-younger recommendation rate. Positive values mean older synthetic applicants are recommended more often; negative values mean younger synthetic applicants are recommended more often. At the three-seat panel grain, representative sortition's live disparity is -0.190, while same-bias selection's live disparity is -0.214. Age bias here *is* simply illegitimate: because true quality is generated independently of age, any age-conditioned shift is a reviewer acting on an irrelevant attribute. That is exactly why age is a useful *probe* — it gives a clean, known-illegitimate signal whose magnitude we can read directly — so the question is not whether age bias is acceptable (it is not) but whether the upstream sampling rule **amplifies or contains** an illegitimate bias the reviewers already carry. The disparities here are negative across the board, meaning this model favors younger synthetic applicants; the experiment measures that illegitimate behavior to compare bias containment across sampling rules (age is a probe; see Ethics).

4.2.3 Analytical and Gemma cells stay juxtaposed, not pooled

The alignment artifact compares analytical prediction signs with live Gemma observations by strategy and panel size. 8 of 8 cells are resolved after zero-sign cells are marked unresolved; the resolved-cell sign-agreement rate is 0.500. This agreement

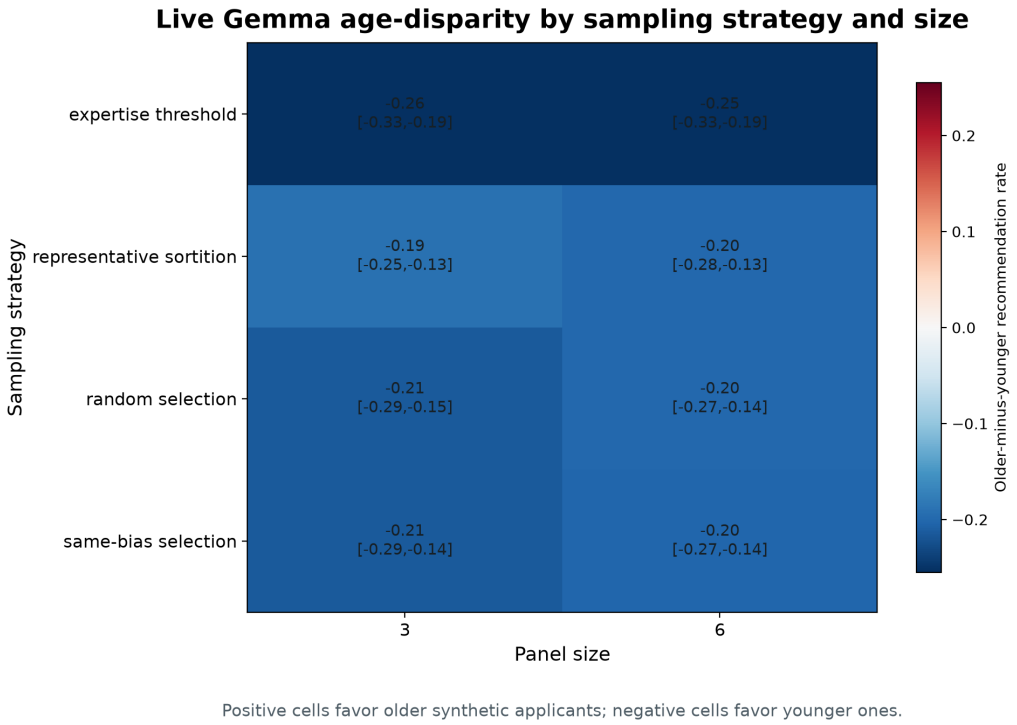


Figure 18: Heatmap of the older-minus-younger recommendation-rate disparity expressed by the live Gemma reviewer panel, by sampling strategy (rows) and panel size (columns), from source `output/data/postdoc_panel_results.json`. Because true latent quality is generated independently of age, any non-zero cell is reviewer age-bias expression rather than signal: positive values mean older synthetic applicants are recommended more often, negative values mean younger applicants are. Metric: age-conditioned recommendation-rate difference, aggregated over 12 seeds; read down a column to compare how each sampling rule amplifies or dampens the irrelevant age signal. Claim: same-bias (single-bloc) sampling is the explicit bias-amplification stress test among the strategies; caveat: all applicants and ages are synthetic, and this figure does not validate Gemma or endorse age-aware real review.

is a **weak directional check, not an independent match**: every live Gemma age-disparity sign in this run is negative (the model uniformly favors younger synthetic applicants), so the rate mostly measures how often the analytical sign is also negative rather than a cell-by-cell coincidence of two freely varying signals. This is the intended bridge between the synthetic and live tracks: same causal question, shared sampling vocabulary, separate evidence levels.

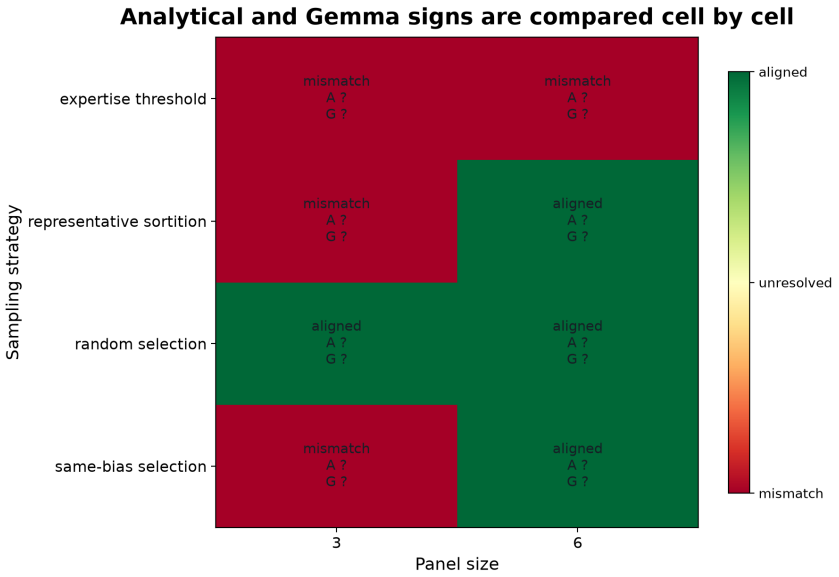


Figure 19: Cell-by-cell alignment grid comparing the analytical age-disparity direction with the live Gemma direction, one cell per strategy x panel-size combination, from source `output/data/postdoc_panel_alignment.json`. Each cell is marked agree, disagree, or unresolved (a zero-sign cell on either track), so the figure functions as the explicit bridge between the controlled and the live track while keeping their uncertainties separate. Statistic: sign agreement between the analytical and live age-disparity directions; resolved-cell agreement 0.500 over 8 resolved cells. Because every live disparity sign is negative in this run, the agreement rate is a weak directional check rather than an independent match, and should be read as such. Claim: the analytical and empirical tracks can be compared cell by cell without pooling their uncertainty; caveat: single-model live evidence is descriptive and n-limited.

4.2.4 Synthetic strategy ranking does not transfer to the live track

H5 (does the synthetic ranking transfer to a live single-model panel?). The synthetic and live tracks are never pooled, and a matched-grain comparison shows why pooling would mislead. At the shared three-seat panel grain (fig. 20) the strategy *rank order* inverts between tracks. The rule with by far the lowest synthetic recovery error — expertise threshold at 0.037 — is the **worst** under the live Gemma panel at 0.347. The other three strategies are bunched on *both* tracks — synthetically at 0.118–0.124 and live at 0.225–0.262 — so they neither clearly invert nor clearly transfer. The robust component of the non-transfer is therefore the expertise-threshold flip alone — it is the lone clear outlier on both tracks (best synthetic, worst live) — so the non-transfer claim rests on the competence-first rule reversing, not on a precisely resolved ordering of the other three (which are statistically indistinguishable on each track). We compare ranks, not magnitudes, because the two tracks own different oracles and different uncertainty, so the figure is a qualitative non-transfer result rather than a pooled effect size.

Why the ranking inverts. The two tracks do genuinely disagree, and for a principled reason. In the synthetic track the generator *sets* each judge's accuracy directly, so competence-first selection seats genuinely higher-accuracy judges and the exact solver recovers them cleanly — the ordering is, in part, built into the data-generating process. Live, "expertise" is only a *prompt instruction*: the local `gemma3:4b` model need not behave more accurately, or with more independent errors, when it is told it is an expert reviewer. Selecting personas by their stated expertise therefore seats no better live judges, and the rule that wins by construction on synthetic data carries no guaranteed live advantage — here it is the worst. Put differently, the synthetic oracle rewards a property (controlled judge accuracy) that the prompted model does not inherit from the persona label. This is a hypothesis about the mechanism, not a measured causal claim, and it is the empirical reason the manuscript keeps the synthetic and live tracks at distinct inference levels rather than reporting a single cross-track strategy winner.

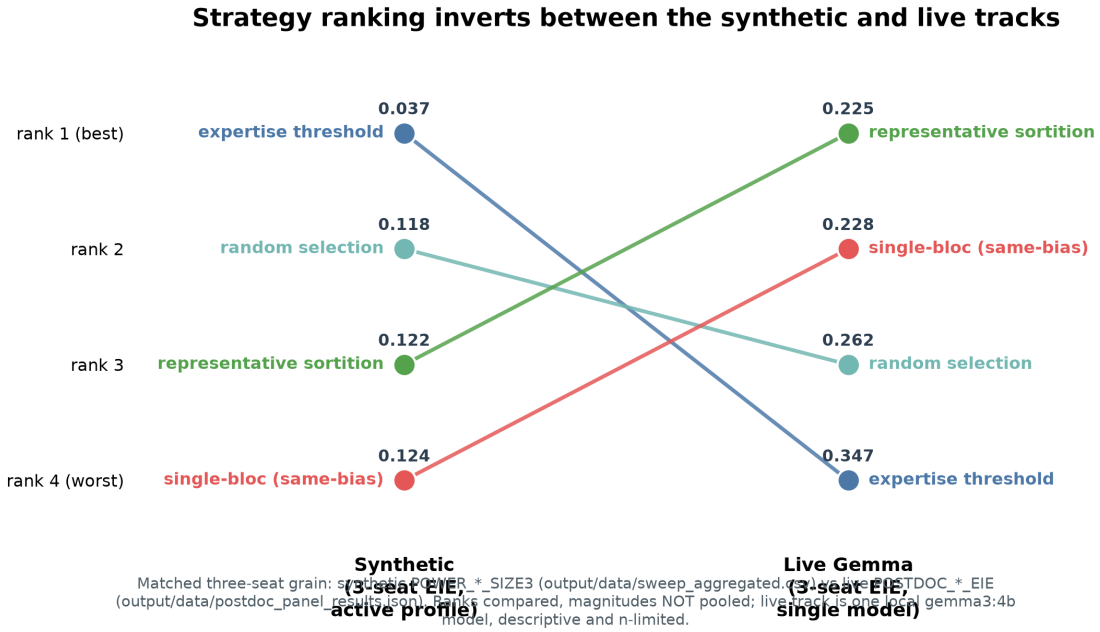


Figure 20: Cross-track strategy-ranking inversion at the matched three-seat grain, shown as a slope chart, from source `output/data/sweep_aggregated.csv` (synthetic `POWER_*_SIZE3`, left column) and source `output/data/postdoc_panel_results.json` (live `POSTDOC_*_EIE`, right column). Each strategy is a coloured line connecting its rank under the synthetic track to its rank under the live track; lines that cross are strategies whose standing reverses, and the steepest crossers are expertise threshold (synthetic best to live worst) and single-bloc selection (synthetic worst to live rank-two), while representative sortition is near-best on both tracks. The y-axis is ordinal rank, with the raw error annotated at each node so the reader can see that the live top three are tightly bunched while expertise threshold is the lone outlier. Statistic: ordinal EIE-error rank per track over 96 synthetic seeds and 12 live seeds; ranks compared, magnitudes not pooled (the two tracks own different oracles and uncertainty). Claim: the formation rule that is best blind on synthetic data is worst under one local Gemma model, so the synthetic ranking does not transfer to the live single-model panel; caveat: single-model live evidence is descriptive and n-limited, not a human-review validation.

5 Discussion: claim boundaries and implications

5.1 Hypothesis verdicts before interpretation

We return an explicit verdict on each pre-stated hypothesis (Introduction, H1–H5) before interpreting it, so the contribution is the adjudication itself rather than a single narrative.

- **H1 (formation strategy is the dominant lever) — supported.** The four panels do not separate into a graded four-way ladder. Competence-first selection (**expertise threshold**) is clearly best at 0.037, while representative sortition, single-bloc selection, and random selection cluster around 0.147-0.148 with overlapping intervals. The resolved result is competence-first versus the bottom cluster, not a precise ordering inside that cluster.
- **H2 (concentrating correlated error degrades recovery) — unresolved on the baseline grid, then RESOLVED by the composition-coupled confound.** On the baseline generator the representative-vs-single-bloc contrast is design-limited, for a structural reason: that generator never realizes H2's premise, because its judges err independently and ideology shifts only marginal accuracy, so the strategies coincide *by construction* and no regime sweep can fan them out. Once a composition-coupled, marginal-accuracy-preserving error confound supplies the missing channel (fig. 9), the contrast resolves cleanly: single-bloc error exceeds representative error in 180/205 matched regimes (paired mean 0.102, 95% CI ±0.012), the gap widening from 0.000 to 0.112 as coupling rises. The advantage is real but **conditional**: an orthogonal-axis negative control removes it, so we claim "balancing the axis the confound rides on preserves recovery," not a universal sortition law.
- **H3 (size is a sampling knob, not guaranteed improvement) — supported only in the bounded sense.** A paired, regime-controlled test resolves a trio-to-six-seat size effect for 3 of the four strategies, and every resolved effect is a tiny increase in error: random selection (+0.015), representative sortition (+0.007), and competence-first selection (+0.004). Single-bloc is within noise. So more experts do not help here, but they also do not create a material penalty; size is essentially neutral at this grid, and the dominant lever is strategy, not size.
- **H4 (error-correlation is measurable, and recovery degrades with it) — diagnostic confirmed; recovery slope resolved by the composition-coupled instrument.** The diagnostic works — the realized correlation tracks the injected coupling (0.019 to 0.123). The *global-injection* tolerance sweep leaves the recovery-vs-correlation slope unresolved (-0.145, 95% CI [-4.335, 0.911] crosses zero), and we now read that as a limitation of that instrument — it couples correlation uniformly to a fixed trio and uses a small grid — rather than as evidence of no effect. The composition-coupled, marginal-accuracy-preserving sweep (fig. 9), with far more observations per point, resolves the recovery half in the affirmative: recovery error rises with within-bloc coupling for the panels that concentrate the confound, and the closed-form Herfindahl account (§Methods) explains the ordering. So the recovery half is a result, not an open question, under a correctly specified confound.
- **H5 (the synthetic ranking transfers to a live single-model panel) — rejected.** At the matched three-seat grain the ranking inverts: the rule that is best blind on synthetic data (expertise threshold, 0.037) is the worst under the live `gemma3:4b` panel (0.347). The robust component is that expertise-threshold flip; the live top three are bunched and the evidence is single-model and n-limited, so we report non-transfer of the best-synthetic rule rather than a precise live winner. The suggested caution, scoped to this one companion model: "choose the most expert judges" most cleanly minimized blind-recovery error on judges of *known* accuracy, yet it was the worst rule once "expert" was merely a prompt label the model need not honor — a hypothesis that self-asserted expertise may be an unsafe blind-evaluation selection criterion, worth testing beyond one model rather than an established result.

The remaining subsections elaborate the mechanism behind each verdict.

5.2 Practical lesson: selection rule before panel size

Three usable lessons follow for anyone selecting judges to be evaluated without an answer key. (i) *Which* rule forms the panel matters far more than how many judges it seats: competence-first selection set the lowest blind-recovery error here, while panel size was essentially neutral (H1, H3). (ii) Representative selection protects unsupervised recovery *only* when the lottery balances the very attribute a shared error rides on; balanced on the wrong axis it gives no protection (H2/H4, negative control), and the exposure it controls is a closed-form concentration (Herfindahl) index you can compute on a *proposed* panel before any votes are cast. (iii) The advantage of picking "expert" judges did not survive when judges were a prompted live model rather than parameterized synthetic ones (H5), so a selection rule validated on controllable judges cannot be assumed to carry over to real ones.

5.3 Formation strategy is the measured lever

Studying NTQR *upstream* — at the panel-formation step rather than at the estimator — reframes ground-truth-free evaluation as a selection problem. The strongest finding is that competence-first selection sets a much lower downstream no-answer-key error floor than the other panel-formation rules. The other three strategies cluster tightly enough that their point-estimate order should not be read as a substantive ranking. The scientific claim is therefore about the competence-first-vs-rest separation, not about naming a bottom-cluster winner or loser.

That framing matters for application review because peer-review scholarship already treats expert judgment as socially situated and panel-dependent rather than mechanically objective Lamont (2009); Lee et al. (2013), and because empirical studies find substantial reviewer disagreement on the same submitted work Cole et al. (1981); Pier et al. (2018), score-model uncertainty large enough to alter the implied funded set Johnson (2008), and limited grant-productivity predictiveness of NIH percentiles Fang et al. (2016). The manuscript's increment is narrower: it instruments one selection mechanism and asks whether different panel draws change an unlabeled evaluator's oracle-referenced error. The claim is about generated artifacts and one local Gemma stress test, not about the global reliability of academic review.

This is, in part, a result *against* the intuitive case for sortition. A representative lottery is the fair, auditable way to form a panel, but on this instrument it does not minimize oracle-referenced EIE error — competence-first does. We report that plainly rather than engineering a narrative in which sortition wins. That is not a general refutation of sortition, deliberative participation, or the "diversity can beat ability" result. Those arguments rely on different objectives and premises: democratic legitimacy and public consultation Fishkin (2009), search diversity Hong and Page (2004), and jury-theorem aggregation under competence and conditional-independence assumptions Grofman et al. (1983). The present result is narrower: in this binary noisy-judge instrument, competence-first sampling gives the lowest oracle-referenced EIE error. Relative to the single-bloc comparator the representative draw's point estimate is now reported over the full active regime grid rather than collapsed to two panel-size means. Some cells resolve in the predicted direction, others remain descriptive or design-limited, so we claim artifact-bounded regime structure, not a general sortition advantage over single-bloc selection.

5.4 Design-limited nulls remain results

Two results are bounded rather than universal, and we do not dress them up.

1. **On the baseline grid, representative vs ideological is design-limited.** Varying expert stringency, bias spread, and panel size jointly, the heatmap reports which regenerated synthetic cells align with the directional prediction, which resolve by descriptive intervals, and which remain uncertain — but the pooled contrast is not resolved there, *by construction*, because the baseline judges err independently. It resolves only once the composition-coupled confound supplies the correlation channel (see the H2 verdict and fig. 9); the null is a property of the baseline design, not of sortition.

2. **Size is not a uniform power knob.** A paired, regime-controlled contrast resolves a trio-to-six-seat size change for 3 of the four strategies, and each resolved change is a small increase in error. The largest delta is +0.015, so more experts do not help and at most very slightly hurt. The clean "more experts always helps" story is rejected, but the result is essentially neutral rather than a material size penalty. That is consistent with peer-review jury-theorem work: adding reviewers helps only under assumptions about competence, dependence, and aggregation that must be checked rather than presumed Arvan et al. (2025).

Reporting these nulls is the point of building a measurement instrument rather than a demonstration.

5.5 Independence explains why strategy ordering changes

NTQR's EIE solver rests on the judges' errors being approximately independent. On the baseline generator, single-bloc selection is indistinguishable from representative and random selection **by construction**: ideology there shifts only each judge's marginal accuracy, every judge errs from an independent stream, and an agreement-only estimator cannot be moved by composition. Single-bloc becomes the *separable* adversarial comparator only once the composition-coupled confound supplies a genuine cross-judge error-correlation channel (fig. 9). Competence-first panels pair high accuracy with whatever independence the population affords, giving the solver the easiest system to invert. The fair-lottery argument should therefore be made from auditability, representation, and bounded empirical performance rather than from an unqualified error-minimization win.

5.6 Error independence must be measured before interpretation

The assumption the whole ordering hangs on — that judges' errors are approximately independent — is, in this instrument, no longer an assumption but a measured quantity. The controlled-correlation sweep confirms the *knob works*: the realized pairwise correlation NTQR reports rises with the injected coupling (0.019 to 0.123). What it does **not** yet resolve, at this grid, is whether that correlation degrades recovery — the fitted slope is **statistically indistinguishable from zero** (-0.145, 95% CI [-4.335, 0.911] spans zero), unresolved rather than absent — and the power layer explains why that is unsurprising rather than disappointing: 12 of 28 contrasts are well-powered at the current seed count (MDE 0.101), 13 reach nominal significance, and 12 survive Holm correction. The honest reading is that the current design resolves the largest separations but leaves smaller neighboring contrasts design-limited. The strategy *ranking* remains an ordering of point estimates; the analysis says exactly how many seeds would let the unresolved contrasts resolve. That global-injection slope is, in any case, the wrong instrument for the recovery question — it couples correlation to a fixed trio regardless of how the panel was formed; the marginal-preserving composition-coupled sweep (fig. 9) resolves the recovery half in the affirmative, as adjudicated in the H4 verdict above.

The Gemma postdoctoral panel is the direct live look at the same sampling mechanism under a real local LLM. It does not ask whether one model family beats another; it asks whether representative, random, same-bias, and expertise-first sampling leave different traces when one `gemma3:4b` model is prompted as reviewers with different expertise and irrelevant age-bias profiles. The live artifact reports 72 fictitious applications per seed and model provenance (digest a2af6cc3eb7f), while the alignment artifact juxtaposes analytical signs with Gemma signs over 8 strategy-size cells. We still deliberately under-claim it: the applications and ages are synthetic, the reviewer personas are prompts rather than humans, and the resolved agreement rate (0.500 over 8 resolved cells) is descriptive companion evidence, not validation that Gemma or any age-aware real review process is appropriate. The competence-first versus representative comparison is likewise gated by an explicit CI-overlap verdict (**separated**, means 0.037 vs 0.122), so "beats" is never asserted across overlapping intervals.

5.7 Scholarship frames the stress test, not the evidence level

The postdoctoral-review setting is intentionally close to a literature where selection, status, and bias are known concerns. Cumulative-advantage accounts of scientific recognition Merton (1968), resubmission experiments showing fragility in journal review Peters and Ceci (1982), blind-review experiments and observational bias studies Tomkins et al. (2017); Helmer et al. (2017), and empirical studies of fellowship or grant outcomes Wennerås and Wold (1997); Ginther et al. (2011) make it reasonable to study reviewer sampling, not only evaluator algebra. The age axis has the same status: ageism and age-discrimination findings motivate it as a protected-attribute stress test North and Fiske (2013); Neumark et al. (2019), but the manuscript does not infer anything about real postdoctoral age discrimination from prompted Gemma votes.

The synthetic and Gemma tracks therefore answer different questions. The synthetic track can make controlled claims because it owns the oracle label and the expert parameters. The live Gemma track can only show whether the same sampling vocabulary produces measurable traces under one local model with serialized provenance. Prompted LLM evaluation is itself an active measurement problem, not a neutral readout Zheng et al. (2023), and language-model risk scholarship cautions against treating model text as a transparent substitute for human judgment Bender et al. (2021), so the correct inference level is empirical feasibility plus directional stress testing. Lottery and collective-allocation proposals in science funding Bollen et al. (2014); Fang and Casadevall (2016), and maverick-science arguments for lotteries Avin (2019) make randomized institutional design a legitimate comparator, but they do not license a claim that lottery-formed reviewer panels optimize NTQR recovery. Scholarship supplies the problem context and the variables worth stress-testing; regenerated artifacts supply the evidence.

The pre-1800 sources sharpen that boundary rather than broadening the claim. Aristotle, Aquinas, Contarini, Montesquieu, Rousseau, Borda, and Condorcet show that lot, choice, mixed selection, and probabilistic group judgment have long been treated as procedural responses to faction, legitimacy, and uncertainty. They do not license a claim that historical sortition "validates" this synthetic NTQR instrument. The contribution here is narrower: a regenerated experiment that keeps historical and modern motivations upstream of the evidentiary claim, then tests how panel formation changes oracle-referenced blind recovery.

The historical sources also draw a useful negative boundary. We exclude gambling lotteries, divinatory lots, and broad political-theory claims that are not about selecting evaluators or aggregating judgments. The manuscript's analogy is procedural: randomness can distribute evaluative authority when deterministic selection is capture-prone or status-weighted. Whether that helps an unlabeled evaluator is not answered by the history; it is answered by the regenerated artifacts above.

5.8 Limitations: synthetic scope, single-model live evidence, historical analogy

The reported `manuscript_contrast` grid fixes prevalence and corpus size while varying mean expertise, bias, panel size, and strategy; it uses 96 experts, 300 as the modal item count in the rendered tokens, and up to 8 trios per panel over 96 seeds. The repository now defines broader sensitivity and finer panel-ladder profiles, but those profiles are configuration surfaces until regenerated and audited as manuscript evidence; the reported results remain bounded to the active profile. The oracle-closest tie-break is deliberately charitable to the unsupervised estimate, so reported errors are a lower bound on what a blind tie-break would incur. The alarm's $O(Q^3)$ cost confines the consistency-alarm track to small corpora ($Q \leq 30$), so the alarm cannot yet serve as a sweep-scale signal. The statistical-power analysis shows most strategy contrasts are underpowered at the current seed count, so several headline comparisons are design-limited rather than settled. The Gemma postdoctoral panel is also bounded: it uses fictitious applications, synthetic age metadata, prompted reviewer personas, and one local model. It tests whether the sampling mechanism is visible under that empirical stress test; it does not establish human-review performance or a policy claim about age. Finally, the synthetic generator is a model of noisy judges, not a guarantee about real ones.

5.9 Synthetic and live tracks operate at different inference levels

This manuscript follows a standard division between controlled experiment and empirical companion evidence. The deterministic synthetic track is the controlled Results spine: it generates the strategy-ranking, panel-size, controlled-correlation, power-budget, alarm-cost, and analytical-alignment numbers from regenerated local artifacts. Those claims are validated against known oracle labels because the generator owns the truth labels.

The real-Ollama track is reported as separate live artifacts and empirical companion evidence, not a pooled extension of the synthetic sweep. It was performed locally with required-live `gemma3:4b` (full provenance in Methods), showing the same sampling vocabulary run on a real local model prompted as different reviewers. It does not validate the full synthetic regime grid, establish a population effect size, or prove that Gemma substitutes for human reviewers.

The combined interpretation is therefore deliberately tiered: synthetic experiments support the controlled mechanism and regime maps; analytical checks test directional expectations against those regenerated artifacts; and live Ollama runs demonstrate empirical feasibility plus n-limited directional support. Wider parameter sweeps and larger empirical panels are the next steps before any stronger general claim.

5.10 Data, code, and generated-artifact availability

All source code, methods, and documentation are openly available at the public repository `docxology/ntqr_allotment`: the deterministic synthetic instrument, the bloc-confound and Herfindahl modules, the live Gemma vote cache with serialized model provenance, every figure, and the manuscript regeneration pipeline. Every reported number is token-injected from `output/data/` by `src/ntqr_allotment/manuscript_variables.py`, so no result is hand-transcribed and the manuscript regenerates from source under a zero-orphan token contract. The synthetic track is fully deterministic under fixed seeds (profile config hash `fda4da941cf0`); the live track reproduces against a local Ollama `gemma3:4b` instance (digest `a2af6cc3eb7f`, config hash `5161ffe474b3`) using the serialized, resumable per-vote cache keyed on the config hash, seed, reviewer, application, model digest, and decode parameters. A steganographic provenance variant of the PDF additionally carries an extractable hash of the current source PDF, verified by `scripts/verify_stego.py`. The archival DOI is 10.5281/zenodo.21083779.

5.11 Ethics, protected attributes, and competing interests

The postdoctoral-review setting is entirely synthetic: the applications, latent quality labels, reviewer personas, and age metadata are all generated, and no human subjects, real applicants, or real review records are involved. True latent quality is generated independently of age; age enters only as a protected-attribute stress test for bias expression under sampling, and its use here is diagnostic, not an endorsement of using age in real admissions, hiring, or fellowship review. The live language-model outputs are treated as an instrumented measurement of a prompted system, not as a substitute for human judgment. The authors declare no competing interests.

6 References

The bibliography below is generated from `manuscript/references.bib` by the render pipeline. This section is intentionally citation-driven rather than a manual numbered list so DOI/URL fields can render as links where the output format supports them.

References

Thomas Aquinas. *Summa Theologiae, Second Part of the Second Part, Question 95, Article 8*. Benziger Brothers, 1920. URL https://www.newadvent.org/summa/3095.htm#article8. Original work composed in the thirteenth century.

Aristotle. *The Athenian Constitution*. Harvard University Press, 1935. URL https://topostext.org/work/99. Original work composed in the fourth century BCE; Loeb Classical Library translation.

Aristotle. *Politics*. Harvard University Press, 1944. URL https://topostext.org/work/100. Original work composed in the fourth century BCE; Loeb Classical Library translation.

Marcus Arvan, Liam Kofi Bright, and Remco Heesen. Jury theorems for peer review. *The British Journal for the Philosophy of Science*, 76(2):319–344, 2025. doi: 10.1086/719117.

Shahar Avin. Mavericks and lotteries. *Studies in History and Philosophy of Science Part A*, 76:13–23, 2019. doi: 10.1016/j.shpsa.2018.11.006.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623, 2021. doi: 10.1145/3442188.3445922.

Johan Bollen, David Crandall, Damion Junk, Ying Ding, and Katy Börner. From funding agencies to scientific agency: Collective allocation of science funding as an alternative to peer review. *EMBO Reports*, 15(2):131–133, 2014. doi: 10.1002/embr.201338068.

Jean-Charles de Borda. Mémoire sur les élections au scrutin. *Histoire de l'Académie Royale des Sciences*, pages 657–665, 1781. URL https://bibbase.org/network/publication/denbspborda-mmoiresurleslectionsauscrutin-1781. Published in the Academy volume for 1781.

Citizen-Infra. allotment: An auditable fair-sortition engine. Software, AGPL-3.0, 2024. URL https://github.com/Citizen-Infra/allotment. Implements the maximin stratified-sortition lottery of Flanigan et al. (2021); used in this work as the representative-sortition panel-formation engine.

Jacob Cohen. *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates, 2 edition, 1988. ISBN 9780805802832.

Stephen Cole, Jonathan R. Cole, and Gary A. Simon. Chance and consensus in peer review. *Science*, 214(4523):881–886, 1981. doi: 10.1126/science.7302566.

Nicolas de Caritat Condorcet. *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix*. Imprimerie Royale, 1785. URL https://archive.org/details/bub_gb_RzAVAAAAQAAJ.

Gasparo Contarini. *The Commonwealth and Government of Venice*. I. Windet for E. Mattes, 1599. URL https://onlinebooks.library.upenn.edu/webbin/book/lookupid?key=olbp14764. English translation of De magistratibus et republica Venetorum.

Andres Corrada-Emmanuel. ntqr: Tools for the logic of evaluation using unlabeled data. Python package version 0.8, 2026. URL https://pypi.org/project/ntqr/. Released May 28, 2026. Documentation: https://ntqr.readthedocs.io/en/latest/.

A. P. Dawid and A. M. Skene. Maximum likelihood estimation of observer error-rates using the em algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 28(1):20–28, 1979. doi: 10.2307/2346806.

Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*. Chapman and Hall/CRC, 1993. doi: 10.1201/9780429246593.

Lawrence J. Emrich and Marion R. Piedmonte. A method for generating high-dimensional multivariate binary variates. *The American Statistician*, 45(4):302–304, 1991. doi: 10.2307/2684786.

Ferric C. Fang and Arturo Casadevall. Research funding: The case for a modified lottery. *mBio*, 7(2):e00422–16, 2016. doi: 10.1128/mBio.00422-16.

Ferric C. Fang, Anthony Bowen, and Arturo Casadevall. Nih peer review percentile scores are poorly predictive of grant productivity. *eLife*, 5:e13323, 2016. doi: 10.7554/eLife.13323.

James S. Fishkin. *When the People Speak: Deliberative Democracy and Public Consultation*. Oxford University Press, 2009. ISBN 9780199604432. URL https://global.oup.com/academic/product/when-the-people-speak-9780199604432.

Bailey Flanigan, Paul Gölz, Anupam Gupta, Brett Hennig, and Ariel D. Procaccia. Fair algorithms for selecting citizens' assemblies. *Nature*, 596:548–552, 2021. doi: 10.1038/s41586-021-03788-6.

Donna K. Ginther, Walter T. Schaffer, Joshua Schnell, Beth Masimore, Faye Liu, Laurel L. Haak, and Raynard Kington. Race, ethnicity, and nih research awards. *Science*, 333(6045):1015–1019, 2011. doi: 10.1126/science.1196783.

Bernard Grofman, Guillermo Owen, and Scott L. Feld. Thirteen theorems in search of the truth. *Theory and Decision*, 15 (3):261–278, 1983. doi: 10.1007/BF00125672.

Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, volume 29, 2016. URL https://papers.nips.cc/paper_files/paper/2016/hash/9d2682367c3935defcb1f9e247a97c0d-Abstract.html.

Markus Helmer, Manuel Schottdorf, Andreas Neef, and Demian Battaglia. Gender bias in scholarly peer review. *eLife*, 6: e21718, 2017. doi: 10.7554/eLife.21718.

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.

Lu Hong and Scott E. Page. Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *Proceedings of the National Academy of Sciences*, 101(46):16385–16389, 2004. doi: 10.1073/pnas.0403723101.

Valen E. Johnson. Statistical analysis of the national institutes of health peer review system. *Proceedings of the National Academy of Sciences*, 105(32):11076–11080, 2008. doi: 10.1073/pnas.0804538105.

David Kaplan, Nicola Lacetera, and Celia Kaplan. Sample size and precision in nih peer review. *PLOS ONE*, 3(7):e2761, 2008. doi: 10.1371/journal.pone.0002761.

David R. Karger, Sewoong Oh, and Devavrat Shah. Budget-optimal task allocation for reliable crowdsourcing systems. *Operations Research*, 62(1):1–24, 2014. doi: 10.1287/opre.2013.1235.

Michèle Lamont. *How Professors Think: Inside the Curious World of Academic Judgment*. Harvard University Press, 2009. doi: 10.4159/9780674054158. URL https://www.hup.harvard.edu/books/9780674057333.

Carole J. Lee, Cassidy R. Sugimoto, Guo Zhang, and Blaise Cronin. Bias in peer review. *Journal of the American Society for Information Science and Technology*, 64(1):2–17, 2013. doi: 10.1002/asi.22784.

Robert K. Merton. The matthew effect in science. *Science*, 159(3810):56–63, 1968. doi: 10.1126/science.159.3810.56.

Charles de Secondat Montesquieu. *The Spirit of Laws*. Online Library of Liberty, 1748. URL https://oll.libertyfund.org/titles/montesquieu-complete-works-vol-1-the-spirit-of-laws. Cited through the Online Library of Liberty edition of the 1777 English translation.

Roger B. Nelsen. *An Introduction to Copulas*. Springer, 2nd edition, 2006.

David Neumark, Ian Burn, and Patrick Button. Is it harder for older workers to find jobs? new and improved evidence from a field experiment. *Journal of Political Economy*, 127(2):922–970, 2019. doi: 10.1086/701029.

Michael S. North and Susan T. Fiske. Act your (old) age: Prescriptive, ageist biases over succession, consumption, and identity. *Personality and Social Psychology Bulletin*, 39(6):720–734, 2013. doi: 10.1177/0146167213480043.

Fabio Parisi, Francesco Strino, Boaz Nadler, and Yuval Kluger. Ranking and combining multiple predictors without labeled data. *Proceedings of the National Academy of Sciences*, 111(4):1253–1258, 2014. doi: 10.1073/pnas.1219097111.

Douglas P. Peters and Stephen J. Ceci. Peer-review practices of psychological journals: The fate of published articles, submitted again. *Behavioral and Brain Sciences*, 5(2):187–195, 1982. doi: 10.1017/S0140525X00011183.

Elizabeth L. Pier, Markus Brauer, Amarette Filut, Anna Kaatz, Joshua Raclaw, Mitchell J. Nathan, Cecilia E. Ford, and Molly Carnes. Low agreement among reviewers evaluating the same nih grant applications. *Proceedings of the National Academy of Sciences*, 115(12):2952–2957, 2018. doi: 10.1073/pnas.1714379115.

Emmanouil Antonios Platanios, Avrim Blum, and Tom M. Mitchell. Estimating accuracy from unlabeled data. In *Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 682–691, 2014.

Vikas C. Raykar, Shipeng Yu, Linda H. Zhao, Gerardo Hermosillo Valadez, Charles Florin, Luca Bogoni, and Linda Moy. Learning from crowds. *Journal of Machine Learning Research*, 11(43):1297–1322, 2010. URL https://jmlr.org/papers/v11/raykar10a.html.

Jean-Jacques Rousseau. *The Social Contract and Discourses*. J. M. Dent and Sons, 1762. URL https://oll.libertyfund.org/titles/cole-the-social-contract-and-discourses. The Social Contract was first published in 1762; cited through the Online Library of Liberty edition.

Lin Shi, Chiyu Ma, Wenhua Liang, Xingjian Diao, Weicheng Ma, and Soroush Vosoughi. Judging the judges: A systematic study of position bias in llm-as-a-judge. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 292–314, 2025. doi: 10.18653/v1/2025.ijcnlp-long.18.

Peter Stone. *The Luck of the Draw: The Role of Lotteries in Decision Making*. Oxford University Press, 2011. doi: 10.1093/acprof:oso/9780199756100.001.0001.

Andrew Tomkins, Min Zhang, and William D. Heavlin. Reviewer bias in single- versus double-blind peer review. *Proceedings of the National Academy of Sciences*, 114(48):12708–12713, 2017. doi: 10.1073/pnas.1707323114.

Christine Wennerås and Agnes Wold. Nepotism and sexism in peer-review. *Nature*, 387:341–343, 1997. doi: 10.1038/387341a0.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://arxiv.org/abs/2306.05685.