# CAN STRATEGIC IDEOLOGUES AGREE ON QUALITY? INCENTIVES ON LATENT FACTOR MODELS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Community Notes ranks fact-checking notes based on ratings from ideologically diverse raters. It fits a low-dimensional model that seeks, for each note, a quality shared by all raters, separate from each rater's ideological agreement with the note.

We analyze this, and related models, from a social choice perspective and discover key structural insights of these mechanisms. We leverage this structure to study both the outcomes compared to different objectives as well as incentive properties. Highlights of our results include: 1) privileging the perspective of the average rater maximizes social welfare, but is subject to simple manipulation strategies; 2) privileging the perspective of the median rater satisfies the Condorcet criterion, and is the only identifiable viewpoint robust to the same manipulations, even by coalitions; 3) more ideologically extreme agents have a larger incentive to participate.

## 1 INTRODUCTION

Machine learning and artificial intelligence are increasingly utilized to aggregate human preferences and opinions, making collective decisions in content moderation (Roberts, 2019), hiring (Tambe et al., 2019), peer review (Noothigattu et al., 2021), and the training of reward models for language models (Siththaranjan et al., 2024; Conitzer et al., 2024). Beyond traditional metrics of prediction performance, we must consider how these systems encode normative trade-offs and mediate between competing values and objectives.

A prominent example, and the motivating object of our study, is X's Community Notes, which decides which fact-checking notes millions of people see beneath viral posts by aggregating the ratings of volunteer reviewers (Wojcik et al., 2022; Allen et al., 2022). The system is among the most visible instances of algorithmic governance at scale. It has been influential enough that other companies have followed suit; Meta has replaced professional fact-checkers in the US with a system built on the same algorithm (Meta, 2025). The algorithm, at its core, can be read as a voting mechanism: each rating is an implicit vote cast by an agent with its own beliefs and motivations. It follows that the algorithm is the rule that decides how those votes are counted.

Regarding strategic behavior on this system, coordinated manipulation has been documented in the wild (Mohammadi et al., 2025) and demonstrated in audits, in which a small fraction of coordinated raters suppress targeted notes (Truong et al., 2025). The algorithm has been modified to detect exactly this coordination (Community Notes, 2026), motivating a more systematic study.

Several aspects differentiate the Community Notes mechanism from classical social choice: many decisions are made simultaneously, a parametric structure is imposed on the agents' preferences via latent representations, the parameters are learned from data, and each user rates only a fraction of the items (Appendix F.6). We provide a principled analysis of such mechanisms, bridging statistical learning and social choice. In particular, we address:

**Social objectives:** Which social objective does the ranking serve, and can it serve the one the platform states, quality?

**Strategic incentives:** Can agents gain influence over the outcome by misreporting their ratings, and what incentive does an agent have to rate at all?



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

Answering either requires a model of what a rating means. While many of our results apply to more general classes of models (Appendix D), here we model each rating as a utility $r_{ij} = q_j + u_i\, v_j$ derived from the note's quality $q_j$ plus the agreement between the rater's *type* $u_i$ and the note's *leaning* $v_j$ (each representing ideology), the form Community Notes fits to its data (up to intercepts). Two properties drive the results. *Strict crossing*: the agent types form a line, and any two notes of differing leaning split it at a single type, with raters on one side preferring one note and those on the other side the other. This is the structure under which median-voter and strategyproofness results hold in social choice (Gans & Smart, 1996; Saporiti, 2009). *Anchoring*: the quality in the model is the evaluation of the notes from the viewpoint of a privileged type, typically called "neutral".

**Social objectives.** In this setting, we study a large class of mechanisms that fit the model's parameters by minimizing error. In particular, when there exists an exact fit to the data, which we call the zero-risk case, a *zero-risk mechanism* is required to recover parameters that exactly fit the data (Definition 5). In such a case, we show that because the model is not identified from the data, the reports themselves (which only contain utilities) do not determine quality. Instead, the neutral position, which we call the *anchor*, must be decided: every zero-risk mechanism ranks the notes from the viewpoint of some type (Theorem 2). We focus on three possible anchors: the neutral $(0)$ type, the mean type, and the median type. We also look at three types of desiderata for social choice: recovering the true underlying quality, maximizing social welfare (the sum of utilities), and the Condorcet criterion, under which the ranking agrees with the pairwise majorities of the raters; each is attained from exactly one of the three anchors (Theorem 1). We establish that no mechanism can attain objective quality on every instance because the neutral anchor, which would achieve it, is not identifiable from the ratings (Proposition 1). In contrast, the mean anchor, which maximizes social welfare, and the median anchor, which satisfies the Condorcet criterion, are both implementable: any fitted model can be re-anchored at either (Theorem 4).

Previous work showed that the regularization of biased matrix factorization, the estimator at the core of Community Notes, places the anchor at a shrunk fraction of the mean ideology (Anonymous, 2026).

**Misreporting by spoofing.** We model the only misreport a zero-risk fit cannot detect, *ideological spoofing*, in which an agent rates as if it held an ideology other than its own (Lemma 8). Any other misreport leaves the ratings with no exact fit, so with exact ratings the platform can detect it and, if fewer than $(n-1)/2$ agents misreport, identify who did (Lemma 9). Community Notes already screens raters by their record and with reputation and diligence models (Community Notes, 2026), and related screens have been proposed or deployed elsewhere: truth discovery (Li et al., 2016b), influence limiters (Resnick & Sami, 2007), and consistency weighting (Goyal et al., 2026; Alimohammadi et al., 2026). These methods are orthogonal and complementary to our analysis.

In contrast, a spoofed report passes every such test, since it is the honest report of some other type. Strategically, the mean and median anchors are opposites (Section 4.1). A single spoofing agent can place the mean anchor anywhere. The median anchor is (uniquely) strategyproof: an agent can move the median by reporting across it, but never *beneficially*. The guarantee extends to coalitions, so no set of spoofers can all profit (Theorem 5), and among identifiable anchors the median is the only anonymous, efficient, strategyproof one (Theorem 6). Simulations on the deployed estimator show the contrast: constrained to the mean, it loses every objective as a coalition extremizes its reports; constrained to the median, it is unaffected (Fig. 3).

**Participation.** Incentives to rate at all are also important. We show agents farther from the anchor gain at least as much from rating as moderates on the same side (Proposition 3). With a cost of rating, when the median is the anchor, agents far from the anchor always rate, so only a band of moderates can abstain (Proposition 13); when the mean is the anchor, few agents rate at all (Proposition 14). Community Notes data support this: the fifth of raters farthest from the median rate about $1.5$ times as much as the fifth nearest to it (Appendix I.6).

**Related work.** Our object of study, a mechanism defined by a model fit to reports from agents who care about its output, is shared with strategyproof estimation (Dekel et al., 2010; Chen et al., 2018), strategic classification (Hardt et al., 2016), and performative prediction (Perdomo et al., 2020). The strategyproofness results sit in the social choice tradition of restricted domains: the Gibbard–

Satterthwaite theorem rules out non-dictatorial strategyproof rules on unrestricted preferences (Gibbard, 1973), Moulin characterized the strategyproof rules on single-peaked domains as generalized medians (Moulin, 1980), and on single-crossing domains the median-voter and strategyproofness results we build on are due to Gans & Smart (1996) and Saporiti (2009). Our turnout analysis follows the costly-voting tradition of rational abstention (Ledyard, 1984). Community Notes' design belongs to bridging-based ranking (Ovadya & Thorburn, 2023), and spoofing parallels shilling (Mehta & Hofmann, 2008) and data poisoning of recommender systems (Li et al., 2016a).

## 2 Utilities, Anchors and Mechanisms

Community Notes fits biased matrix factorization to its rating matrix and ranks notes by the fitted quality (Wojcik et al., 2022; Anonymous, 2026). The fit decomposes each rating into the note's quality and the alignment between the rater's ideology and the note's leaning. Our starting point is to read this decomposition as a model of utility: the rater cares which note the platform displays, and a rating expresses how much the rater values the note. Consider $n$ agents and $m$ items, the notes proposed for one post. The results hold for a more general class of utility models, defined in Appendix D. We work with the linear form, the simplest that satisfies the properties the proofs use.

**Definition 1** (Latent factor utility model, LFUM)**.** Each agent $i$ has a *type* $u_i \in \mathbf{R}$, its position on the ideological line; each item $j$ has a quality $q_j \in \mathbf{R}$ and a leaning $v_j \in \mathbf{R}$ on the ideological line. The utility of agent $i$ for item $j$ is

$$U(u_i, j) \;=\; q_j + u_i\, v_j. \tag{1}$$

An *instance* is $E = (u, q, v)$ with $u \in \mathbf{R}^n$ and $q, v \in \mathbf{R}^m$ (recall $n$ and $m$ are the number of agents and items, respectively).

A type profile $u$ is *nondegenerate* if not all types are equal. Throughout, $u$ is nondegenerate and $v$ is nonzero, so two agents rate some item differently. Two instances with the same numbers of agents and items are *utility-equivalent* if every agent has the same utility for every item in both, $U(u_i, j) = U(u'_i, j)$ for all $i$ and $j$.

Two properties of the model are immediate. Every agent shares the quality term, a common value of the item. And an agent of type 0 values item $j$ at exactly $q_j$, so quality is the evaluation of the items from the viewpoint of one privileged type, the one declared neutral. More interesting is that this type can be changed: for any $\gamma \in \mathbf{R}$ the transformation

$$E_\gamma \;=\; (u - \gamma,\; q + \gamma v,\; v) \tag{2}$$

leaves every utility the same, since $(q_j + \gamma v_j) + (u_i - \gamma)v_j = q_j + u_i v_j$, while the neutral type is now $\gamma$: the qualities of $E_\gamma$ are the utilities of an agent at $\gamma$ in $E$. One set of utilities is therefore represented by a whole family of instances, one for each choice of neutral type, the gauge freedom of the companion paper (Anonymous, 2026). We call the type declared neutral the *anchor*.

**Running example.** Five raters value three notes $A$, $B$, $C$; their utilities are listed in Appendix C. These utilities have the form (1) with types $u = (-0.5,\ 0,\ 0.5,\ 1.5,\ 2.5)$ and $(q_j, v_j)$ equal to $(1.2, -1)$, $(0.6, 1)$ and $(0.9, 0.5)$ (Fig. 1). Note A has the highest quality and leans left. Notes B and C lean right. The pool skews right, with mean $\bar{u} = 0.8$ and median $\tilde{u} = 0.5$. Details in Appendix C.

**Mechanisms and objectives.** We now turn to how the ratings are aggregated into a single ranking of the items, which we call a social outcome. We view this aggregation as a mechanism. Each agent $i$ submits a *report* $r_i \in \mathbf{R}^m$, one number per item; the *report matrix* for all agents and items is thus $R = (r_{ij})$. A report is *truthful* when it is the agent's utility row, $r_{ij} = U(u_i, j)$ for every $j$. Until Section 4.1 reports are truthful and complete, every agent rating every item, so the instance determines the report matrix, written $R(E)$.

**Definition 2** (Mechanism)**.** A *mechanism* is a function $f$ from a report matrix $R \in \mathbf{R}^{n \times m}$ to a social outcome: a ranking of the $m$ items or, for a fixed selection size $k \geq 1$, a set of $k$ of them.

The mechanism has access only to the report matrix and nothing else. Ranking by quality is Community Notes' stated aim, a ground truth that exists independently of the raters, as in epistemic

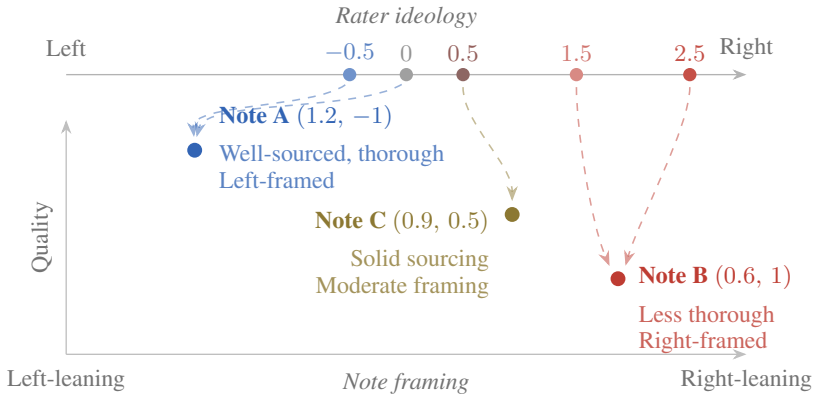



Figure 1: Running example: which note should the platform display? Five raters on the ideological line (top) and three candidate fact-check notes, placed by leaning $v_j$ against quality $q_j$ (bottom); parameters in Table 1.

social choice (List & Goodin, 2001). Nonetheless, recovering quality rankings need not be the only objective of a mechanism, and we define two others:

**Definition 3** (Social objectives)**.** Fix an instance. A ranking of the items *attains*:

**Objective Quality (OQ)** if it orders items by quality, $j$ above $j'$ whenever $q_j > q_{j'}$.
**Social Welfare (SW)** if it orders items by total utility, $j$ above $j'$ whenever $\sum_i U(u_i, j) > \sum_i U(u_i, j')$.
**Condorcet Criterion (CC)** if it orders items by majority, $j$ above $j'$ whenever $j$ beats $j'$, where $j$ beats $j'$ when $\left|\{i : U(u_i, j) \geq U(u_i, j')\}\right| > n/2$.

A set of $k$ items attains an objective if it is the top $k$ of a ranking that does, and a mechanism attains an objective on an instance if its output on $R(E)$ does.

The three can disagree. For instance, in the running example, by quality the notes rank $A \succ C \succ B$. By total utility they rank $B \succ C \succ A$, most of $B$'s coming from the right-leaning majority. By majority they rank $C \succ B \succ A$, Note C winning every pairwise comparison.

We study mechanisms whose output is the ranking from the perspective of one type. For a type $\gamma \in \mathbf{R}$, let $\pi_\gamma(E)$ be the ranking of the items by the utility $U(\gamma, j)$ of that type. Note that the ranking $\pi_\gamma$ may not be a realizable mechanism since it has access to the types and utilities of the instance, whereas a 'regular' mechanism sees only reports.

**Definition 4** (Type-anchored mechanism)**.** Let $\gamma \in \mathbf{R}$ be a type. A mechanism $f$ is *anchored at* $\gamma$ on an instance $E$ if its outcome on $R(E)$ agrees with $\pi_\gamma(E)$, as a ranking or as its top $k$, ties in utility broken by a fixed priority order on items. For a statistic $a : \mathbf{R}^n \to \mathbf{R}$ of the types, $f$ is *$a$-anchored* if it is anchored at $a(u)$ on every instance.

For instance, in the running example, a *median-anchored* mechanism, $a(u) = \tilde{u}$, outputs $\pi_{0.5}(E) : C \succ B \succ A$ the ranking of the median rater at 0.5; a *mean-anchored* one, $a(u) = \bar{u}$, outputs $\pi_{0.8}(E) : B \succ C \succ A$, a type no individual rater occupies; and a *neutral-anchored* mechanism, the constant statistic $a \equiv 0$, outputs $\pi_0(E) = A \succ C \succ B$. We write $S(\gamma)$ for the top $k$ of $\pi_\gamma(E)$, the selection of a mechanism anchored at $\gamma$, and an agent of type $u_i$ values a selection $S$ at $\sum_{j \in S} U(u_i, j)$.

Each objective of Definition 3 is attained by anchoring at a type read off the instance, the neutral type, the mean or the median, so the choice of anchor is a choice of objective.

**Theorem 1** (Anchor–objective correspondence)**.** *Let $E = (u, q, v)$ be an instance in which no two items tie in quality, in total utility, or beat each other, and let $f$ be a mechanism applied to $R(E)$. Then:*

*(i) $f$ attains OQ on $E$ if and only if it is anchored at $0$ on $E$;*
*(ii) $f$ attains SW on $E$ if and only if it is anchored at $\bar{u}$ on $E$;*

*(iii) for odd $n$, $f$ attains CC on $E$ if and only if it is anchored at $\tilde{u}$ on $E$.*

*Proof sketch.* Quality is $U(0, \cdot)$ and total utility is $n\,U(\bar{u}, \cdot)$. For a pair $j, j'$ the difference $U(t, j) - U(t, j')$ is affine in $t$, so the agents preferring $j$ lie on one side of those preferring $j'$, and $j$ beats $j'$ exactly when the median agent prefers $j$. Without ties each attaining ranking is unique (proof in Appendix E). □

## 3 Learning Anchors

At its core, Community Notes reconstructs the parameters of (1) from the rating matrix and ranks by the reconstructed quality. Every mechanism of this kind is type-anchored.

**Definition 5** (Zero-risk models as mechanisms)**.** Given a rating matrix $R \in \mathbf{R}^{n \times m}$, a *model* of the family (1) is a parameter set $\theta = (\hat{u}, \hat{q}, \hat{v}) \in \mathbf{R}^n \times \mathbf{R}^m \times \mathbf{R}^m$ with predicted ratings $\hat{q}_j + \hat{u}_i \hat{v}_j$, and its *risk* is the squared error $\sum_{i,j} (r_{ij} - \hat{q}_j - \hat{u}_i \hat{v}_j)^2$. The *zero-risk models* $\Theta(R)$ are those that reproduce every rating. A mechanism is *zero-risk* if, for every instance $E$, its output on $R(E)$ is the ranking by $\hat{q}$ of some zero-risk model in $\Theta(R(E))$, or the top $k$ of that ranking, ties broken by the fixed priority order of Definition 4.

Biased matrix factorization, the estimator Community Notes uses, fits a model by minimizing the risk plus an $L_2$ penalty on the parameters (Anonymous, 2026); the deployed version adds rater and global intercepts. It is a zero-risk mechanism (as the regularization weight tends to zero): it ranks by the reconstructed quality of a model that reproduces the ratings, and its regularization, a penalty added to the risk, selects which of the zero-risk models it is. The results of this section extend to noisy ratings, where minimum risk replaces zero risk and the anchor is a type of the fitted instance (Appendix F.5); here we present the zero-risk case for simplicity and clarity. They also hold on partial ratings under the condition of Appendix F.6.

**Lemma 1** (Zero-risk models, Anonymous, 2026)**.** *Let $E = (u, q, v)$ be an instance. Then*

$$\Theta(R(E)) \;=\; \big\{\, (c\,(u - \gamma),\; q + \gamma v,\; v/c) : \gamma, c \in \mathbf{R},\; c \neq 0 \,\big\},$$

*and the reconstructed quality of the model with parameters $(\gamma, c)$ is the utility of type $\gamma$, $U(\gamma, \cdot)$. Equivalently, the zero-risk models are the instances utility-equivalent to $E$: the shifts $E_\gamma$ of* (2) *combined with the scalings $(cu,\ q,\ v/c)$, which also preserve every utility.*

**Theorem 2** (Zero-risk mechanisms are anchored)**.** *A mechanism $f$ is zero-risk if and only if for every instance $E$ there is a type $\gamma \in \mathbf{R}$ such that $f$ is anchored at $\gamma$ on $E$.*

By Lemma 1 the zero-risk models of a rating matrix form one family, indexed by $\gamma$, so a zero-risk mechanism has to choose an anchor. In biased matrix factorization the regularization makes that choice, placing the anchor near the mean type of the raters (Anonymous, 2026). Not every mechanism is of this kind. Plurality, ranking notes by how many raters place them first, puts Note C last in the running example, which no type does.

We focus on three anchors, the neutral type, the mean and the median, since by Theorem 1 they attain the three objectives. The question is which of them a mechanism can be anchored at, given that it sees only the ratings. In the running example the model at $\gamma = 1$ has qualities $q + v = (0.2,\ 1.6,\ 1.4)$, so its neutral type ranks $B \succ C \succ A$ where the truth ranks $A \succ C \succ B$, while the median rater ranks $C \succ B \succ A$ in both.

**Definition 6** (Identifiable anchor)**.** A statistic $a$ of the types is *identifiable* if the oracle ranking from it is the same on all utility-equivalent instances: $\pi_{a(u)}(E) = \pi_{a(u')}(E')$ whenever $E = (u, q, v)$ and $E' = (u', q', v')$ have the same utilities.

**Theorem 3** (Identifiable anchor means implementable mechanism)**.** *A statistic $a$ is identifiable if and only if some mechanism is $a$-anchored on every instance. The mechanism ranks any zero-risk model of $R(E)$ from $a$, that is, by $\hat{q}_j + a(\hat{u})\,\hat{v}_j$.*

By Lemma 1 the instances utility-equivalent to $E$ are the family $E_{\gamma,c} = (c(u - \gamma),\ q + \gamma v,\ v/c)$, so identifiability is constancy of $\pi_{a(u)}$ along that family. Identifiability is a property of the ranking from $a$: the value $a(u)$ itself changes along the family.

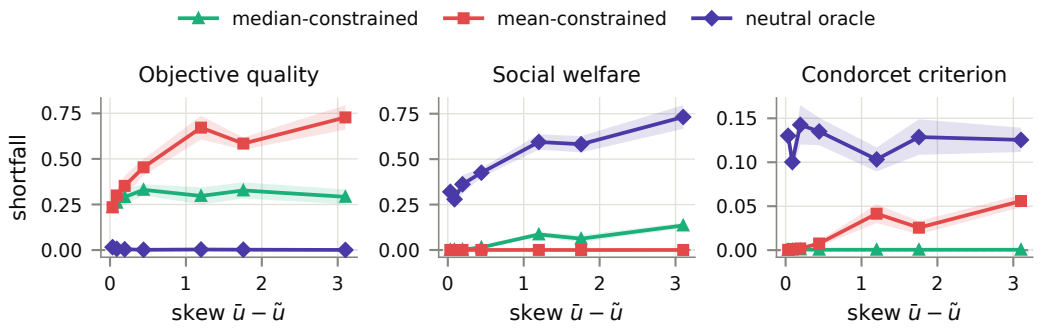



Figure 2: Shortfall of the top-10 selection from each objective (zero means attained) as the population skews with its median fixed, for BMF on noisy ratings with its anchor constrained to the mean or the median. Each anchor attains its own objective, and quality only the oracle, since the neutral type is not identifiable (Proposition 1).

**Proposition 1** (Which anchors are identifiable)**.** *(a) The neutral anchor, the constant statistic $a \equiv 0$, is not identifiable on instances in which two items differ in leaning.*

*(b) A statistic $a$ of the types is identifiable if and only if $a(c\,u + b) = c\,a(u) + b$ for all nondegenerate $u \in \mathbf{R}^n$, $b \in \mathbf{R}$ and $c \neq 0$. The mean and, for odd $n$, the median are identifiable. Fixed types and order statistics other than the median are not.*

Intuitively, the ratings establish relative comparisons between agents, but not absolute types. An absolute type would require a point of reference, and Lemma 1 shows that the point of reference can be chosen arbitrarily.

**Corollary 1** (Objectives, anchors, and identifiability)**.** *Under the hypotheses of Theorem 1, by Proposition 1 and the proof of its part (a), and by Theorem 3:*

*(i) OQ is attained only from the neutral anchor, which is not identifiable, so no mechanism attains OQ on every instance;*

*(ii) SW is attained only from the mean anchor, which is identifiable, so a mechanism attaining SW exists;*

*(iii) for odd $n$, CC is attained only from the median anchor, which is identifiable, so a mechanism attaining CC exists.*

Finally, we give the mechanism of Theorem 3 explicitly for the mean and the median.

**Theorem 4** (Re-anchoring)**.** *Let $E$ be an instance, let $(\hat{u}, \hat{q}, \hat{v}) \in \Theta(R(E))$ be any zero-risk model, and let $a$ be the mean or, for odd $n$, the median. Then $\hat{q}_j + a(\hat{u})\,\hat{v}_j = U(a(u), j)$ for every item $j$. Thus, ranking by this* re-anchored *quality attains SW, respectively CC, under the hypotheses of Theorem 1.*

For biased matrix factorization, re-anchoring amounts to a constraint on the reconstructed types: requiring their mean, or their median, to be zero places the anchor at the mean or the median of the types, since $\hat{u} = c(u - \gamma)$. The constraint can be imposed during the fit or after it (Appendix I). Figure 2 shows the constrained fits on noisy ratings and a skewed population: constrained to the mean, BMF attains SW; constrained to the median, it attains CC.

## 4 Strategic Incentives

We now turn to strategic incentives under anchored mechanisms. We examine two cases: an agent can change what it reports (Section 4.1), and whether to report at all (Section 4.2).

### 4.1 Misreporting by Spoofing

In spoofing, an agent reports as if it were of a different type. For instance, an agent on the left can rate as if it stood farther left, weighting its alignment with each note more heavily.

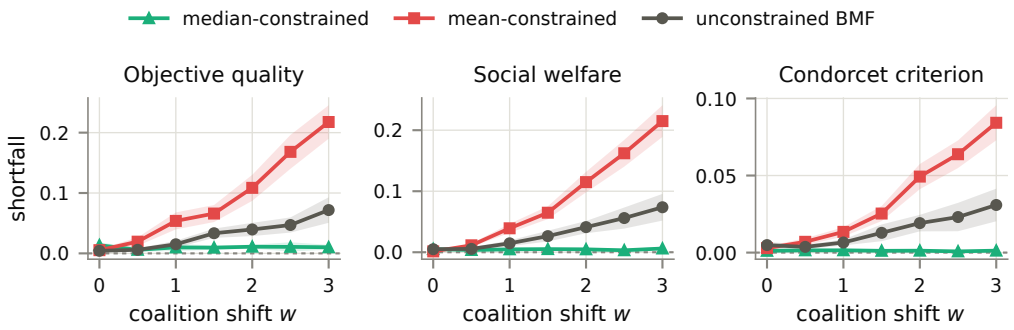



Figure 3: Shortfall of the top-10 selection from each objective (zero means attained) as the top half of the agents with positive type report shifted outward by $w$, for biased matrix factorization constrained to the median, constrained to the mean, and unconstrained (Appendix I). The mean-constrained fit loses all three objectives as the coalition moves its anchor. The median-constrained fit holds all three. Plain BMF lies between.

**Definition 7** (Spoofing)**.** A *report profile* $w \in \mathbf{R}^n$ assigns each agent a reported type: agent $i$ reports the row $r_{ij} = q_j + w_i\, v_j$, truthfully when $w_i = u_i$, and *spoofs* when $w_i \neq u_i$. The *reported instance* is $E_w = (w, q, v)$. Report profiles are nondegenerate.

The mechanism sees only the report matrix, which under the profile $w$ is $R(E_w)$, the truthful matrix of the reported instance. An $a$-anchored mechanism is therefore anchored at $a(w)$ under $w$ and ranks the items by $U(a(w), \cdot)$. By Corollary 1 the mean anchor is the mechanism that attains social welfare and the median the one that attains the Condorcet criterion, so the two results below say which attainable objective survives spoofing. Welfare does not: one agent moves the mean anywhere. The Condorcet criterion does, even against coalitions.

**Proposition 2** (Mean-anchored mechanisms are manipulable)**.** *Let $f$ be mean-anchored and let the agents other than $i^*$ report truthfully. Then $i^*$ can move the anchor to any target $u^t \in \mathbf{R}$ by reporting as $w_{i^*} = n\,u^t - \sum_{i \neq i^*} u_i$. With target $u^t = u_{i^*}$ the selection is the top $k$ items under its own utility.*

The incentive extends to biased matrix factorization. Its anchor is a fraction $s \in (0, 1)$ of the mean of the reported types (Anonymous, 2026), so a spoof that shifts the reported mean shifts the anchor with it, and the spoofing agent benefits (Lemma 7).

**Theorem 5** (Median-anchored mechanisms are group strategyproof)**.** *Let $n$ be odd and let $f$ be median-anchored. Then:*

- *(a) truthful reporting is a weakly dominant strategy against spoofing: whatever the other agents report, no agent improves its selection by spoofing;*
- *(b) no coalition of agents can jointly spoof so that every member is strictly better off;*
- *(c) if no two items tie in utility for the truthful median type, no coalition can jointly spoof so that every member is weakly better off and at least one strictly.*

*Proof sketch.* A spoof moves the median only by reporting across it. Every change of selection this triggers is a swap of two items at a crossing on the far side of the deviating agent, and the item that enters is the one preferred beyond that crossing, which the deviator likes less than the item that leaves (Lemma 6). Moving the median forces some coalition member to report across it from the far side, so that member is harmed, strictly under the no-tie condition (proof in Appendix G). □

An agent can move the median, by reporting across it, but every such move pushes the median away from its own type. The theorem covers biased matrix factorization re-anchored at the median (Theorem 4), whose re-anchored quality under $w$ is $U(\tilde{w}, \cdot)$. Figure 3 shows the contrast with biased matrix factorization constrained to each anchor: as a coalition extremizes its reports, the mean-constrained fit loses quality, welfare and the Condorcet criterion, while the median-constrained fit holds all three.

Recall for agent type $\gamma \in \mathbf{R}$ the outcomes of a type-anchored mechanism are the rankings $\pi_\gamma$. By Lemma 7 in Appendix G each agent's utility over them is single-peaked in $\gamma$ with its peak at the agent's own type. On such a domain Moulin's median rules are strategyproof, and among the identifiable anchors only the median is:

**Theorem 6** (The median is the only identifiable strategyproof anchor)**.** *Let $n$ be odd and let $a$ be an identifiable statistic. If the $a$-anchored mechanism is strategyproof against spoofing, anonymous, and efficient*[1]*, then $a$ is the median.*

*Proof sketch.* By Lemma 7 the mechanism is a peaks-only voting rule on the line, so Moulin's characterization (Moulin, 1980) makes it a generalized median of the reports and $n-1$ fixed phantom positions. Identifiability requires $a(c\,w+b) = c\,a(w)+b$ (Proposition 1(b)), which a finite phantom fails under a shift of the reports and an order statistic other than the median fails under a reflection (proof in Appendix G). □

The guarantee is proved for exact ratings, and Fig. 3 shows it holding empirically on noisy ones.

### 4.2 Participation

Another strategic decision is whether to rate at all. We model this as a participation game, as in the calculus of costly voting (Ledyard, 1984): each agent decides whether to rate at a cost $c > 0$, everyone, participant or not, receives the utility of the selected set, and the anchored mechanism is computed over the agents who rate, whom we denote as the rating pool. By participating, agent $i$ enters the pool and pulls the anchor toward $u_i$. A move of the anchor toward $u_i$ weakly improves $i$'s selection (by Lemma 7), so participating never hurts. Write the benefit as

$$B_i \;=\; \sum_{j \in S(\gamma')} U(u_i, j) \;-\; \sum_{j \in S(\gamma)} U(u_i, j), \tag{3}$$

the utility of the selection after $i$ joins, at anchor $\gamma'$, minus the utility before, at anchor $\gamma$.

**Proposition 3** (Participation gains grow with distance)**.** *Let $f$ be mean- or median-anchored, let $Q$ be a nonempty pool with anchor $\gamma = \gamma(Q)$, the mean or median of $Q$, and let $a, b \notin Q$ be agents with $u_a > u_b > \gamma$. Then $B_a \geq B_b \geq 0$, where $B_i$ is* (3) *with $\gamma' = \gamma(Q \cup \{i\})$. The symmetric statement holds for $u_a < u_b < \gamma$.*

*Proof sketch.* Under the median, a joining agent moves the anchor at most to a neighboring order statistic, however extreme it is, so the size of the move does not favor extreme agents. Each selection change induced by the move is a swap that promotes an item preferred by the joiner (Lemma 6). An agent farther from the anchor values each such swap at least as much, since utility differences are affine in the type. Under the mean, the farther agent also moves the anchor farther, which adds swaps it values (see Appendix H for proof). □

Because the gain from rating grows with distance, a cost of rating filters out moderates: for any pool of raters, the agents outside it who would join at cost $c$ are those beyond a threshold on each side of the anchor (Proposition 3). The benefit depends on the anchor and the anchor on who participates, so turnout is a fixed point.

**Definition 8** (Turnout equilibrium)**.** Given an eligible population with types $\{u_i\}$ and a cost $c$, a set of participants $P$ is a *turnout equilibrium* if every agent in $P$ has benefit at least $c$ from being counted and every agent outside $P$ has benefit below $c$ from joining, with the anchor $\gamma(P)$ the mean or median of $P$.

In the examples of Appendix H.4 and in simulation, two kinds of turnout equilibrium recur. In the first, the raters are from the tails of the population, and the center abstains, up to a band at the boundary. In the second, a moderate agent rates, while a tail agent on the same side does not, since the moderate's rating holds the anchor across a change in selection, benefiting the tail agent.

[1] Anonymous: invariant to permuting the reports. Efficient: on every instance, no other type is weakly better for every agent and strictly better for one; holds when the anchor lies between the smallest and largest types.

The anchor greatly influences how many rate: Under the median, a rating moves the anchor across the empty center, regardless of the pool's size; thus, a rater's influence does not shrink as more agents rate. Agents far from the center always rate, so only a band of moderates around the anchor can abstain, and the band narrows as rating gets cheaper (Proposition 13). Under the mean, a rating moves the anchor by the rater's distance divided by the number of raters; thus, a rater's influence shrinks as more agents rate. Few agents, therefore, rate in a pure equilibrium, when one exists, and varying the cost or adding agents to the population does not help, as the pool remains small regardless (Appendix I). Which few rate depends on where the selection changes relative to the anchor (Proposition 14). Enumerating the equilibria in simulations confirms these patterns (Appendices H.4 and I). The theory is also consistent with Community Notes data: on US-politics notes, the fifth of raters farthest from the median rate about $1.5$ times as much as the fifth nearest to it (Appendix I.6).

## 5 Discussion, Limitations, and Conclusion

**Choosing the anchor.** According to our results, we suggest that anchoring at the median is likely superior to the currently deployed mechanisms, which, because they use $L_2$ penalties (Appendix F.4), largely anchor at the mean (Anonymous, 2026). The deployed mechanism and Meta's copy of it implicitly set the anchor via $L_2$ regularization (Anonymous, 2026; Meta, 2025). Note that $L_2$ (or any other) regularization could still be used to learn the parameters while using additional constraints to ensure the anchor remains the mean or median of the constructed parameters.

In practice, setting the anchor is more complicated than a mere median. As mentioned in the introduction, there are many proposals to adjust the weight of raters (e.g. downweighting low-quality responses). Yet, it still fundamentally requires choosing a way to anchor. In the regime where not every agent rates every note, these weights can also provide additional incentives (e.g. rating many notes, diverse notes, or notes that do not have many ratings). Ideas from bridging designs that act on the rating graph (Ovadya & Thorburn, 2023; Wojcik et al., 2022) could also be incorporated.

**Connections to other systems.** In ideal-point estimation, which is widely used in political science, there is a model similar to the LFUM. The choice of scale parameter is usually treated as a technical identification; ideal points are constrained to mean zero and unit variance, or two legislators are fixed at chosen positions (Clinton et al., 2004; Bafumi et al., 2005). From our framework's perspective, mean-zero normalization is the welfare anchor, and fixing two legislators anchors at their midpoint. In LLM reward models fitted to heterogeneous annotators, a similar choice arises: a single fitted reward aggregates the annotators implicitly (Siththaranjan et al., 2024). Best-of-$n$ selection ranks responses by a fitted reward. When annotators differ along a latent axis, as in ideal-point reward models (Chen et al., 2025), that reward is the utility of some annotator type. Extending the anchoring results to the pairwise comparisons such models are trained on is a promising direction.

**Limitations.** Turnout moves every anchor that ratings can identify, pure turnout equilibria need not exist, and moderates are the first to abstain (Appendix H.4). Designing more stable conditions is an open mechanism-design problem. The results do not need the exact latent factor form, only the anchored utility models of Appendix D: agents ordered on a line, quality as the evaluation of one type, and strict crossing, the valence assumption of spatial models of politics (Stokes, 1963) together with the structure that underlies median-voter results. The requirement is substantive. When what counts as quality depends on the rater's type, no single type defines quality, so there is no anchor to choose. A second limit is dimensional: the model assumes a single shared axis, along which two communities disagree on a common scale. The incommensurable case, in which communities measure different things, lies outside it. The analysis is also mostly noiseless. Its statistical version, where zero risk becomes risk minimization, needs its own machinery, developed in part in the companion paper (Anonymous, 2026). Under noise, a report that no type could produce can pass as noise, so strategyproofness must hold against a larger class of deviations.

**Conclusion.** The anchor a system recovers, and how robust that anchor is, depend as much on who rates what as on the rule that aggregates the ratings. Weighted anchors offer a lever on that structure: a statistic of the reported types that favors raters who cover particular, many, or thinly rated items is still identifiable (Proposition 1). The anchor and the turnout it induces can therefore be designed together, which we leave to future work.

### Ethics statement

This work analyzes a deployed rating-aggregation mechanism theoretically and in simulation, and checks one prediction on the public Community Notes data release. No human subjects were involved and no data was collected; the public release identifies raters only by pseudonymous identifiers. The normative implications of the choice of anchor, and of the incentives each choice creates, are discussed in Section 5.

### Reproducibility statement

All theoretical claims are stated with their assumptions in the main text and proved in full in the appendix. The simulation setup behind every figure, including the data-generating process, the mechanisms compared, and the metrics, is described in Appendix I, and the Community Notes analysis uses the public data release described in Appendix I.6; code to reproduce the figures will be released with the paper.

### AI use statement

In this work, we used generative AI tools, primarily Claude (Anthropic), for writing and editing (improving clarity, grammar, and flow; in some cases generated text was edited and incorporated), coding support (refactoring experiment code, adding evaluation metrics, and generating plotting utilities), proof assistance (checking intermediate steps and suggesting proof strategies), and ideation (brainstorming and exploratory discussion). All theoretical results were developed and verified by the authors, and all conclusions are the authors' own. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

## References

Yeganeh Alimohammadi, Karissa Huang, Christian Borgs, and Jennifer Chayes. Auditing the auditors: Does community-based moderation get it right?, March 2026.

Jennifer Allen, Cameron Martel, and David G Rand. Birds of a feather don't fact-check each other: Partisanship and the evaluation of news in Twitter's Birdwatch crowdsourced fact-checking program. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, pp. 1–19, New York, NY, USA, April 2022. Association for Computing Machinery. ISBN 978-1-4503-9157-3. doi: 10.1145/3491102.3502040.

Anonymous. Can ideologues agree on quality? From non-identifiable latent factors to collective outcomes. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2026. To appear.

Joseph Bafumi, Andrew Gelman, David K. Park, and Noah Kaplan. Practical issues in implementing and understanding Bayesian ideal point estimation. *Political Analysis*, 13(2):171–187, 2005. doi: 10.1093/pan/mpi010.

Craig Boutilier, Ioannis Caragiannis, Simi Haber, Tyler Lu, Ariel D. Procaccia, and Or Sheffet. Optimal social choice functions: A utilitarian view. *Artificial Intelligence*, 227:190–213, October 2015. ISSN 0004-3702. doi: 10.1016/j.artint.2015.06.003.

Felix Brandt, Vincent Conitzer, Ulle Endriss, Jérôme Lang, and Ariel D. Procaccia. Introduction to Computational Social Choice. In Felix Brandt, Vincent Conitzer, Ulle Endriss, Jérôme Lang, and Ariel D. Procaccia (eds.), *Handbook of Computational Social Choice*, pp. 1–29. Cambridge University Press, 2016. doi: 10.1017/CBO9781107446984.

Lindsey Brown, Hoang Ha, and Jonathan K. Hodge. Single-peaked preferences over multidimensional binary alternatives. *Discrete Applied Mathematics*, 166:14–25, March 2014. ISSN 0166-218X. doi: 10.1016/j.dam.2013.11.006.

Giulio Giacomo Cantone, Venera Tomaselli, and Valeria Mazzeo. Review bombing: Ideology-driven polarisation in online ratings: The case study of The Last of Us (part II). *Quality & Quantity*, September 2024. ISSN 1573-7845. doi: 10.1007/s11135-024-01981-z.

Daiwei Chen, Yi Chen, Aniket Rege, Zhi Wang, and Ramya Korlakai Vinayak. PAL: Sample-efficient personalized reward modeling for pluralistic alignment. In *International Conference on Learning Representations (ICLR)*, 2025.

Yiling Chen, Chara Podimata, Ariel D. Procaccia, and Nisarg Shah. Strategyproof Linear Regression in High Dimensions. In *Proceedings of the 2018 ACM Conference on Economics and Computation*, EC '18, pp. 9–26, New York, NY, USA, June 2018. Association for Computing Machinery. ISBN 978-1-4503-5829-3. doi: 10.1145/3219166.3219175.

Paul-Alexandru Chirita, Wolfgang Nejdl, and Cristian Zamfir. Preventing shilling attacks in online recommender systems. In *Proceedings of the 7th Annual ACM International Workshop on Web Information and Data Management*, WIDM '05, pp. 67–74, New York, NY, USA, November 2005. Association for Computing Machinery. ISBN 978-1-59593-194-8. doi: 10.1145/1097047.1097061.

Joshua Clinton, Simon Jackman, and Douglas Rivers. The Statistical Analysis of Roll Call Data. *American Political Science Review*, 98(2):355–370, May 2004. ISSN 0003-0554, 1537-5943. doi: 10.1017/S0003055404001194.

Community Notes. Note ranking algorithm. https://communitynotes.x.com/guide/en/under-the-hood/ranking-notes, 2026.

Vincent Conitzer, Rachel Freedman, Jobst Heitzig, Wesley H. Holliday, Bob M. Jacobs, Nathan Lambert, Milan Mossé, Eric Pacuit, Stuart Russell, Hailey Schoelkopf, Emanuel Tewolde, and William S. Zwicker. Social Choice for AI Alignment: Dealing with Diverse Human Feedback, April 2024.

Ofer Dekel, Felix Fischer, and Ariel D. Procaccia. Incentive compatible regression learning. *Journal of Computer and System Sciences*, 76(8):759–777, 2010.

Yashar Deldjoo, Tommaso Di Noia, and Felice Antonio Merra. A Survey on Adversarial Recommender Systems: From Attack/Defense Strategies to Generative Adversarial Networks. *ACM Computing Surveys*, 54(2):1–38, March 2022. ISSN 0360-0300, 1557-7341. doi: 10.1145/3439729.

Chiara Patricia Drolsbach and Nicolas Pröllochs. Diffusion of Community Fact-Checked Misinformation on Twitter. *Proc. ACM Hum.-Comput. Interact.*, 7(CSCW2):267:1–267:22, October 2023. doi: 10.1145/3610058.

Sara Fish, Paul Gölz, David C. Parkes, Ariel D. Procaccia, Gili Rusak, Itai Shapira, and Manuel Wüthrich. Generative Social Choice, September 2023.

Joshua S. Gans and Michael Smart. Majority voting with single-crossing preferences. *Journal of Public Economics*, 59(2):219–237, 1996. doi: 10.1016/0047-2727(95)01503-5.

Allan Gibbard. Manipulation of Voting Schemes: A General Result. *Econometrica*, 41(4):587–601, 1973. ISSN 0012-9682. doi: 10.2307/1914083.

William Godel, Zeve Sanderson, Kevin Aslett, Jonathan Nagler, Richard Bonneau, Nathaniel Persily, and Joshua A. Tucker. Moderating with the Mob: Evaluating the Efficacy of Real-Time Crowdsourced Fact-Checking. *Journal of Online Trust and Safety*, 1(1), October 2021. ISSN 2770-3142. doi: 10.54501/jots.v1i1.15.

Mohak Goyal, Nishka Arora, and Ashish Goel. Quality-Sensitive Matrix Factorization for Community Notes: Towards Sample Efficiency and Manipulation Resistance, April 2026.

Nick Gravin and Jianhao Jia. Approximation guarantees of Median Mechanism in $\mathbb{R}^d$, February 2025.

Moritz Hardt, Nimrod Megiddo, Christos Papadimitriou, and Mary Wootters. Strategic classification. In *Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science (ITCS)*, pp. 111–122, 2016. doi: 10.1145/2840728.2840730.

Jinyuan Jia, Yupei Liu, Yuepeng Hu, and Neil Zhenqiang Gong. PORE: Provably Robust Recommender Systems against Data Poisoning Attacks, March 2023.

Franziska B. Keller, David Schoch, Sebastian Stier, and JungHwan Yang. Political Astroturfing on Twitter: How to Coordinate a Disinformation Campaign. *Political Communication*, 37(2): 256–280, March 2020. ISSN 1058-4609. doi: 10.1080/10584609.2019.1661888.

Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization, January 2017.

John O. Ledyard. The pure theory of large two-candidate elections. *Public Choice*, 44(1):7–41, January 1984. ISSN 1573-7101. doi: 10.1007/BF00124816.

Bo Li, Yining Wang, Aarti Singh, and Yevgeniy Vorobeychik. Data Poisoning Attacks on Factorization-Based Collaborative Filtering. *ArXiv*, August 2016a.

Yaliang Li, Jing Gao, Chuishi Meng, Qi Li, Lu Su, Bo Zhao, Wei Fan, and Jiawei Han. A Survey on Truth Discovery. *SIGKDD Explor. Newsl.*, 17(2):1–16, February 2016b. ISSN 1931-0145. doi: 10.1145/2897350.2897352.

Christian List and Robert E. Goodin. Epistemic democracy: Generalizing the Condorcet jury theorem. *Journal of Political Philosophy*, 9(3):277–306, 2001. doi: 10.1111/1467-9760.00128.

Benjamin Lubin and D. Parkes. Quantifying the Strategyproofness of Mechanisms via Metrics on Payoff Distributions. In *Conference on Uncertainty in Artificial Intelligence*, June 2009.

Dina Mayzlin, Yaniv Dover, and Judith Chevalier. Promotional Reviews: An Empirical Investigation of Online Review Manipulation. *American Economic Review*, 104(8):2421–2455, August 2014. ISSN 0002-8282. doi: 10.1257/aer.104.8.2421.

B. Mehta and Thomas Hofmann. A Survey of Attack-Resistant Collaborative Filtering Algorithms. *IEEE Data Eng. Bull.*, 2008.

Meta. Testing Begins for Community Notes on Facebook, Instagram and Threads, March 2025.

Saeedeh Mohammadi, Narges Chinichian, Hannah Doyal, Anna Bertani, Kristina Skutilova, Hao Cui, Michele d'Errico, Siobhan Grayson, and Taha Yasseri. From birdwatch to community notes, from twitter to x: Four years of community-based content moderation, October 2025.

H. Moulin. On strategy-proofness and single peakedness. *Public Choice*, 35(4):437–455, 1980. ISSN 0048-5829, 1573-7101. doi: 10.1007/BF00128122.

Ritesh Noothigattu, Snehalkumar 'Neil' S. Gaikwad, Edmond Awad, Sohan Dsouza, Iyad Rahwan, Pradeep Ravikumar, and Ariel D. Procaccia. A Voting-Based System for Ethical Decision Making. In *THIRTY-SECOND AAAI CONFERENCE ON ARTIFICIAL INTELLIGENCE / THIRTIETH INNOVATIVE APPLICATIONS OF ARTIFICIAL INTELLIGENCE CONFERENCE / EIGHTH AAAI SYMPOSIUM ON EDUCATIONAL ADVANCES IN ARTIFICIAL INTELLIGENCE*, pp. 1587–1594, Palo Alto, 2018. Assoc Advancement Artificial Intelligence. ISBN 978-1-57735-800-8.

Ritesh Noothigattu, Nihar B. Shah, and Ariel D. Procaccia. Loss functions, axioms, and peer review. *Journal of Artificial Intelligence Research*, 70:1481–1515, 2021.

Aviv Ovadya and Luke Thorburn. Bridging Systems: Open Problems for Countering Destructive Divisiveness across Ranking, Recommenders, and Governance, July 2023.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An imperative style, high-performance deep learning library. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, number 721, pp. 8026–8037. Curran Associates Inc., Red Hook, NY, USA, December 2019.

B. Paudel and A. Bernstein. Cross-Cutting Political Awareness through Diverse News Recommendations. *ArXiv*, September 2019.

Juan Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. Performative prediction. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 7599–7609. PMLR, 2020.

W. Peremans, H. Peters, H. v. d. Stel, and T. Storcken. Strategy-proofness on Euclidean spaces. *Social Choice and Welfare*, 14(3):379–401, June 1997. ISSN 1432-217X. doi: 10.1007/s003550050074.

Keith T. Poole and Howard Rosenthal. A Spatial Model for Legislative Roll Call Analysis. *American Journal of Political Science*, 29(2):357–384, 1985. ISSN 0092-5853. doi: 10.2307/2111172.

Nicolas Pröllochs. Community-based fact-checking on Twitter's Birdwatch platform. *Proceedings of the International AAAI Conference on Web and Social Media*, 16:794–805, 2022. doi: 10.1609/icwsm.v16i1.19335.

Tianyi Qiu. Representative Social Choice: From Learning Theory to AI Alignment, October 2025.

Paul Resnick and Rahul Sami. The influence limiter: Provably manipulation-resistant recommender systems. In *Proceedings of the 2007 ACM Conference on Recommender Systems*, RecSys '07, pp. 25–32, New York, NY, USA, October 2007. Association for Computing Machinery. ISBN 978-1-59593-730-8. doi: 10.1145/1297231.1297236.

Sarah T. Roberts. *Behind the Screen: Content Moderation in the Shadows of Social Media*. Yale University Press, 2019. ISBN 978-0-300-23588-3. doi: 10.2307/j.ctvhrcz0v.

Alejandro Saporiti. Strategy-proofness and single-crossing. *Theoretical Economics*, 4(2):127–163, 2009.

Anand Siththaranjan, Cassidy Laidlaw, and Dylan Hadfield-Menell. Distributional Preference Learning: Understanding and Accounting for Hidden Context in RLHF. In *International Conference on Learning Representations (ICLR)*, 2024.

Donald E. Stokes. Spatial models of party competition. *American Political Science Review*, 57(2):368–377, 1963. doi: 10.2307/1952828.

Prasanna Tambe, Peter Cappelli, and Valery Yakubovich. Artificial Intelligence in Human Resources Management: Challenges and a Path Forward. *California Management Review*, 61(4):15–42, August 2019. ISSN 0008-1256. doi: 10.1177/0008125619867910.

Bao Tran Truong, Siqi Wu, Alessandro Flammini, Filippo Menczer, and Alexander J. Stewart. Community notes are vulnerable to rater bias and manipulation, November 2025.

Chenlong Wang and Pablo Lucas. Efficiency of Community-Based Content Moderation Mechanisms: A Discussion Focused on Birdwatch. *Group Decision and Negotiation*, 33(3):673–709, June 2024. ISSN 1572-9907. doi: 10.1007/s10726-024-09881-1.

Stefan Wojcik, Sophie Hilgard, Nick Judd, Delia Mocanu, Stephen Ragain, M. B. Fallin Hunzaker, Keith Coleman, and Jay Baxter. Birdwatch: Crowd Wisdom and Bridging Algorithms can Inform Understanding and Reduce the Spread of Misinformation. *arXiv preprint arXiv:2210.15723*, 2022. doi: 10.48550/ARXIV.2210.15723.

# A NOTATION TABLE

For convenience, we collect the paper-specific notation used across sections. Standard mathematical notation, and symbols defined at their only use site, are not listed. The model and its instances are defined in Definition 1, mechanisms and anchoring in Definitions 2 and 4, zero-risk models in Definition 5, and the general class in Definition 9.

| Symbol | Meaning |
|---|---|
| *Utility model (Section 2)* | |
| $n, m$ | Number of agents and of items. |
| $u_i$ | Type of agent $i$, its position on the ideological line; $u = (u_1, \dots, u_n)$. |
| $q_j, v_j$ | Quality and leaning of item $j$; $q, v \in \mathbf{R}^m$. |
| $U(t, j)$ | Utility $q_j + t\, v_j$ of type $t$ for item $j$ (1); $U(t, \cdot)$ its utility row. |
| $E = (u, q, v)$ | Instance. Throughout, two agents rate some item differently. |
| $E_\gamma$ | $(u - \gamma,\ q + \gamma v,\ v)$: declares type $\gamma$ neutral, same utilities as $E$ (2). |
| $E_{\gamma,c}$ | $(c(u - \gamma),\ q + \gamma v,\ v/c)$, $c \neq 0$: the instances utility-equivalent to $E$ (Lemma 1). |
| $\bar{u}, \tilde{u}$ | Mean and median of the types. |
| *Mechanisms and anchors (Section 2)* | |
| $r_i, R = (r_{ij})$ | Report of agent $i$ (one number per item) and the report matrix. |
| $R(E)$ | Report matrix under truthful, complete reporting, $r_{ij} = U(u_i, j)$. |
| $f$ | Mechanism: a function of $R$ to a ranking or a selection (Definition 2). |
| $k$ | Selection size; ties broken by a fixed priority order on items. |
| $\pi_\gamma(E)$ | Oracle ranking of the items by $U(\gamma, \cdot)$, computed from the instance. |
| $S(\gamma)$ | Top-$k$ selection at anchor $\gamma$. |
| $V_u(\gamma)$ | Utility $\sum_{j \in S(\gamma)} U(u, j)$ of type $u$ from the selection at anchor $\gamma$. |
| $a(u)$ | Statistic of the types; $f$ is $a$-anchored if anchored at $a(u)$ on every $E$. |
| OQ, SW, CC | Objective quality, social welfare, Condorcet criterion (Definition 3). |
| *Learning (Section 3)* | |
| $\theta = (\hat{q}, \hat{u}, \hat{v})$ | Model of the family (1); predicted ratings $\hat{q}_j + \hat{u}_i \hat{v}_j$. |
| $\Theta(R)$ | Zero-risk models of $R$, those that reproduce every rating (Definition 5). |
| $(\gamma, c)$ | Parameters of a zero-risk model, $\hat{q} = q + \gamma v$, $\hat{u} = c(u - \gamma)$, $\hat{v} = v/c$. |
| $\hat{q}_j + a(\hat{u})\hat{v}_j$ | Re-anchored quality of a zero-risk model (Theorem 4). |
| *Strategic behavior (Section 4)* | |
| $w$ | Report profile: agent $i$ reports the row of type $w_i$; a spoof if $w_i \neq u_i$ (Definition 7). |
| $E_w$, $\bar{w}$, $\tilde{w}$ | Reported instance $(w, q, v)$; mean and median of the reported types. |
| $g(t)$, $\gamma^*$ | Utility difference $U(t, j) - U(t, j')$ of two items, and the crossing where it vanishes. |
| $B_i$, $B_i(P)$ | Benefit of participating (3), and its unified form (7). |
| $c$ | Cost of rating in the participation game (Section 4.2). |
| $P$, $\gamma(P)$ | Pool of raters and its anchor, the mean or median of its types. |
| $W(\gamma)$ | Total leaning $\sum_{j \in S(\gamma)} v_j$ of the selection at $\gamma$ (8). |
| *General class (Appendices D and F.2)* | |
| $\mathcal{T}$ | Ordered space of types. |
| $E = (A, X, U, q)$ | General instance: agents $A$, items $X$, utility $U$, quality $q$. |
| $E_t$ | Designation of type $t$ as neutral, $q := U(t, \cdot)$. |
| $\mathcal{C}$, $\mathcal{H}$ | Class of anchored utility models (Definition 9) and a subclass. |
| $\pi_t(E)$, $\alpha$ | Oracle ranking from type $t$; anchor rule assigning each instance a type. |
| *Experiments (Appendix I)* | |
| $\mu_u$ | Mean type of the population in the data-generating process. |
| $\Delta$ | Noise variance of the ratings; the theory takes $\Delta = 0$. |
| $s \in (0, 1)$ | Fraction of the mean type at which BMF's regularization anchors (Anonymous, 2026). |

**Overloaded symbols.** $c$ is the scale parameter of a zero-risk model in Section 3 and the cost of rating in Section 4.2; the two never meet. $\gamma^*$ is the crossing of two items in the proofs of Section 4; in the companion paper it denotes the anchor selected by regularization.

## B Extended Related Work

Our strategyproofness results sit in the social choice tradition of restricted domains. The Gibbard–Satterthwaite theorem rules out non-dictatorial strategyproof rules on unrestricted preferences (Gibbard, 1973). On single-peaked domains Moulin (1980) characterized the strategyproof rules as generalized medians, with extensions to Euclidean and multidimensional settings (Peremans et al., 1997; Brown et al., 2014). On single-crossing domains, the median-voter and strategyproofness results of Gans & Smart (1996) and Saporiti (2009) are the ones we build on. Quantitative treatments of these guarantees, approximation bounds for median mechanisms (Gravin & Jia, 2025) and measures of manipulability (Lubin & Parkes, 2009), are close in spirit to our comparison of the welfare the median gives up (Appendix G.1), and strategyproof estimation (Chen et al., 2018) shares our object of study, a mechanism defined by a model fit to reports. Our turnout equilibrium follows the costly-voting tradition of rational abstention (Ledyard, 1984). More broadly, the paper joins work in computational social choice (Brandt et al., 2016; Boutilier et al., 2015) and the growing line treating learned systems as preference aggregators (Conitzer et al., 2024; Fish et al., 2023; Qiu, 2025; Noothigattu et al., 2018).

On the platform side, Community Notes has been studied empirically for its rating behavior, diffusion, and efficiency (Wojcik et al., 2022; Allen et al., 2022; Pröllochs, 2022; Drolsbach & Pröllochs, 2023; Wang & Lucas, 2024; Godel et al., 2021), and its design belongs to the bridging-based ranking program (Ovadya & Thorburn, 2023; Paudel & Bernstein, 2019). The manipulations we model have documented counterparts: promotional reviewing (Mayzlin et al., 2014), ideologically driven review bombing (Cantone et al., 2024), political astroturfing (Keller et al., 2020), and coordinated rating on Community Notes itself (Mohammadi et al., 2025; Truong et al., 2025). They parallel shilling and data-poisoning attacks on recommender systems (Chirita et al., 2005; Mehta & Hofmann, 2008; Deldjoo et al., 2022; Li et al., 2016a), against which manipulation-resistant designs have been proposed (Resnick & Sami, 2007; Jia et al., 2023). For Community Notes specifically, two recent designs re-weight raters within the factorization: by sensitivity to the learned quality (Goyal et al., 2026), and by consistency rather than agreement, so a rater may disagree with the consensus and keep its influence (Alimohammadi et al., 2026); both remain anchored mechanisms in our sense (Section 5). The latent factor utility model itself descends from ideal-point and valence models in political science (Poole & Rosenthal, 1985; Clinton et al., 2004; Stokes, 1963); what BMF recovers under it is the subject of the companion paper (Anonymous, 2026).

## C Running Example: Details

A platform displays one fact-checking note beneath a viral post. Three notes have been proposed and five raters rate them, few enough that every quantity below can be checked by hand.

**Agents.** Five raters with types

$$u_1 = -0.5, \quad u_2 = 0, \quad u_3 = 0.5, \quad u_4 = 1.5, \quad u_5 = 2.5.$$

The population is right-skewed: mean $\bar{u} = 0.8$, median $\tilde{u} = 0.5$.

**Alternatives.** Three notes, one high-quality left-leaning and two moderate-quality right-leaning (Table 1). Ordered by quality alone, $A \succ C \succ B$.

Table 1: Running example: three candidate fact-check notes.

| Note | $q_j$ | $v_j$ | Character |
|---|---|---|---|
| A | 1.2 | $-1.0$ | high quality, left-leaning |
| B | 0.6 | 1.0 | moderate, right-leaning |
| C | 0.9 | 0.5 | good, mildly right-leaning |

**Utilities.** Applying $U(u_i, j) = q_j + u_i v_j$ (Definition 1):

Agent 1 strongly favors Note A; Agents 4 and 5 strongly favor Note B; Agent 3 (the median rater) slightly prefers Note C.

Table 2: Utilities $U(u_i, j) = q_j + u_i v_j$. Each agent's most-preferred note is **bold***.

| | Note A ($q$=1.2, $v$=−1) | Note B ($q$=0.6, $v$=1) | Note C ($q$=0.9, $v$=0.5) |
|---|---|---|---|
| Agent 1 ($u$=−0.5) | $\mathbf{1.70}^*$ | 0.10 | 0.65 |
| Agent 2 ($u$=0) | $\mathbf{1.20}^*$ | 0.60 | 0.90 |
| Agent 3 ($u$=0.5) | 0.70 | 1.10 | $\mathbf{1.15}^*$ |
| Agent 4 ($u$=1.5) | −0.30 | $\mathbf{2.10}^*$ | 1.65 |
| Agent 5 ($u$=2.5) | −1.30 | $\mathbf{3.10}^*$ | 2.15 |

### C.1 Three objectives, three different winners

**Objective Quality.** Ranking by $q_j$: $q_A = 1.2 > q_C = 0.9 > q_B = 0.6$. **Note A is selected.**

**Social Welfare.** The average utility for note $j$ is $q_j + \bar{u} v_j$ with $\bar{u} = 0.8$:

$$\text{A}: \ 0.40, \qquad \text{B}: \ 1.40, \qquad \text{C}: \ 1.30.$$

**Note B is selected.** The totals over the five agents are 2.00, 7.00 and 6.50, five times the averages.

**Condorcet Criterion.** Pairwise comparisons using Table 2:

- B vs. A: Agents 3, 4, 5 prefer B. **B wins 3–2.**
- C vs. A: Agents 3, 4, 5 prefer C. **C wins 3–2.**
- C vs. B: Agents 1, 2, 3 prefer C. **C wins 3–2.**

Note C is the Condorcet winner. Equivalently, Agent 3 ($u_3 = 0.5$, the median) ranks the notes $C \succ B \succ A$, confirming Theorem 1. **Note C is selected.**

**Three winners.** The three objectives select three different notes. Note A has the highest quality. Note B has the most total utility, most of it from the right-leaning majority. Note C wins every pairwise contest although only one rater ranks it first, against two each for A and B; it is the kind of note bridging-based ranking aims to surface (Ovadya & Thorburn, 2023).

### C.2 Gauge symmetry: quality is ambiguous

Apply the shift (2) with $\gamma = 1$:

$$u'_i = u_i - 1 = -1.5, \ -1, \ -0.5, \ 0.5, \ 1.5,$$
$$q'_A = 1.2 + (1)(-1) = 0.2, \quad q'_B = 0.6 + (1)(1) = 1.6, \quad q'_C = 0.9 + (1)(0.5) = 1.4,$$
$$v'_j = v_j \quad \text{(unchanged)}.$$

Every utility is preserved, for instance $U'(u'_1, A) = 0.2 + (-1.5)(-1) = 1.70$, but the quality ranking is now $B \succ C \succ A$, the complete reverse. The shift declares the type 1, between Agents 3 and 4, neutral, and from that type's viewpoint Note B is best.

### C.3 Anchoring determines the outcome

The utility of type $\gamma$ is $U(\gamma, j) = q_j + \gamma v_j$ (Definition 4). In our example:

$$U(\gamma, A) = 1.2 - \gamma, \qquad U(\gamma, B) = 0.6 + \gamma, \qquad U(\gamma, C) = 0.9 + 0.5\gamma.$$

Notes A and C tie at $\gamma = 0.2$ and notes C and B at $\gamma = 0.6$, so the selected note changes twice (Table 3).

The neutral anchor ($\gamma = 0$) falls in Regime I, the median ($\gamma = 0.5$) in Regime II, and the mean ($\gamma = 0.8$) in Regime III.

Table 3: Anchor regimes. Each regime selects a different note.

| Regime | Anchor range | Selected note | Corresponding objective |
|---|---|---|---|
| I | $\gamma < 0.2$ | A | Quality-oriented |
| II | $0.2 < \gamma < 0.6$ | C | Condorcet / majority |
| III | $\gamma > 0.6$ | B | Welfare-oriented |

**What $L_2$ regularization selects.** The companion paper (Anonymous, 2026) shows that $L_2$-regularized BMF anchors implicitly at $\gamma^* = s\,\bar{u}$ for a shrinkage factor $s \in (0, 1)$ set by the regularization weights; in this example, with equal weights and no intercepts, $\gamma^* \approx 0.49$, which falls in Regime II and selects Note C. Penalizing quality more pushes $\gamma^*$ toward 0 and Note A; penalizing it less pushes toward $\bar{u} = 0.8$ and Note B.

**The anchor need not be any agent's type.** The anchor $\gamma^* \approx 0.49$ is not any agent's type. The mechanism evaluates quality *as if* a hypothetical agent at type $0.49$ were the neutral observer, but no such agent exists in the population. The mean anchor, $0.8$, is not an agent's type either; for odd $n$ only the median anchor always is.

### C.4 Strategic manipulation

**The mean anchor is manipulable (Proposition 2).** Under the mean anchor ($\gamma = \bar{u} = 0.8$), Note B is selected. Agent 1 ($u_1 = -0.5$) prefers Note A (utility $1.70$ vs. $0.10$ for B). By reporting $u_1^{\text{report}} = w$, the mean becomes

$$\bar{u}' = \frac{4.5 + w}{5}.$$

For Note A to be selected we need $\bar{u}' < 0.2$ (Regime I), i.e. $w < -3.5$. A single agent, by exaggerating from $-0.5$ to below $-3.5$, moves the mean from $0.8$ to below $0.2$, switching the outcome across two regime boundaries.

**The median anchor cannot be moved beneficially (Theorem 5).** Under the median anchor ($\gamma = \tilde{u} = 0.5$), Note C is selected. The median is the third value in the sorted list $\{-0.5,\ 0,\ \mathbf{0.5},\ 1.5,\ 2.5\}$. Agent 1 can move the median, but only away from itself: any report at or below $0.5$ leaves the median at $0.5$, while a report above $0.5$ raises it to as much as $1.5$, moving the selection from Note C toward Note B, which Agent 1 values even less. The symmetric statement holds for Agent 5. No deviation is profitable, singly or in coalition.

## D The Anchored Utility Model

The body states every result for the latent factor utility model. This appendix defines the class of utility models on which the results hold, those that split utility into a shared valence term and a positional term as spatial models of politics do (Stokes, 1963; Poole & Rosenthal, 1985; Clinton et al., 2004), and lists which result uses which property.

An *instance* is $E = (A, X, U, q)$. The agents $A = (u_1, \ldots, u_n)$ are points of an ordered space $\mathcal{T}$ of *types*, two agents possibly at the same point. The items $X$ form a set of size $m$. The utility $U : \mathcal{T} \times X \to \mathbf{R}$ assigns utility to each type-item pair, and $q : X \to \mathbf{R}$ assigns quality to each item. For a type $t$ the *designation* of $t$ as neutral is the instance $E_t := (A, X, U, U(t, \cdot))$, the same agents, items and utility with $t$'s row as the quality. For a type $t$, $\pi_t(E)$ is the ranking of the items by $U(t, \cdot)$, ties broken by the fixed priority order, and a mechanism is *anchored at $t$ on $E$* if its output on $R(E)$ is $\pi_t(E)$ or its top $k$, as in Definition 4.

**Definition 9** (Anchored utility model)**.** An instance $E = (A, X, U, q)$ is an *anchored utility model* (AUM) if it satisfies:

Alignability (AL). Quality is some type's utility row: a *neutral type* $t_0 \in \mathcal{T}$ has $U(t_0, j) = q_j$ for every item $j$.

Strict crossing (SC). For any two items $j, j'$ with distinct utility profiles, the difference $g(t) = U(t, j) - U(t, j')$ vanishes at most at one type and $t \mapsto \operatorname{sgn} g(t)$ is monotone along $\mathcal{T}$.

Instances satisfying both axioms form the class $\mathcal{C}$, and such an instance is *admitted*. Two further conditions are used by single results. An instance has *monotone differences* if every difference $g$ is monotone in the type, and is *cardinally linear* if $\mathcal{T} = \mathbf{R}$ and each $U(\cdot, j)$ is affine in the type.

The neutral type need not be unique, since two types can share a row. For the same reason, the body's standing convention reads on the class as a condition on rows: two agents have different utility rows, which in the LFUM with $v \neq 0$ means two distinct types. Cardinal linearity implies strict crossing and monotone differences, and it is exactly the form (1) after recentering the types at a neutral type, so the latent factor utility model is an admitted instance with both further conditions. Monotone differences and strict crossing are independent. Neither axiom is verified by a mechanism; they delimit the instances over which a guarantee is stated. Strict crossing can be tested in part by checking whether the reported rows admit an order in which it holds. Alignability cannot be tested at all by the following closure property.

**Proposition 4** (Designation-closure)**.** *For every $E \in \mathcal{C}$ and every type $t \in \mathcal{T}$, the designation $E_t$ is in $\mathcal{C}$, and it has the same agents and the same utility as $E$.*

*Proof.* Strict crossing is a condition on the utility alone, which $E_t$ shares with $E$. Alignability holds in $E_t$ with $t$ as the neutral type. □

**Remark 1** (Strict and single crossing)**.** Strict crossing strengthens *single crossing*, the condition that $g$ changes sign at most once. Single crossing suffices for Proposition 5, Lemma 2, Proposition 4 and the general form of Proposition 1(a). Two examples show that strict crossing cannot be weakened to single crossing elsewhere. The rows $(0, 0), (1, 3), (2, 1)$ on two items have pair differences $0, -2, 1$; under single crossing both $(0, -2, 1)$ and $(0, 1, -2)$ are admissible type orders, and their median rows order the pair oppositely, so the median anchor is not identifiable. And two items whose difference is nonnegative everywhere and vanishes only at the truthful median, with the tie there broken toward the item preferred nowhere, give every agent below the median a profitable spoof: reporting far above moves the median off the tie and selects the item it strictly prefers, so the median is not strategyproof (Appendix G).

**Remark 2** (Instances beyond the linear model)**.** The spatial-voting literature supplies two further instances. The quadratic ideal-point model, standard in roll-call estimation (Clinton et al., 2004), gives a type at $t$ the utility $U(t, j) = q_j - (t - x_j)^2$ for an item at position $x_j$ with valence $q_j$. Expanding leaves a $-t^2$ term common to all items, which cancels from every comparison, so this model makes the same comparisons as the LFUM with quality $\tilde{q}_j = q_j - x_j^2$ and leaning $v_j = 2x_j$; its ratings differ by the rater-specific constant $-t^2$, which a factorization with rater intercepts absorbs. The linear-loss variant, $U(t, j) = q_j - |t - x_j|$, is not cardinal-linear. It has monotone differences and, for generic parameters, strict crossing, so the median and participation results hold for it.

**Proposition 5** (Size of the producible set)**.** *Let $E$ be an admitted instance with $m$ items, write $\Pi(E) = \{\pi_t(E) : t \in \mathcal{T}\}$ for the rankings its types produce, and fix a selection size $k$. Then:*

1. *$\Pi(E)$ contains at most $\binom{m}{2} + 1$ of the $m!$ strict rankings;*
2. *the anchors at which $U(\gamma, \cdot)$ selects a given $k$-set form an interval of $\mathcal{T}$, up to anchors where a selected and an unselected item tie.*

*Proof.* Traverse the anchors in the order of $\mathcal{T}$. For a pair $j, j'$, single crossing lets the strict order of $U(\gamma, j)$ and $U(\gamma, j')$ change at most once, so across anchors where no pair ties the ranking changes at most $\binom{m}{2}$ times, and at most $\binom{m}{2} + 1$ strict rankings occur. If a set $S$ is selected at $\gamma_1$ and $\gamma_3$ but not at some $\gamma_2$ between them, then at $\gamma_2$ some $a \in S$ is unselected and some $c \notin S$ is selected. Then $a$ is above $c$ at $\gamma_1$ and $\gamma_3$ and below it at $\gamma_2$, two sign changes of one difference, against single crossing. □

The running example attains the bound: the crossings at $\gamma = 0.2, 0.3, 0.6$ give four rankings, and $\binom{3}{2} + 1 = 4$.

The general forms follow the linear results they extend. Appendix E states the correspondence of Theorem 1 on the general class. Appendix F.2 defines compatible mechanisms, the general form of

Table 4: Which property each result of the body uses on the general class. Results not listed use the linear form itself.

| Result | Property used |
|---|---|
| Proposition 5, Lemma 2, Proposition 4 | single crossing |
| Theorem 1 for OQ; Proposition 1(a), general form | alignability and closure |
| Proposition 1, median case; Theorem 1 for CC | strict crossing |
| Theorem 1 for SW; the mean anchor | cardinal linearity |
| Theorem 5; the Moulin step of Theorem 6 | strict crossing |
| Proposition 3 | monotone differences |
| Lemma 1, Theorem 4, phantom step of Theorem 6 | linear form |

zero-risk mechanisms, and states which anchors are identifiable on the class. The lemmas of Appendix G are stated for admitted instances, so Theorem 5 and the characterization step of Theorem 6 hold there. The proof of Proposition 3 in Appendix H uses only monotone differences; the turnout results of Appendix H.4 use the linear form. Table 4 lists the property each result uses.

## E Proofs for Section 2

**Theorem 1** (Anchor–objective correspondence)**.** *Let $E = (u, q, v)$ be an instance in which no two items tie in quality, in total utility, or beat each other, and let $f$ be a mechanism applied to $R(E)$. Then:*

*(i) $f$ attains OQ on $E$ if and only if it is anchored at $0$ on $E$;*
*(ii) $f$ attains SW on $E$ if and only if it is anchored at $\bar{u}$ on $E$;*
*(iii) for odd $n$, $f$ attains CC on $E$ if and only if it is anchored at $\tilde{u}$ on $E$.*

*Proof. Quality.* $U(0, j) = q_j$, so $\pi_0(E)$ orders items by quality and, with no ties in quality, is the only ranking that does. Hence $f$ attains OQ on $E$ if and only if $f(R(E))$ is $\pi_0(E)$ or its top $k$, which is anchoring at $0$.

*Welfare.* $\sum_i U(u_i, j) = n\, q_j + \big(\sum_i u_i\big) v_j = n\, U(\bar{u}, j)$, so $\pi_{\bar{u}}(E)$ orders items by total utility and, with no ties, is the only ranking that does; likewise $f$ attains SW if and only if it is anchored at $\bar{u}$.

*Condorcet.* List the agents by type. For a pair $j, j'$, $g(t) = U(t, j) - U(t, j') = (q_j - q_{j'}) + t\,(v_j - v_{j'})$ is affine in $t$, so it is increasing, decreasing or constant, and the agents with $g > 0$ lie on one side of those with $g < 0$. If the median agent strictly prefers $j$, the agents strictly preferring $j'$ lie on one side of the median, number at most $(n-1)/2$, and more than $n/2$ agents weakly prefer $j$: $j$ beats $j'$. If the median agent is indifferent, $g$ vanishes at $\tilde{u}$, and by the same count more than $n/2$ agents weakly prefer $j$ and more than $n/2$ weakly prefer $j'$, so the two beat each other, which the hypothesis excludes. Hence $j$ beats $j'$ exactly when the median agent strictly prefers $j$, the majority relation is $\pi_{\tilde{u}}(E)$, it is the only ranking attaining CC, and likewise $f$ attains CC if and only if it is anchored at $\tilde{u}$. □

The correspondence holds on the general class in the following form.

**Lemma 2** (One-sided agreement)**.** *Let $E$ be an admitted instance and $j, j'$ two items. Then every type in $\{t \in \mathcal{T} : U(t, j) > U(t, j')\}$ lies below every type in $\{t \in \mathcal{T} : U(t, j) < U(t, j')\}$, or every type in the first set lies above every type in the second.*

*Proof.* Write $g(t) = U(t, j) - U(t, j')$. Suppose some type with $g > 0$ lies below some type with $g < 0$ and some type with $g < 0$ lies below some type with $g > 0$, say $t_a < t_b$ with $g(t_a) > 0 > g(t_b)$ and $t_c < t_d$ with $g(t_c) < 0 < g(t_d)$. Then $t_a < t_b < t_d$ if $t_b < t_d$, and $t_c < t_d < t_b$ otherwise, three types whose signs of $g$ alternate, against single crossing. Hence every type with $g > 0$ lies below every type with $g < 0$, or every type with $g > 0$ lies above every type with $g < 0$. □

**Proposition 6** (Anchor–objective correspondence, general form)**.** *Let $E = (A, X, U, q)$ be an admitted instance with neutral type $t_0$, and let $f$ be a mechanism applied to its report matrix*

$R(E)$ *under truthful, complete reporting. Suppose no two items of* $X$ *tie in quality, in total utility* $\sum_i U(u_i, j)$*, or beat each other. Then:*

1. *$f$ attains OQ on $E$ if and only if it is anchored at $t_0$ on $E$;*
2. *if $E$ is cardinal-linear with mean type $\bar{u}$, $f$ attains SW on $E$ if and only if it is anchored at $\bar{u}$ on $E$;*
3. *if $n$ is odd, $f$ attains CC on $E$ if and only if it is anchored at the median type $\tilde{u}$ on $E$.*

*Proof. Quality.* By alignability $U(t_0, j) = q_j$, so the ranking of $U(t_0, \cdot)$ orders items by quality and, with no ties in quality, is the only ranking that does; its top $k$ is then the only $k$-set attaining OQ. Hence $f$ attains OQ on $E$ if and only if its output is that ranking, or that top $k$, which is anchoring at $t_0$. This uses alignability alone.

*Welfare.* In a cardinal-linear instance each $U(\cdot, j)$ is affine, so $\sum_i U(u_i, j) = n\,U(\bar{u}, j)$, so the ranking of $U(\bar{u}, \cdot)$ orders items by total utility and, with no ties, is the only ranking that does; the argument of the quality case gives the equivalence with anchoring at $\bar{u}$.

*Condorcet.* Let $n$ be odd, and list the agents in the order of $\mathcal{T}$. For a pair $j, j'$, Lemma 2 places the agents strictly preferring $j$ on one side of those strictly preferring $j'$. If the median type strictly prefers $j$, the agents strictly preferring $j'$ all lie on one side of the median, so they number at most $(n-1)/2$ and more than $n/2$ agents weakly prefer $j$: $j$ beats $j'$. If the median type is indifferent, strict crossing makes the signs monotone along the roster, so the agents strictly preferring $j'$ again lie on one side of the median and number at most $(n-1)/2$; more than $n/2$ agents weakly prefer $j$, symmetrically more than $n/2$ weakly prefer $j'$, so $j$ and $j'$ beat each other, which the hypothesis excludes. Hence $j$ beats $j'$ exactly when the median type strictly prefers $j$, so the majority relation is the median type's order, a strict linear order. The ranking of $U(\tilde{u}, \cdot)$ is then the only ranking attaining CC, and the equivalence with anchoring at $\tilde{u}$ follows as in the quality case. This uses strict crossing and the order alone. □

# F Proofs for Section 3

## F.1 Zero-risk models and identifiable anchors

**Lemma 1** (Zero-risk models, Anonymous, 2026)**.** *Let $E = (u, q, v)$ be an instance. Then*

$$\Theta(R(E)) \;=\; \big\{\,(c\,(u-\gamma),\; q+\gamma v,\; v/c) : \gamma, c \in \mathbf{R},\; c \neq 0\,\big\},$$

*and the reconstructed quality of the model with parameters $(\gamma, c)$ is the utility of type $\gamma$, $U(\gamma, \cdot)$. Equivalently, the zero-risk models are the instances utility-equivalent to $E$: the shifts $E_\gamma$ of* (2) *combined with the scalings $(cu,\, q,\, v/c)$, which also preserve every utility.*

*Proof.* Let agents $i, i'$ rate item $j_0$ differently, so $u_i \neq u_{i'}$ and $v_{j_0} \neq 0$. Subtracting the zero-risk identities of $i$ and $i'$ at an arbitrary item $j$ gives $(\hat{u}_i - \hat{u}_{i'})\,\hat{v}_j = (u_i - u_{i'})\,v_j$. At $j_0$ the right-hand side is nonzero, so $\hat{u}_i \neq \hat{u}_{i'}$; setting $c = (\hat{u}_i - \hat{u}_{i'})/(u_i - u_{i'}) \neq 0$ yields $\hat{v}_j = v_j/c$ for every $j$. Zero risk at a pair $(k, j)$ then reads $\hat{q}_j = q_j + (u_k - \hat{u}_k/c)\,v_j$. Applied at $j_0$, it forces $u_k - \hat{u}_k/c = (\hat{q}_{j_0} - q_{j_0})/v_{j_0}$, a constant $\gamma$ independent of $k$; hence $\hat{u}_k = c\,(u_k - \gamma)$ for every agent and $\hat{q}_j = q_j + \gamma\,v_j = U(\gamma, j)$ for every item. Conversely, any $(\gamma, c)$ reproduces every rating, since $\hat{q}_j + \hat{u}_i\hat{v}_j = q_j + \gamma v_j + (u_i - \gamma)v_j = r_{ij}$. □

**Theorem 2** (Zero-risk mechanisms are anchored)**.** *A mechanism $f$ is zero-risk if and only if for every instance $E$ there is a type $\gamma \in \mathbf{R}$ such that $f$ is anchored at $\gamma$ on $E$.*

*Proof.* Let $f$ be zero-risk and $E$ an instance. The output $f(R(E))$ is the ranking by $\hat{q}$ of some model in $\Theta(R(E))$, and by Lemma 1 that model is $(c(u-\gamma),\, q+\gamma v,\, v/c)$ for some $\gamma$ and $c \neq 0$, with $\hat{q} = U(\gamma, \cdot)$. So $f(R(E)) = \pi_\gamma(E)$, and $f$ is anchored at $\gamma$ on $E$. Conversely, let $f$ be anchored at $\gamma$ on $E$. Then $f(R(E)) = \pi_\gamma(E)$ is the ranking by $U(\gamma, \cdot)$, which by Lemma 1 is the reconstructed quality $\hat{q}$ of every zero-risk model $(c(u-\gamma),\, q+\gamma v,\, v/c)$, $c \neq 0$. So the output is the ranking by $\hat{q}$ of a zero-risk model, and $f$ is zero-risk. □

**Theorem 3** (Identifiable anchor means implementable mechanism)**.** *A statistic $a$ is identifiable if and only if some mechanism is $a$-anchored on every instance. The mechanism ranks any zero-risk model of $R(E)$ from $a$, that is, by $\hat{q}_j + a(\hat{u})\,\hat{v}_j$.*

*Proof.* Utility-equivalent instances have the same report matrix, so a mechanism outputs the same ranking on both. If some mechanism $f$ is $a$-anchored, then on utility-equivalent $E$ and $E'$ the common output $f(R(E))$ equals both $\pi_{a(u)}(E)$ and $\pi_{a(u')}(E')$, so the two oracle rankings agree and $a$ is identifiable. Conversely, let $a$ be identifiable. For a report matrix $R$ generated by some instance, define $f(R)$ as $\pi_{a(u')}(E')$ for any instance $E' = (u', q', v')$ with $R(E') = R$, and arbitrarily on other matrices. Instances with the same report matrix are utility-equivalent, so by identifiability the choice does not matter, and $f(R(E)) = \pi_{a(u)}(E)$ on every instance. The instances generating $R$ are its zero-risk models, so $f$ ranks any zero-risk model $(\hat{u}, \hat{q}, \hat{v})$ of $R$ by $\hat{q}_j + a(\hat{u})\,\hat{v}_j$. □

**Proposition 1** (Which anchors are identifiable)**.** *(a) The neutral anchor, the constant statistic $a \equiv 0$, is not identifiable on instances in which two items differ in leaning.*

*(b) A statistic $a$ of the types is identifiable if and only if $a(c\,u + b) = c\,a(u) + b$ for all nondegenerate $u \in \mathbf{R}^n$, $b \in \mathbf{R}$ and $c \neq 0$. The mean and, for odd $n$, the median are identifiable. Fixed types and order statistics other than the median are not.*

*Proof.* (a) On the instance $E_{\gamma,c}$ the neutral type is $\gamma$ of $E$, so $\pi_0(E_{\gamma,c}) = \pi_\gamma(E)$ ranks by $U(\gamma, \cdot)$. For items $j, j'$ with $v_j \neq v_{j'}$, $U(\gamma, j) - U(\gamma, j') = (q_j - q_{j'}) + \gamma(v_j - v_{j'})$ takes both signs as $\gamma$ ranges over $\mathbf{R}$, so two instances of the family order the pair oppositely. Choosing $\gamma$ with the difference strictly positive in one instance and strictly negative in the other, the two instances order $j$ and $j'$ strictly oppositely by quality, so a mechanism, which gives one output on their common report matrix, attains OQ on at most one of them.

(b) On $E_{\gamma,c}$ the statistic is $a(c(u-\gamma))$ and the oracle ranks by $q_j + \gamma v_j + a(c(u-\gamma))\,v_j/c$. If $a(c(u-\gamma)) = c\,(a(u) - \gamma)$, this equals $q_j + a(u)v_j = U(a(u), j)$ on the whole family, so $a$ is identifiable. Conversely, suppose $\gamma + a(c(u-\gamma))/c \neq a(u)$ for some nondegenerate $u$, some $\gamma$ and some $c \neq 0$. The oracle ranks $E$ by $U(a(u), \cdot)$ and $E_{\gamma,c}$ by $U(\gamma', \cdot)$ with $\gamma' \neq a(u)$; take an instance with these agents and two items whose crossing $(q_{j'} - q_j)/(v_j - v_{j'})$ lies strictly between $\gamma'$ and $a(u)$, and the two order the pair oppositely. The mean commutes with $u \mapsto cu + b$; for odd $n$ so does the median, since a decreasing relabeling reverses the order of the reports and fixes the middle one. A constant fails at $b \neq 0$. The $k$-th order statistic becomes the $(n{+}1{-}k)$-th under $c < 0$, so it fails unless $k = (n{+}1)/2$. □

**Theorem 4** (Re-anchoring)**.** *Let $E$ be an instance, let $(\hat{u}, \hat{q}, \hat{v}) \in \Theta(R(E))$ be any zero-risk model, and let $a$ be the mean or, for odd $n$, the median. Then $\hat{q}_j + a(\hat{u})\,\hat{v}_j = U(a(u), j)$ for every item $j$. Thus, ranking by this* re-anchored *quality attains SW, respectively CC, under the hypotheses of Theorem 1.*

*Proof.* By Lemma 1, $\hat{q} = q + \gamma v$, $\hat{u} = c\,(u - \gamma)$ and $\hat{v} = v/c$; by Proposition 1(b), $a(\hat{u}) = c\,(a(u) - \gamma)$; so $\hat{q}_j + a(\hat{u})\hat{v}_j = q_j + \gamma v_j + (a(u) - \gamma)v_j = U(a(u), j)$, and Theorem 1 gives the objective. □

### F.2 General forms

On the general class there are no parameters to fit, and the condition that generalizes zero risk is stated on the output.

**Definition 10** (Compatible mechanism)**.** A mechanism is *compatible* with a class $\mathcal{H} \subseteq \mathcal{C}$ if, for every $E \in \mathcal{H}$, its output on $R(E)$ is the quality ranking, or its top $k$, of some instance $E' \in \mathcal{H}$ with $R(E') = R(E)$.

In the linear model a fit generates the ratings exactly when its risk is zero, so compatible and zero-risk coincide. The following statement is the general form of Theorem 2.

**Proposition 7** (Compatible and anchored mechanisms)**.** *Let $\mathcal{H} \subseteq \mathcal{C}$ be closed under designation.*

*(i) A mechanism anchored at some type on every instance of $\mathcal{H}$ is compatible with $\mathcal{H}$.*

*(ii) Every mechanism compatible with $\mathcal{H}$ is anchored on every instance of $\mathcal{H}$ if and only if instances of $\mathcal{H}$ with the same reports have the same set of type rankings, $R(E) = R(E') \Rightarrow \{\pi_t(E) : t \in \mathcal{T}\} = \{\pi_t(E') : t \in \mathcal{T}\}$.*

*(iii) The linear model satisfies the condition of (ii) at every report matrix in which two agents rate some item differently.*

*(iv) The class $\mathcal{C}$ does not, even with monotone differences.*

*Proof.* (i) If the output on $R(E)$ is $\pi_t(E)$, it is the quality ranking of the designation $E_t$, which lies in $\mathcal{H}$ and generates $R(E)$.

(ii) Suppose the condition holds and $f$ is compatible. On $E$ the output is the quality ranking of some $E'$ with $R(E') = R(E)$, which by alignability is $\pi_{t'}(E')$ for a neutral type $t'$ of $E'$, hence equals $\pi_t(E)$ for some $t$ by the condition. Suppose the condition fails: $R(E) = R(E')$ and some ranking $\rho = \pi_{t'}(E')$ is held by no type of $E$. The mechanism that outputs $\rho$ on this matrix, elsewhere the quality ranking of any instance generating the reports, and anything on matrices no instance generates, is compatible, since $\rho$ is the quality ranking of $E'_{t'}$, and it is not anchored on $E$.

(iii) By Lemma 1, two linear instances with the same reports have utilities $U'(\varphi(t), j) = U(t, j)$ for the affine bijection $\varphi(t) = c\,(t - \gamma)$, so $\pi_{\varphi(t)}(E') = \pi_t(E)$ for every $t$.

(iv) Take $\mathcal{T} = [0, 1]$, two agents at 0 and 1, and items $A, B, C$ with $U(t, A) = 3 - 2t$ and $U(t, C) = 1 + 2t$ in both instances. In $E$ let $U(t, B)$ be the piecewise linear function through $(0, 2)$, $(\frac{1}{4}, 2.5)$, $(\frac{3}{4}, 2.5)$, $(1, 2)$; in $E'$ let it pass through $(0, 2)$, $(\frac{1}{4}, 1.5)$, $(\frac{3}{4}, 1.5)$, $(1, 2)$. Both instances report the rows $(3, 2, 1)$ and $(1, 2, 3)$. In $E$ the differences $A - B$, $B - C$ and $A - C$ are each nonincreasing with one zero, at $\frac{1}{4}$, $\frac{3}{4}$ and $\frac{1}{2}$, so the types of $E$ rank $ABC$, $BAC$, $BCA$, $CBA$ in turn. In $E'$ the zeros of $A - B$ and $B - C$ are exchanged and the types rank $ABC$, $ACB$, $CAB$, $CBA$. Both instances satisfy strict crossing and have monotone differences, and any row may be designated as quality. A mechanism that outputs $ACB$ on these reports is compatible, by (ii), and is not anchored when $E$ is the true instance. □

**Proposition 8** (Zero risk is undominated)**.** *Restrict to linear instances in which two agents rate some item differently and no two items tie in quality. No mechanism attains OQ on a strictly larger set of instances than $f$ if and only if, on every report matrix $R(E)$, $f$ outputs the quality ranking of a zero-risk model of $R(E)$ in which no two items tie in quality.*

*Proof.* Write $S_f$ for the set of instances on which $f$ attains OQ; without ties, $f$ attains OQ on $E$ exactly when $f(R(E))$ is the quality ranking of $E$. Suppose $f$ has the stated property and $S_g \supsetneq S_f$. Take $E \in S_g \setminus S_f$ and $R = R(E)$. By hypothesis $f(R)$ is the quality ranking of some zero-risk model of $R$ without quality ties, which by Lemma 1 has the quality of $E_\gamma = (u - \gamma,\, q + \gamma v,\, v)$ for some $\gamma$; so $E_\gamma$ is in the domain, $R(E_\gamma) = R$, and $E_\gamma \in S_f \subseteq S_g$. Then $g(R)$ is the quality ranking of $E_\gamma$, so $g(R) = f(R)$ and $g$ attains OQ on $E$ exactly when $f$ does, a contradiction. Suppose that on some $R = R(E)$ the output is the quality ranking of no zero-risk model of $R$ without quality ties. Every instance of the domain generating $R$ has the quality of some $E_\gamma$ by Lemma 1, with no ties, so $f$ attains OQ on none of them. Let $g$ agree with $f$ except on $R$, where it outputs the quality ranking of $E$; then $S_g \supseteq S_f \cup \{E\}$. □

On the general class an anchor is a map $\alpha$ from instances to types, such as the median of the agents or the neutral type, and it is *identifiable* on $\mathcal{H}$ if $R(E) = R(E')$ implies $\pi_{\alpha(E)}(E) = \pi_{\alpha(E')}(E')$ for $E, E' \in \mathcal{H}$, as in Definition 6. The neutral anchor is not identifiable on any class closed under designation in which two types order some pair strictly oppositely: if $t$ reverses a pair relative to the neutral type $t_0$ of $E$, then $E$ and its designation $E_t$ have the same reports (Proposition 4) and order the pair oppositely by quality, so no mechanism attains OQ on both. The median is identifiable, and the following statements say what else is.

**Lemma 3** (Identifiable anchor means implementable mechanism, general form)**.** *Let $\mathcal{H} \subseteq \mathcal{C}$ and let $\alpha$ be an anchor. Then $\alpha$ is identifiable on $\mathcal{H}$ if and only if some mechanism is anchored at $\alpha(E)$ on every $E \in \mathcal{H}$.*

*Proof.* Instances with the same report matrix give a mechanism the same input. If $f$ is anchored at $\alpha(E)$ on every $E \in \mathcal{H}$, then for $E, E' \in \mathcal{H}$ with $R(E) = R(E')$ the common output equals both

$\pi_{\alpha(E)}(E)$ and $\pi_{\alpha(E')}(E')$, so $\alpha$ is identifiable. Conversely, let $\alpha$ be identifiable. For a report matrix $R$ generated by some instance of $\mathcal{H}$, define $f(R)$ as $\pi_{\alpha(E')}(E')$, or its top $k$, for any $E' \in \mathcal{H}$ with $R(E') = R$, and arbitrarily on other matrices. By identifiability the choice of $E'$ does not matter, so $f(R(E)) = \pi_{\alpha(E)}(E)$ on every $E \in \mathcal{H}$. □

The mechanism of the proof is re-anchoring in general form: find any instance of the class that reproduces the reports and rank from the anchor computed in that instance. The next lemma computes it from the reports for the median and the mean.

**Lemma 4** (The median and mean anchors are identifiable)**.** *Let $R$ be a report matrix with $n$ rows, $n$ odd. (i) If an admitted instance generates $R$, every such instance gives its median type the same ranking, whose sign on each pair $j, j'$ is the median of the $n$ signs of $r_{ij} - r_{ij'}$. Hence the median oracle is implementable among admitted instances. (ii) If a cardinal-linear instance generates $R$, the column averages of $R$ are the mean type's utility row. Hence the mean oracle is implementable among cardinal-linear instances.*

*Proof.* (i) Let $E$ be an admitted instance generating $R$: listing its agents' types in the order of $\mathcal{T}$ as $t_1 \le \cdots \le t_n$ and renumbering the rows accordingly, $r_{ij} = U(t_i, j)$. Fix items $j, j'$ and write $g_i = r_{ij} - r_{ij'} = U(t_i, j) - U(t_i, j')$. If the two items have identical utility profiles, every $g_i$ is zero; otherwise strict crossing makes the signs of $t \mapsto U(t, j) - U(t, j')$ monotone along $\mathcal{T}$, so the sequence $(\operatorname{sgn} g_1, \ldots, \operatorname{sgn} g_n)$ is monotone, duplicate types included. A monotone sequence is a sorted arrangement of its values, so its middle element is the median of the multiset $\{\operatorname{sgn}(r_{ij} - r_{ij'})\}_i$, a quantity computed from $R$ alone. The median type of $E$ is $t_{(n+1)/2}$, and its sign on the pair is that middle element. Hence every admitted instance generating $R$ gives its median type the same sign on every pair; the signs are one utility row's, so they induce a ranking, and all such instances induce the same one: the median-anchored ranking is identifiable at $R$. The mechanism that outputs the ranking whose pair signs are the medians of the report signs is therefore well defined from the reports, and coincides with the median-anchored mechanism on every admitted instance. Strict crossing is not dispensable (Remark 1). (ii) By cardinal linearity, $\frac{1}{n}\sum_i r_{ij} = \frac{1}{n}\sum_i U(u_i, j) = U(\bar{u}, j)$, so the column averages are the mean type's row and their ranking is the mean-anchored ranking. □

**Proposition 9** (Only the agents' own rankings are identifiable)**.** *Let $\mathcal{H}$ be $\mathcal{C}$ or its subclass with monotone differences, and let $\alpha$ be identifiable on $\mathcal{H}$. Then on every $E \in \mathcal{H}$ in which no agent is indifferent between two items, $\pi_{\alpha(E)}(E)$ is the ranking of one of the agents.*

*Proof.* Let $E = (A, X, U, q) \in \mathcal{H}$ have no agent indifferent between two items, with agents $u_{(1)} \le \cdots \le u_{(n)}$. For $t \in \mathcal{T}$ let $\rho(t)$ be the largest agent type at or below $t$, and $\rho(t) = u_{(1)}$ when there is none. Let $E^s = (A, X, U^s, q^s)$ with $U^s(t, \cdot) = U(\rho(t), \cdot)$ and $q^s = U(u_{(1)}, \cdot)$. Then $U^s(u_i, \cdot) = U(u_i, \cdot)$, so $R(E^s) = R(E)$. Since $\rho$ is nondecreasing, each difference $g^s = g \circ \rho$ runs through the values of $g$ at the agents in order, with repeats: it never vanishes, its signs are monotone, and it is monotone whenever $g$ is. So $E^s \in \mathcal{H}$. Identifiability gives $\pi_{\alpha(E)}(E) = \pi_{\alpha(E^s)}(E^s)$, and every type of $E^s$ ranks as the agent $\rho(\alpha(E^s))$ does. □

The linear form is what makes positions between the agents available: Proposition 1(b) admits the mean, which Proposition 9 excludes on the general class.

### F.3 Mechanisms anchored by direct computation

**Proposition 10** (Per-item aggregates are anchored)**.** *In the LFUM, under truthful complete reports: (i) a rule that scores item $j$ by an affine combination of its ratings, $\sum_i \lambda_i r_{ij}$ with $\sum_i \lambda_i = 1$ and weights fixed across items, is anchored at $\gamma = \sum_i \lambda_i u_i$; (ii) for odd $n$, the per-item median is anchored at the median type $\tilde{u}$; (iii) conversely, provided two agents' types differ, every $\gamma \in \mathbf{R}$ arises as in (i).*

*Proof.* (i) $\sum_i \lambda_i r_{ij} = \sum_i \lambda_i (q_j + u_i v_j) = q_j + \big(\sum_i \lambda_i u_i\big) v_j = U(\gamma, j)$. (ii) Per item, $u \mapsto q_j + u v_j$ is monotone in $u$, increasing, decreasing, or constant, and for odd $n$ a monotone map preserves the

middle element, so $\mathrm{med}_i\, r_{ij} = q_j + \tilde{u}\, v_j = U(\tilde{u}, j)$. (iii) Any $\gamma$ is an affine combination of two distinct types. □

Not every per-item statistic qualifies: the per-item maximum tracks $\max_i u_i$ on items with $v_j > 0$ but $\min_i u_i$ on items with $v_j < 0$, so no single anchor reproduces it, and the same failure affects every quantile except the median.

**Remark 3** (Even $n$)**.** For even $n$ the agents have two middle types, $u_{(n/2)}$ and $u_{(n/2+1)}$, writing $u_{(k)}$ for the $k$-th smallest. Either one is a generalized median in Moulin's sense (Moulin, 1980), hence strategyproof in the sense of Section 4.1. The Condorcet claim of Theorem 1 weakens to weak Condorcet consistency: the favorite of the agent at $u_{(n/2)}$ beats or ties every other item, by one-sided agreement. Theorem 4 needs one more input, the orientation of the fitted axis. Under Lemma 1 with $c < 0$ the fitted order reverses, $\hat{u}_{(k)} = c\,(u_{(n+1-k)} - \gamma)$, so a fixed order statistic of $\hat{u}$ is the mirror-image statistic of $u$. The central statistic $k = (n+1)/2$ is its own mirror image, so odd $n$ needs no sign condition in the proof of Theorem 4. With $n$ even, re-anchoring at $\hat{u}_{(n/2)}$ anchors at $u_{(n/2)}$ when $c > 0$ and at $u_{(n/2+1)}$ when $c < 0$. The ratings do not determine the sign of $c$, since the reflection $(\hat{u}, \hat{v}) \mapsto (-\hat{u}, -\hat{v})$ of any zero-risk fit is a zero-risk fit; fixing the orientation takes information beyond the ratings. The midpoint of the two middle types is its own mirror image, so re-anchoring at it needs no orientation, but it is not a generalized median and is manipulable; for even $n$ identifiability and strategyproofness come apart.

**Remark 4** (Modified factorizations)**.** The argument covers a functional form other than the LFUM's whenever it admits a rating-preserving shift of the neutral position. Quality-sensitive matrix factorization (Goyal et al., 2026) scales each rater's response to quality by a sensitivity $\rho_i \geq 0$, predicting $\rho_i \hat{q}_j + \hat{u}_i \hat{v}_j$ up to intercepts. The shift $\hat{u}_i \mapsto \hat{u}_i - g\rho_i$, $\hat{q}_j \mapsto \hat{q}_j + g\hat{v}_j$ preserves every predicted rating for every $g \in \mathbf{R}$, so its quality is determined only up to an anchor. Its $L_2$ penalties select one, a shrunk sensitivity-weighted combination of the reported types (Appendix F.4). Re-anchoring at the mean (Theorem 4) applies to its fit, since the shift moves the mean fitted type by $g$ times the mean sensitivity, which the normalization sets to one; re-anchoring at the median does not, since the shift moves raters by different amounts.

### F.4 Quality-sensitive factorization: gauge and anchor

Quality-sensitive matrix factorization (Goyal et al., 2026) predicts $\hat{r}_{ij} = \mu + \alpha_i + \rho_i \hat{q}_j + \hat{u}_i \hat{v}_j$ with $\rho_i \geq 0$, and minimizes squared error plus $L_2$ penalties, among them $\lambda_u \sum_i \hat{u}_i^2$ and $\lambda_q \sum_j \hat{q}_j^2$, with $\rho$ penalized toward one and normalized to mean one after the fit. For every $g \in \mathbf{R}$ the map $\hat{u}_i \mapsto \hat{u}_i - g\rho_i$, $\hat{q}_j \mapsto \hat{q}_j + g\hat{v}_j$ leaves every $\hat{r}_{ij}$ unchanged, since $\rho_i(\hat{q}_j + g\hat{v}_j) + (\hat{u}_i - g\rho_i)\hat{v}_j = \rho_i \hat{q}_j + \hat{u}_i \hat{v}_j$, and leaves $\rho$, the intercepts, and the sign constraint unchanged as well. Along this orbit only two penalty terms vary, $\lambda_u \sum_i (\hat{u}_i - g\rho_i)^2 + \lambda_q \sum_j (\hat{q}_j + g\hat{v}_j)^2$, whose minimizer is

$$g^* = \frac{\lambda_u \sum_i \rho_i \hat{u}_i - \lambda_q \sum_j \hat{v}_j \hat{q}_j}{\lambda_u \sum_i \rho_i^2 + \lambda_q \sum_j \hat{v}_j^2}.$$

In the regime of the companion paper's analysis, where the regularized solution converges to the minimum-penalty point of the zero-risk orbit and item quality is uncorrelated with item leaning (Anonymous, 2026), the second term of the numerator vanishes at the true parameters, and at their scale

$$g^* = s\, \frac{\sum_i \rho_i u_i}{\sum_i \rho_i^2}, \qquad s = \frac{\lambda_u \sum_i \rho_i^2}{\lambda_u \sum_i \rho_i^2 + \lambda_q \sum_j v_j^2} \in (0, 1),$$

which has the form $s\,\bar{u}$ of the companion paper when $\rho \equiv 1$. The penalties also select the scale, through the rescalings $(\rho, \hat{q}) \mapsto (c\rho, \hat{q}/c)$ and $(\hat{u}, \hat{v}) \mapsto (c\hat{u}, \hat{v}/c)$, and the companion paper's shrinkage factor accounts for both; in the running example the shift alone gives $0.57$ and the joint optimum $0.49$ (Appendix C).

### F.5 Noisy ratings

With noisy ratings no fit has zero risk. Let the loss of a fit $\theta$ against a rating matrix $R$ be $\ell(P(\theta), R)$ for a function $\ell$ of the predicted matrix $P(\theta) = \mathbf{1}\hat{q}^\top + \hat{u}\hat{v}^\top$ alone, as squared error is. Let $P^* \in \arg\min\{\ell(P, R) : P = P(\theta) \text{ for some } \theta\}$ be a best prediction.

**Proposition 11** (Minimum-risk fits)**.** *If the best prediction $P^*$ is unique, the minimum-risk fits of $R$ are exactly the zero-risk fits of $P^*$. If two rows of $P^*$ differ on some column, they form the family of Lemma 1 for the instance $(u^*, q^*, v^*)$ of any one of them, so a mechanism that ranks by the fitted quality of a minimum-risk fit is anchored at some type of the fitted instance, and Theorem 2 and Theorem 4 hold with $P^*$ in place of $R$.*

*Proof.* The loss depends on $\theta$ only through $P(\theta)$, so $\theta$ minimizes it exactly when $P(\theta) = P^*$, which is zero risk against $P^*$. The rest is Lemma 1 applied to $P^*$. □

For complete ratings and squared loss, center the fitted types at zero by moving their mean into $\hat{q}$. The residual then splits as $\|\mathbf{1}(\bar{r} - \hat{q})^\top\|^2 + \|R_c - \hat{u}\hat{v}^\top\|^2$, where $\bar{r}$ is the vector of column means and $R_c$ the column-centered matrix, since the columns of $R_c - \hat{u}\hat{v}^\top$ are orthogonal to $\mathbf{1}$. So $q^* = \bar{r}$, and $u^*v^{*\top}$ is the best rank-one approximation of $R_c$, unique when the two largest singular values of $R_c$ differ. Two consequences hold under every noise model. Proposition 1(a) holds at every sample size, since two fits with the same predicted ratings have the same loss. And the anchoring is on the fitted parameters, which coincide with the true ones only without noise; the distance between the two is the estimation question of the companion paper (Anonymous, 2026). With partial ratings the uniqueness of $P^*$ on the observed entries is an assumption, stated together with the condition of Appendix F.6.

### F.6 Partial ratings

Completeness of the ratings is used only for identification, in Lemma 1 and the results that rest on it. It matters in a different way for the mechanisms of Proposition 10. With raters $N_j$ of item $j$, the per-item average scores $j$ at $q_j + \bar{u}_{N_j} v_j$, a position that depends on the item, and no single anchor reproduces it unless every $N_j$ has the same mean type.

The condition that replaces completeness rests on one observation. For a non-neutral item, zero risk makes the reconstructed types of its raters one affine function of their true types. Two items that share two raters of distinct type share that function.

**Definition 11** (Framed ratings)**.** Start from a non-neutral item with two raters of distinct type and call its raters *framed*. Any non-neutral item with two framed raters of distinct type frames all of its raters. The ratings are *framed* if this closure reaches every agent, every item has two framed raters with distinct types, and every agent rates at least one non-neutral item.

**Lemma 5** (Zero-risk models under framed ratings)**.** *Let $E = (u, q, v)$ be an instance whose agents rate only some items, and let the ratings be framed. Then the zero-risk models of the observed ratings are exactly the models $(c(u - \gamma),\ q + \gamma v,\ v/c)$ with $\gamma \in \mathbf{R}$ and $c \neq 0$, as in Lemma 1.*

*Proof.* A model $(\hat{u}, \hat{q}, \hat{v})$ has zero risk when $\hat{q}_j + \hat{u}_i\hat{v}_j = q_j + u_i v_j$ for every rated pair $(i, j)$. Let $j$ be a non-neutral item with raters $i, i'$ of distinct type. Subtracting their two conditions gives $(\hat{u}_i - \hat{u}_{i'})\hat{v}_j = (u_i - u_{i'})v_j \neq 0$, so $\hat{v}_j \neq 0$, and every rater $i$ of $j$ satisfies $\hat{u}_i = A_j(u_i)$ for the affine map $A_j(t) = \big(q_j - \hat{q}_j + tv_j\big)/\hat{v}_j$, whose slope $v_j/\hat{v}_j$ is nonzero. Let $A$ be the map of the item that starts the closure in Definition 11, so that every framed agent so far satisfies $\hat{u}_i = A(u_i)$. An item that frames its raters has two framed raters of distinct type, at which its map agrees with $A$; two affine maps that agree at two points are equal, so the item's map is $A$ and its raters satisfy $\hat{u}_i = A(u_i)$. Since the closure reaches every agent, $\hat{u} = A(u)$ for all agents. Write $A(t) = c(t - \gamma)$ with $c \neq 0$. Every item $j$ has two raters of distinct type, and for its raters the zero-risk condition reads $\hat{q}_j + c(u_i - \gamma)\hat{v}_j = q_j + u_i v_j$. Both sides are affine in $u_i$ and agree at two distinct values, so $c\hat{v}_j = v_j$ and $\hat{q}_j - c\gamma\hat{v}_j = q_j$, that is, $\hat{v}_j = v_j/c$ and $\hat{q}_j = q_j + \gamma v_j$. Conversely, every such model reproduces every rating, observed or not (Lemma 1). □

By Lemma 5, Theorem 2 and Theorem 4 hold under framed ratings with the same proofs. The zero-risk conditions over the rated pairs leave exactly the two free parameters $\gamma$ and $c$. Framed ratings therefore need at least $n + 2m - 2$ ratings, against $nm$ for complete ratings. The count is attained when two agents of distinct type rate every item and every other agent rates one non-neutral item.

Connectivity of the rating graph is not enough. Suppose item $j$ is rated by agents 1 and 2, and item $j'$ by agents 2 and 3. The graph is connected, but $\hat{u}_3$ is free and $\hat{q}_{j'} - q_{j'} = u_2 v_{j'}\big(1 - \frac{u_3 - u_2}{\hat{u}_3 - u_2}\big)$ can

take any value, so the zero-risk model is not anchored. The same failure occurs when two blocks of raters overlap only on neutral items, or when the raters of an item all share one type.

**Remark 5** (Spoofing under partial ratings)**.** Suppose the ratings remain framed when a deviating agent, or a coalition, is removed. A spoofed report on the items the deviator rates is the truthful partial report of the reported instance $E_w$, and those ratings are framed as well, since each non-neutral item the deviator rates has two framed raters of distinct type among the others. By Lemma 5 a mechanism that re-anchors at the median is then anchored at the median of the reported types, and Theorem 5 applies as stated. The hypothesis rules out items rated only by the coalition, on which the median offers no guarantee.

Under partial ratings the per-item aggregates are not anchored, so the mean and the median are reached only through a zero-risk model followed by re-anchoring, which gives factorization a role even without noise. A truthful agent is placed at its type no matter which items it rates, as long as the ratings stay framed. Choosing which items to rate therefore moves the anchor only when it breaks the framing condition.

## G Proofs for Section 4.1

The median anchor is not a dictatorship: no single agent's report determines the outcome. On unrestricted preferences the Gibbard–Satterthwaite theorem states that a strategyproof rule with at least three possible outcomes must be a dictatorship (Gibbard, 1973), so a non-dictatorial strategyproof rule can survive only on a restricted domain. Strict crossing is that restriction: it makes each agent's stake in the anchor single-peaked (Lemma 7), the loophole Gibbard–Satterthwaite leaves open.

Single-peakedness makes the spoofing game tractable: an agent's only interest is to pull the anchor toward its own type, and an anchor rule is strategyproof exactly when no agent can do so by misreporting. Moulin's characterization of strategyproof rules on single-peaked domains, combined with the gauge freedom the ratings leave open (Proposition 1), then determines which identifiable anchors are strategyproof (Theorem 6).

Throughout, a *tie-breaking rule* is a fixed strict priority order on items, applied whenever scores tie.

**Lemma 6** (Exchanged items cross exactly once)**.** *In an anchored utility model, fix the items, the selection size $k$, and a tie-breaking rule. Let $\gamma < \gamma'$ be anchors and let $j' \in S(\gamma) \setminus S(\gamma')$ and $j \in S(\gamma') \setminus S(\gamma)$ be any pair exchanged between the selections. Then $g(t) = U(t, j) - U(t, j')$ is not identically zero, satisfies $g(t) < 0$ for every $t < \gamma$ and $g(t) > 0$ for every $t > \gamma'$, and vanishes at no more than one type, a crossing $\gamma^* \in [\gamma, \gamma']$, with $\gamma^* = \gamma$ when $g$ has no zero; in the LFUM, $g$ is affine and nonconstant with unique zero $\gamma^*$. In particular an agent with $u_i \le \gamma$ weakly prefers the leaving item $j'$ to the entering item $j$, strictly if $u_i < \gamma^*$; the statement for $u_i \ge \gamma'$ is the mirror image.*

*Proof.* At anchor $\gamma$ the item $j'$ is selected and $j$ is not, so the tie-broken ranking places $j'$ above $j$: either $g(\gamma) < 0$, or $g(\gamma) = 0$ and the priority order favors $j'$. At $\gamma'$ the ranking places $j$ above $j'$: either $g(\gamma') > 0$, or $g(\gamma') = 0$ and the priority order favors $j$. If $g$ were identically zero, the pair would be ordered by the fixed priority alone, identically at both anchors, contradicting the flip; by strict crossing, $g$ has at most one zero and its signs are monotone along the type order. The signs are nondecreasing: $g(\gamma) \le 0 \le g(\gamma')$, and not both are zero, since that would be two zeros and the one priority order cannot favor $j'$ at $\gamma$ and $j$ at $\gamma'$. Monotone signs with a single zero give $g < 0$ strictly below the zero and $g > 0$ strictly above it, and place the zero in $[\gamma, \gamma']$; set $\gamma^*$ to the zero if one exists, and $\gamma^* = \gamma$ otherwise, so that in every case $g < 0$ strictly below $\gamma$ and $g > 0$ strictly above $\gamma'$. For $u_i \le \gamma \le \gamma^*$ this gives $g(u_i) \le 0$, so $U(u_i, j') \ge U(u_i, j)$, strictly when $u_i < \gamma^*$. □

**Lemma 7** (Single-peakedness over the anchor)**.** *Let $E$ be an admitted instance (Definition 9), fix a selection size $k$ and a tie-breaking rule, and for an agent of type $u_i$ let $V_i(\gamma) = \sum_{j \in S(\gamma)} U(u_i, j)$ be its utility from the selection $S(\gamma)$ at anchor $\gamma$. Then $V_i$ is single-peaked in $\gamma$ with peak at $\gamma = u_i$: weakly increasing for $\gamma \le u_i$ and weakly decreasing for $\gamma \ge u_i$. The same holds under monotone differences in place of strict crossing.*

*Proof.* Let $u_i \le \gamma < \gamma'$. The sets $S(\gamma) \setminus S(\gamma')$ and $S(\gamma') \setminus S(\gamma)$ have the same size; pair them by any bijection. By Lemma 6 agent $i$ weakly prefers each leaving item to its partner, so $V_i(\gamma) \ge V_i(\gamma')$.

The case $\gamma < \gamma' \leq u_i$ is the mirror image. Under monotone differences each pair's difference $g = U(\cdot, j) - U(\cdot, j')$ is nondecreasing with $g(\gamma) \leq 0$ (Lemma 10), so $g(u_i) \leq 0$ and the same comparison holds. □

Parts (a) and (b) of Theorem 5 use only these weak comparisons and need no tie hypothesis.[2]

**Lemma 8** (Spoofs are the undetectable deviations)**.** *Let the agents other than $i^*$ report truthfully, with two of them of distinct type and some $v_{j_0} \neq 0$. The reported matrix admits a zero-risk fit if and only if $i^*$'s row is $q_j + w\, v_j$ for some $w \in \mathbf{R}$. The same holds for a coalition when the agents outside it include two distinct types.*

*Proof.* Restrict a zero-risk fit to the rows $i \neq i^*$. It is a zero-risk fit of a truthful complete instance, so by Lemma 1 $\hat{q}_j = q_j + \gamma v_j$ and $\hat{v}_j = v_j/c$. Zero risk on row $i^*$ reads $r'_j = \hat{q}_j + \hat{u}_{i^*}\hat{v}_j = q_j + (\gamma + \hat{u}_{i^*}/c)v_j$, the row of $w = \gamma + \hat{u}_{i^*}/c$. Conversely a spoof of $w$ is the truthful matrix of the instance with $u_{i^*}$ replaced by $w$. □

**Lemma 9** (Detecting and identifying misreports)**.** *Let the agents have distinct types and $v \neq 0$, and let a set $K$ of $k$ agents report rows that are not the row of any type, the other $n - k \geq 2$ agents reporting truthfully. Then:*

- *(a) the reported matrix has no zero-risk model;*
- *(b) if $k < (n-1)/2$, the truthful agents form the unique largest set of agents whose rows have a zero-risk model.*

*The bound in (b) is tight.*

*Proof.* A set of rows has a zero-risk model exactly when the rows lie on one line $\{\hat{q} + t\hat{v} : t \in \mathbf{R}\}$ of $\mathbf{R}^m$. The truthful rows are distinct points of the line $L = \{q + tv\}$, since the types are distinct and $v \neq 0$, and the misreported rows lie off $L$. (a) Two truthful rows determine $L$, so no line contains every row. (b) The truthful agents have a zero-risk model, the instance itself. Let $G$ be a set of agents whose rows lie on a line $L'$. If $L' = L$, then $G$ contains no misreporter. Otherwise $L'$ meets $L$ in at most one point, so $G$ contains at most one truthful agent and $|G| \leq k + 1 < n - k$. Hence every set of size at least $n - k$ with a zero-risk model lies on $L$ and consists of truthful agents only, and the truthful agents are the unique largest such set. For tightness, at $k = (n-1)/2$ the misreporters can place their rows on a line through one truthful row, which gives a second set of size $k + 1 = n - k$ with a zero-risk model. □

**Proposition 2** (Mean-anchored mechanisms are manipulable)**.** *Let $f$ be mean-anchored and let the agents other than $i^*$ report truthfully. Then $i^*$ can move the anchor to any target $u^t \in \mathbf{R}$ by reporting as $w_{i^*} = n\,u^t - \sum_{i \neq i^*} u_i$. With target $u^t = u_{i^*}$ the selection is the top $k$ items under its own utility.*

*Proof.* The anchor is the reported mean $\bar{w} = \frac{1}{n}\big(\sum_{i \neq i^*} u_i + w_{i^*}\big) = u^t$, and $U(u_{i^*}, \cdot)$ is the agent's own utility. The gain over truthful reporting is strict whenever $i^*$ values the selections at $\bar{u}$ and at $u_{i^*}$ differently, for instance with $k = 1$ and two items whose utilities cross strictly between $\bar{u}$ and $u_{i^*}$. □

**Theorem 5** (Median-anchored mechanisms are group strategyproof)**.** *Let $n$ be odd and let $f$ be median-anchored. Then:*

- *(a) truthful reporting is a weakly dominant strategy against spoofing: whatever the other agents report, no agent improves its selection by spoofing;*
- *(b) no coalition of agents can jointly spoof so that every member is strictly better off;*
- *(c) if no two items tie in utility for the truthful median type, no coalition can jointly spoof so that every member is weakly better off and at least one strictly.*

[2] Strict crossing cannot be relaxed to single crossing. Take two items whose difference $g(t) = U(t, j) - U(t, j')$ is nonnegative for every type and vanishes only at the truthful median $\tilde{u}$, with the tie there broken toward $j'$: truthfully $j'$ is selected, yet every agent below $\tilde{u}$ strictly prefers $j$ and secures it by reporting far above, moving the median off the tie. Median strategyproofness then fails, so Theorem 5 needs the strict crossing of Definition 9; in the LFUM, where differences are affine, no such touching-without-crossing difference exists.

*Proof.* The anchor is the median of the reported types. Fix the reports of the agents outside the deviating group: arbitrary types in (a), truthful in (b) and (c). Write $\tilde{u}$ for the median of the reported types when the deviators report truthfully and $\tilde{u}'$ for the median after the deviation; by symmetry we may assume $\tilde{u}' \geq \tilde{u}$.

*Step 1: who moved the median.* Before the deviation at most $(n-1)/2$ reports exceed $\tilde{u}$. If $\tilde{u}' > \tilde{u}$ then at least $(n+1)/2$ reports are now at or above $\tilde{u}'$, so some deviator $i^*$ with $u_{i^*} \leq \tilde{u}$ now reports above $\tilde{u}$. In every case $u_{i^*} \leq \tilde{u} \leq \tilde{u}'$.

*Step 2: what the move changes.* If $\tilde{u}' = \tilde{u}$ the selection is unchanged. Otherwise the selections at $\tilde{u}$ and at $\tilde{u}'$ differ by exchanged pairs, an item $j'$ that leaves and an item $j$ that enters. By Lemma 6 the utility difference $U(t, j) - U(t, j')$ is negative for every type $t < \tilde{u}$ and positive for every $t > \tilde{u}'$, with its crossing $\gamma^* \in [\tilde{u}, \tilde{u}']$.

*Step 3: the mover is harmed.* Since $u_{i^*} \leq \tilde{u} \leq \gamma^*$, agent $i^*$ weakly prefers every leaving item to the entering one, so the deviation weakly lowers its utility. This is (a) for a single deviator and (b) for a coalition, since $i^*$ is a member that does not strictly gain. For (c), if $u_{i^*} < \tilde{u}$ the preference is strict by Lemma 6; if $u_{i^*} = \tilde{u}$ it is strict because no exchanged pair ties in utility at $\tilde{u}$. A strict gain by any member requires at least one exchange, which then strictly harms $i^*$. The discussion after this proof shows the no-tie condition cannot be dropped. □

The no-tie condition of (c) cannot be dropped. Suppose two items tie in score exactly at $\tilde{u}$, the priority order favors the item preferred below $\tilde{u}$, and $k = 1$. The median agent and any agent above $\tilde{u}$ form a coalition: the median agent reports far above its type, the median moves up to the next report, the selection flips to the item preferred above $\tilde{u}$, the upper member strictly gains, and the median agent, indifferent between the tied items, is exactly as well off. Every member is weakly better off and one strictly, so (c) fails at the tie; (b) survives, since the median agent does not strictly gain.

**Theorem 6** (The median is the only identifiable strategyproof anchor)**.** *Let $n$ be odd and let $a$ be an identifiable statistic. If the $a$-anchored mechanism is strategyproof against spoofing, anonymous, and efficient, then $a$ is the median.*

*Proof. Step 1: the mechanism is a voting rule on peaks.* By Lemma 7 each agent's utility over the anchor is single-peaked with its peak at the agent's type, and a spoof declares exactly a peak. So the statistic $a$ is a voting rule that maps reported peaks to a point on the line, and we show that strategyproofness against spoofing is equivalent to the strategyproofness of $a$ on the single-peaked domain. A move of the anchor away from the deviator never profits, by single-peakedness. Suppose instead that an agent at $u_i < \gamma$ can move the anchor from $\gamma$ to some $\gamma' < \gamma$. Take an instance with these types and two items whose crossing $\gamma^*$ lies in $(\max(\gamma', u_i), \gamma)$: at $\gamma$ the mechanism selects the item preferred above $\gamma^*$, at $\gamma'$ the item preferred below, and the deviator, below $\gamma^*$, strictly gains. The case $u_i > \gamma$ is symmetric. Hence the mechanism is strategyproof against spoofing on every instance exactly when no agent can move $a$ from $\gamma$ to another point on the same side of $\gamma$ as its type, which is strategyproofness in Moulin's sense.

*Step 2: Moulin's characterization.* The mechanism's anonymity and efficiency on every instance are the same properties of $a$: if $a(w) \neq a(\sigma w)$ for a permutation $\sigma$, an instance with two items crossing between the two values makes the outputs differ, and if $a(w) > \max_i w_i$, two items crossing between $\max_i w_i$ and $a(w)$ make $\max_i w_i$ strictly better for every agent (symmetrically below). Setting $a(t, \ldots, t) = t$ on profiles with all reports equal, which spoofing never produces, keeps $a$ strategyproof, anonymous and efficient: from such a profile a deviator already obtains its own type, and a deviation into one yields the others' common type $t$, while by efficiency the truthful outcome lies between the deviator's type and $t$. Moulin (1980) shows that the strategyproof, anonymous, and efficient voting rules on peaks are the *generalized medians*, $a(w) = \mathrm{med}(w_1, \ldots, w_n, y_1, \ldots, y_{n-1})$ for fixed *phantom* positions $y_1 \leq \cdots \leq y_{n-1}$ in $[-\infty, \infty]$ counted alongside the reports. The mean is not of this form; Proposition 2 shows it is manipulable.

*Step 3: identifiability leaves the median.* By Proposition 1(b), $a(c\,w + b) = c\,a(w) + b$ for all nondegenerate $w$, all $b$ and all $c \neq 0$. Suppose $f \geq 1$ phantoms are finite, with $l$ at $-\infty$ and $n - 1 - l - f$ at $+\infty$. On a profile of distinct reports all below the finite phantoms the rule returns the $(n - l)$-th smallest report; on the same profile shifted above them it returns the $(n - l - f)$-th

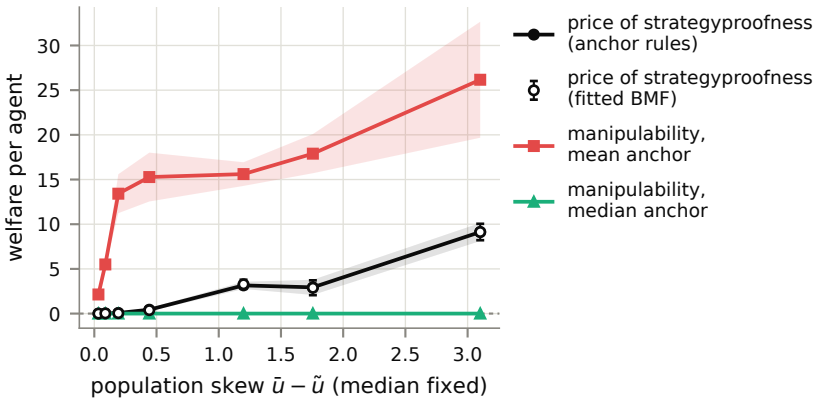



Figure 4: The price of strategyproofness against population skew $\bar{u} - \tilde{u}$, median held fixed. PoSP and manipulability are computed exactly on the anchored selections $S(\gamma)$, in per-agent utility units; the open markers report the same welfare difference for the fitted mean- and median-constrained factorizations. At $\bar{u} = \tilde{u}$ the anchors coincide and the PoSP is zero; as the mean pulls away the PoSP grows, while the median anchor's manipulability, the largest benefit a single spoofer can achieve, is exactly zero throughout (Theorem 5) and the mean anchor's is positive throughout (Proposition 2).

smallest. The two differ, so a shift ($c = 1$) breaks the identity. With no finite phantom the rule is an order statistic, the $k$-th smallest report for some $k$. Order statistics commute with every shift and every $c > 0$, but a reflection $c < 0$ exchanges the $k$-th for the $(n+1-k)$-th, so the identity forces $k = (n+1)/2$, the median. Only this step uses the linear form; Steps 1 and 2 need only strict crossing. □

**Remark 6** (Rankings)**.** The results of Section 4 are stated for a selected set. They hold for a ranking under any utility that weights the item in position $r$ by $\alpha_r$ with $\alpha_1 \geq \alpha_2 \geq \cdots \geq \alpha_m \geq 0$, since such a utility equals $\sum_{k=1}^{m}(\alpha_k - \alpha_{k+1})$ times the utility of the top-$k$ set, with $\alpha_{m+1} = 0$, a nonnegative combination of the set utilities. A deviation that weakly lowers every top-$k$ utility weakly lowers the ranking utility (Theorem 5), and a benefit that is larger for every $k$ is larger for the ranking (Proposition 3).

### G.1 The welfare the median gives up

Choosing the median over the mean gives up welfare whenever the two selections differ, and the difference can be measured as

$$\mathrm{PoSP} \;=\; \mathrm{SW}(S(\bar{u})) - \mathrm{SW}(S(\tilde{u})), \qquad \mathrm{SW}(S) = \sum_i \sum_{j \in S} U(u_i, j),$$

the welfare lost when the median anchor's selection $S(\tilde{u})$ replaces the welfare-optimal mean anchor's $S(\bar{u})$; it is nonnegative, and zero when the two selections agree (Fig. 4 reports it per agent, dividing by $n$), and in the spirit of the price-of-anarchy and price-of-fairness quantities of mechanism design and of quantitative measures of manipulability (Lubin & Parkes, 2009). Figure 4 traces the price against population skew: as the mean pulls away from the fixed median the PoSP grows, while the median's manipulability stays exactly at zero. The median's protection covers spoofing only: the anchor is still exposed through who reports at all (Section 4.2).

## H Proofs for Section 4.2

### H.1 Characterizing selection changes

We first establish which items enter and exit the selection when the anchor shifts.

**Lemma 10** (Exchanged items under monotone differences)**.** *In an anchored utility model with monotone differences (Definition 9), suppose the anchor shifts from $\gamma$ to $\gamma' > \gamma$ and items $j$*

*and $j'$ swap positions in the ranking, with item $j$ moving up and item $j'$ moving down. Then $g(t) = U(t, j) - U(t, j')$ is nondecreasing, nonpositive for $t \le \gamma$, and nonnegative for $t \ge \gamma'$: the entering item is the one preferred by high types. In the LFUM this is $v_j > v_{j'}$.*

*Proof.* Ranking $j'$ above $j$ at $\gamma$ means $g(\gamma) \le 0$, ties decided by the fixed priority, and ranking $j$ above $j'$ at $\gamma'$ means $g(\gamma') \ge 0$; since a fixed priority cannot flip a tie on its own, at least one inequality is strict. By monotone differences $g$ is monotone, and a nonincreasing $g$ with $g(\gamma) \le 0 \le g(\gamma')$ would force the excluded double tie $g(\gamma) = g(\gamma') = 0$; so $g$ is nondecreasing, whence $g \le 0$ below $\gamma$ and $g \ge 0$ above $\gamma'$. In the LFUM $g$ is affine, and a constant $g$ would make both inequalities equalities, so its slope $v_j - v_{j'}$ is positive. □

### H.2 Participation gains grow with distance

**Proposition 3** (Participation gains grow with distance)**.** *Let $f$ be mean- or median-anchored, let $Q$ be a nonempty pool with anchor $\gamma = \gamma(Q)$, the mean or median of $Q$, and let $a, b \notin Q$ be agents with $u_a > u_b > \gamma$. Then $B_a \ge B_b \ge 0$, where $B_i$ is* (3) *with $\gamma' = \gamma(Q \cup \{i\})$. The symmetric statement holds for $u_a < u_b < \gamma$.*

*Proof.* We give the proof for the mean anchor in four steps; the median case follows it.

**Step 1: Anchor movements.** When agent $i$ joins a mean-anchored mechanism with $n$ existing participants, the anchor shifts by

$$\Delta\gamma_i = \frac{u_i - \gamma}{n+1}. \tag{4}$$

Since $u_a > u_b > \gamma$,

$$\Delta\gamma_a = \frac{u_a - \gamma}{n+1} > \frac{u_b - \gamma}{n+1} = \Delta\gamma_b > 0. \tag{5}$$

Both anchors shift rightward, but agent $a$ induces a strictly larger shift: $\gamma_a = \gamma + \Delta\gamma_a > \gamma_b = \gamma + \Delta\gamma_b > \gamma$. Since $\Delta\gamma_a < u_a - \gamma$ and $\Delta\gamma_b < u_b - \gamma$, also $\gamma_a < u_a$ and $\gamma_b < u_b$. The remaining steps use only $\gamma \le \gamma_b \le \gamma_a \le u_a$ and $\gamma_b \le u_b$.

**Step 2: Splitting the benefit.** Write $V_t(\gamma) = \sum_{j \in S(\gamma)} U(t, j)$ for the value of the selection at anchor $\gamma$ to type $t$. Split agent $a$'s move at $\gamma_b$:

$$B_a = \big[V_a(\gamma_b) - V_a(\gamma)\big] + \big[V_a(\gamma_a) - V_a(\gamma_b)\big], \qquad B_b = V_b(\gamma_b) - V_b(\gamma).$$

**Step 3: The move to $\gamma_b$ is worth more to $a$.** The sets $S(\gamma_b) \setminus S(\gamma)$ and $S(\gamma) \setminus S(\gamma_b)$ have the same size; pair them by any bijection. For a pair with $j'$ leaving and $j$ entering, Lemma 10 makes $g(t) = U(t, j) - U(t, j')$ nondecreasing under monotone differences, so $V_t(\gamma_b) - V_t(\gamma)$, the sum of $g(t)$ over the pairs, is nondecreasing in $t$. Since $u_a > u_b$, the first bracket is at least $B_b$. In the LFUM each $g$ is affine with slope $v_j - v_{j'} > 0$, so the first bracket exceeds $B_b$ by $(u_a - u_b)\sum(v_j - v_{j'})$, which is positive whenever $b$'s joining changes the selection.

**Step 4: The rest of $a$'s move.** Since $\gamma_b \le \gamma_a \le u_a$ and $V_a$ is nondecreasing below $u_a$ (Lemma 7), the second bracket is nonnegative. Since $\gamma \le \gamma_b \le u_b$, also $B_b \ge 0$. Hence $B_a \ge B_b \ge 0$. □

For median-anchored mechanisms, Step 1 is replaced by the following fact.

**Lemma 11** (The median with one report added)**.** *Let $Q$ be a pool with order statistics $u_{(1)} \le \cdots \le u_{(n)}$ and let $x$ be a report added to it. Under the midpoint convention and under either middle-element convention of Remark 3, the median of $Q \cup \{x\}$ is nondecreasing in $x$ and lies weakly between the median of $Q$ and $x$.*

*Proof.* Write $u_{(0)} = -\infty$ and $u_{(n+1)} = +\infty$. The $r$-th order statistic of $Q \cup \{x\}$ is $x$ clamped to $[u_{(r-1)}, u_{(r)}]$: it equals $u_{(r-1)}$ if $x \le u_{(r-1)}$, equals $x$ if $u_{(r-1)} \le x \le u_{(r)}$, and equals $u_{(r)}$ if $x \ge u_{(r)}$. Each clamp is nondecreasing in $x$, so every middle element of the enlarged pool, and the midpoint of two of them, is nondecreasing in $x$. For $n = 2m+1$ the median of $Q$ is $u_{(m+1)}$, and the two middle elements of the enlarged pool are $x$ clamped to $[u_{(m)}, u_{(m+1)}]$ and to $[u_{(m+1)}, u_{(m+2)}]$;

for $x \le u_{(m+1)}$ the first lies in $[x, u_{(m+1)}]$ and the second equals $u_{(m+1)}$, and for $x \ge u_{(m+1)}$ the first equals $u_{(m+1)}$ and the second lies in $[u_{(m+1)}, x]$. For $n = 2m$ the median of $Q$ lies in $[u_{(m)}, u_{(m+1)}]$ and the median of the enlarged pool is $x$ clamped to that interval, which lies between the median of $Q$ and $x$ in each of the three cases. □

With $\gamma_b$ and $\gamma_a$ the medians after $b$ and after $a$ joins, Lemma 11 gives $\gamma \le \gamma_b \le \gamma_a$, $\gamma_b \le u_b$ and $\gamma_a \le u_a$, which is all that Step 1 supplies, and Steps 2 to 4 apply unchanged, including the strict comparison of Step 3 in the LFUM. Unlike the mean, two joiners can move the median to the same order statistic.

### H.3 Geometric intuition: contour rotation

At anchor $\gamma$ items are ranked by $U(\gamma, j) = q_j + \gamma v_j$. In the $(v, q)$ space, items with equal score lie on lines satisfying

$$q = c - \gamma v, \tag{6}$$

for constant $c$. These *iso-score contours* have slope $-\gamma$.

When the anchor shifts from $\gamma$ to $\gamma' > \gamma$, the contours rotate: their slopes change from $-\gamma$ to $-\gamma'$, becoming steeper. This rotation causes items to swap positions in the ranking, and the magnitude of rotation is $|\gamma' - \gamma|$: larger shifts induce more rotation, causing more swaps.

Figure 5 illustrates the mechanism. Items $A$ and $B$ have the same score under anchor $\gamma$, but when a more extreme agent joins and shifts the anchor to $\gamma_a$, the contour rotates substantially, elevating item $B$ (with higher $v$) above item $A$. A less extreme agent shifting the anchor only to $\gamma_b$ induces a smaller rotation. The picture is for the mean, where the farther joiner moves the anchor farther. Under the median both joiners can move the anchor to the same point, the rotations then coincide, and the comparison of Proposition 3 comes from Step 3 alone.

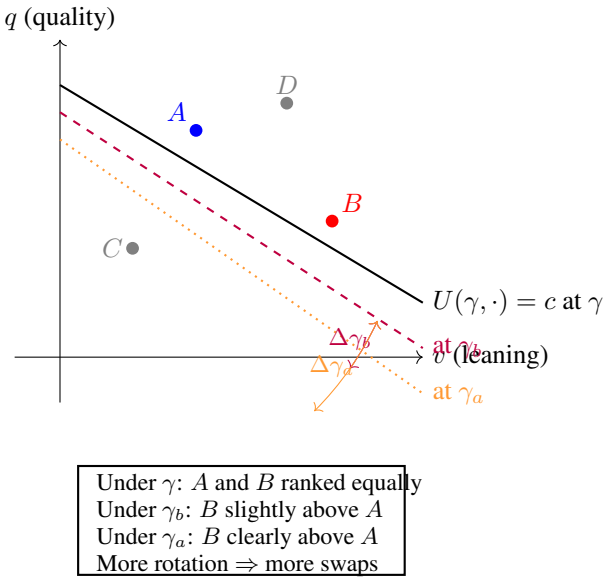



Figure 5: Geometric intuition for participation incentives. Items lie in the $(v, q)$ space. Iso-score contours rotate when the anchor changes; the shift $\Delta\gamma_a > \Delta\gamma_b$ induces greater rotation, causing more pairs of items to swap rankings. Items with higher leaning (like $B$) rise under rightward shifts.

### H.4 Turnout equilibria

Throughout this subsection the instance is linear, and $\gamma(P)$ is the anchor the mechanism computes from the pool $P$, the mean of the members' types or their median, with the median of an even pool

the midpoint of its two middle types. For every agent $i$, member or not, write

$$B_i(P) = V_i\big(\gamma((P \setminus \{i\}) \cup \{i\})\big) - V_i\big(\gamma(P \setminus \{i\})\big), \tag{7}$$

the value of being counted with the other members of $P$ fixed. For $i \notin P$ this is (3), and for $i \in P$ it is the benefit from being counted in Definition 8. A sole rater stays, so $B_i(\{i\}) \geq c$ by convention and a pool is never empty. A pool $P$ is a turnout equilibrium exactly when $i \in P$ if and only if $B_i(P) \geq c$. Let

$$W(\gamma) = \sum_{j \in S(\gamma)} v_j \tag{8}$$

be the total leaning of the items selected at anchor $\gamma$. By Lemma 10 it is nondecreasing in $\gamma$, and it is a step function whose steps are the anchors at which the selection changes.

**Lemma 12** (Value of an anchor move)**.** *For a type $u$ and anchors $\gamma < \gamma'$,*

$$V_u(\gamma') - V_u(\gamma) = \int_{(\gamma, \gamma']} (u - s)\, dW(s).$$

*Proof.* As the anchor increases from $\gamma$ to $\gamma'$, the selection changes only at anchors $s$ where the items in the $k^{\text{th}}$ and $(k{+}1)^{\text{th}}$ positions tie, $q_j + sv_j = q_{j'} + sv_{j'}$, and there $j$ replaces $j'$ with $v_j > v_{j'}$ (Lemma 10). The change replaces $q_{j'} + uv_{j'}$ by $q_j + uv_j$ in $V_u$, a difference of $(q_j - q_{j'}) + u(v_j - v_{j'}) = (u - s)(v_j - v_{j'})$, and increases $W$ by $v_j - v_{j'}$. Summing over the changes along the path gives the integral. □

**Lemma 13** (Members farther from the anchor gain more from being counted)**.** *Let $P$ be a pool with $|P| \geq 2$ and anchor $\gamma = \gamma(P)$, mean or median, and let $a, b \in P$ with $u_a \geq u_b > \gamma$. Then $B_a(P) \geq B_b(P) \geq 0$. The mirror statement holds for $u_a \leq u_b < \gamma$.*

*Proof.* Write $\gamma_{-a} = \gamma(P \setminus \{a\})$ and $\gamma_{-b} = \gamma(P \setminus \{b\})$. Under the mean, with $n = |P|$, $\gamma_{-a} = (n\gamma - u_a)/(n-1)$ and $\gamma_{-b} = (n\gamma - u_b)/(n-1)$, so $\gamma_{-a} \leq \gamma_{-b} < \gamma < u_a$. Split

$$B_a(P) = \big[V_a(\gamma) - V_a(\gamma_{-b})\big] + \big[V_a(\gamma_{-b}) - V_a(\gamma_{-a})\big].$$

The second bracket is nonnegative by Lemma 7, since $\gamma_{-a} \leq \gamma_{-b} \leq u_a$. The first bracket and $B_b(P) = V_b(\gamma) - V_b(\gamma_{-b})$ are the values of the same upward move $\gamma_{-b} \to \gamma$ for the types $u_a \geq u_b$; by Lemma 12 the value of a fixed upward move is nondecreasing in the type, so the first bracket is at least $B_b(P)$, and $B_b(P) \geq 0$ by Lemma 7 since $\gamma_{-b} < \gamma < u_b$.

Under the median, removing any member strictly above the median lands on the same anchor. Write $u_{(1)} \leq \cdots \leq u_{(n)}$ for the order statistics of $P$. If $n = 2m + 1$, a member above $u_{(m+1)}$ occupies a position at least $m{+}2$ in every sorted order, so its removal leaves positions 1 to $m{+}1$ unchanged and the median of the remaining even pool is determined by $u_{(m)}$ and $u_{(m+1)}$ under each convention, hence is at most $u_{(m+1)}$. If $n = 2m$, a member above the median lies at a position at least $m + 1$, its removal leaves positions 1 to $m$ unchanged, and the median of the remaining odd pool is $u_{(m)}$. Hence $\gamma_{-a} = \gamma_{-b} \leq \gamma < u_b \leq u_a$, and $B_a(P)$ and $B_b(P)$ are the values of the same upward move $\gamma_{-a} \to \gamma$ for the types $u_a \geq u_b$, so the conclusion follows from Lemma 12 as before. □

The hypothesis $u_b > \gamma$ is needed for the median. In the pool $\{0, 1, 5\}$ with one change of selection at anchor 2 and $k = 1$, removing the member at 1 moves the midpoint median to 2.5 and its benefit is positive, while removing the member at 5 moves it to 0.5 and its benefit is zero.

**Proposition 12** (Boundary test)**.** *A pool $P$ is a turnout equilibrium if and only if, on each side of $\gamma(P)$, the member closest to $\gamma(P)$ has $B_i(P) \geq c$ and the non-member farthest from $\gamma(P)$ has $B_i(P) < c$, and every member at $\gamma(P)$ has $B_i(P) \geq c$.*

*Proof.* Each condition is an instance of the definition, which gives necessity. For sufficiency fix one side. Every member on that side is at least as far from $\gamma(P)$ as the closest member, so its benefit is at least $c$ by Lemma 13; every non-member on that side is at most as far as the farthest non-member, so its benefit is below $c$ by Proposition 3, both applied to the pool $P$. A non-member at $\gamma(P)$ leaves the mean unchanged and, by Lemma 11, leaves the median unchanged, so its benefit is $0 < c$. □

The test decides whether a pool is an equilibrium from four benefit evaluations, or five when a member sits at $\gamma(P)$, as in a median pool of odd size, and makes it feasible to enumerate the pools of a given shape at $n = 101$. It makes no comparison between a member and a non-member on the same side, where monotonicity fails. A member's benefit is measured over the move of the anchor when it leaves, which lies on the far side of $\gamma(P)$, and a non-member's over the move when it joins, which lies on its own side. Take $k = 1$, items $(q, v) = (1, -1)$, $(0.4, 1)$, $(-0.5, 2)$, agents at $-1$, 1, 1, 1.5, the mean anchor and $c = 1$. The pool $\{-1, 1, 1\}$ has anchor $1/3$ and selects the second item; removing the agent at $-1$ moves the anchor to 1 and the selection to the third item, a benefit of 1.9; removing an agent at 1 moves it to 0 and the selection to the first item, a benefit of 1.4; the agent at 1.5 joining moves it to 0.625 and changes nothing. The pool is an equilibrium in which agents at distance $2/3$ rate and an agent at distance $7/6$ on the same side does not.

**Proposition 13** (Turnout under the median)**.** *Let $f$ be median-anchored, $P$ a turnout equilibrium with $|P|$ even and anchor $\gamma$, and $u_- < \gamma < u_+$ its two middle members. Then every agent with $u_i \geq u_+ + c/\big(W(u_+) - W(\gamma)\big)$ and every agent with $u_i \leq u_- - c/\big(W(\gamma) - W(u_-)\big)$ is in $P$, with $c/0 = \infty$.*

*Proof.* Let $i \notin P$ with $u_i > u_+$. The pool $P \cup \{i\}$ has odd size and its median is $u_+$, so by Lemma 12 the benefit of $i$ is $\int_{(\gamma, u_+]} (u_i - s)\, dW(s) \geq (u_i - u_+)\big(W(u_+) - W(\gamma)\big)$. In equilibrium this is below $c$, so $u_i < u_+ + c/(W(u_+) - W(\gamma))$. The other side is the mirror image. □

**Proposition 14** (Turnout under the mean)**.** *Let $f$ be mean-anchored, $P$ a turnout equilibrium with $N = |P| \geq 2$ and anchor $\gamma$, $u_+$ the smallest member above $\gamma$, and $x_+ = u_+ - \gamma$. Then*

$$x_+ \Big[ W(\gamma) - W\Big(\gamma - \frac{x_+}{N-1}\Big)\Big] \;\geq\; \Big(1 - \frac{1}{N}\Big) c,$$

*so the selection changes at some anchor in $(\gamma - x_+/(N-1), \gamma]$. The mirror inequality holds for the largest member below $\gamma$.*

*Proof.* Removing $u_+$ lowers the mean by $\ell = x_+/(N-1)$. By Lemma 12 its benefit is $\int_{(\gamma-\ell, \gamma]} (u_+ - s)\, dW(s) \leq (x_+ + \ell)\big(W(\gamma) - W(\gamma - \ell)\big)$, and $x_+ + \ell = x_+ N/(N-1)$. In equilibrium the benefit is at least $c$. □

A turnout equilibrium need not exist. Take $k = 1$, items $(0, 0)$, $(2, 1)$, $(0, 2)$, whose selection changes at anchors $-2$ and 2, agents at $-5$, 5.5 and 6, and $c = 1$. Under the mean, each of the seven pools has a profitable deviation: from $\{-5\}$ the agent at 6 joins (the anchor moves from $-5$ to 0.5, benefit 8); from $\{5.5\}$ or $\{6\}$ the agent at $-5$ joins (benefit 7); from $\{-5, 5.5\}$ the agent at 6 joins and from $\{-5, 6\}$ the agent at 5.5 joins (benefits 4 and 3.5); from $\{5.5, 6\}$ either member leaves, since its removal changes nothing; from $\{-5, 5.5, 6\}$ the agent at $-5$ leaves, since its removal moves the anchor from $13/6$ to 5.75 and changes nothing. Under the midpoint median the two-member pools have the same anchors, the three-member pool has anchor 5.5, and the same deviations apply. The agents at 5.5 and 6 make each other dispensable.

**Beyond fixed costs.** The protection of Theorem 5 also ends where participation costs can themselves be manipulated. Mobilizing one tail can move the median away from the population's. An interested party can shift the anchor by manipulating the costs of different agents to change who finds it worth rating.

# I Experimental Details

## I.1 Data-generating process

All experiments instantiate the LFUM (Definition 1). The default DGP draws agent types $u_i \overset{\text{iid}}{\sim} \mathcal{N}(\mu_u, 1)$ (with $\mu_u = 0$ unless varied) and item parameters $(q_j, v_j) \overset{\text{iid}}{\sim} (\mathcal{N}(0,1), \mathcal{N}(0,1))$ independently. Continuous ratings are generated as $r_{ij} = q_j + u_i v_j + \varepsilon_{ij}$ with $\varepsilon_{ij} \sim \mathcal{N}(0, \Delta)$ and $\Delta = 0.25$ (standard deviation 0.5); the theory takes the noiseless limit. The default sample sizes are $n = 101$

agents (odd, so the median results apply) and $m = 100$ items. Every experiment refers to this default and describes only the parameters that differ.

In the strategic experiments the reporting process deviates from truth in a controlled way: in spoofing experiments, an agent of type $u_i$ reports the ratings of some other type $u'_i \neq u_i$ (Definition 7), and in participation experiments agents choose whether to report at all.

### I.2 Mechanisms and oracles

We fit BMF with a `PyTorch` (Paszke et al., 2019) implementation that makes it straightforward to swap regularizers and explicit gauge constraints. The factors are initialized Xavier-uniform and the qualities and intercepts at zero, and the model is trained full-batch with Adam (Kingma & Ba, 2017) at learning rate $0.05$ for at most $500$ epochs, stopping when the training loss improves by less than $10^{-5}$ for $10$ epochs; the objective is mean squared error with $L_2$ regularization weight $0.01$ unless stated. The *re-anchored variants* apply the ex-post reparameterization of Theorem 4 to the fitted parameters, moving the anchor to the mean or the median of the fitted types; the *constrained variants* impose it during fitting instead: after $100$ warm-up epochs, every $10$ epochs and at the end, a gauge shift (2) moves the mean or the median of the fitted types to $0$.

Outcomes are compared against three oracles with access to ground truth, one per objective of Definition 3: the *quality oracle* selects the top items by true quality $q_j$; the *average-utility oracle* by average utility $q_j + \bar{u}v_j$; and the *median-utility oracle* by the median agent's utility $q_j + \tilde{u}v_j$, which for odd $n$ selects the Condorcet-consistent set. The quality oracle is unachievable from the ratings (Corollary 1); the other two are the targets of the mean- and median-anchored mechanisms. The metrics below are normalized against the oracles.

### I.3 Estimating the realized anchor

By Lemma 1 a zero-risk fit has $\hat{q} = q + \gamma v$ and $\hat{u} = c\,(u - \gamma)$; with access to the true simulation parameters we read the realized anchor $\gamma$ off either, in two ways: the *item-based estimator* fits the regression $\hat{q}_j = \beta_0 + \beta_1 v_j + \epsilon_j$ and takes $\hat{\gamma}_{\text{item}} = \hat{\beta}_1$, capturing the covariation between fitted quality and item leaning; the *user-based estimator* fits $\hat{u}_i = \beta_0 + \beta_1 u'_i + \epsilon_i$ on the *reported* types $u'_i$ (equal to the true ones under truthful reporting) and takes $\hat{\gamma}_{\text{user}} = -\hat{\beta}_0/\hat{\beta}_1$, measuring the recentering of the fitted types. Regressing on the reports rather than the truth matters in the spoofing experiments, where the fit can only recenter what was reported. Under correct specification, the two agree, and we use them to track how far a manipulation has moved a mechanism's implicit anchor.

### I.4 Evaluation metrics

For a mechanism $f$, let $S(f)$ be the top-$k$ items by its scores; we use $k = 10$. Each objective of Definition 3 has a metric, normalized against its oracle where applicable:

**Objective Quality gap.** $\mathrm{OQ}_{\text{gap}}(f) = \big(\mathrm{OQ}(f^*_q) - \mathrm{OQ}(f)\big)/\mathrm{OQ}(f^*_q)$ with $\mathrm{OQ}(f) = \sum_{j \in S(f)} q_j$ and $f^*_q$ the quality oracle.

**Social Welfare gap.** $\mathrm{SW}_{\text{gap}}(f) = 1 - \mathrm{SW}(f)/\mathrm{SW}(f^*_{\text{avg}})$, with $\mathrm{SW}(f) = \sum_i \sum_{j \in S(f)} U(u_i, j)$ and $f^*_{\text{avg}}$ the average-utility oracle. The price of strategyproofness of Appendix G.1 is the per-agent difference $\big(\mathrm{SW}(S(\bar{u})) - \mathrm{SW}(S(\tilde{u}))\big)/n$, so it shares utility units with the single-spoofer manipulability in Fig. 4.

**Condorcet Compliance.** The fraction of selected–unselected pairs where a weak majority prefers the selected item,

$$\mathrm{CC}(f) = \frac{1}{k(m-k)} \sum_{j \in S(f)} \sum_{j' \notin S(f)} \mathbf{1}\Big[\,\big|\{i : U(u_i, j) \geq U(u_i, j')\}\big| > \tfrac{n}{2}\Big]\,,$$

with $\mathrm{CC} = 1$ attained by the median-utility oracle for odd $n$.

Unless stated, metrics are means over 10 replications with $\pm 1$ standard-error bands.

### I.5 Experiment setups

**E1: Ideological extremization (Fig. 3).** Starting from the default DGP with a symmetric population ($\mu_u = 0$, so the truthful mean and median nearly coincide and manipulation is the only mover of the anchor), a fixed coalition, the top half of the agents with positive type (about the top quartile), spoofs outward: each member reports as if at $u_i + w$, for $w \in \{0, 0.5, \ldots, 3\}$. We fit the mean- and median-constrained factorizations and $L_2$-regularized BMF, with $L_2$ weights $0.01$ and $0.1$ pooled and 25 replications for each, and record the three gaps; the bands in Fig. 3 are $95\%$ confidence intervals. The prediction of Proposition 2 and Theorem 5: the mean-constrained fit's anchor moves with the coalition's offset and its objectives degrade; BMF's anchor, a shrunk multiple of the reported mean (Anonymous, 2026), moves less; the median-constrained fit is unchanged.

**E2: The price of strategyproofness (Fig. 4).** Types are drawn from Gamma distributions with median 1 and (shape, scale) equal to $(10, 0.1)$, $(4, 0.27)$, $(2, 0.6)$, $(1, 1.44)$, $(0.5, 4.4)$, $(0.4, 6.89)$ and $(0.3, 13.67)$, which sweep the mean–median gap $\bar{u} - \tilde{u}$ from 0.03 to 3.1 with the median fixed. PoSP and the single-agent manipulability (the largest benefit one spoofed report can achieve) are computed exactly on the anchored selections $S(\gamma)$ and reported per agent; the open markers in Fig. 4 report the same welfare difference for the fitted mean- and median-constrained factorizations. All of them use the same 70 populations (7 skews, 10 replications), shared across the constraint arms. The predictions: PoSP grows with the gap; the median's manipulability is zero (Theorem 5) and the mean's is positive (Proposition 2).

**E3: Turnout equilibria (Fig. 6).** Ratings are exact and the anchor is computed from the pool directly, with the conventions of Appendix H.4: a sole rater stays, the median of an even pool is the midpoint of its two middle types, and ties are broken by item index. We take $n = 101$, $m = 100$, $k = 10$ and the default item parameters, and three populations with median 0: the standard normal, and Gamma distributions of shape 2 and scale 1 and of shape $1/2$ and scale 2, shifted to median 0, both with variance 2 and the heavier tail on the right (mean minus median $0.32$ and $0.55$). Each population has 20 instances, and the costs are $c \in \{0.03, 0.1, 0.3, 1, 3\}$. Simultaneous best response from full participation does not find equilibria: it converges in $2\%$ of runs under the mean and $0.3\%$ under the median, mostly alternating between full and empty turnout. We therefore enumerate the pools that are tail sets relative to their own anchor, including those with a member at the anchor, test each with Proposition 12, and search locally for equilibria that are not tail sets.

Under the median a tail equilibrium exists in 7 to 9 of 20 instances at $c = 0.03$ and in 19 or 20 at $c = 3$. Its pool has a median of 96 raters at $c = 0.03$ and 58 to 72 at $c = 3$, the abstaining band around the anchor widens from about $0.15$ to about 1, and the anchor stays within $0.2$ of the population median. Under the mean a tail equilibrium exists in 2 of 20 symmetric instances and in 4 to 18 of 20 skewed ones. Its pool is a few raters from the tails, with a median of 4 to 17 and roughly independent of $c$, and in the skewed populations its anchor sits $1.2$ to $2.1$ into the heavier tail. For comparison, Fig. 6 also shows a smoothed best response, in which each agent rates with probability $\sigma\big((\mathbb{E}[B_i] - c)/\tau\big)$, $\sigma$ the logistic function, and the temperature $\tau$ is lowered from $4c$ to $0.01c$. It converges at high temperature and, as $\tau$ falls, either reaches a pure equilibrium, almost always a tail equilibrium, or stops converging. Propositions 13 and 14 hold at all $1{,}451$ equilibria found, and the boundary test agrees with the definition on every pool checked, about 1.3 million.

**E4: Constrained anchoring (Fig. 2).** On the populations of E2, with draws shared across arms, we fit the mean- and median-constrained factorizations and, as the third arm, re-anchor an unconstrained fit ex post at the neutral type, using the true parameters to locate it (Lemma 1); ratings carry the default noise, so no arm attains zero risk. All three metrics are reported as gaps, so zero means the objective is attained. The prediction of Theorems 1 and 4: each variant attains its own objective up to the noise, at every skew, while the neutral target is met only with the oracle.

### I.6 Evidence from Community Notes data

Proposition 3 predicts that raters farther from the anchor gain more from rating. We test the implication that they rate more on the public Community Notes data (snapshot of 23 September 2026, ratings of notes created since 14 April 2025). Partisan rating on the platform is documented (Allen et al., 2022; Pröllochs, 2022); the test here is of the gradient in distance from the anchor. We fit

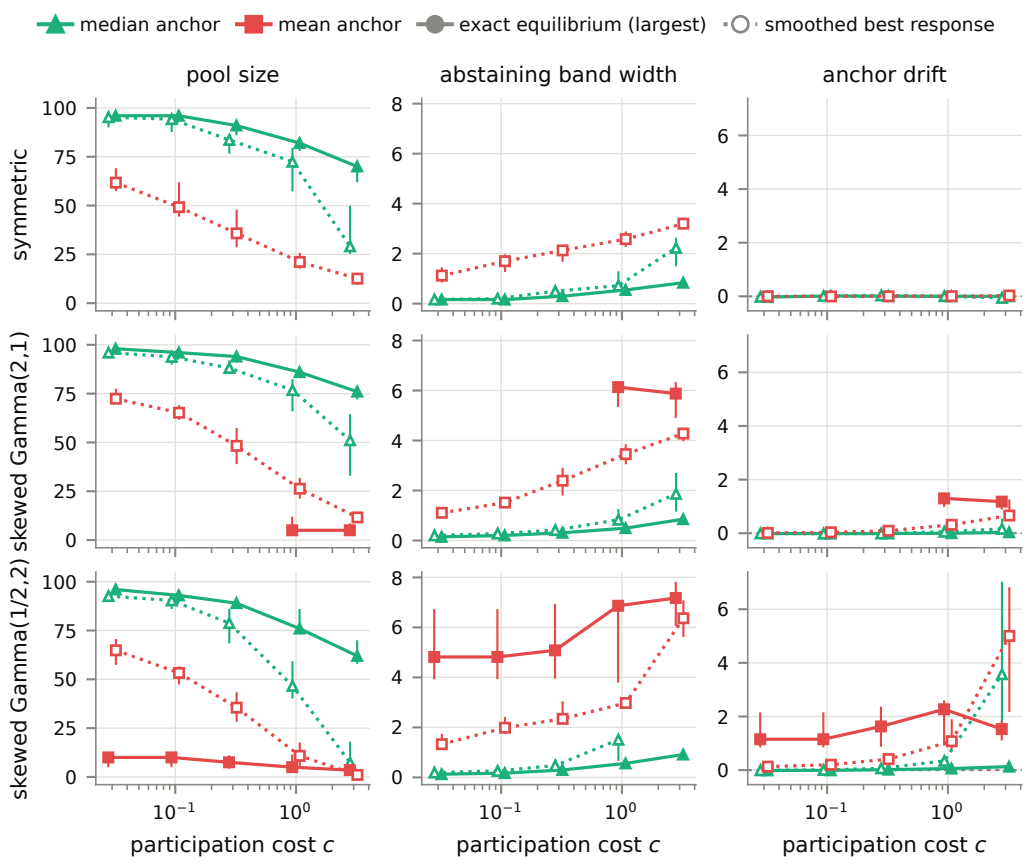



Figure 6: Turnout equilibria at $n = 101$ against the cost $c$, for three populations (rows): pool size, width $d_- + d_+$ of the abstaining band, and drift of the pool's anchor from the population's. Filled markers are the largest exact tail equilibrium of each instance; hollow markers are the smoothed best response at its last converged temperature, started from full participation. Bars are interquartile ranges over instances. A tail equilibrium exists in 2 of the 20 instances under the mean anchor in the symmetric population, 4 to 7 and 6 to 10 in the two skewed ones, and 6 to 20 under the median anchor; filled markers are shown only where at least 5 instances have one.

the scorer's one-factor model (helpful $= 1$, somewhat helpful $= 0.5$, not helpful $= 0$, with its $L_2$ weights) to the ratings of one six-month window, 26 September 2025 to 25 March 2026, and count each rater's ratings in the next six months. Regularization shrinks the factors of light raters toward zero, and that alone makes activity rise with distance from the center: with the factor and the activity taken from the same window, the Spearman correlation between them is $0.43$. We remove this confound by re-estimating each rater's factor from exactly $K$ random ratings of the first window, with the note parameters fitted on the other half of the raters; on synthetic data in which activity does not depend on type, this estimator shows no gradient.

Activity rises with distance from the median on both sides (Table 5 and Fig. 7), more steeply on the right-leaning side. The gradient is in how much raters rate: the probability of rating at all in the later window hardly varies with distance ($\rho \leq 0.015$). It survives stratifying by enrollment state and by the information in each rater's sampled notes. Two caveats limit the reading. The data are observational, so the gradient is consistent with the participation incentive without establishing it. And a rater who reports as a more extreme type (Section 4.1) produces the same pattern in the fitted factors, since only reported types are observed.

Table 5: Rating activity in the later window against distance from the median factor, per side of the median: Spearman correlation and the ratio of mean activity between the top and bottom quintiles of distance, with 95% bootstrap intervals. Global: all ratings, $K = 20$, 266,793 raters. US politics: notes matched by a keyword list, a topic-specific factor with its own median, topic ratings only, $K = 10$, 70,064 raters.

| Sample | Side | Spearman $\rho$ | Top/bottom quintile |
|---|---|---|---|
| Global | left | $0.062\ [0.057, 0.068]$ | $1.26\times\ [1.22, 1.33]$ |
| Global | right | $0.085\ [0.081, 0.090]$ | $1.53\times\ [1.47, 1.58]$ |
| US politics | Democrat-leaning | $0.053\ [0.043, 0.063]$ | $1.44\times\ [1.32, 1.56]$ |
| US politics | Republican-leaning | $0.070\ [0.061, 0.079]$ | $1.51\times\ [1.42, 1.58]$ |

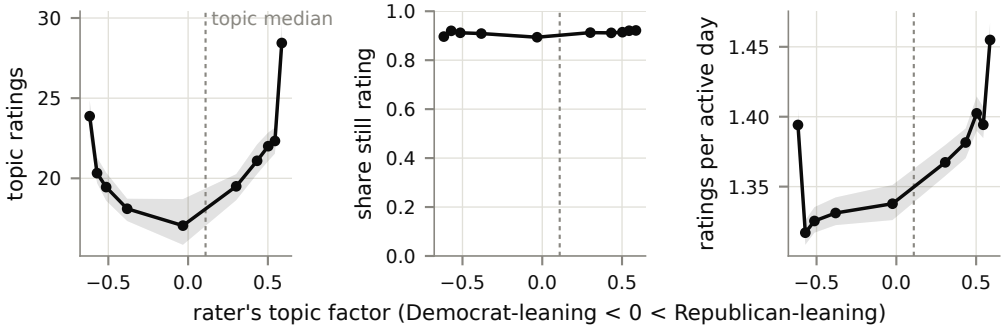



Figure 7: Community Notes, US-politics notes: activity in the later window against the rater's topic factor (fixed $K = 10$, cross-fitted; 70,064 raters; ten quantile bins with 95% confidence intervals). Left: mean number of topic ratings; middle: probability of rating the topic at all; right: ratings per active day. The dashed line is the topic median.