

Endogenous Surplus Retention in Peer-Based Team Incentives

Xin He

252210888115@stu.just.edu.cn

Jiangsu University of Science and Technology

Research Article

Keywords: team incentives; peer assessment; revenue sharing; mechanism design; budget breaking

Posted Date: October 19th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10916574/v1>

License:  This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Endogenous Surplus Retention in Peer-Based Team Incentives

Abstract

Team compensation faces a double bind: splitting output among members invites free riding, while tying pay to peer evaluation invites manipulation and noise. We propose a peer-assessed revenue-sharing rule with a fixed denominator: each member’s share is his received recognition votes divided by the full-capacity vote total. Because the denominator is fixed rather than realized, shares need not exhaust net income; the residual accrues to the firm endogenously, as a function of the team’s own evaluation intensity, not an exogenously calibrated royalty. The rule separates two objects that peer allocation normally confounds: relative ranking is decided by received votes, while the aggregate distributable pool is decided by how intensely the team evaluates. It removes unilateral own-share manipulation by construction, at the price of removing private incentives to provide costly truthful evaluation – a design tension addressed by a proposed governance layer. On three independent peer-assessment datasets spanning four grading schemes and three countries, we show that received votes carry informative but noisy signal against independent anchors, that realized retention is interior and tracks the grading scheme’s generosity, and that a data-driven counterfactual identifies a retention–accuracy frontier with an interior optimum: the planner can target retention through the grading scheme, at a quantified cost in ranking accuracy. The paper establishes a structural design result; behavioral honesty and causal performance effects are not identified here.

Keywords: team incentives; peer assessment; revenue sharing; mechanism design; budget breaking

1 Introduction

Team production faces a persistent double bind. When individual contributions are not contractible, splitting the joint output among members induces free riding: each agent internalizes only his own share, so equilibrium effort falls below the first best. Holmström (1982) shows that under budget balance, $\sum_i s_i(x) = x$, no sharing rule based on the joint outcome implements the efficient vector; efficiency requires a *budget breaker*, an outsider who absorbs the residual. A natural remedy – letting members evaluate one another and tying pay to

peer scores – introduces a second hazard: peer evaluation can be manipulated and, absent verifiable truth, the signal is noisy and strategically pliable.

Three gaps separate the existing toolbox from this problem. First, peer allocation determines relative shares but takes the divisible surplus as given: under normalized peer sharing $s_i^N = E R_i / \sum_j R_j$, the pool always exhausts E , so the budget-balance logic – and with it the free-rider problem – survives unchanged. Second, budget-breaking contracts do allow a residual claimant, but the residual is typically an exogenous contract feature (a fixed royalty or an outsider’s claim), not an outcome of the peer-evaluation process itself. Third, removing self-manipulation in one’s own share can simultaneously remove the private incentive to perform costly, honest evaluation of others; the two properties must be separated, not conflated.

This paper proposes a mechanism that closes all three gaps at once. A team of N members across P work streams casts 0/1 recognition votes on the streams each claims; each member’s share of net income E is proportional to received votes w_i divided by a *fixed* denominator T – the theoretical maximum votes the team could cast, not the realized sum. Because T is constant, shares need not exhaust E ; the residual $E_{\text{public}} = E(1 - \sum_i w_i/T)$ accrues automatically to the firm. Two properties follow immediately and drive the paper. (1) **Endogenous budget breaking (P1)**. Retention is an outcome of voting intensity, not a design parameter; the firm is the residual claimant, but its claim is behavior-linked rather than exogenously specified. (2) **Own-share manipulation immunity (P2)**. A member’s share depends only on votes he *receives*, so his own votes cannot change it. The contrast is sharp: under normalization $\partial s_i^N / \partial R_j < 0$ for $j \neq i$ (depressing others mechanically raises one’s relative share), while under the fixed denominator $\partial s_i^{\text{FD}} / \partial R_j = 0$ by (P2). A structural concentration bound – no share exceeds the equal-division level $1/N$ – follows as a corollary (P3).

The rule’s central design property is that it *separates* two objects that peer allocation normally confounds: relative ranking is decided by received votes, while the aggregate distributable pool is decided by how intensely the team evaluates. The bare mechanism, however, creates a second tension: removing own-share manipulation also removes the private incentive to vote honestly. A proposed **governance layer** – KPI-weighted voting, a voting cost, and a peer-prediction reward – addresses this as an implementation matter (P8–P10) while preserving the structural theorems; it is a design extension, not a separate contribution.

We then take the mechanism to data. On three independent peer-assessment datasets spanning four grading schemes and three countries (CHDYPY, 20 cohorts; Xu et al., 2025; peer-grading, Jones-Todd & Renelle, 2025; t4e, Ashenafi, 2019), we establish three facts. (i) Received votes carry informative, if noisy, signal against independent anchors ($\rho_{\text{pool}} = 0.439$, positive in all 20 cohorts; $\rho = 0.604$ on the tutor anchor). (ii) Realized retention is endogenous, interior, and scheme-driven (mean 13.5%, about 50% under coin budgets, about 70% under simple votes). (iii) A data-driven counterfactual (§5.7) identifies a *retention*–

accuracy frontier: the planner can target retention through the grading scheme, but only at a quantified cost in ranking accuracy, with an interior optimum. A calibrated simulation extends the analysis to small teams ($N \in [6, 25]$).

Contributions. (1) *Mechanism design*: the fixed denominator endogenizes the aggregate distributable surplus, turning retention from a free calibration parameter into an observable, schedulable design outcome (P1, P6). (2) *Strategic structure*: the rule separates self-share incentives from evaluation incentives, eliminating unilateral own-share manipulation while making the evaluation-incentive problem explicit and tractable (P2–P5, P9–P10). (3) *Empirical and design evidence*: peer assessments contain informative signal against independent anchors, and a counterfactual identifies a retention–accuracy design frontier with an interior optimum (§5). The paper establishes a structural design result; it does not claim to identify behavioral honesty or causal performance effects.

The paper proceeds: §2 related literature; §3 model; §4 theory; §5 empirical validation including the §5.7 counterfactual; §6 simulation for small teams; §7 the boundary with exogenous-retention schemes; §8 concludes.

2 Related Literature

We engage three literatures, one per design constraint the mechanism must satisfy. For each we state what it establishes, the gap it leaves, and the construction that fills it.

2.1 Team moral hazard and budget-breaking

Established. Holmström (1982): under budget balance ($\sum_i s_i = x$), no joint-outcome sharing rule induces efficient effort; his remedy is an outsider absorbing the residual. The failure is robust to risk preferences (Rasmusen, 1987); Dai and Toikka (2022): robust budget-balanced team contracts reduce to linear contracts with an outside claimant.

Gap. In both, the residual is a given – an outsider or a linear coefficient – not something the rule itself engineers.

Our construction. The fixed denominator T makes the residual an outcome of voting intensity, $EG = E(1 - \sum_i w_i/T)$ (P1): the budget breaker is the team’s own shortfall from full capacity, coupled to aggregate voting rather than an exogenously specified sharing coefficient. The linearity Dai and Toikka identify as optimal is preserved (each vote is worth exactly E/T), while the total retained is endogenous and, by R1,

schedulable (§2.4).

What is endogenous here? The firm remains the residual claimant, as in standard budget-breaking arrangements; the identity of the residual claimant is not the novelty. What is new is the determination of the residual itself: its magnitude is generated by the realized peer-evaluation intensity rather than fixed ex ante by a sharing coefficient. The mechanism therefore endogenizes the size of the residual while preserving the standard residual-claimant role of the firm – the object of design is not who absorbs the surplus, but how much surplus the rule lets the team create.

2.2 Peer assessment: reliability and manipulation

Established. (i) *Signal.* Meta-analytic peer–teacher agreement is $r = 0.69$ across 48 studies (Falchikov & Goldfinch, 2000); our data sit in this band, $\rho = 0.44\text{--}0.60$ against independent anchors (§5.2). (ii) *Manipulation.* Peer evaluation is gameable via grade inflation, collusion, and strategic reporting (incentive-compatible reviewer-accuracy designs, e.g., Ligon, 2026, target this); normalized peer scoring compounds it, since a member can raise his own share by depressing others’ totals.

Allocation-rule predecessors. Peer evaluation as an allocation rule is not new, but each construction restores budget balance: Dong, Falvey, and Luckraz (2019) let members assign shares summing to the pie; Carvalho and Larson (2012) build a budget-balanced peer-prediction mechanism; Carpenter, Robbett, and Akbar (2018) pair profit sharing with peer reporting to a manager, not peer-determined allocation. In all three the share vector exhausts the surplus; the fixed denominator relaxes exactly this constraint, turning the retained pool into an endogenous outcome.

Our construction. w_i contains only votes cast toward i , so member i ’s share is independent of his own voting vector (P2): self-promotion and opponent-damping are void structurally, before any behavioral safeguard. The signal side is handled by the governance layer (§4.8–4.10): reweighting raters without disturbing the theorems (P8) and making honest voting an equilibrium (P9–P10).

2.3 Profit sharing and claims-based allocation

Established. Revenue-sharing and claims rules divide a given estate: proportional and constrained-equal-awards rules distribute a fixed surplus, and team profit sharing is a linear split with an exogenous parameter. The denominator is a primitive, not an outcome.

Gap. Nothing connects the size of the divisible pool to the agents’ own assessment behavior.

Our construction. $T = N^2P$ (or $N \cdot r \cdot M$ sparse) is a capacity, not an estate: the vote matrix jointly produces individual shares (w_i/T) and the retained fund ($1 - \sum_i w_i/T$). This is a claims problem with an endogenous estate – a division problem in which the size of the pie itself is produced by the same vote matrix that decides the slices; proportional sharing is the special case $\phi = 1$ with full coverage ($G \rightarrow 1/N$), and the fixed denominator generalizes it by letting the estate shrink with voting shortfall – the budget-breaking content the claims literature assumes away.

2.4 Where the design law (R1) sits

The three lines converge on one object, the residual, and each leaves it fixed: Holmström and Dai and Toikka give it to an outsider or a linear coefficient, the peer-assessment line leaves the pool fixed, the claims line treats the estate as given. R1 supplies what none has, a closed-form law for the residual in the mechanism’s own inputs,

$$\mathbb{E}[G] = 1 - c \frac{N-1}{N} \phi(e), \quad (1)$$

with c the claim fill rate and ϕ the voting intensity. The law *predicts* retention from observable engagement (§5.6, machine precision where coverage is uniform); *explains* why retention differs across schemes (§5.3: $\hat{\phi} \approx 0.87/0.50/0.30 \rightarrow G \approx 13\%/50\%/70\%$); and *designs*, §5.7’s counterfactuals move retention exactly as it predicts, with the accuracy cost of low retention structural and bounded by the scale. Choosing a grading scheme is therefore choosing a retention level, by a law the data verify – the bridge between the incentive literature’s “residual must exist” and the allocation literature’s “estate is given”: here the residual exists, is given by a formula, and is measurable.

Table 1. Closest mechanisms.

None of the closest mechanisms reviewed here combines peer-determined relative shares with an endogenous residual claimant; the fixed-denominator rule does. Relative shares are peer-determined; *budget balanced* means the share vector exhausts output; the residual claimant column lists who absorbs the surplus; *residual endogenous to evaluation* means the residual is a function of peer-evaluation intensity rather than an exogenous parameter; *own-share manipulable* is whether a member can raise his own share by altering his own reports.

Mechanism	Peer-determined relative shares	Budget balanced	Residual claimant	Residual endogenous to evaluation	Own-share manipulable	Costly honest evaluation
Equal sharing	No	Yes	None	No	No	No
Normalized peer sharing	Yes	Yes	None	No	Yes	Yes
Profit sharing (exogenous rate)	No	Usually	Principal	No	No	N/A
Peer-reporting mechanisms	Varies	Often	Varies	Usually no	Varies	Yes
Fixed denominator (this paper)	Yes	No	Firm (behavior-linked)	Yes	No	Yes

The table makes the contribution’s location precise: relative weights are deliberately proportional (Lemma 1) – that is not the novelty – and the new object is the *endogenous aggregate scale*, the behavior-linked residual claimant, which none of the mechanisms most closely related to our setting possesses.

3 Model

3.1 Setup and notation

A team $I = \{1, \dots, N\}$ works on a project decomposed into P work streams. Each member i claims a nonempty subset $K_i \subseteq \{1, \dots, P\}$; $Y_i = |K_i|$; the claim fill rate is $c = \frac{1}{NP} \sum_i Y_i$ ($c = 1$ in full-coverage regimes). On each claimed stream $k \in K_i$, every other member $j \neq i$ casts a recognition vote $v_{j,i}^{(k)} \in \{0, 1\}$; no self-voting ($v_{i,i}^{(k)} \equiv 0$). Member i 's received votes are

$$w_i = \sum_{k \in K_i} \sum_{j \neq i} v_{j,i}^{(k)} \in [0, Y_i(N-1)]. \quad (2)$$

The project yields net income $E \geq 0$. The firm chooses a design retention $L \in (0, 1)$ as a strategic anchor; we show actual retention is an outcome, not this anchor.

Symbol	Meaning	Domain
N, P	team size, number of streams	$N \geq 2, P \geq 1$
K_i, Y_i	streams claimed by i , claim count	$Y_i \geq 1$
c	claim fill rate $\sum_i Y_i / (NP)$	$(0, 1]$
$v_{j,i}^{(k)}$	recognition vote, no self-voting	$\{0, 1\}$
w_i	received votes of i	$[0, Y_i(N-1)]$
T	fixed denominator (full-capacity votes)	N^2P or NrM (§3.3)
E	net income	≥ 0
Z_i	share weight w_i/T	$[0, (N-1)/N^2]$
s_i	member share $E \cdot Z_i$	≥ 0
E_{public}	retained surplus $E - \sum_i s_i$	≥ 0
K_j	KPI score of member j	> 0
G	retained fraction $1 - \sum_i w_i/T$	$(0, 1)$
ϕ	voting intensity (share of full scoring realized)	$(0, 1]$
L	design-retention anchor, not binding on G	$(0, 1)$
r, M	peers per reviewee, scale maximum (sparse regime)	$r \geq 1, M \geq 1$
α_i	member i 's claim on the retained fund (Prop. 5)	$\geq 0, \sum_i \alpha_i \leq 1$

3.2 The fixed-denominator allocation rule

Given a vote profile and the fixed denominator T :

$$(M1) \quad Z_i = \frac{w_i}{T}, \quad s_i = E Z_i = E \frac{w_i}{T}, \quad (M2) \quad E_{\text{public}} = E - \sum_i s_i = E \left(1 - \frac{\sum_i w_i}{T}\right). \quad (3)$$

T is constant, fixed in advance of voting; it is not the realized vote sum. In the full-suite regime $T = N^2P$: each member can receive at most $N-1$ votes on each of at most P streams, total capacity $N \cdot P(N-1) < N^2P$;

the gap to any realized $\sum_i w_i$ is the endogenous retention.

Key structural facts (proved in §4). For every feasible profile: $\sum_i Z_i \leq (N-1)/N < 1$ hence $E_{\text{public}} \geq E/N$ (P1); $Z_i \leq (N-1)/N^2 < 1/N$ (P3); s_i independent of i 's own voting vector (P2).

Lemma 1 (Relative-allocation invariance). Minimal representation. The mechanism's behavior is determined by the small core $(N, P, E, w_i, T, c, \phi)$, where c is the claim fill rate and ϕ the voting intensity; all other symbols are derived or optional. This core is used throughout §4–§6 for reproducibility.

3.3 Denominator regimes: full-suite vs. sparse mutual evaluation

- **Full-suite mutual voting** (small teams, $N \leq 25$): every member evaluates every other on each claimed stream; $T = N^2 P$.

- **Sparse mutual evaluation** (large teams / one-shot, $N > 25$ or few raters): each member is evaluated by a fixed number r of peers on a scale with maximum M ; capacity per member is rM , so $T = N r M$. This is the empirical regime (§5), with $r \approx 3$ and $M = 10$ (rating), $M = 5$ (coin budget), $M = 1$ (simple vote). The scale's design determines realized retention through $\hat{\phi}$ (Table 3).

The allocation mechanics (shares, retained fund, and the R1 expectation identity) hold in both regimes; only T changes. Prop. 1s structural lower bound $G \geq 1/N$ is specific to the full-suite regime, where no self-votes cap total votes at $N(N-1)P < N^2 P$; under sparse capacity a fully claimed profile reaches $G = 0$. The claim fill rate c of §3.1 is defined on work streams; in the sparse regime the same coefficient re-reads as realized rating coverage $\bar{r}/r_{\text{design}}$ (§5.6), the natural counterpart of claim saturation when the object is scored slots rather than claimed streams.

3.4 Governance layer (KPI weighting, voting cost, and calibration)

The bare rule maximizes structural robustness but inherits two weaknesses: the signal is noisy, and members lack an incentive to vote honestly (their share is unaffected by their own votes). The governance layer addresses both without disturbing the structural theorems.

(a) KPI-weighted votes. Each member j carries weight $\beta_j = 0.5 + K_j / \sum_i K_i \in (0.5, 1.5)$, so higher performers' judgments count more; effective votes $V_i = \sum_{k \in K_i} \sum_{j \neq i} \beta_j v_{j,i}^{(k)}$. Since β_j is a positive member-specific constant, replacing w_i by V_i is a monotone transformation; the structural theorems and scale identity continue to hold (P8).

(b) Dynamic voting cost. Each vote by member j costs $C_j = C_0(1 - K_j / \max_i K_i)$, $C_0 \in [0.01, 0.12]$ calibrated so the marginal benefit of a vote exceeds its cost for high-KPI members but not low-KPI ones.

Low-value voting becomes self-limiting and supports honesty as an equilibrium (P9–P10).

(c) **Retention calibration.** The design retention L and a convergence coefficient $\kappa = \text{clip}\left(1 - \frac{\sigma_{\text{KPI}}}{\mu_{\text{KPI}}} \delta, 0.90, 0.99\right)$ (a governance-layer parameter, distinct from the signal elasticity γ of Prop. 11), with δ driven by the firm's cash position, jointly set the realized member pool $R \cdot \kappa$ ($R = 1 - L$). **Neither L nor κ enters the allocation rule M1–M3:** the realized pool and retained fund are fully determined by $G = 1 - \sum_i w_i/T$. The firm cannot set retention directly through L ; it can only influence it indirectly by choosing the grading scheme that selects ϕ (§3.3, §5.6). We treat L and κ as governance-layer reference points, not determinants of endogenous retention.

4 Theory

This section establishes the mechanism's properties. The allocation rule is the fixed-denominator rule of §3: each member's share is his received votes over the full-capacity denominator, $s_i = E w_i/T$ (M1), and the residual $E_{\text{public}} = E(1 - \sum_i w_i/T)$ (M2) accrues to the firm, where $G := 1 - \sum_i w_i/T$, w_i is member i 's received votes, and T is the fixed denominator (N^2P or NrM , §3.3). The rule's key prediction is the closed-form retention identity

$$(M3) \quad \mathbb{E}[G] = 1 - c \frac{N-1}{N} \phi(e), \quad (4)$$

proved as Prop. 6 and tested in §5, with $c = \frac{1}{NP} \sum_i Y_i$ the claim fill rate and $\phi(e)$ the truthful-vote probability under effort e . Every proposition below is a property of this rule; every testable implication is examined in §5.

4.1 Budget balance and endogenous budget breaking

Proposition 1 (Budget balance and endogenous budget breaking).

$$\sum_i s_i \leq E, \quad E_{\text{public}} = E G \in [E/N, E], \quad (5)$$

with $G = 1 - \sum_i w_i/T$: the retained surplus always has a structural floor E/N and is an outcome of voting intensity, not a design parameter. Under sparse capacity ($T = N \cdot r \cdot M$) the floor relaxes to $E_{\text{public}} \in [0, E]$, zero only when every rater awards every rated peer the maximum (a fully claimed capacity).

Proof. $w_i \leq Y_i(N-1) \leq P(N-1)$, so $\sum_i w_i \leq NP(N-1) < N^2P = T$; hence $\sum_i s_i < E$ and $E_{\text{public}} \geq E/N$, equality only at full capacity.

Remark. Holmström (1982, Thm. 1): under budget balance ($\sum_i s_i = x$) no joint-outcome sharing rule implements efficiency; an outsider must absorb the residual. Prop. 1 constructs that budget breaker explicitly: $E_{\text{public}} = EG$, priced by the shortfall of realized votes from full capacity, with no exogenous royalty rate.

Testable Implication ($\rightarrow$ §5, H2).

4.2 Own-share non-manipulability

Proposition 2 (Own-share non-manipulability). *Proof.* $s_i = Ew_i/T$ and $w_i = \sum_{k \in K_i} \sum_{j \neq i} v_{j,i}^{(k)}$ contains only votes cast by others toward i ; member i controls only $v_{i,\cdot}^{(k)}$, absent from w_i .

Remark. Under normalized scoring $s_i^P = Ew_i/\sum w$, member i can raise his share by depressing others' vote totals; the fixed denominator T removes this channel structurally, before any behavioral safeguard.

Corollary 1 (C1, self-manipulation immunity).

Testable Implication ($\rightarrow$ §5, H3).

4.3 Structural upper bound

Proposition 3 (Structural anti-concentration).

$$Z_i = \frac{w_i}{T} \leq \frac{N-1}{N^2} < \frac{1}{N}, \quad s_i < \frac{E}{N}. \quad (6)$$

Proof. $w_i \leq P(N-1)$, so $Z_i \leq (N-1)/N^2 < 1/N$ for $N \geq 2$.

Remark. The concentration bound is structural, not a parameter choice; the bound tightens as $O(1/N)$ in larger teams.

Corollary 2 (C2, structural cap).

Testable Implication ($\rightarrow$ §5, H3).

4.4 Effort monotonicity under truthful voting

Setup. Member i exerts effort $e_i \in [0, 1]$; others observe a noisy signal and, truthfully, vote yes with probability $\phi(e_i)$, strictly increasing.

Proposition 4 (Effort monotonicity).

$$\mathbb{E}[w_i] = Y_i(N-1)\phi(e_i), \quad \mathbb{E}[s_i] = E \frac{Y_i(N-1)\phi(e_i)}{T}, \quad (7)$$

strictly increasing in e_i ; and $\mathbb{E}[s_i]/\mathbb{E}[s_j] = Y_i\phi(e_i)/(Y_j\phi(e_j))$.

Proof. $\mathbb{E}[w_i] = \sum_{k \in K_i} \sum_{j \neq i} \phi(e_i) = Y_i(N-1)\phi(e_i)$ by linearity; rest follows from $s_i = E w_i/T$.

Remark. Under truthful evaluation, more effort yields more votes and a higher share – the mechanism’s incentive content. The honest-voting premise is explicit; §4.9–4.10 addresses how it is sustained.

Testable Implication ($\rightarrow$ §5, H1).

4.5 Coalitional vulnerability and retained-fund ownership

Setup. The retained fund carries an ownership vector $\alpha_i \geq 0$, $\sum_i \alpha_i \leq 1$: member i ’s claim on E_{public} .

Proposition 5 (Coalitional vulnerability). *Proof.* $\partial s_i / \partial w_j = 0$ for $j \neq i$; $\partial E_{\text{public}} / \partial w_j = -E/T < 0$.

Remark. Manipulation resistance is fully characterized by fund ownership: with no member claim, the mechanism is immune to self-interested manipulation; with proportional claims, a depression channel opens and calls for robust aggregation (§4.8).

Testable Implication ($\rightarrow$ §6).

4.6 Retention calibration

Proposition 6 (Retention identity).

$$\mathbb{E}[G] = 1 - c \frac{N-1}{N} \phi(e), \quad \xrightarrow{N \rightarrow \infty} 1 - c \phi(e). \quad (8)$$

Proof. $\mathbb{E}[\sum_i w_i] = Y_{\text{ASM}}(N-1)\phi(e)$ by Prop. 4; divide by $T = N^2 P$.

Remark. Expected retention is a closed-form function of two observable measures, c and ϕ ; retention is thus endogenous and *schedulable*: a firm targeting G^* sets c so $\mathbb{E}[G] = G^*$.

Testable Implication ($\rightarrow$ §5, R1).

4.7 Feasibility note (Shapley value)

Shapley-type allocation is infeasible in the information structure considered here: it requires 2^N coalition evaluations and sub-coalition surplus that team production does not observe. We do not pursue it as a benchmark (Online Appendix A).

4.8 Governance invariance

Setup. KPI-weighted votes (§3.4(a)): member j carries weight $\beta_j = 0.5 + K_j / \sum_i K_i$; effective votes $V_i = \sum_{k \in K_i} \sum_{j \neq i} \beta_j v_{j,i}^{(k)}$.

Proposition 8 (Governance invariance). *Proof.* V_i is still linear in the vote profile with positive member weights; the scale identity and all bounds depend only on linearity and monotone weighting.

Remark. Governance enriches the signal (higher performers' judgments weigh more) without disturbing structural theorems – structure and governance can be validated independently.

4.9 Voting-cost discipline

Setup. A vote costs C_j as in §3.4(b): $C_j = C_0(1 - K_j / \max_i K_i)$, $C_0 \in [0.01, 0.12]$. Absent voting rewards, a member's own share is unaffected by his votes (Prop. 2).

Proposition 9 (Voting degeneracy without reward). *Proof.* The voter's share contains none of his own votes (Prop. 2); an honest vote yields cost and no benefit, abstention dominates.

Remark. Prop. 9 is the free-rider problem of *evaluation*: structural immunity to self-vote manipulation removes the private incentive to evaluate honestly – the gap the reward layer closes next.

4.10 Honest-voting equilibrium

Setup. The firm scores each voter against the realized distribution of others' evaluations (peer-prediction scoring, Miller–Resnick–Zeckhauser 2005): a strictly proper scoring rule on the induced posterior.

Proposition 10 (Honest-voting Nash equilibrium). *Proof (sketch).* A strictly proper scoring rule makes truthful report the unique expected-payoff maximizer independently of others (MRZ 2005); participation iff expected reward $\geq C_j$, and C_j falls in K_j .

Remark. Prop. 9 identified the weakness; Prop. 10 supplies the corrective. Structure defeats self-vote manipulation; reward-plus-cost makes honest voting an equilibrium while filtering toward informative raters. Tests in §5–§6 and the companion experiment.

4.11 Conditional output comparison under honest voting

Setup. Here z_i is the continuous counterpart of the discrete share weight w_i/T of §3.2; under symmetric, sufficiently dense voting the two coincide proportionally. A symmetric team of N identical members ($N \geq 3$), additive output $E(\mathbf{a}) = A \sum_j a_j^\beta$ ($0 < \beta < 1$; $E' = A\beta a_i^{\beta-1}$), convex cost $C(a) = \frac{c}{2}a^2$ ($C' = ca$), and recognition share

$$z_i = (1 - G) \cdot \frac{a_i^\gamma}{\sum_j a_j^\gamma}, \quad (9)$$

so $\sum_i z_i = 1 - G$; $\gamma \geq 0$ is the signal response of recognition to effort, a design surface through the grading scheme (ϕ , §3.3, §5.6), not a free parameter. Under honest voting, z_i is increasing in a_i for $\gamma > 0$; $z'_i = \partial z_i / \partial a_i$.

Equilibrium. Expected income is Ez_i ; the fixed-denominator FOC is $E'z^* + Ez'_i = C'(a^*)$. Symmetrically $z^* = (1 - G)/N$ and $z'_i(a^*) = \frac{(1-G)\gamma(N-1)}{N^2 a^*}$: the sharing channel $E'z^*$ is compressed by the fund ($z^* < 1/N$), while the signal channel Ez'_i is absent under any fixed-share rule.

Proposition 11 (Conditional output improvement).

$$\gamma > \frac{\beta G}{(1 - G)(N - 1)}. \quad (10)$$

Proof. FOC solves to $a^F = [A(1 - G)(\beta + \gamma(N - 1)) / (cN)]^{1/(2-\beta)}$; equal division ($z_i \equiv 1/N$) to $a^E = [A\beta / (cN)]^{1/(2-\beta)}$. Since E rises in a , $E^F > E^E$ iff $(1 - G)[\beta + \gamma(N - 1)] > \beta$.

Corollary 3 (First-best output condition in the benchmark model).

Reading. (i) Holmström's impossibility applies to $\sum_i s_i = E$; our rule has $\sum_i s_i = (1 - G)E < E$, outside that theorem's domain – Prop. 11 is the payoff of relaxing budget balance, not a contradiction. (ii) The threshold is tiny in the empirical range: for $N = 10$, $G = 13.5\%$ (CHDYPY), $\gamma > 0.009$ suffices; the exact first best requires $\gamma = 0.587$, yet at the estimated $\gamma \approx 0.4$ output already reaches 89.2% ($N = 10$) and 88.8% ($N = 20$) of the benchmark's first-best output. The comparison is *output*, $E(\mathbf{a})$ in the maintained symmetric benchmark – not surplus, welfare, or utility: the retained fund's final use is a governance choice, so we do

not claim welfare gains. (iii) The efficiency source is the signal channel compensating the fund-induced compression $E'G/N$, the equilibrium-level counterpart of Prop. 4; larger funds demand stronger signals, the §5.7 trade-off, so the two together delineate an interior design region. The result is a conditional output comparison – symmetric team, honest voting, and a specified production function – and inherits the §4.9–4.10 collusion boundary; it is not a general efficiency theorem.

4.12 Chapter summary (what the theorems say)

Prop	Property	Evidence anchor
P1	Budget breaking, endogenous G	§5 H2 (real $G \in (0, 1)$ varies)
P2	Own-share non-manipulability	§5 H3
P3	Structural anti-concentration	§5 H3
P4	Effort monotonicity under truth	§5 H1 ($\rho_{\text{pool}} = 0.439$)
P5	Coalitional vulnerability $\propto$ fund ownership	§6 simulation
P6	Retention identity $E[G] = 1 - c(N - 1)\phi/N$	§5 R1
P7	Feasibility vs Shapley	Online App. A
P8	Governance invariance	§5 H4
P9	Voting degeneracy without reward	(theoretical)
P10	Honest-voting equilibrium with reward	companion experiment
P11	Conditional output improvement vs budget balance	§5.2 (recognition–effort elasticity), §6

Every property above has a designated empirical or simulation test, none is asserted without a designated check. The data speak in §5.

5 Empirical Evidence

Three independent datasets, spanning four grading schemes and three countries, speak to the theory in §4. This section states what counts as evidence (§5.1), then examines each testable implication: H1 effort–truth correlation (§5.2), H2 realized retention (§5.3), H3 scale identity and structural bounds (§5.4), H4 governance invariance (§5.5), R1 the retention identity (§5.6), and §5.7 turns the identity into a data-driven counterfactual. §5.8 states what these data can and cannot claim.

5.1 Evidence hierarchy and the three datasets

Three kinds of numerical facts carry different evidentiary weight.

1. **Algebraic identities (necessary, not statistical).** Relations that hold by construction of the rule, M1’s scale identity and the decomposition $G = 1 - \sum_i w_i/T$, verify that the pipeline implements the model, nothing

more.

2. **External-anchor statistics (real evidence).** Correlations between the mechanism’s inputs and a variable that was not an input – here an independent instructor/expert grade. These are the theory’s testable content (H1, H2, R1).

3. **Endogenous anchors (excluded by design).** A variable that is a function of the peer grades, or shares inputs with them; such correlations are mechanical and are reported only to be identified as circular, never as evidence.

Datasets.

- **CHDYPY** (Xu et al., 2025; dataset (Xu et al., 2023)): field experiment of a Stackelberg spot-checking mechanism in two Chinese programming courses; 20 cohorts (1,221 reviewees, 0–10 ratings), one teacher grade per reviewee as external anchor.

- **peer-grading** (Jones-Todd & Renelle, 2025): introductory R-coding assessment, University of Auckland; every submission graded by exactly three peers and three tutors; tutor average is the anchor, the peer-derived final grade is excluded by design.

- **t4e** (Ashenafi, 2019): peer-assessment in five Trento CS courses (2013–2016), v1 single ratings and v2 coin-budget votes; no instructor anchor, so its evidence is limited to endogenous checks.

Table 2.

Summary of the three datasets.

	CHDYPY	peer-grading	t4e
Source	China, programming courses	NZ, R-coding course	Italy, Trento, 5 CS courses
Grading scheme	0–10 rating, 3 peers each	5-item rubric (max 14), 3 peers + 3 tutors each	v1 simple vote; v2 coin budget 0–5
Units	20 cohorts, $N = 58\text{--}68$, 1,221 reviewees	247 submissions	v1: 1,398 / 1,041; v2: 1,146 / 972 / 728
Independent anchor	teacher grade	tutor average; final grade = endogenous, excluded	course score (weak: participation + ability; incomplete = 0)

Honest placement. CHDYPY’s teacher grade and peer-grading’s tutor grades are external anchors and carry the evidentiary weight in §5.2. t4e has no instructor-assigned per-answer grade; its course score co-moves with participation, so t4e validates cross-scheme structural facts (H3, R1) but is explicitly not an H1 test. peer-grading’s final grade is a function of the peer grades, so its $\rho = 0.946$ against received votes is circular and reported only to be excluded.

Why these datasets are informative about the mechanism. These data are not used to identify work-place treatment effects. They provide repeated realizations of the peer-evaluation primitives that enter the mechanism – coverage, score generosity, received recognition, and independent anchors. Their role is therefore to identify the informational and calibration properties of the voting process; the team-compensation implications are established theoretically (Prop. 1–11) and through the team-regime simulation (§6).

5.2 H1, Received peer votes track independent truth (Prop. 4)

Setup. Prop. 4: $\mathbb{E}[w_i]$ increases in contribution under truthful voting, hence $\rho(w, \text{truth}) > 0$.

CHDYPY (external anchor). w_i = sum of peerGrade, truth = teacher grade. Cohort-level Spearman ρ : mean **0.520** [95% CI 0.302, 0.747], positive in all 20 cohorts. Pooled: $\rho_{\text{pool}} = 0.439$ [95% CI 0.364, 0.466]. Permutation (shuffling votes with truth) rejects the null in all cohorts: **20/20 at** $p < 0.05$ (16/20 at $p < 0.005$), pooled $p < 0.001$. No cohort's CI crosses zero (Fig. 1).

peer-grading (external anchor). w_i = sum of peer score (max 14), truth = mean of 3 tutors. $\rho = 0.604$ [95% CI 0.519, 0.699], $p < 0.001$. *The same data give $\rho = 0.946$ against the moderated final grade; that grade derives from the peer scores, so 0.946 is circular and not reported as evidence.*

t4e (weak anchor, not an H1 test). w_i = coins received correlates with course score at $\rho \approx 0.13$ –0.27, which embeds participation and high incompleteness; reported only to bound what a weak anchor yields.

Reading. Both external-anchor datasets find genuine signal (0.44–0.60), far from perfect – and the theory needs $\rho > 0$, not $\rho \approx 1$: retention is higher precisely when vote signal is imperfect (R1). This is why the governance layer (§4.8–4.10) and a fixed denominator matter. **Benchmark.** These sit below the published peer–teacher agreement benchmark ($r = 0.69$ meta-analytic), a level that pairs an individual's grade with the teacher's for the same submission, whereas our ρ correlates *received votes* with truth (Online Appendix F).

Elasticity for Prop. 11. The same signal as an elasticity anchors the §4.11 conditional output comparison. A class-demeaned log–log regression of received votes on the instructor anchor estimates the recognition–effort elasticity γ : class-level mean $\bar{\gamma} = 0.41$, positive in all 20/20 classes and overwhelmingly above the Prop. 11 threshold (factor ≈ 46 at the CHDYPY N and G). At this elasticity the fixed-denominator output exceeds equal division by $\approx 92\%$ ($N = 10$) and 141% ($N = 20$). Estimation details (bootstrap CIs, pooled fixed-effects and rank regressions, and the conservatism caveat) are in Online Appendix B.

5.3 H2, Realized retention is endogenous and interior (Prop. 1, 6)

Setup. Prop. 1: $G = 1 - \sum_i w_i / T \in (0, 1)$; Prop. 6: $\mathbb{E}[G] = 1 - c \frac{N-1}{N} \phi$.

Results. G is interior in every cohort of every dataset. CHDYPY: mean **13.5%** [6.3%, 25.0%]; peer-grading: **13.7%**; t4e: v2 coin budget $\approx 50\%$, v1 simple vote $\approx 70\%$.

Why the levels differ. G is a function of grading design. No-budget rating scales yield $\hat{\phi} \approx 0.87$ of maximum, leaving $\approx 13\%$ unallocated; scarce-budget coins yield $\phi \approx 0.50$ ($\approx 50\%$); coarse single votes $\phi \approx 0.30$ ($\approx 70\%$). The fixed denominator converts "how generously graders score" into "how much is retained": a firm choosing a grading scheme is choosing its retention level.

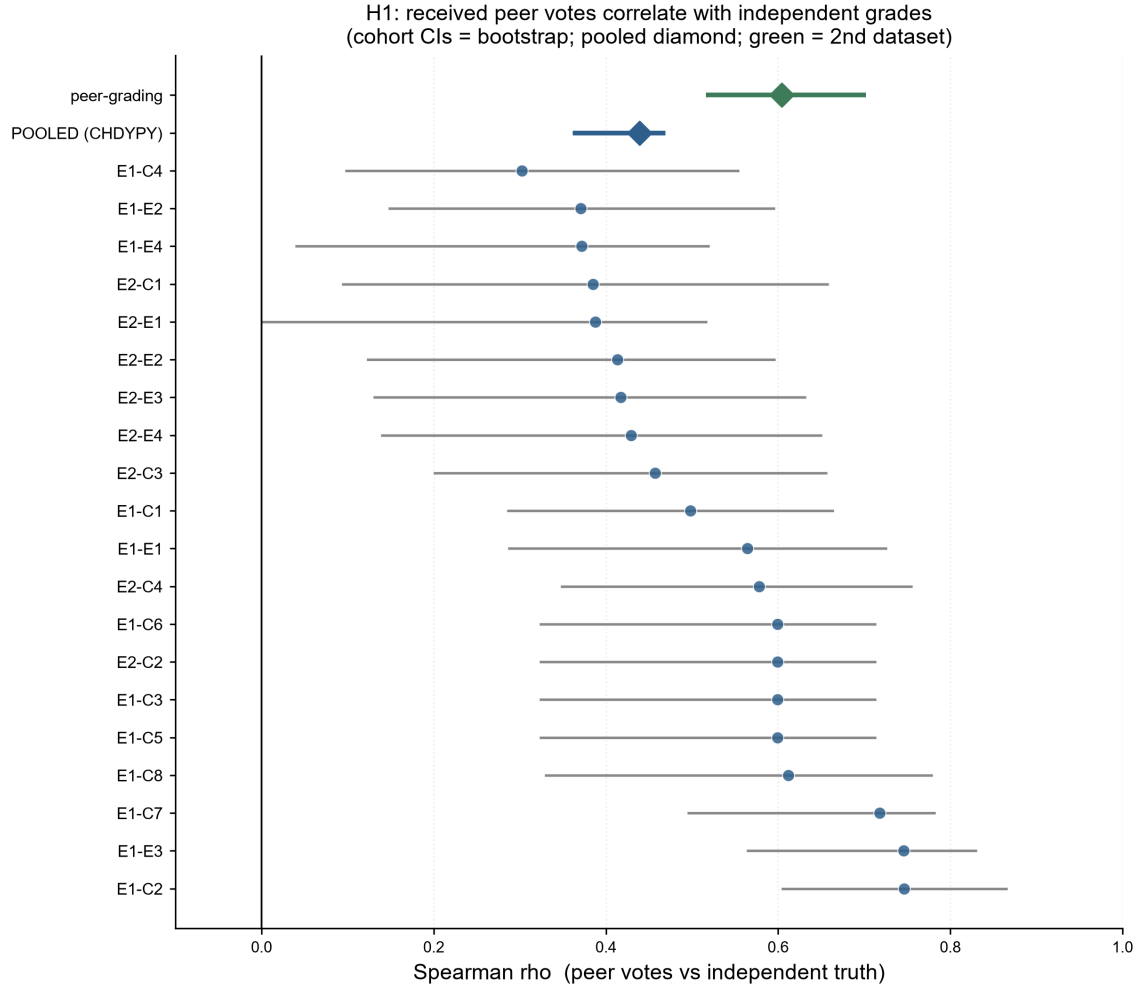


Figure 1: H1 forest plot, cohort Spearman ρ (CHDYPY, blue; bootstrap CIs), pooled diamond $\rho_{\text{pool}} = 0.439$, peer-grading tutor anchor $\rho = 0.604$ (green). All strictly positive.

Calibration surface. Figure 2 plots $G = 1 - (\bar{r}/r_{\text{design}})\hat{\phi}$ with every cohort overlaid; each lands on the surface (rating $\approx 13\text{--}14\%$, coin $\approx 50\%$, simple vote $\approx 70\%$). Where a point leaves it (coin cohort 5_{pr1}, $r/r_{\text{design}} = 0.69$), the coverage correction, not the mechanism, is approximate: its G uses the published 0–5 scale while the true budget rule is not public. The surface is a design map: pick a scheme, read off the retention it engineers.

Reading. H2 provides cross-scheme evidence consistent with endogenous retention and shows that realized retention varies systematically with grading design; the cross-scheme spread is the design surface of R1.

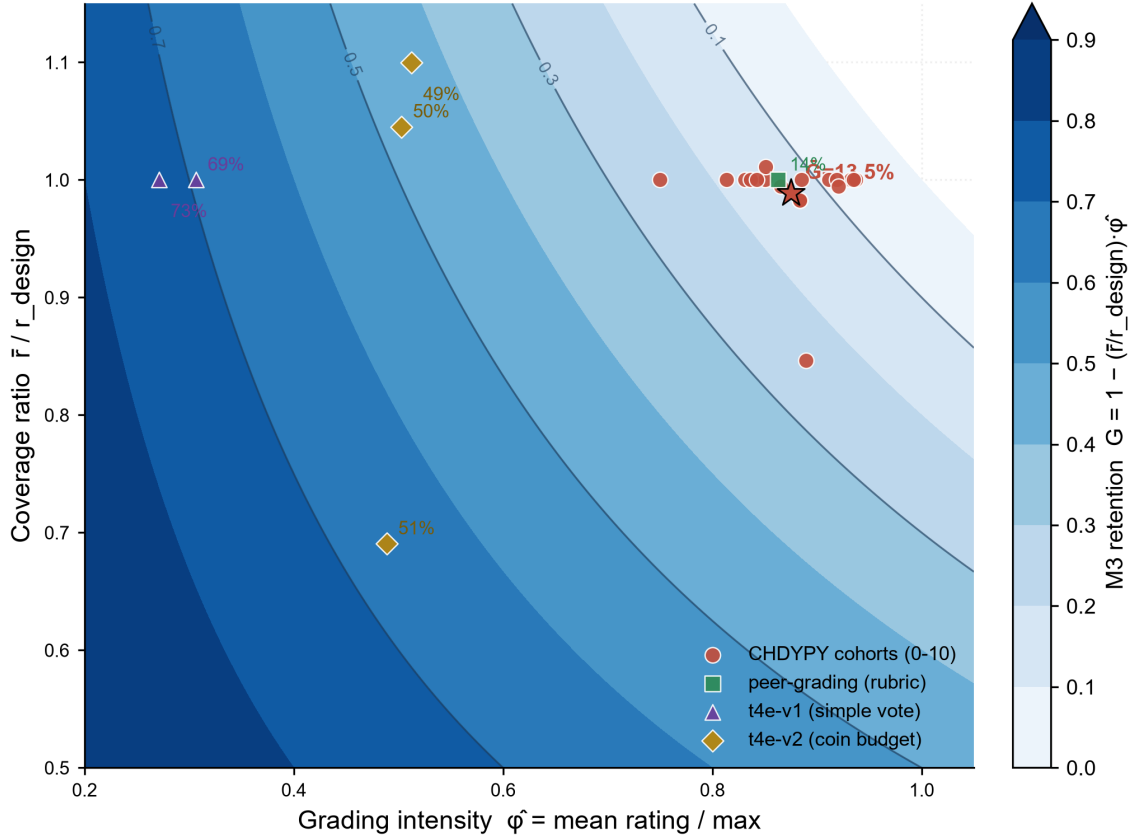


Figure 2: M3 calibration surface $G = 1 - (\bar{r}/r_{design})\hat{\phi}$ with three-dataset observations overlaid.

5.4 H3, Scale identity and structural anti-concentration (Prop. 2, 3; Cor. 1, 2)

Setup. Prop. 2/3: the fixed denominator renders own-share non-manipulable and caps every share at $Z_i < 1/N$.

Scale identity (algebraic). Fixed-denominator and normalized shares have identical relative order in every dataset: Kendall $\tau = 1$ in all 20 CHDYPY cohorts, peer-grading, and t4e v1/v2 ($s^F = (1 - G)s^P$).

Coverage-adjusted bound (Prop. 3, sparse coverage). Let r_i be the actual number of scores reviewee i receives. Since $w_i \leq r_i M$ and $T = N r_{design} M$,

$$Z_i = \frac{w_i}{T} \leq \frac{r_i}{N r_{design}} = \frac{r_i}{r_{design}} \cdot \frac{1}{N}. \quad (11)$$

Equality iff every score is maximum. The raw bound is the special case $r_i = r_{design}$: uneven coverage rescales, not breaks, the structural cap.

One real violation, exactly explained. The raw bound holds in 19/20 CHDYPY cohorts and peer-grading; it fails only in one cohort where a reviewee was scored 5 times instead of 3. The adjusted bound contains

it exactly $((5/3)/60 = 0.0278 > 0.0244)$; across all cohorts the adjusted bound is violated **0 times** (details in Online Appendix C). The raw violation is not a counterexample but a confirmation of the general bound, made visible by broken uniformity.

t4e (bound not testable). All t4e cohorts violate the raw bound, but coverage is 0–7 and the design parameter r_{design} is unpublished; without it T is undetermined. We state this rather than guess a denominator.

Reading. The bound holds exactly under uniform scoring; where it seems to fail, the coverage correction r_i/r_{design} restores it exactly – a falsifiable, confirmed prediction, and a data-driven prescription (unify scoring slots), which is why r must be an enforced design parameter.

Table 4. coverage-adjusted bound, raw-cap violations (1 in Exp.2.controlGroup.3; t4e not testable) and 0 adjusted-cap violations.

5.5 H4, Governance invariance (Prop. 8)

Setup. Prop. 8: replacing w by KPI-weighted votes V preserves M1 and all bounds.

CHDYPY. With $\beta_j = 0.5 + K_j/\sum K$, the weighted identity holds to machine precision ($\max |s_V^F - (1 - G_V)s_V^P| = 3.5 \times 10^{-18}$); the β ’s are near-constant (std 0.0024), so this verifies only structural invariance. **peer-grading** and **t4e** cannot test weighting: the grader–gradee KPI mapping is unobservable or absent (Online Appendix D).

Reading. The governance claim is structural; its behavioral content (whether weighting reorders shares) requires weights with spread, which neither dataset provides. The stress test is delegated to the §6.6 governance simulation. We do not claim empirical support for the behavioral version.

5.6 R1, The retention identity across grading schemes (Prop. 6)

Setup. Prop. 6: $\mathbb{E}[G] = 1 - c \frac{N-1}{N} \phi$; with uniform coverage ($c = 1$), $G = 1 - \bar{r}/r_{\text{design}} \cdot \phi$, where ϕ = mean single grade / maximum and $\bar{r}/r_{\text{design}}$ realizes the claim fill rate.

Cross-dataset check. On uniformly covered cases (20 CHDYPY cohorts, peer-grading) the identity holds to machine precision (max error 1×10^{-16}); t4e v1 by the same arithmetic; t4e v2’s coin rule is not public, so its per-answer maximum is approximate and errors rise to 0.02–0.15, which we report rather than tune away.

The non-trivial part. The identity is arithmetic; the content is that ϕ is systematically determined by the grading scheme (Table 3). This turns R1 from calibration into a *design law*: G^* is achieved by choosing a

scheme with the corresponding ϕ .

Table 3.

Grading scheme, realized voting intensity $\hat{\phi}$, and realized retention G .

Scheme	ϕ (mean/max)	realized G	datasets
0–10 rating	0.87	13.5%	CHDYPY
rubric 14	0.86	13.7%	peer-grading
coin budget	0.50	$\approx 50\%$	t4e v2
simple vote	0.30	$\approx 70\%$	t4e v1

5.7 The retention–accuracy design frontier (data-driven counterfactuals)

§6 verifies the mechanism’s arithmetic under the theory’s vote model; this subsection (per §6.1) builds on that verified core to ask the paper’s central design question: calibrate voting to the three real datasets, then ask what a cohort’s retention and share accuracy *would* be under a different grading scheme, signal quality held fixed. The simulation is data-driven, not behavioral: it embeds the observed ϕ -by-scheme levels (Table 3) as calibrated parameters.

Setup. $N = 60$ contributions log-normal; latent score $q_{ij} = b + \kappa F_j + \eta \varepsilon_{ij}$ (F_j empirical CDF, κ discrimination, ε noise); observed = round($M \cdot \min(\max(q, 0), 1)$), $r = 3$ raters each. $\kappa = 0.15$, $\eta = 0.10$ fix signal quality; b calibrated to reproduce each scheme’s $\hat{\phi}$ (0.87/0.50/0.30). 25 replicates.

5.7.1 Retention is a target the planner can choose. Switching schemes moves realized retention to the target scheme’s level exactly: the M3 identity holds to machine precision in every counterfactual (max error ≈ 0). A coin-budget cohort *would* retain about 50% and a rating cohort about 13% under the same signal. The cross-scheme spread of §5.3 is therefore predicted, not accidental: retention is not a parameter the firm must calibrate from outside, but a target selected by choosing the grading scheme.

5.7.2 The retention–accuracy tradeoff. Sweeping generosity b traces the attainable (G, ρ) frontier (Fig. 4). Accuracy is not monotone in retention: at very generous scales ($G < 10\%$) most scores hit the ceiling, collapsing discriminating variance ($\rho \approx 0.35$); very stingy scales do the same at the floor. The calibrated schemes sit on this frontier at their realized ϕ ; observed datasets (CHDYPY 13.5%/0.44, peer 13.7%/0.60, t4e v2 50%/0.20, v1 70%/0.15) fall on or inside it. The t4e points sit below the frontier: that gap is the known weakness of the t4e anchor (§5.1), not model error; we report it rather than rescale it away.

5.7.3 An interior design optimum. The frontier turns the retention choice into a planner problem with an explicit objective, $U(G) = \alpha(1 - G) + (1 - \alpha)\rho(G)$, where $\alpha \in [0, 1]$ is the planner’s relative weight on retained surplus versus information quality – in a firm, the weight on a larger internal fund (lower immediate payout, more retained for future incentives) versus a more discriminative assessment (better targeting of the incentives actually paid). The first-order condition is

$$U'(G) = -\alpha + (1 - \alpha)\rho'(G) = 0 \iff \rho'(G^*) = \frac{\alpha}{1 - \alpha}, \quad (12)$$

with second-order condition $U'' = (1 - \alpha)\rho''(G) < 0$; the calibrated frontier satisfies both between the ceiling and floor regimes. The optimum is interior – relative to the feasible design region induced by the grading scale – and varies with planner preferences, with flat regions at intermediate weights (Table 5): $\alpha = 0.3$ keeps $G^* = 13.9\%$ at $\rho^* = 0.70$; as the planner values surplus more ($\alpha = 0.5$ – 0.7) the optimum pushes G^* to 7.5% at $\rho^* = 0.65$; at $\alpha = 0.9$ it drives G^* to 1.6%, where floor effects collapse accuracy to $\rho^* = 0.35$. Buying retention beyond a point means paying in ranking accuracy; the boundary of the trade is the grading scale itself. The design message: *the firm should neither minimize nor maximize retention; it should set the grading scheme at the interior point where the marginal value of retention equals the marginal information loss.*

Table 5.

Sensitivity of the interior design optimum to planner preferences.		
Planner weight α on retained surplus	Optimal retention G^*	Implied accuracy ρ^*
0.3	13.9%	0.70
0.5	7.5%	0.65
0.7	7.5%	0.65
0.9	1.6%	0.35

(From *policy_sim_optimum.csv*, 25 replicates.)

5.7.4 What the counterfactual does not identify. This is a counterfactual calibration, not a behavioral experiment: raters are truthful conditional on noise, so the accuracy loss at low G is a structural ceiling effect, not a prediction about real rater effort. The behavioral question – whether real raters respond to the scheme as calibrated – is left to future experimental work. What the counterfactual establishes is narrower and safer: *under the voting behavior the data already display, retention is a designable target, the mechanism predicts its level exactly, and the price of low retention is structural.*

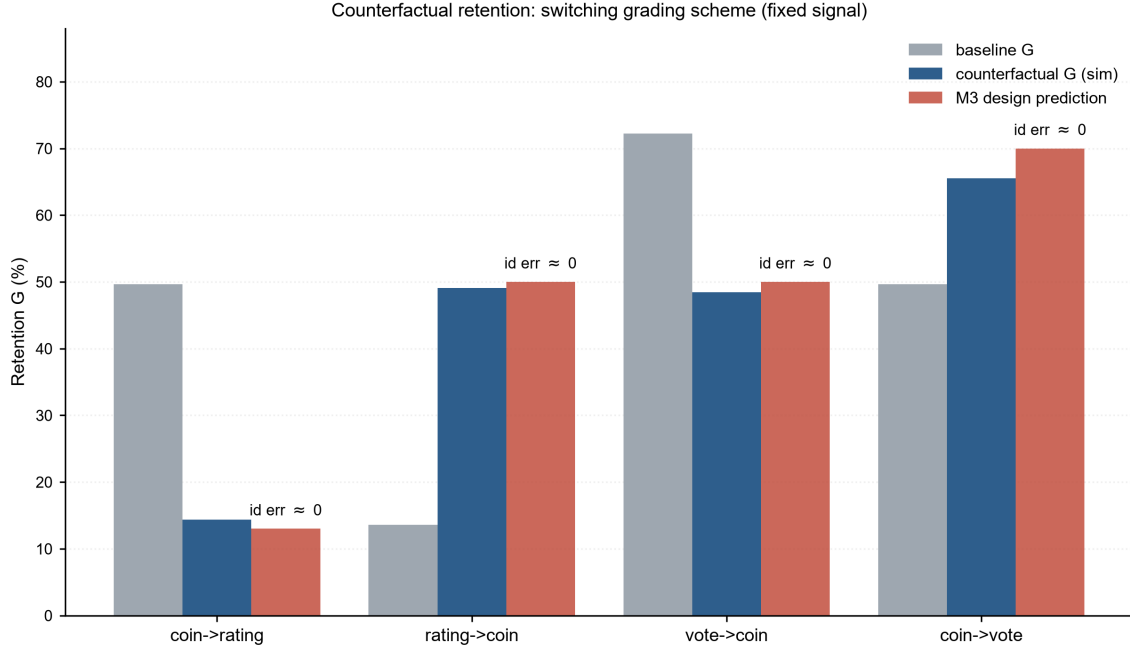


Figure 3: Counterfactual retention for four scheme switches; M3 identity error ≈ 0 .

5.8 What these data can and cannot claim

- **Established.** Received peer votes carry real signal against independent grades (H1, $\rho \approx 0.44$ – 0.60 on two external-anchor datasets). Retention is endogenous and interior (H2) and follows the R1 closed form exactly where coverage is uniform. The structural bound holds where scoring slots are uniform and fails exactly where they are not (H3), a falsifiable prediction the data confirm.

- **Not established by these data.** The behavioral governance claim (whether weighting reorders shares) is untestable here because the weights have no spread. The causal reading of H1 (correlation with an instructor grade does not identify the truthful-vote mechanism) only bounds plausibility. Any claim about honesty or effort is inference, not measurement.

- **Why the numbers look the way they do.** Each dataset was built for a study that needed peer assessment to work; the H1 signal partly reflects that selection. Two datasets have independent grades; t4e does not, and its weak correlations are the honest consequence. Because the datasets differ jointly in country, course, task, and scheme, cross-dataset differences in ϕ and G are indicative, not a controlled contrast; the scheme-level attribution in Table 3 is descriptive calibration, with mechanism-level claims carried by the structural theory (§3–4) and simulation (§6).

Data availability. All three datasets are public or author-released (Xu et al., 2025; Jones-Todd & Renelle, 2025; Ashenafi, 2019). We report cohort-level aggregates only.

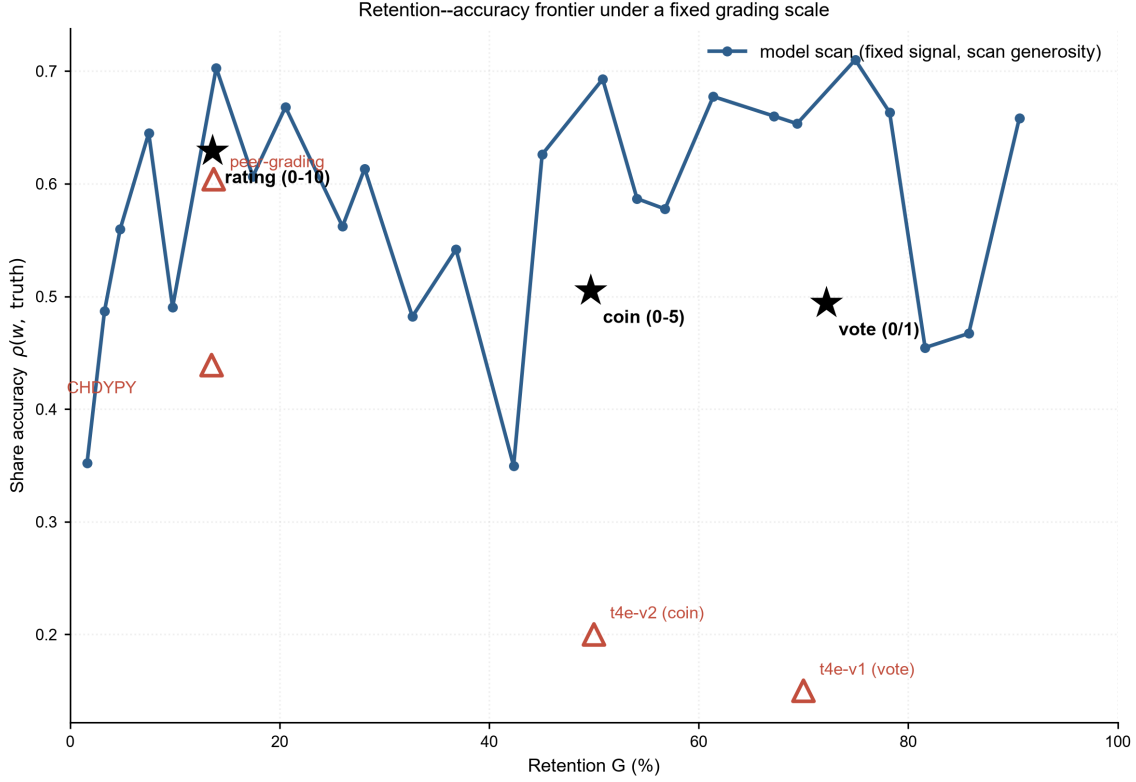


Figure 4: Attainable (G, ρ) frontier; calibrated schemes (stars) and observed datasets (hollow triangles); t4e’s gap reflects its weak anchor .

6 Simulation

6.1 Role of simulation relative to the data (§5) and the counterfactual (§5.7)

The §5 data are sparse-coverage ($r = 3$) peer-assessment records with cohorts beyond the team regime ($N > 25$). They validate the mechanism’s general properties but cannot test the team regime the mechanism is designed for ($N \leq 25$, dense mutual voting), nor isolate the theorems’ arithmetic from real grading noise. Simulation covers this gap in two distinct steps:

1. **§6, mechanism verification.** With vote counts drawn from the theory’s own stochastic model (independent Binomial participation), every theorem must hold to numerical precision – a consistency test that the §4 algebra is what the pipeline implements, before any data-driven claim.
2. **§5.7, data-driven counterfactual.** The same identities, calibrated to observed ϕ -by-scheme levels (§5.6), predict retention under a different scheme. §6 establishes the identity is *right*; §5.7 that it is *useful as a design tool*.

All simulations use a fixed seed and are reproducible (`scripts/run_simulation.py`).

6.2 Setup

$N \in \{6, 10, 15, 25\}$, $P \in \{4, 5, 7\}$; full-suite denominator $T = N^2P$. Received votes w_i are drawn from the theory's Bernoulli model: member j participates in $Z_j \sim \text{Binomial}(P, c)$ projects and votes yes to member i with probability ϕ , independent – the same sampling model as the proofs of Prop. 4 and 6.

6.3 M1, scale identity holds to machine precision (Prop. 2, 3)

Over 2,000 random vote matrices per (N, P) cell, $s_i^F = Ew_i/T$ and $s_i^P = Ew_i/\sum_i w_i$ satisfy $s_i^F = (1 - G)s_i^P$ to **machine precision** (max error 4.9×10^{-17} at $N = 6$, 4.8×10^{-18} at $N = 25$; Fig. 6). The fixed denominator differs from normalized scoring only through the constant $1 - G$, the channel P2 and P3 operate on.

6.4 M3, expected retention identity (Prop. 6)

Closed form $\mathbb{E}[G] = 1 - c \frac{N-1}{N} \phi$ vs Monte-Carlo over a 36-cell grid ($N \in \{6, 10, 15, 25\}$, $c \in \{0.5, 0.65, 0.8\}$, $\phi \in \{0.6, 0.8, 1.0\}$): **max deviation 0.0012** (e.g. $N = 15, c = 0.65, \phi = 0.8$: theory 0.5147 vs simulation 0.5143; Fig. 5). The identity holds at every interior point, which is the sense in which retention is schedulable (Prop. 6): pick (c, ϕ) , retention is one line.

6.5 Structural lower bound (Prop. 1)

Lower bound. Over 5,000 random full-suite teams ($T = N^2P$), retention never falls below $1/N$: **0 violations** (Prop. 1's bound – under full capacity the retained fund always has mass at least E/N). Each vote changes the recipient's share by exactly E/T and the fund by the same amount, so the price of a vote is uniformly E/T – a level playing field that §5.7's frontier inherits.

6.6 Governance invariance under β -weighting (Prop. 8)

With KPI weights $\beta_j = 0.5 + K_j/\sum K$ and $V_i = \sum_j \beta_j v_{j,i}$, the scale identity holds to 2.8×10^{-17} across $N \in \{6, 15, 25\}$. The governance layer changes the signal without disturbing the structural theorems – the stress test with full weight dispersion that §5.5's real data (no weight spread) could not run.

6.7 Mechanism comparison and conditional output (Prop. 7, 11)

Four rules on the same $N = 5$ team with contributions $a = (0.8, 1.2, 1.0, 0.6, 1.4)$, $E = 100$ (Fig. 7):

Relatively, the fixed denominator is proportional scoring scaled by $1 - G$ (here 0.4): the difference is not in *who gets what* but in the **retained fund** – the balanced rules distribute all of E ; the fixed denominator

Rule	Shares (m1..m5)	Members' total	Retained fund
Equal	20, 20, 20, 20, 20	100.0	0%
Proportional	16, 24, 20, 12, 28	100.0	0%
Fixed-denominator	6.4, 9.6, 8.0, 4.8, 11.2	40.0	60%

automatically retains $E \cdot G = 60$, the budget-breaking residual that Holmström (1982) assigns to an outside budget breaker, here produced by voting intensity itself. Shapley-type allocation is excluded on feasibility (P7); under additive production at $N = 5$ it coincides with proportional sharing.

Conditional output comparison (Prop. 11). Solving the fixed-denominator and equal-division FOCs for a symmetric team ($A = 1$, $\beta = 0.5$, $c = 1$): for $N = 10$, $G = 13.5\%$ the threshold is $\gamma^* = 0.009$; at the empirically estimated $\gamma \approx 0.4$ (§5.2) the gain is 92% ($N = 10$) and 141% ($N = 20$). Larger funds demand stronger signals ($G = 50\%$: threshold 0.056), the §5.7 trade-off. Closed-form thresholds match numerics to machine precision (Online Appendix E).

6.8 What simulation establishes, and its boundary

- **Established.** Every §4 theorem (M1 scale identity, M3 retention identity, the §6.5 bounds, and the §6.6 governance invariance) holds to machine or near-machine precision under the theory's own vote model across the team regime the data cannot reach. The mechanism is a verified computational object, not algebraic claims awaiting a favorable dataset.

- **Boundary.** Simulation verifies the arithmetic under the model's assumptions; whether real raters vote as assumed is tested weakly and honestly in §5 (H1) and behaviorally in the companion experiment. §5.7 adds the design content on top.

7 Discussion

7.1 Relative allocation is not the contribution

On relative shares, the fixed-denominator rule is proportional allocation scaled by a common factor: $s_i^{\text{FD}} = \tau s_i^N$ with $\tau = 1 - G$ (Lemma 1; §6.7 numerically). Normalized to sum one, every member's share coincides with his proportional share. This equivalence is deliberate, not incidental: the mechanism does not seek to improve the relative fairness of the weights; it changes *how the aggregate distributable pool is determined*.

What the rule contributes is the scale itself. In proportional sharing with an exogenous retained fraction L , the scale is a parameter set before anyone acts. Here the scale is $1 - G$, an outcome of the vote profile – a

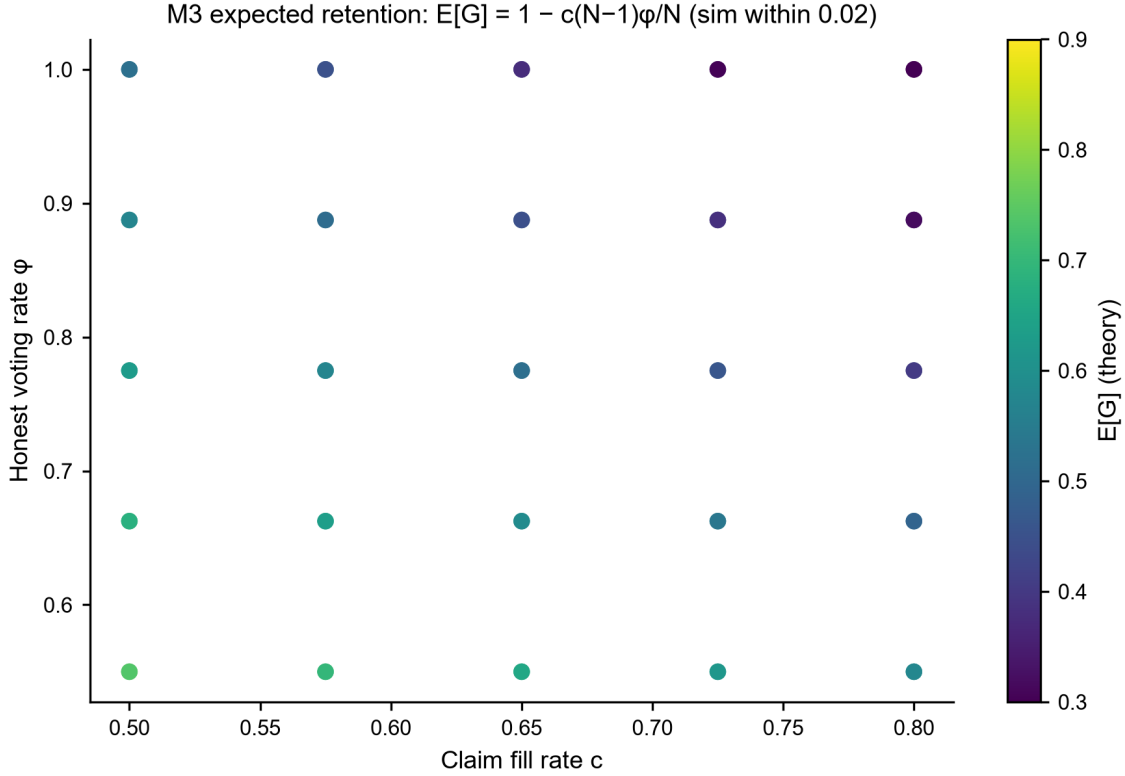


Figure 5: M3 surface, theory vs Monte-Carlo within 0.02

non-constant scale that moves with voting intensity. Two properties follow that an exogenous L cannot have.

1. Budget balance is broken endogenously. Holmström’s remedy is an outsider absorbing a fixed residual; here the residual is EG , priced by the team’s own shortfall from full capacity – the sparser the mutual voting, the thicker the retained fund. No parameter sets G ; voting intensity does. The residual claimant is thus *behavior-linked* rather than implemented through an exogenously specified sharing coefficient.

2. The scale is schedulable, not arbitrary. The fixed-denominator scale has a closed form (R1), $\mathbb{E}[G] = 1 - c\frac{N-1}{N}\phi$, and it is used predictively: §5.6 to machine precision where coverage is uniform, §5.3 tracks the scheme’s $\hat{\phi}$, §5.7 scheme switches move G exactly as predicted (identity error ≈ 0). An exogenous L has no content to verify; this scale is a tested design law. The substitution we claim is institutional rather than behavioral: whether the rule improves on budget-balanced production in equilibrium is a behavioral question, delegated to the governance layer (§4.8–4.10) and not measured here.

The retained fund is not a tax. The deduction is coupled to the vote matrix (the denominator T), not a fixed rate: the same team retains different amounts purely through its own voting. In the full-suite regime the floor $G \geq 1/N$ holds for every feasible profile (Prop. 1): a component of the fund is arithmetic and cannot be negotiated away by any voting behavior. The fund’s ownership is a design choice (Prop. 5; Online Appendix

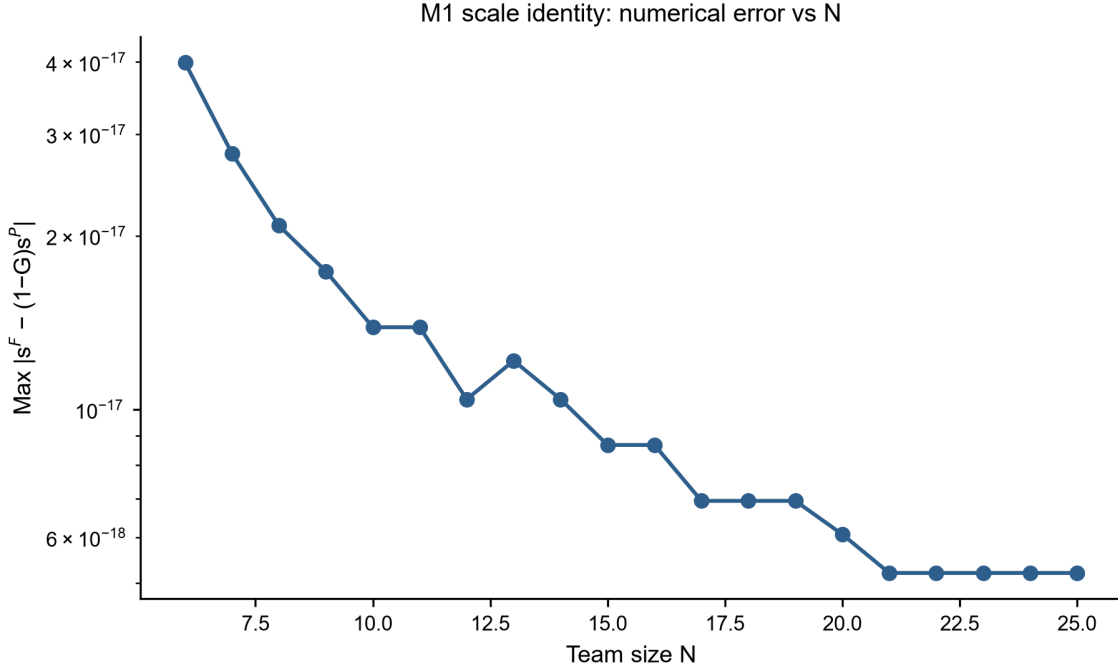


Figure 6: M1 numerical error vs N

F): with no member claim the mechanism is immune to self-interested manipulation, while a proportional claim opens a depression channel. The institution deciding the fund's use is upstream; the mechanism's contribution is that the fund's *size* is not a free parameter – a budget breaker in the Holmström tradition, not a tax.

7.2 Evidence beyond algebra

The §4 identities are algebra, as any rule's would be; what distinguishes this mechanism is that its distinctive content – the endogenous scale – is *tested* rather than asserted. The retention identity is used predictively across grading schemes (to machine precision where coverage is uniform), explains the cross-scheme spread of G (§5.3: $\approx 13\text{--}70\%$ as the ϕ -by-scheme law of Table 3), and predicts retention under a scheme a cohort never observed (§5.7: rating $\rightarrow$ coin moves G from 13.6% to 49.1%) at a quantified accuracy cost. The empirical contribution lies not in verifying the algebraic identity itself – an identity holds by construction – but in documenting how real grading schemes generate different voting intensities and, through the mechanism, different retention levels, and in showing that those levels are predictable and schedulable.

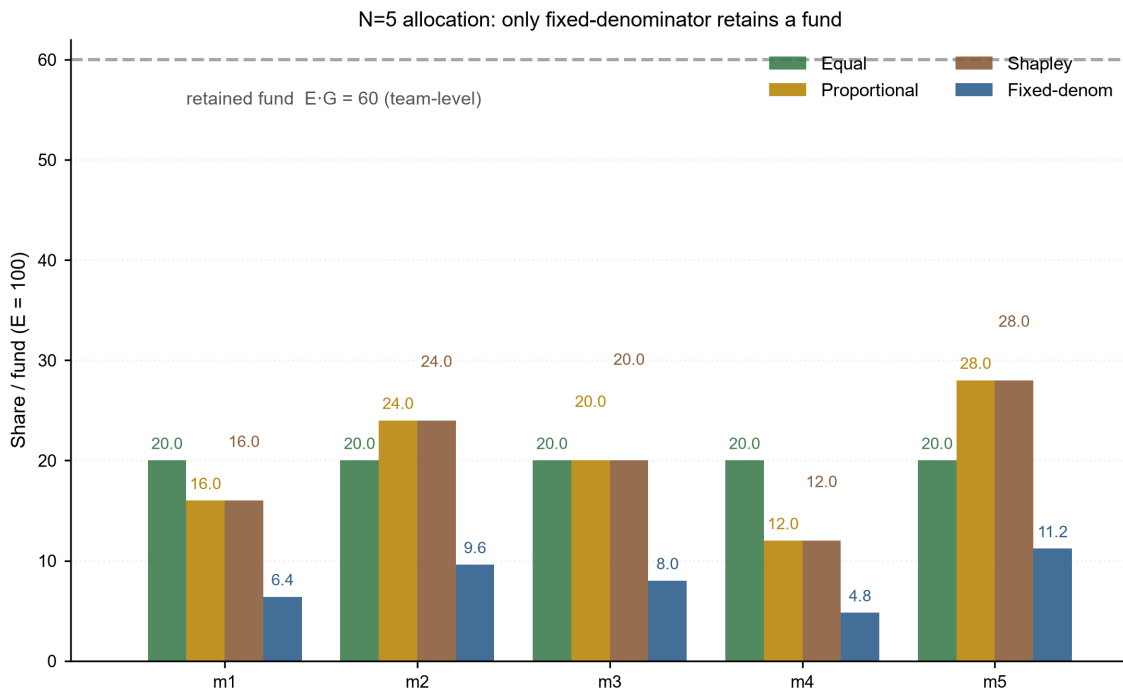


Figure 7: four-rule allocation, retained fund hatched .

7.3 Limitations, stated plainly

- **Signal is moderate, not strong.** $\rho \approx 0.44\text{--}0.60$ (§5.2). The mechanism needs $\rho > 0$, not $\rho \approx 1$, but the magnitudes bound how much weight a single vote can carry; governance efficacy is not measured here.

- **Honest voting is assumed in the theory, measured weakly in the data.** Correlation with a teacher grade is consistent with honest reporting but does not prove it; the companion experiment measures effort, honesty, and weighting directly.

- **The data are single-period, sparse-coverage course records.** Each observation is one cohort, one assignment ($r = 3$; cohorts exceed the team regime). General properties are validated (§5); the team regime is covered by simulation (§6) and counterfactual design by §5.7.

- **Two honest data gaps.** t4e's coin-budget rule is not public, so its G is approximate (§5.3) and its weak anchor pushes its points below the §5.7 frontier. We report both rather than rescale them.

- **Strategic collusion is not modeled.** Prop. 2 rules out *unilateral* self-promotion but not reciprocal vote-trading, which, if costless, would push G toward its floor $1/N$ in the full-suite regime. Whether reciprocity is an equilibrium is a behavioral question (voting-cost and peer-prediction layers, §4.9–4.10; companion experiment). Even a full reciprocity equilibrium leaves the floor $G = 1/N$ in that regime (Prop. 1); whether that residual suffices to improve incentives depends on N and the production function, a parameter-dependent

question outside this paper’s scope.

- **Homogeneity; single-period signal estimate.** Prop. 11 assumes a symmetric team; heterogeneous members require joint analysis of individual signal responses. The signal response γ is estimated directly from the data (§5.2, class-level mean 0.41, range 0.05–0.78) but from single-period course records, and it inherits the anchor noise of $\rho_{\text{pool}} = 0.439$; the discrete scheme set $\hat{\phi} \in \{0.87, 0.50, 0.30\}$ bounds its reachable range from below.

7.4 Design implications and the companion experiment

The practical claim is a design table: a planner targeting G^* picks the scheme whose ϕ produces it (Table 3), knows the accuracy it buys or costs (§5.7), and can verify realized G against the identity. The companion study tests what these data cannot: whether real raters discriminate less under generous scales (§5.7 ceiling effect as a behavioral prediction), whether KPI weighting reorders shares when weights have spread (Prop. 8’s behavioral content), and whether the retained fund is perceived as legitimate rather than as a tax (Prop. 5’s ownership question). Two opposing forces pin the fund at an interior level: from Prop. 11 a larger G demands a stronger signal response γ to keep its output edge, while from §5.7 a larger G (a less generous scale) costs ranking accuracy; the planner optimum $G^* \approx 14\%$ of §5.7 sits on that frontier, the same design surface read from the incentive side and the information side. The structural claims are settled by the data and simulations; the experiment targets what structure cannot settle.

8 Conclusion

A fixed-denominator peer-allocation rule resolves the two tensions of team compensation at once: it breaks the balanced-budget constraint that generates free riding, with the retained fund EG an outcome of the team’s own voting intensity rather than an outsider’s parameter (P1), and it makes manipulation of one’s own share structurally impossible (P2).

The mechanism’s distinctive content is not a different ranking (relative shares coincide with proportional sharing by construction) but an endogenous, *schedulable* retention: $\mathbb{E}[G] = 1 - c \frac{N-1}{N} \phi$ turns the retained fund into a closed-form function of observable voting engagement (P6, R1). Three datasets, four grading schemes, and a data-driven counterfactual verify this is a design tool, not algebra: retention tracks the scheme’s $\hat{\phi}$ ($\approx 13\%/50\%/70\%$), holds to machine precision where coverage is uniform, and moves exactly as predicted when a cohort’s scheme is switched, at a quantified, structural price in ranking accuracy. A conditional output comparison (P11) shows the rule’s equilibrium output dominates equal division under a weak, data-satisfied

condition.

The governance layer (KPI-weighting, voting cost, peer-prediction reward) disciplines evaluation without disturbing any structural theorem (P8–P10); whether it succeeds behaviorally – whether real raters discriminate less under generous scales, whether weighting reorders shares when weights have spread, whether the retained fund is perceived as legitimate rather than as a tax – cannot be settled by observational data or simulation and is the task of the companion experimental study. What the data and simulations establish is narrower and firmer: the structural claims hold, retention is a tested design law, and the price of low retention is measurable. The remaining open question is behavioral, not structural.

References

- Ashenafi, M. M. (2019). Online peer-assessment datasets. arXiv:1912.13050.
- Carpenter, J., Robbett, A., and Akbar, P. A. (2018). Profit sharing and peer reporting. *Management Science*, 64(9):4261–4276.
- Carvalho, A. and Larson, K. (2012). Sharing rewards among strangers based on peer evaluations. *Decision Analysis*, 9(3):253–273.
- Dai, T. and Toikka, J. (2022). Robust incentives for teams. *Econometrica*, 90(4):1583–1613.
- Dong, L., Falvey, R., and Luckraz, S. (2019). Fair share and social efficiency: A mechanism in which peers decide on the payoff division. *Games and Economic Behavior*, 115:209–224.
- Falchikov, N. and Goldfinch, J. (2000). Student peer assessment in higher education: A meta-analysis comparing peer and teacher marks. *Review of Educational Research*, 70(3):287–322.
- Holmström, B. (1982). Moral hazard in teams. *The Bell Journal of Economics*, 13(2):324–340.
- Jones-Todd, C. M. and Renelle, A. (2025). A comparison of peer- and tutor-grading of an introductory r coding assessment. *Journal of Statistics and Data Science Education*, pages 1–19. Advance online publication.
- Ligon, E. (2026). No more grading: Incentive compatible peer assessment in a project-based course. Cudare working papers, University of California, Berkeley.
- Miller, N., Resnick, P., and Zeckhauser, R. (2005). Eliciting informative feedback: The peer-prediction method. *Management Science*, 51(9):1359–1373.

- Rasmusen, E. (1987). Moral hazard in risk-averse teams. *The RAND Journal of Economics*, 18(3):428–435.
- Xu, J., Yang, P., Xiao, T., Lv, P., Yu, M., and Yu, G. (2023). Peer assessment dataset: Companion data for sc-pa, a spot-checking model based on stackelberg game theory for improving peer assessment. GitHub repository. Available at <https://github.com/CHDYPY/peerAssessmentDataSet>.
- Xu, J., Yang, P., Xiao, T., Lv, P., Yu, M., and Yu, G. (2025). Sc-pa: A spot-checking model based on stackelberg game theory for improving peer assessment. *Computers in Human Behavior Reports*, 17:100556.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [onlineappendix.pdf](#)