The Informativeness of Frequency-Report Scoring Rules

Jesper Armouti-Hansen*

June 17, 2026

Abstract

An experimenter elicits a subject's latent multinomial beliefs through an incentivized count report under a scoring rule. We recast the inference as a partial-identification problem: each rule maps the report to the set of beliefs under which it is optimal, from which coordinate and linear-functional bounds follow. We characterize three rules—squared-distance scoring (closed-form coordinate bounds and linear-program means), frequency-guessing (the known closed-form fixed-prize rule), and Manhattan distance (sharp one-dimensional bounds via threshold root-finds)—unified by a single structural condition: separability and discrete convexity of a per-coordinate cost. No rule dominates: squared-distance for concentrated beliefs, frequency-guessing for balanced, with the two closed-form rules' coordinate widths crossing in the number of positive-report coordinates—a crossover the design comparison locates in the Dirichlet concentration. Manhattan is rarely tightest but barely moves across regimes, the robust choice when the regime is unknown. The body assumes risk neutrality; a binary-lottery implementation extends to risk-averse subjects.

Keywords: belief elicitation; scoring rules; frequency reports; partial identification; informativeness; multinomial beliefs.

JEL classification: C44; C81; C90; D81; D83.

*Email: jesper@armoutihansen.xyz.

1 Introduction

Belief elicitation in experiments often asks subjects to report probabilities directly. Frequency reports ask instead for counts: out of n realizations, how many times will each outcome occur? This format is natural when the object of interest is a distribution of choices, signals, or outcomes in a finite group. It may also be easier for subjects than reporting abstract probabilities under a standard scoring rule.[1] Many probability scoring rules, including the quadratic scoring rule, are well understood theoretically (Savage, 1971; Gneiting and Raftery, 2007; Selten, 1998), but they require subjects to understand how probability reports map into payments. That mapping can be cognitively demanding, especially when subjects reason more naturally in frequencies than in abstract probabilities (Hogarth, 1975; Gigerenzer and Hoffrage, 1995; Schlag et al., 2015). Direct monetary scoring rules also raise the familiar risk-aversion problem (Armantier and Treich, 2013).

The practical problem is that the researcher observes a count report r, not the subject's latent belief vector p. The report is an incentivized message generated by a scoring rule. In other words, the question is inverse: given the rule and the observed report, which beliefs are consistent with optimal reporting? We follow the chain

$$S \;\to\; R_S(p) \;\to\; P_S(r) \;\to\; \big[\underline{p}_{S,i}(r), \overline{p}_{S,i}(r)\big].$$

Here S is a frequency-report scoring rule, $R_S(p)$ is the set of reports that are optimal for a subject with beliefs p, and

$$P_S(r) = \{p \in \Delta^k : r \in R_S(p)\}$$

is the identified set induced by the observed report. This recasts incentivized belief elicitation as a partial-identification problem (Manski, 2003): the scoring-rule mechanism, rather than a sampling distribution over repeated observations, supplies the identifying restriction. The final object of the chain is the bound $[\underline{p}_{S,i}(r), \overline{p}_{S,i}(r)]$ it yields on each individual probability p_i. A researcher may also care about functionals of the belief vector, such as a mean or another linear index; bounds on these are obtained from the same identified set $P_S(r)$.

We use the term *informativeness* in an operational sense. A rule is more informative for a given design objective if the observed report leads to narrower coordinate bounds or narrower mean and functional bounds. These inferential gains must be weighed against implementation costs: exact-match payment probabilities, cognitive simplicity, computational burden, and robustness to risk aversion.

Our main analytical contribution is for squared-distance frequency scoring. Under this rule, a subject reports the feasible count vector closest to the expected count vector np. This gives a linear identified set and closed-form bounds for each p_i. Mean bounds and other linear functional bounds are then simple linear programs over the same region.

[1] This cognitive-ease claim is a behavioral premise we inherit from the frequency-format literature (Hogarth, 1975; Gigerenzer and Hoffrage, 1995) rather than test ourselves. The claim is that subjects reason more readily in counts than in abstract probabilities and find a distance-based payment more transparent than a probability score—an empirical regularity the literature has documented in adjacent settings but that our exercise does not itself establish. Our informativeness comparison conditions on the report itself: it asks, given that the subject reports under a given count-loss rule, how informative the report is about latent beliefs. Whether the count format also raises the realism of the optimal-reporting assumption is a separate empirical question, taken up in Section 5.2 as a scope condition on the comparison.

Extending the partial-identification program of Schlag and Tremewan (2021), we compare this rule to two other count-loss mechanisms along a spectrum of analytical tractability. Frequency-guessing scoring is their known exact-match mechanism: it elicits multinomial modes, gives closed-form bounds, and is robust to risk aversion under a fixed-prize implementation. Manhattan-distance scoring has an exact identified set and sharp, semi-analytical coordinate bounds.

The contribution is not a new exact-match mechanism and not a general theory of scoring rules; Schlag and Tremewan already derive belief bounds for the frequency-guessing rule. It is, first, the characterizations themselves: a closed-form identified set and coordinate bounds for squared-distance scoring, the centerpiece result, and sharp semi-analytical bounds for Manhattan distance, alongside the known frequency-guessing bounds. These three characterizations are not separate facts: a single structural observation—that an optimal report solves a separable resource-allocation problem—organizes them, and a classical convexity criterion then says which rules admit closed-form bounds and which only semi-analytical ones. It is, second, the contingent decision rule those characterizations make rigorous. No rule is uniformly best: squared-distance scoring is most informative when the subject's beliefs are concentrated on a few categories and frequency-guessing scoring when beliefs are balanced, with a crossover as concentration varies—first established analytically, where the two closed-form rules' average coordinate widths cross, then quantified and extended to Manhattan distance by the design comparison. Manhattan is rarely the most informative rule, but its bounds vary little with the regime, making it the robust choice when belief concentration cannot be anticipated. Partial identification is the lens that holds these together—each scoring rule maps an observed report to an identified set of latent beliefs—and is what lets the comparison be posed as a sharp, finite-sample question rather than an informal one. Two further count-loss rules, Hamming and Chebyshev distance, mark the boundary of the approach; Section 3.6 discusses them as rules whose identified sets are exact but resist tractable sharp bounds.

Related literature. Our paper relates to four literatures.

The closest is the work on *frequency guessing*. The frequency-guessing rule studied here is due to Schlag and Tremewan (2021): the subject reports a count vector and wins a fixed prize if it matches the realized counts. Schlag and Tremewan already establish, for this rule, the three steps we rely on—the optimal report is a multinomial mode, the beliefs consistent with a report form a region of the simplex, and that region yields bounds on individual probabilities and on means and variances through linear programs. That inferential program, for this one rule, is theirs; the frequency-guessing results in Section 3 restate it so the comparison is self-contained. The contribution here is to carry the program beyond the rule for which it is known—to squared-distance scoring, where it delivers a closed-form identified set and closed-form coordinate bounds, and to Manhattan distance—and to turn the collection of rules into a comparison.[2] Schlag and Tremewan also show that, for a binary event, frequency guessing

[2]The extension is not a routine generalization. Schlag and Tremewan's derivations for the frequency-guessing rule turn on the multinomial-mode structure: a single-count transfer changes the multinomial probability by a closed-form ratio, and the optimality inequalities are linear in p. For a general count loss, neither step survives—the expected loss is not a product, and single-count transfers can fail to be sufficient for optimality even when they are necessary. What unifies the three rules is a structural observation we identify in Section 3: each rule's expected loss reduces to a separable resource-allocation problem in the report r, and discrete convexity of the per-coordinate cost is exactly the condition under which single-count transfers are both necessary and sufficient for optimality. The same condition determines whether the bounds are closed-form (linear in p) or

is the most precise rule eliciting a single frequency; the question here is the different one of which rule is sharpest across belief regimes, where no rule wins outright. The boundary with their contribution is therefore three-fold. Their binary-event dominance result is uniform within their class of single-frequency-eliciting rules; the present comparison is contingent across multinomial belief regimes and rules, and a uniform winner is shown not to exist. What the partial-identification framework adds beyond restated bounds is a unified treatment of linear-functional inference across rules: mean and linear-functional bounds on a single polytope of single-transfer inequalities for the closed-form rules, and on a single threshold-indexed box-simplex slice for Manhattan, without per-rule bespoke derivations. Schlag et al. (2015) survey belief-elicitation methods broadly; we study a specific within-method comparison—frequency-report rules under count losses—that the survey identifies as understudied.

A second literature is the theory of *elicitability*. The contrast drawn here between the mean-oriented squared-distance rule and the median-oriented Manhattan rule is an instance of a classical correspondence: squared-error loss elicits the mean and absolute-error loss elicits the median (Gneiting, 2011), and, more generally, a functional of a distribution is elicitable exactly when its level sets are convex (Lambert et al., 2008). That correspondence is not our contribution; its use is. The reports studied here are integer count vectors generated against a multinomial law, and each rule is inverted into an identified set of beliefs rather than evaluated as a point forecast. The frequency-guessing rule sits at the edge of the correspondence: it elicits the multinomial mode, and the mode is not directly elicitable for general distributions—it has the maximal elicitation complexity of the full distribution (Lambert et al., 2008) and is not elicitable at all in the continuous case (Heinrich, 2014). What makes mode elicitation possible here is the discreteness of the count outcome space, which renders the relevant identified set a polytope.

A third literature concerns what a coarse belief report identifies. A point forecast does not reveal whether it reports the mean, median, or mode of the underlying distribution (Gneiting, 2011; Engelberg et al., 2009), and Engelberg et al. (2009) accordingly bound the means, medians, and modes of forecasters' subjective distributions from their coarse interval-probability reports. Set-valued inference about a belief distribution from a coarse report is, in that sense, not new; the difference here is the identifying restriction. Engelberg, Manski, and Williams take the reported interval probabilities at face value, whereas the identified set $P_S(r)$ is cut out by the scoring rule's optimality condition—the report is informative only because it was incentive-generated. The count format itself is also established: survey research elicits subjective probability distributions directly (Manski, 2004), often by having respondents allocate a fixed number of tokens or beans across bins (Delavande et al., 2011), which is a frequency report in all but name.

A fourth literature is the experimental study of elicitation methods, surveyed in Schlag et al. (2015), Schotter and Trevino (2014), and Charness et al. (2021). Two of its recurring concerns are robustness to risk attitudes—direct monetary scoring rules distort reports for risk-averse subjects in ways that depend on stakes and hedging opportunities (Armantier and Treich, 2013)—and behavioral simplicity. Trautmann and van de Kuilen (2015) run a "horse race" among elicitation mechanisms on the accuracy of the beliefs they recover, and Offerman et al. (2009) keep a standard scoring rule but correct the elicited report ex post for risk attitudes and probability weighting; both aim to recover a point belief, by mechanism choice

only semi-analytical (threshold-determined). Hamming and Chebyshev distance are the rules where one of the two hypotheses fails—the limits of the approach, marked in Section 3.6.

or by calibration. The exercise here differs in object: it holds optimal reporting fixed and compares count-loss rules by the sharpness of the belief *set* each report identifies, with no calibration step. That comparison is conditional on optimal reporting, an assumption that can fail behaviorally—Danz et al. (2022) find that the binarized scoring rule, a leading risk-robust mechanism, produces systematically center-biased reports. The informativeness ranking obtained here is thus an upper bound on what a rule delivers, not a behavioral prediction; Section 5.2 returns to this as a scope condition.

2 Setup

The environment is a finite belief-elicitation problem. There are k possible outcomes, and the subject has latent beliefs

$$p = (p_1, \ldots, p_k) \in \Delta^k, \quad p_i \geq 0, \quad \sum_{i=1}^{k} p_i = 1.$$

The elicitation task asks about $n \in \mathbb{N}$ independent repetitions. If the subject's beliefs are p, the realized count vector is

$$\omega = (\omega_1, \ldots, \omega_k) \sim \text{Mult}(n, p).$$

The set of feasible count vectors is

$$\Omega = \left\{ \omega \in \mathbb{N}_0^k : \sum_{i=1}^{k} \omega_i = n \right\},$$

and the multinomial probability mass function is

$$\Pr_p(\omega) = \frac{n!}{\omega_1! \cdots \omega_k!} \prod_{i=1}^{k} p_i^{\omega_i}.$$

The subject reports another feasible count vector

$$r = (r_1, \ldots, r_k) \in \Omega.$$

Definition 1 (Frequency-report scoring rule)**.** A frequency-report scoring rule is a function

$$S : \Omega \times \Omega \to \mathbb{R},$$

where $S(r, \omega)$ is the score assigned to report r when the realized count vector is ω. We study rules that reward reports close to realized counts. A convenient form is

$$S_D(r, \omega) = a - bD(r, \omega), \quad b > 0,$$

where $D : \Omega \times \Omega \to \mathbb{R}_+$ is a count-error index. The function D need not be a metric: squared distance is a divergence-style count loss, not a metric, because it fails the triangle inequality.

Two clarifications of terminology are in order. A *frequency report* $r \in \Omega$ is a vector of counts—how many of the n realizations fall in each category—not a vector of relative frequencies; dividing by n recovers the latter. And these are not *proper* scoring rules in the sense of Savage (1971) and Gneiting and Raftery (2007): they are not designed to make a truthful report of the belief vector p optimal, and generally do not. As the next section shows, the optimal report is instead a functional of the multinomial sampling distribution that p induces—its mode, its mean, or its coordinate-wise medians, depending on the rule—so recovering p calls for the inverse, partial-identification step developed below.

The constants a and b do not matter for expected-score rankings as long as $b > 0$. They matter for implementation. If scores are paid directly as money and nonnegative payments are desired, one can choose

$$a \geq b \max_{r,\omega \in \Omega} D(r, \omega).$$

How the score is implemented as a payment affects robustness to the subject's risk attitude; our discussion of risk aversion and implementation takes up this question separately.

Definition 2 (Optimal reports)**.** Under risk neutrality, the optimal-report correspondence induced by a scoring rule S is

$$R_S(p) = \arg\max_{r \in \Omega} \mathbb{E}_p[S(r, \omega)].$$

Because Ω is finite, $R_S(p)$ is nonempty.

Risk neutrality is the maintained assumption throughout the body of the paper; the discussion of risk aversion and implementation separately shows how the analysis extends to risk-averse expected-utility subjects.

This is the forward incentive object. It answers the question: if the subject's latent beliefs are p, which reports are optimal? The econometric object is the inverse question. After observing r, the researcher asks which beliefs could have generated that report.

Definition 3 (Identified set)**.** For an observed report $r \in \Omega$, the identified set induced by S is

$$P_S(r) = \{p \in \Delta^k : r \in R_S(p)\}.$$

Ties are included: if r is one of several optimal reports under p, then $p \in P_S(r)$.

We call $P_S(r)$ the *mechanism-induced identified set*: the identifying restriction is the optimal-report condition $r \in R_S(p)$ imposed by the scoring-rule mechanism, not a sampling distribution over repeated observations. By construction $P_S(r)$ is exactly the set of beliefs consistent with that restriction.

The most direct summaries of $P_S(r)$ are coordinate-wise identification bounds:

$$\underline{p}_{S,i}(r) = \inf_{p \in P_S(r)} p_i, \quad \overline{p}_{S,i}(r) = \sup_{p \in P_S(r)} p_i,$$

with interval

$$P_{S,i}(r) = [\underline{p}_{S,i}(r), \overline{p}_{S,i}(r)].$$

When the scoring rule is clear from context, the subscript S is omitted.

Many experiments require more than coordinate-by-coordinate probability bounds. If outcomes have numerical values $x_1, \ldots, x_k$, the latent mean implied by beliefs p is

$$\mu(p; x) = \sum_{i=1}^{k} p_i x_i.$$

This could represent an expected outcome, an expected treatment share, a predicted group average, or another linear index of the belief distribution. The sharp mean bounds after observing r are

$$\underline{\mu}_S(r; x) = \inf_{p \in P_S(r)} \sum_{i=1}^{k} p_i x_i, \quad \overline{\mu}_S(r; x) = \sup_{p \in P_S(r)} \sum_{i=1}^{k} p_i x_i.$$

The corresponding mean interval is

$$[\underline{\mu}_S(r; x), \overline{\mu}_S(r; x)].$$

These bounds generally require optimizing over the full joint region $P_S(r)$. Coordinate intervals alone do not generally give sharp bounds for a mean, because they ignore dependence across coordinates imposed by the simplex and the scoring rule.

The design comparison later summarizes the informativeness of a rule using coordinate widths,

$$\overline{p}_{S,i}(r) - \underline{p}_{S,i}(r),$$

and functional widths,

$$\overline{\mu}_S(r; x) - \underline{\mu}_S(r; x).$$

Exact-match payment probability is reported only for the frequency-guessing rule, as an implementation diagnostic of that fixed-prize mechanism; it is not a cross-rule comparison metric.

The next section applies this setup to the three headline rules—squared-distance, frequency-guessing, and Manhattan—characterizing the two closed-form rules first and then developing the structural lemma that links them and carries the analysis to Manhattan distance.

3 Frequency-Report Scoring Rules

This section characterizes the three headline rules—squared-distance, frequency-guessing, and Manhattan-distance scoring—and then locates the boundary of the approach. The rules differ in inferential orientation—squared-distance scoring is mean-oriented, frequency-guessing elicits multinomial modes, and Manhattan distance is median-oriented—and they span a spectrum of analytical tractability. The two rules with closed-form identified sets and closed-form coordinate bounds come first, followed by an analytical comparison of their bounds. A structural subsection then explains why the two characterizations take the same form—each optimal report solves a separable resource-allocation problem—and supplies the tool that carries the analysis to Manhattan-distance scoring, whose bounds are sharp but semi-analytical: exact, yet located by a one-dimensional numerical step rather than a formula. The closing subsection characterizes Hamming and Chebyshev distance, the two rules whose count losses fail the structural requirements and which mark where tractable sharp bounds stop.

3.1 Squared Distance: Mean Projection and Closed-Form Bounds

The analytical centerpiece is squared-distance frequency scoring. It pays less when the reported count vector is far from the realized count vector in squared Euclidean distance:

$$S_Q(r,\omega) = a - b\|r-\omega\|_2^2, \quad b>0.$$

This rule is connected in spirit to quadratic scoring rules, but it is not a standard probability score. It is a frequency-report count-loss rule: subjects report counts, and the penalty is the squared difference between reported and realized counts. Write R_Q and P_Q for the report correspondence and identified set induced by this rule.

Given beliefs p, the expected count vector is np. The key forward fact is that squared-distance scoring asks the subject to report the feasible integer count vector closest to np. The inverse question is therefore geometric: for which beliefs is the observed report r the closest feasible count vector to np?

Proposition 1 (Squared distance: reports, regions, and bounds)**.** *For every $p \in \Delta^k$,*

$$R_Q(p) = \arg\min_{s\in\Omega} \|s-np\|_2^2.$$

For an observed report $r \in \Omega$, the identified set is

$$P_Q(r) = \left\{ p \in \Delta^k : n(p_i - p_j) \le r_i - r_j + 1 \quad \forall i,j \text{ with } r_j > 0 \right\}.$$

Let

$$m(r) = |\{j : r_j > 0\}|$$

be the number of categories with positive reported counts. The sharp coordinate intervals are

$$P_{Q,i}(r) = \begin{cases} \left[0, \dfrac{m(r)}{n(m(r)+1)}\right], & r_i = 0, \\ \left[\dfrac{r_i-1}{n} + \dfrac{1}{nk}, \dfrac{r_i+1}{n} - \dfrac{1}{n\,m(r)}\right], & r_i > 0. \end{cases}$$

Projection ties are included.

The inequalities compare the observed report r with reports obtained by moving one count from a positive reported category j to another category i. If all such single-count moves weakly increase squared distance from np, then r is a closest feasible report. Thus $P_Q(r)$ is a polytope in the belief simplex—a region cut out by finitely many linear inequalities—so sharp bounds on means and other linear functionals $\sum_i p_i x_i$ are linear programs over the same region that yields the closed-form coordinate intervals. The proof is in Appendix A.1.

Practical interpretation. Squared distance is mean-oriented: it links a report r to beliefs for which r is the integer projection of np—the count-report counterpart of squared-error loss eliciting the mean (Gneiting, 2011). Its main implementation cost is that direct monetary squared-distance payments generally require risk neutrality; the binary-lottery implementation discussed below removes the requirement.

3.2 Frequency Guessing: Multinomial Modes and Closed-Form Bounds

The second closed-form rule is the frequency-guessing mechanism of Schlag and Tremewan (2021). Its count-error index is the count-vector discrete metric,

$$D_0(r, \omega) = \mathbf{1}\{r \neq \omega\},$$

so, in the form $S_D = a - bD$ of Section 2, the score is

$$S_0(r, \omega) = a - b\,\mathbf{1}\{r \neq \omega\}, \quad b > 0.$$

Equivalently, $S_0(r, \omega) = (a - b) + b\,\mathbf{1}\{r = \omega\}$: the subject receives an extra b whenever the realized count vector equals the report. This fixed-prize reading is useful for implementation and risk aversion; the two formulations induce the same optimal reports. Write R_0 and P_0 for the report correspondence and identified set induced by this rule.

The mechanism is simple: a subject with beliefs p maximizes the probability of an exact match,

$$\Pr_p(\omega = r) = \frac{n!}{r_1! \cdots r_k!} \prod_{i=1}^{k} p_i^{r_i}.$$

It follows that the forward object is the multinomial mode set. The inverse object is the set of beliefs for which the observed report is one of those modes.

Proposition 2 (Frequency guessing: reports, regions, and bounds)**.** *Let $r \in \Omega$, and allow ties in the multinomial mode. Then r is optimal under frequency-guessing scoring if and only if r is a multinomial mode:*

$$R_0(p) = \arg\max_{s \in \Omega} \Pr_p(\omega = s).$$

The identified set is

$$P_0(r) = \left\{ p \in \Delta^k : r_j p_i \leq (r_i + 1) p_j \quad \forall i, j \text{ with } r_j > 0 \right\}.$$

For each coordinate, the sharp identification interval is

$$P_{0,i}(r) = \left[\frac{r_i}{n + k - 1}, \frac{r_i + 1}{n + 1} \right].$$

The inequalities say that no single-count transfer from a reported category j to another category i can increase the multinomial probability of the report. Weak inequalities include ties. The formula also handles boundaries: if $r_i = 0$, the lower bound is 0; if $r_i = n$, the upper bound is 1. These are the frequency-guessing bounds in Schlag and Tremewan (2021, Proposition 1); they are restated here to make the informativeness comparison self-contained. $P_0(r)$ is likewise a polytope, so sharp mean and linear-functional bounds are again linear programs over the joint region rather than combinations of the coordinate intervals. The proof is in Appendix A.2.

Practical interpretation. Frequency-guessing scoring is robust to risk aversion under direct payment because winning is a fixed-prize event, as the discussion of risk aversion below makes precise. Its implementation cost is that the probability of an exact match can be small when n is large or the outcome space is rich; the design comparison quantifies this. It is a prominent known rule in the comparison, not a new mechanism introduced here.

3.3 Comparing the Two Closed-Form Rules

The closed-form coordinate bounds of Propositions 1 and 2 can be compared directly, before any simulation. For a rule S and report r, call the mean of the coordinate widths $\overline{p}_{S,i}(r) - \underline{p}_{S,i}(r)$ over the k coordinates the *average coordinate width*.

Corollary 1 (Average coordinate width of the closed-form rules)**.** *Fix $n \geq 2$ and $k \geq 2$, and for a report r let $m = m(r)$ be the number of categories with positive reported counts.*

1. *The frequency-guessing average coordinate width equals the constant*

$$W_0(n,k) = \frac{(k-1)(2n+k)}{k(n+1)(n+k-1)},$$

   the same for every report r.

2. *The squared-distance average coordinate width depends on r only through m:*

$$W_Q(m;n,k) = \frac{1}{nk}\left[2m - 1 - \frac{m}{k} + \frac{m(k-m)}{m+1}\right].$$

3. *The two widths cross: $W_Q(1;n,k) < W_0(n,k) < W_Q(k;n,k)$.*

The proof—elementary algebra on the two coordinate-bound formulas—is in Appendix A.3.

Parts (1) and (2) together say something clean: the frequency-guessing average width is a constant, while the squared-distance average width is a function of m alone, so the comparison between the two closed-form rules is decided entirely by $m = m(r)$, the number of categories the report touches. Part (3) shows that comparison is not one-sided—the squared-distance average width falls below the frequency-guessing constant at the most concentrated report and rises above it at $m = k$—so which rule gives the narrower average coordinate bound genuinely turns on how concentrated the report is. The design comparison of the next section ties this report-level contrast to belief concentration, quantifies it, and extends it to Manhattan distance.

3.4 The Common Structure: Optimal Reports as Resource Allocation

The characterizations of Propositions 1 and 2 run in parallel: each identified set is cut out by one linear inequality per single-count transfer. The parallel is structural. For each headline rule, an optimal report minimizes a *separable* objective over the feasible count vectors,

$$R_S(p) = \arg\min_{r \in \Omega} \sum_{i=1}^{k} \ell(r_i; p_i),$$

for a per-coordinate cost ℓ. For squared-distance scoring the expected loss is itself additively separable, and ℓ is the expected per-coordinate loss; the same holds for the Manhattan-distance rule $S_M(r,\omega) = a - b\sum_i |r_i - \omega_i|$, taken up in the next subsection. Frequency-guessing scoring is not an additively separable loss—the multinomial probability is a product—but maximizing $\Pr_p(\omega = r)$ is equivalent to minimizing $\sum_i \{\log(r_i!) - r_i \log p_i\}$, so it has the same form, with $\ell(t;p_i) = \log(t!) - t\log p_i$.

This is the integer *separable resource-allocation problem*: distribute n units across k categories under separable costs. Its solution structure is classical (Fox, 1966; Federgruen and Groenevelt, 1986; Ibaraki and Katoh, 1988). The intuition for what follows is that, under discrete convexity, no local single-count transfer being profitable rules out any larger transfer too: each successive count moved along a path to a competing report is at least as costly as the first. In other words, the local no-improvement condition is also the global one.

Lemma 1 (Single-transfer characterization)**.** *Suppose an optimal report minimizes a separable objective $\sum_i \ell(r_i; p_i)$ over Ω, and each $\ell(\cdot; p_i)$ is* discrete-convex*: its successive differences $\ell(t+1; p_i) - \ell(t; p_i)$ are nondecreasing in t. Then r is optimal if and only if no single-count transfer lowers the objective,*

$$\ell(r_i + 1; p_i) - \ell(r_i; p_i) \;\geq\; \ell(r_j; p_j) - \ell(r_j - 1; p_j) \qquad \forall\, i, j \text{ with } r_j > 0,$$

and the identified set $P_S(r)$ is exactly the set of beliefs satisfying these $O(k^2)$ inequalities.

Necessity of the no-transfer condition is immediate: a transfer produces a feasible competing report. Sufficiency is a monotone-path exchange argument—any competing report is reached by single-count transfers along which discrete convexity makes each step at least as costly as the corresponding transfer evaluated at r; it is proved in Appendix A.4. That the equivalence holds precisely when the per-coordinate costs are discrete-convex is the classical greedy-optimality characterization for separable resource allocation (Fox, 1966; Federgruen and Groenevelt, 1986), with discrete convexity the standard diminishing-differences condition (Murota, 2003).

The three headline rules satisfy the hypothesis: the squared-distance cost is quadratic in r_i, the Manhattan cost $\mathbb{E}_p|r_i - \omega_i|$ inherits discrete convexity from the absolute value, and the frequency-guessing cost $\log(r_i!) - r_i \log p_i$ is discrete-convex because $\log(t!)$ has nondecreasing differences. At boundary beliefs the frequency-guessing cost is read in the extended reals—it equals $+\infty$ when $p_i = 0$ and $r_i > 0$, so optimal reports place no count outside the support of p (any report concentrated on the support has finite loss); restricted to such reports the lemma applies verbatim, its successive-difference comparisons remaining well defined with the differences extended by continuity from $p_i \in (0, 1)$.

Propositions 1 and 2 are therefore instances of Lemma 1, and their single-transfer inequalities are linear in p—which is why the identified sets are polytopes and the coordinate bounds closed form. For Manhattan scoring the inequalities are monotone but transcendental in p—binomial distribution functions rather than linear forms—so no closed form is available; the next subsection shows that the identified set is nonetheless exact and the sharp coordinate bounds reduce to one-dimensional computations. The lemma thus plays two different roles: for the closed-form rules it organizes what their self-contained proofs already establish, while for Manhattan distance it does real work—sufficiency of the single-transfer condition is exactly what makes the identified set of the next subsection exact rather than an outer approximation.

The lemma's reach is not confined to these three: any count loss of the location form $\sum_i \psi_i(r_i - \omega_i)$ with convex ψ_i is governed by it in the same way, because taking expectations preserves discrete convexity coordinate by coordinate—asymmetric-absolute losses, for instance, which elicit quantiles other than the median—so the headline rules are representatives of a wider tractable family rather than a closed list. The ranked probability score, a squared distance on *cumulative* counts, lies outside this family as stated: its expected loss

is separable in the cumulative rather than the per-category counts, and the lemma's transfer argument does not carry over without modification.

The lemma also marks the limits of the approach: its two hypotheses—separability and discrete convexity—are exactly what a count loss can fail, and Section 3.6 characterizes the two rules that do. The optimization content of the lemma is classical; what we add is the observation that incentivized frequency-report elicitation *is* a resource-allocation problem of this kind, and the inversion of the resulting optimal-report correspondence into an identified set of latent beliefs.

3.5 Manhattan Distance: Binomial Medians and Semi-Analytical Bounds

Unlike the two closed-form rules, Manhattan-distance scoring has no closed-form coordinate bounds; its optimal-report correspondence and identified set are nonetheless characterized exactly, and its sharp bounds reduce to a one-dimensional computation. The Manhattan-distance rule is

$$S_M(r,\omega) = a - b\sum_{i=1}^{k} |r_i - \omega_i|, \quad b > 0.$$

It rewards total absolute count accuracy rather than squared count accuracy, and its inferential interpretation is median-oriented rather than mean-oriented—the count-report counterpart of absolute-error loss eliciting the median (Gneiting, 2011).

For category i the realized count has binomial marginal $W_i \sim \mathrm{Bin}(n, p_i)$; write $F_i(t; n, p_i) = \Pr(W_i \le t)$ for its distribution function; by convention $F_i(-1; n, p_i) = 0$ and $F_i(n; n, p_i) = 1$. By linearity of expectation the expected Manhattan loss is separable across categories, $\mathbb{E}_p \sum_i |r_i - \omega_i| = \sum_i \mathbb{E}_p |r_i - W_i|$, while the simplex constraint $\sum_i r_i = n$ couples the coordinates.

Proposition 3 (Manhattan distance: optimal reports and identified set)**.** *Optimal reports minimize the separable expected Manhattan loss,*

$$R_M(p) = \arg\min_{r \in \Omega} \sum_{i=1}^{k} \mathbb{E}_p |r_i - W_i|,$$

and r is optimal if and only if no single-count transfer lowers that loss, equivalently

$$F_i(r_i; n, p_i) \ge F_j(r_j - 1; n, p_j) \qquad \forall\, i, j \text{ with } r_i < n,\ r_j > 0.$$

This condition is sufficient as well as necessary, so the identified set is exactly

$$P_M(r) = \left\{ p \in \Delta^k :\ F_i(r_i; n, p_i) \ge F_j(r_j - 1; n, p_j)\ \ \forall\, i, j \text{ with } r_i < n,\ r_j > 0 \right\}.$$

Equivalently, $p \in P_M(r)$ if and only if some threshold $c \in [0,1]$ satisfies

$$F_i(r_i - 1; n, p_i) \le c \le F_i(r_i; n, p_i) \qquad \forall\, i.$$

Each coordinate loss $\mathbb{E}_p|t - W_i|$ is discrete-convex in the reported count t—its successive increments are nondecreasing—so Lemma 1 applies and $P_M(r)$ is the exact identified set rather than an outer approximation. Weak inequalities include ties, and there is no receiver constraint when $r_i = n$ nor sender constraint when $r_j = 0$. The proof is in Appendix A.5.

Identification bounds. For $k = 2$, write $r = (t, n-t)$; the Manhattan loss is $2|t - W|$ with $W \sim \text{Bin}(n, p_1)$, and the report identifies the binomial-median interval

$$P_{M,1}(r) = \{p_1 \in [0,1] : F(t-1; n, p_1) \leq 1/2 \leq F(t; n, p_1)\}.$$

This binary case is clean: for $n = 10$ and the report $r = (6, 4)$, inverting the binomial CDF at $1/2$ at the integer boundaries $t - 1 = 5$ and $t = 6$ gives $P_{M,1}(r) \approx [0.55, 0.64]$—the set of beliefs at which 6 is a median of $\text{Bin}(10, p_1)$.

For $k > 2$ the threshold representation of Proposition 3 drives the bounds. At a fixed threshold c the inequalities $F_i(r_i - 1; n, p_i) \leq c \leq F_i(r_i; n, p_i)$ invert, coordinate by coordinate, to a box of beliefs through inverse binomial CDFs; intersecting that box with the simplex gives the feasible beliefs at that threshold. Each sharp coordinate bound is then the optimum, over $c \in [0, 1]$, of a function that is unimodal in c, so it is the solution of a single monotone scalar equation; see Lemma 2 in Appendix A.5. The coordinate bounds are therefore sharp and semi-analytical: not closed form, because the inverse binomial CDF is not elementary, but exact and one-dimensional rather than grid-limited—each coordinate bound is the unique root of the monotone scalar equation of Lemma 2 and is located by bisection. The sharp *mean* bounds traverse the same threshold parameterization: the identified mean is a linear functional of p optimized over the same threshold-indexed box-simplex slice, computed on a fine grid of c-values (with reported tolerance) rather than by a one-dimensional root-find.

Practical interpretation. Manhattan distance is median-oriented, as the binomial-median interval above makes concrete. Its multi-category coordinate bounds are sharp and reduce to a one-dimensional root-find, and its mean bounds are computed on the threshold parameterization to a stated tolerance—exact or tolerance-controlled, not approximations, though not closed form. The design comparison of the next section gives the rule its practical role: rarely the most informative, but nearly invariant to the belief regime.

Other functionals of the induced outcome distribution. The identified sets of all three rules yield bounds on a broader inferential menu—variance, median, other quantiles—when a numerical or ordinal outcome vector is specified. The available functionals depend on what the outcome vector represents (no x: coordinate bounds only; ordinal x: median and quantiles; numerical x: mean, variance, and higher moments). For squared-distance and frequency-guessing the polytope identified sets reduce each functional to a linear program (mean, quantile), a convex quadratic program (variance upper bound), or a vertex search (variance lower bound). For Manhattan-distance the threshold-indexed slice family supports the same operations on each slice; the supremum or infimum is then taken over the threshold parameter, as established in Lemma 3 of Appendix A.6, which also reports a worked example. Schlag and Tremewan (2021) derive the corresponding bounds for frequency-guessing in their setting; the appendix carries that program to all three rules.

3.6 Two Rules Outside the Structure: Hamming and Chebyshev Distance

The three rules characterized above do not exhaust the family of count losses, and the partial-identification approach is not a general theory of discrete scoring rules. Two further rules—Hamming and Chebyshev distance—are worth stating precisely, because they mark where the approach stops yielding tractable sharp bounds. They are, in fact, the two ways the hypotheses

of Lemma 1 can fail: Hamming distance meets its separability hypothesis but not its discrete-convexity hypothesis, and Chebyshev distance fails separability itself.

Hamming distance. The Hamming rule $S_H(r,\omega) = a - b\sum_i \mathbf{1}\{r_i \neq \omega_i\}$ penalizes the number of categories whose reported count differs from the realized count. By linearity of expectation its expected loss is $\sum_i\{1-\Pr(\mathrm{Bin}(n,p_i)=r_i)\}$, so an optimal report maximizes the separable objective $\sum_i \Pr(\mathrm{Bin}(n,p_i)=r_i)$ subject to $\sum_i r_i = n$, and $P_H(r)$ is the set of beliefs at which r attains that maximum—cut out by one expected-loss inequality per alternative report. For $k=2$ the two coordinate indicators $\mathbf{1}\{r_i \neq \omega_i\}$ are equal, so the Hamming count-error index $\sum_i \mathbf{1}\{r_i \neq \omega_i\}$ is exactly twice the discrete metric $\mathbf{1}\{r \neq \omega\}$; the rule coincides with frequency-guessing scoring and inherits its closed-form bounds. For $k > 2$ it does not. Each term $\Pr(\mathrm{Bin}(n,p_i)=r_i)$ is unimodal but not discrete-concave in the reported count, so the exchange argument that makes single-count-transfer optimality sufficient for Manhattan distance fails here: the condition remains necessary but is no longer sufficient, because a transfer of two or more counts can raise the separable objective where no single-count transfer does. This failure is not confined to the boundary of the simplex. At $n=3$, $k=5$, and uniform beliefs $p=(1/5,\ldots,1/5)$, for instance, the report $(3,0,0,0,0)$ admits no improving single-count transfer, yet $(0,0,1,1,1)$ yields a strictly higher expected score. The instance also certifies that the failure is not an artifact of the particular separable representation: a strictly increasing transformation of the expected *loss* preserves both rankings—$(3,0,0,0,0)$ attains weakly lower transformed loss than each of its single-transfer neighbours while $(0,0,1,1,1)$ attains strictly lower transformed loss—so if any such transformation admitted a separable representation with discrete-convex per-coordinate costs, Lemma 1 would force $(3,0,0,0,0)$ to be globally optimal at these beliefs; none therefore exists. The logarithmic transformation that separates the frequency-guessing score has no Hamming counterpart. The identified set $P_H(r)$ is therefore a non-convex set defined by polynomial inequalities, not a polytope. The modal box $\prod_i[r_i/(n+1),(r_i+1)/(n+1)]$ collects the beliefs at which every r_i is a mode of its binomial marginal; there the separable objective is maximized term by term, so r is optimal, and the modal box is a closed-form inner bound for $P_H(r)$, while the single-count-transfer inequalities give an outer bound. The sharp coordinate bounds lie between these two and require optimizing a coordinate directly over the non-convex set—a global optimization whose feasible region is cut out by one polynomial inequality for each of the $\binom{n+k-1}{k-1}$ feasible reports, a constraint count that grows combinatorially and exceeds 10^{10} at the largest cell of the design comparison of Section 4 ($n=50$, $k=10$). A direct feasibility check bore out the obstruction: computing the sharp Hamming bounds proved both prohibitively slow and numerically unreliable for $k=10$, while the closed-form bounds of the other rules are immediate. Hamming distance is for this reason characterized here rather than included in the comparison.

Chebyshev distance. The Chebyshev rule $S_\infty(r,\omega) = a - b\max_i |r_i - \omega_i|$ penalizes the largest single-category count error. Unlike the other count losses, its expected loss does not separate across coordinates: the maximum couples the category errors within every realization, so there is no per-coordinate transfer condition to exploit. For $k=2$ the two absolute errors $|r_i - \omega_i|$ are equal, so the count-error index $\max_i |r_i - \omega_i|$ is half of $\sum_i |r_i - \omega_i|$; the rule coincides with Manhattan scoring and inherits the binomial-median interval. For $k > 2$ neither half of the pipeline is tractable at scale. The non-separable objective admits no transfer or exchange argument, so no characterization of the optimal report beyond search is known:

a certified optimal report must be found by comparing all $\binom{n+k-1}{k-1}$ feasible reports. The identified set, in turn, is cut out only by the full system of expected-loss inequalities, one per alternative report, with no transfer-polytope or threshold reduction—the same kind of non-convex semialgebraic set as $P_H(r)$, whose sharp bounds the Hamming feasibility check above already found intractable at $k = 10$. Chebyshev distance is therefore computable by enumeration for small n and k but, like Hamming, cannot be carried across the design grid; it is characterized here rather than included in the comparison.

These two rules are not pathological—their identified sets are well defined and finite to check—but they show that the closed-form and semi-analytical tractability of the headline rules is a property of those rules, not a generic feature of count-loss scoring.

4 Informativeness: A Design Comparison

The analytical regions are useful because they tell an experimenter what can be inferred after observing a report. This section reports a design comparison generated by the reproducible script `scripts/design_efficiency.py`. The exercise draws latent beliefs

$$p \sim \text{Dirichlet}(\alpha, \ldots, \alpha)$$

and, for each of the three rules with tractable sharp bounds—frequency-guessing, squared-distance, and Manhattan distance—computes an optimal report $r_S(p)$ and the bounds induced by $P_S(r_S(p))$. Hamming and Chebyshev distance fall outside this comparison; Section 3.6 characterized them as count-loss rules whose identified sets are exact but resist tractable sharp bounds. The final run uses 5,000 belief draws for each cell in

$$n \in \{5, 10, 20, 50\}, \qquad k \in \{2, 3, 5, 10\}, \qquad \alpha \in \{0.1, 0.3, 1, 3, 10\}.$$

The Dirichlet parameter α sets belief concentration: small α draws sparse beliefs near a vertex of the simplex, large α draws balanced, interior beliefs.[3] For each rule the comparison records two headline inferential targets—the average width of the coordinate bounds, and the width of the bound on the linear functional $\sum_i p_i x_i$ with outcome values $x_i = (i-1)/(k-1)$. Manhattan coordinate bounds are sharp, computed by a threshold root-find; Manhattan mean bounds are threshold-computed with tolerance 10^{-4}.

The comparison is conditional on the theoretical characterizations above. It is not evidence of incentive compatibility, and it is not an optimality theorem. It plays three roles: it confirms the analytic crossover between the two closed-form rules established in Section 3, quantifies that crossover through win shares and regret, and extends the comparison to Manhattan distance, whose lack of a closed-form bound rules out a comparable analytic treatment. It is, in short, a diagnostic for the practical question: which rule gives a researcher tighter inference for the object of interest?

[3] Symmetric Dirichlet draws are standard in belief-elicitation simulations because they parameterize dispersion through a single concentration scalar—small α places mass near a vertex, large α near the center of the simplex—and a scalar concentration is what an experimenter typically has a prior over. The cost is restrictiveness: a symmetric Dirichlet is exchangeable across categories, so its draws cannot capture geometries in which the subject's beliefs favor particular outcomes more than others. We check this directly in Appendix B with nine asymmetric concentration vectors at $n = 20$, $k = 5$. The regime story carries over, with one exception the appendix discusses—squared-distance scoring extends its dominance into the high-α-spread cells of the linear-mean target. The asymmetric runs are reported separately because the main exercise is organized around the scalar α, which has no direct counterpart in the asymmetric case.

The ranking is governed by belief concentration. Corollary 1 showed, from the closed-form bounds alone, that which of the two closed-form rules gives the narrower average coordinate bound turns on report concentration—the number m of categories with positive reported counts. The design comparison shows that report concentration tracks belief concentration. In other words, the crossover is governed by how concentrated the subject's beliefs are. Averaged over the whole grid the three rules barely differ—mean coordinate width 0.103, 0.104, and 0.102 for frequency-guessing, squared-distance, and Manhattan scoring—because each rule's favorable and unfavorable regimes cancel. Table 1 reports the win shares—the share of belief draws on which a rule gives the narrowest interval—broken out by α.[4]

	Average-coordinate			Linear-mean		
α	Guessing	Squared	Manhattan	Guessing	Squared	Manhattan
0.1	0.08	0.84	0.07	0.08	0.83	0.10
0.3	0.28	0.58	0.14	0.22	0.60	0.18
1	0.76	0.16	0.08	0.58	0.27	0.15
3	0.98	0.01	0.01	0.83	0.13	0.04
10	1.00	0.00	0.00	0.92	0.06	0.01

Table 1: Win share by belief concentration α: the share of belief draws on which a rule gives the narrowest interval, averaged over all n and k at each α (5,000 symmetric-Dirichlet draws per design cell). Squared-distance scoring wins for sparse beliefs (small α); frequency-guessing scoring for balanced beliefs (large α).

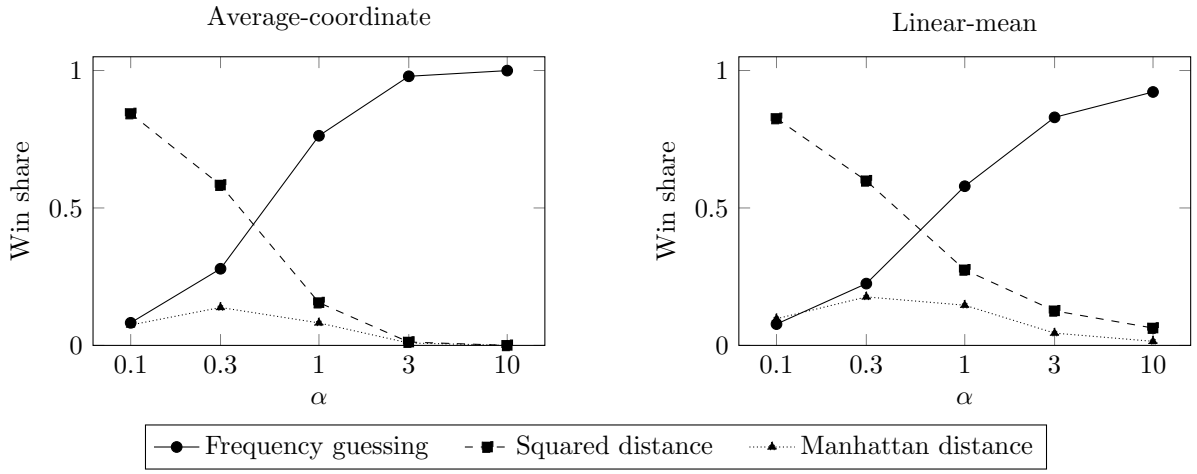

Figure 1: Win share by belief concentration α (log scale), averaged over all n and k: the share of belief draws on which a rule gives the narrowest interval, for the average-coordinate target (left) and the linear-mean target (right). Same data as Table 1.

Squared-distance scoring is the most informative rule when beliefs are sparse and frequency-guessing scoring when they are balanced, with a crossover near $\alpha = 1$; Figure 1 traces the two win-share crossings. At $\alpha = 0.1$ squared-distance gives the narrowest average-coordinate in-

[4]Each row of Table 1 averages over 5,000 draws per cell and over the 4×4 (n, k) grid, so 80,000 draws per row; the binomial standard error on a single-rule win share is at most $\sqrt{0.25/80{,}000} \approx 0.0018$, well under the two displayed decimals.

terval on 0.84 of draws; by $\alpha \geq 3$ frequency-guessing does so on essentially every draw. The linear-mean target shows the same crossover. The number of categories k and the number of trials n are second-order, as Table 4 in Appendix B documents: for the coordinate target the ranking is nearly flat in both, while for the mean target larger n mildly favors squared-distance and the binary case $k = 2$ favors frequency-guessing. The worst-coordinate target follows the same pattern shifted in squared-distance's favor—squared-distance gives the narrowest maximum-coordinate interval on a majority of draws for every $\alpha \leq 1$. The ranking is robust to the choice of outcome vector and is unaffected by optimal-report ties, which are negligible under the Dirichlet draws; Appendix B reports both checks, along with nine asymmetric concentration vectors at $n = 20$, $k = 5$ (Table 5). Squared-distance dominates the three concentrated-asymmetric cells and the transition cell (at least one component $\alpha_i \leq 0.5$); on the average-coordinate target frequency-guessing dominates all five balanced-asymmetric cells (all components $\alpha_i \geq 2$), with the margin shrinking as the α-spread grows; on the linear-mean target the ranking reverses within the balanced regime, with squared-distance taking over at the steepest α-spreads. The governing variable is therefore not only marginal concentration but also, within the balanced regime, the steepness of the gradient.

The crossover has a mechanism in the projection interpretation of squared-distance scoring. A subject reports the integer count vector closest to np; when beliefs are sparse, np lies near a vertex of the simplex, the feasible reports around it are extreme count vectors, and the projection polytope is tight, so the induced bounds are narrow. As beliefs become balanced, np moves into the interior, the projection polytope widens, and frequency-guessing scoring—which elicits the multinomial mode—becomes the more informative rule.

The cost of the wrong rule, and Manhattan as a hedge. Win shares count only first places. Cell-best regret—a rule's interval width minus the narrowest width available in the same cell—measures how much precision is forfeited by not using the best rule. Table 2 reports regret by α.

	Average-coordinate			Linear-mean		
α	Guessing	Squared	Manhattan	Guessing	Squared	Manhattan
0.1	0.032	0.002	0.011	0.047	0.002	0.013
0.3	0.016	0.005	0.007	0.024	0.005	0.009
1	0.003	0.012	0.007	0.006	0.013	0.008
3	0.000	0.019	0.011	0.001	0.020	0.012
10	0.000	0.023	0.014	0.000	0.024	0.016

Table 2: Mean cell-best regret by belief concentration α: a rule's interval width minus the smallest width among the three rules, averaged over all n and k at each α (5,000 symmetric-Dirichlet draws per design cell). Lower is better. Frequency-guessing and squared-distance scoring each swing between near-zero and substantial regret; Manhattan's regret is nearly flat.

Frequency-guessing and squared-distance scoring each have a regime of near-zero regret and a regime in which regret is substantial; Figure 2 shows the two swinging profiles against Manhattan's flat one. Frequency-guessing's average-coordinate regret runs from 0.000 at $\alpha = 10$ to 0.032 at $\alpha = 0.1$; squared-distance's runs the other way, reaching 0.024 on the mean target at $\alpha = 10$. Against typical interval widths near 0.10 to 0.15, the worst-regime values are a quarter or more of an interval width—a real loss of precision from choosing the rule that does not match the belief regime. Manhattan distance behaves differently. It is never

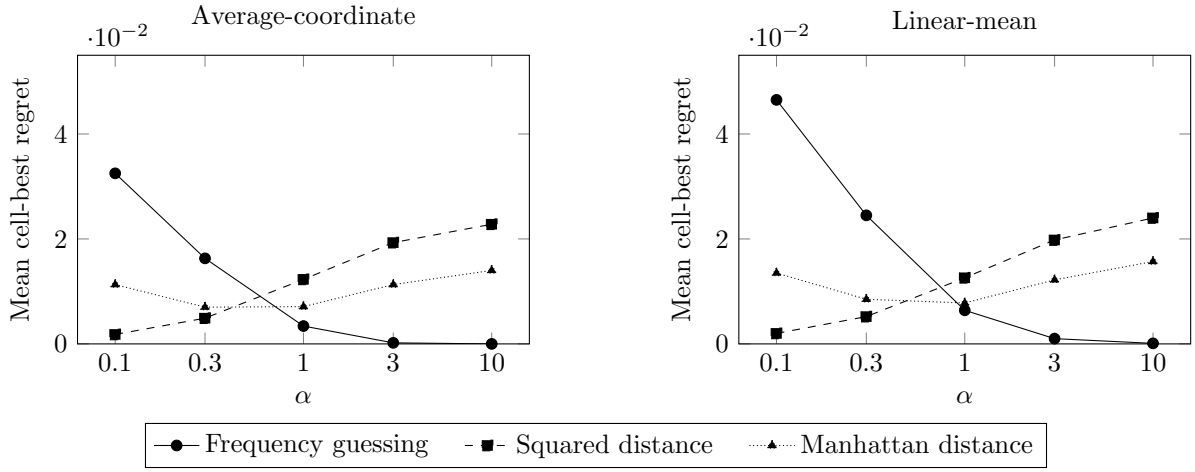

Figure 2: Mean cell-best regret by belief concentration α (log scale), averaged over all n and k: a rule's interval width minus the smallest width among the three rules, for the average-coordinate target (left) and the linear-mean target (right). Same data as Table 2; lower is better. Manhattan's nearly flat profile is the regime-robustness property discussed in the text.

the narrowest rule, but its regret is nearly constant in α—between 0.007 and 0.016 at every concentration and for both targets—so it is never far from the best rule either. Manhattan is therefore the regime-robust choice: a researcher who cannot anticipate how concentrated subjects' beliefs are minimizes worst-case regret by using it, whereas committing in advance to frequency-guessing or squared-distance scoring risks the unfavorable regime. This is the precise sense in which Manhattan is a low-regret rule: it hedges the belief-concentration uncertainty to which the other two rules are exposed.

Payment probability. The frequency-guessing rule has a fixed-prize implementation, which is attractive under risk aversion. Its implementation cost is that the exact-match payment probability can be small. In the final run, the average exact-match probability is high for very sparse binary beliefs, but it falls sharply as n, k, and belief balance increase; for instance, it is essentially zero in the $n = 50, k = 10, \alpha \in \{1, 3, 10\}$ cells. Thus narrow bounds under frequency-guessing scoring should be weighed against the chance that the subject rarely wins the fixed prize.

The comparison thus delivers a contingent recommendation, not a single winner: the most informative rule is the one matched to the anticipated belief regime and inferential target, a choice the discussion turns into practical guidance.

5 Discussion

We study frequency-report scoring rules as inferential devices. The report r is not the belief vector. It is evidence about p, and the scoring rule determines how informative that evidence is. The relevant chain is therefore

$$S \;\to\; R_S(p) \;\to\; P_S(r) \;\to\; \big[\underline{p}_{S,i}(r), \overline{p}_{S,i}(r)\big].$$

The design comparison shows that the most informative rule depends on how concentrated the subject's beliefs are; three practical recommendations follow. First, use frequency-guessing scoring when the subject is expected to spread probability fairly evenly across the outcomes, with no single outcome carrying most of the probability. It is the most informative rule in that regime, has closed-form coordinate bounds, and is the only rule robust to risk aversion under direct payment; its implementation cost is that exact-match payment probabilities can be low away from concentrated reports.

Second, use squared-distance frequency scoring when the subject is expected to place most of the probability on one or a few outcomes. It is the most informative rule in that regime—for the average-coordinate, worst-coordinate, and mean targets alike—and it offers a transparent projection rule, a linear identified set, closed-form coordinate bounds, and linear-program bounds for any linear functional.

Third, use Manhattan distance when it cannot be anticipated in advance whether the subject's beliefs will be concentrated on a few outcomes or spread across many. Its identified set is exact and its coordinate bounds are sharp—semi-analytical, reducing to a one-dimensional threshold computation rather than a closed form. Manhattan rarely gives the single narrowest interval, but the design comparison shows its regret is nearly constant across belief concentrations. In other words, it is never far from the best rule, and so minimizes worst-case regret for a researcher who does not know the regime. Its inferential interpretation is also median-oriented, which suits settings where a median-type count prediction is substantively appropriate.

5.1 Risk Aversion and Implementation

The body of the paper assumes risk neutrality. Whether that assumption can be relaxed depends on how the scoring rule is implemented as a payment. The distinction is between paying the score directly as money and using the score to determine the probability of a fixed prize.

Direct monetary scores. If the subject is paid the score itself, the subject solves

$$\arg\max_{r\in\Omega} \mathbb{E}_p[u(S(r,\omega))].$$

For the frequency-guessing rule this does not change incentives. The payment is a fixed prize B if $r=\omega$ and zero otherwise, so

$$\mathbb{E}_p[u] = u(0) + \Pr_p(\omega = r)\{u(B) - u(0)\},$$

and any strictly increasing u preserves the same optimal reports as maximizing the exact-match probability. Direct monetary distance scores do not have this robustness: squared-distance and Manhattan distance generate several possible payment levels, so nonlinear utility can change the ranking of reports. This is the standard risk-aversion problem for monetary scoring rules (Armantier and Treich, 2013; Schlag et al., 2015). One response in the literature keeps the scoring rule unchanged and corrects the elicited report ex post for the subject's risk attitude and probability weighting (Offerman et al., 2009); the route taken here is instead to choose a rule, or an implementation, under which no such correction is needed.

Probability implementation. The distance rules can instead be implemented as a binary lottery. Let $q(r, \omega) \in [0, 1]$ be the score normalized into a probability, and pay a fixed prize B with probability $q(r, \omega)$ and zero otherwise. If the prize and outside payment are fixed, utility depends only on final money, and the randomization is objective and independent, then

$$\mathbb{E}_p[q(r, \omega)u(B) + (1 - q(r, \omega))u(0)] = u(0) + (u(B) - u(0))\,\mathbb{E}_p[q(r, \omega)],$$

so maximizing expected utility is equivalent to maximizing the expected normalized score, regardless of the curvature of u. A positive affine normalization of the bounded score into $[0, 1]$ preserves the expected-score ranking of reports, so the body's risk-neutral analysis carries over unchanged. This is the binary-lottery, or binarized-scoring-rule, technique (Roth and Malouf, 1979; Berg et al., 1986; Karni, 2009; Hossain and Okui, 2013); the contribution here is to apply it, not to introduce it.

The algebra above establishes only *theoretical* robustness, and two qualifications bound what it buys in practice. First, the robustness is to risk aversion *within* expected utility. Subjects who depart from expected utility—for instance through probability weighting—can report non-truthfully even under the binary-lottery implementation, and we do not address such departures. Second, the implementation asks subjects to reason about a compound lottery layered on top of the count report, which partially offsets the cognitive simplicity that motivates frequency reports in the first place. The empirical record reflects this tension. Harrison et al. (2014) find that a binary-lottery procedure does induce approximately risk-neutral reporting in a subjective-probability elicitation task, evidence that the technique can deliver its theoretical promise. But Danz et al. (2022), testing the binarized scoring rule directly, find that informing subjects of the quantitative incentives induces systematic center-biased misreporting—a failure they trace not to risk attitudes, which the binary lottery already neutralizes, but to the incentive information itself. The binary-lottery implementation should therefore be read as resolving the risk-attitude problem the body sets aside while leaving a separate, behavioral one open.[5]

The implementation choice also interacts with a cost noted in the design comparison. The frequency-guessing prize is paid only on an exact match, and the exact-match probability falls sharply as n, k, and belief balance increase, so at larger designs the rule becomes a lottery with a vanishing chance of any payment. The frequency-guessing rule is thus robust to risk aversion even under direct payment but can become nearly unwinnable; the distance rules avoid that under the binary-lottery implementation, whose payment probability is the expected normalized score, at the cost of relying on expected utility.

5.2 An Open Empirical Question

The case for frequency reports rests partly on a behavioral premise we do not test: that subjects reason more readily in counts than in abstract probabilities, and find a distance-based payment more transparent than a probability score. That premise is taken here on the

[5]The body's risk-neutrality assumption is not a substantive constraint on the distance-rule comparison—under the binary-lottery implementation any strictly increasing utility leaves the ranking of reports unchanged, so the risk-neutral analysis transfers verbatim. The frequency-guessing rule does not need the lottery: a fixed-prize event is already robust under direct payment. What the binary-lottery technique buys for the distance rules is the same robustness, at the cost of a compound-lottery implementation whose behavioral implications we take up below. The deviation Danz et al. (2022) document is not a problem of risk attitudes—it survives the binary-lottery correction—but of incentive comprehension, a separate channel orthogonal to the analysis here.

strength of the cited literature. A second, more basic premise maintained throughout the body is that subjects report optimally given the rule—a premise that, as the risk-aversion discussion above noted, the evidence on incentive-compatible mechanisms does not uniformly support (Danz et al., 2022). The informativeness comparison is therefore best read as a statement about the mechanism conditional on optimal reporting, not as a behavioral prediction.

The conditional matters differently for different rules, and the partial-identification framework sharpens the prediction. Consider center-biased reporting—the deviation Danz et al. (2022) document under the binarized scoring rule. Subjects pull reports toward the interior of the simplex. This moves the squared-distance identified set by a single facet-translation when one count transfers between categories, because that set is the projection polytope of the reported count vector; the frequency-guessing identified set, by contrast, jumps between mode cells with the same single-count change. The two geometries differ, however, only in how far the inferred set moves, not in whether it remains correct. Under any scoring rule the identified sets of distinct reports cover the belief simplex and overlap only at beliefs where the two reports tie as optimal: every belief has an optimal report because the report set is finite, and joint membership in two identified sets is precisely a tie. Away from ties, each belief rationalizes exactly one report, so any reporting error that changes the report places the latent belief outside the sharp identified set—under every count-loss rule alike. Robustness to misreporting is therefore a question of the magnitude of the induced error, never of its absence. Hedging-style behavior aimed at balancing payoffs across realizations is structurally similar. Lottery confusion under the binary-lottery implementation (Hossain and Okui, 2013; Harrison et al., 2014) is orthogonal: it affects the distance rules, whose risk-aversion robustness relies on the lottery, but not frequency-guessing, which retains its risk-aversion robustness under direct payment. These are testable predictions in the induced-beliefs design below, not assertions.

The partial-identification framework itself, however, suggests how both premises could be examined. If beliefs are induced by the experimenter—through draws from an urn of known composition, or a signal structure with a computable posterior—the true belief p is known, so an elicited report r can be checked against the identified set $P_S(r)$ it generates: under optimal reporting $P_S(r)$ contains p, and the empirical coverage rate, with the direction of any systematic miss, measures how far subjects depart from optimal reporting. The partition observation above is what gives this check its power: coverage fails for any deviation that changes the report, so the test detects misreporting in principle whenever it is effective, not only when it is large. Eliciting the same induced beliefs under a frequency-report rule and under a standard probability scoring rule isolates the effect of the report format; crossing that with direct versus binary-lottery payment isolates the effect of the implementation, which the risk-aversion analysis above predicts to matter only for subjects who are not risk neutral. Response times, incentivized comprehension questions, and the rate of weakly dominated reports would complement the coverage measure. Such a design would test, rather than presume, the behavioral premise on which the case for frequency reports partly rests—a premise separate from the present contribution, which is that different frequency-report scoring rules induce different identified sets, and those sets determine the informativeness of the mechanism.

References

Armantier, O. and Treich, N. (2013). Eliciting beliefs: Proper scoring rules, incentives, stakes and hedging. *European Economic Review*, 62:17–40.

Berg, J. E., Daley, L. A., Dickhaut, J. W., and O'Brien, J. R. (1986). Controlling preferences for lotteries on units of experimental exchange. *The Quarterly Journal of Economics*, 101(2):281–306.

Charness, G., Gneezy, U., and Rasocha, V. (2021). Experimental methods: Eliciting beliefs. *Journal of Economic Behavior & Organization*, 189:234–256.

Danz, D., Vesterlund, L., and Wilson, A. J. (2022). Belief elicitation and behavioral incentive compatibility. *American Economic Review*, 112(9):2851–2883.

Delavande, A., Giné, X., and McKenzie, D. (2011). Measuring subjective expectations in developing countries: A critical review and new evidence. *Journal of Development Economics*, 94(2):151–163.

Engelberg, J., Manski, C. F., and Williams, J. (2009). Comparing the point predictions and subjective probability distributions of professional forecasters. *Journal of Business & Economic Statistics*, 27(1):30–41.

Federgruen, A. and Groenevelt, H. (1986). The greedy procedure for resource allocation problems: Necessary and sufficient conditions for optimality. *Operations Research*, 34(6):909–918.

Fox, B. L. (1966). Discrete optimization via marginal analysis. *Management Science*, 13(3):210–216.

Gigerenzer, G. and Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review*, 102(4):684–704.

Gneiting, T. (2011). Making and evaluating point forecasts. *Journal of the American Statistical Association*, 106(494):746–762.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378.

Harrison, G. W., Martínez-Correa, J., and Swarthout, J. T. (2014). Eliciting subjective probabilities with binary lotteries. *Journal of Economic Behavior & Organization*, 101:128–140.

Heinrich, C. (2014). The mode functional is not elicitable. *Biometrika*, 101(1):245–251.

Hogarth, R. M. (1975). Cognitive processes and the assessment of subjective probability distributions. *Journal of the American Statistical Association*, 70(350):271–289.

Hossain, T. and Okui, R. (2013). The binarized scoring rule. *The Review of Economic Studies*, 80(3):984–1001.

Ibaraki, T. and Katoh, N. (1988). *Resource Allocation Problems: Algorithmic Approaches*. MIT Press, Cambridge, MA.

Karni, E. (2009). A mechanism for eliciting probabilities. *Econometrica*, 77(2):603–606.

Lambert, N. S., Pennock, D. M., and Shoham, Y. (2008). Eliciting properties of probability distributions. In *Proceedings of the 9th ACM Conference on Electronic Commerce (EC '08)*, pages 129–138.

Manski, C. F. (2003). *Partial Identification of Probability Distributions*. Springer Series in Statistics. Springer, New York.

Manski, C. F. (2004). Measuring expectations. *Econometrica*, 72(5):1329–1376.

Murota, K. (2003). *Discrete Convex Analysis*. SIAM Monographs on Discrete Mathematics and Applications. Society for Industrial and Applied Mathematics, Philadelphia.

Offerman, T., Sonnemans, J., van de Kuilen, G., and Wakker, P. P. (2009). A truth serum for non-Bayesians: Correcting proper scoring rules for risk attitudes. *The Review of Economic Studies*, 76(4):1461–1489.

Roth, A. E. and Malouf, M. W. K. (1979). Game-theoretic models and the role of information in bargaining. *Psychological Review*, 86:574–594.

Savage, L. J. (1971). Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, 66(336):783–801.

Schlag, K. H. and Tremewan, J. (2021). Simple belief elicitation: An experimental evaluation. *Journal of Risk and Uncertainty*, 62(2):137–155.

Schlag, K. H., Tremewan, J., and van der Weele, J. J. (2015). A penny for your thoughts: A survey of methods for eliciting beliefs. *Experimental Economics*, 18:457–490.

Schotter, A. and Trevino, I. (2014). Belief elicitation in the laboratory. *Annual Review of Economics*, 6:103–128.

Selten, R. (1998). Axiomatic characterization of the quadratic scoring rule. *Experimental Economics*, 1:43–62.

Trautmann, S. T. and van de Kuilen, G. (2015). Belief elicitation: A horse race among truth serums. *The Economic Journal*, 125(589):2116–2135.

A Proofs

A.1 Squared-Distance Scoring

Proof of Proposition 1. Maximizing S_Q is equivalent to minimizing $\mathbb{E}_p\|r-\omega\|_2^2$. For fixed r,

$$\mathbb{E}_p\|r-\omega\|_2^2 = \|r-\mathbb{E}_p[\omega]\|_2^2 + \sum_{i=1}^{k}\mathrm{Var}_p(\omega_i).$$

The variance term is independent of r, and $\mathbb{E}_p[\omega] = np$. Hence optimal reports are exactly the feasible integer projections of np.

The full inverse projection condition is

$$\|r-np\|_2^2 \le \|s-np\|_2^2 \quad \forall s \in \Omega,$$

or equivalently

$$2np\cdot(s-r) \le \|s\|_2^2 - \|r\|_2^2.$$

Taking $s = r + e_i - e_j$, feasible when $r_j > 0$, gives

$$n(p_i - p_j) \le r_i - r_j + 1.$$

Conversely, suppose all these single-count inequalities hold. For any $s \in \Omega$, write $d = s - r$. Pair each positive unit of d_i with a negative unit of d_j. Each coordinate j with $d_j < 0$ has $s_j < r_j$ and hence $r_j > 0$, so the single-count inequality $n(p_i - p_j) \le r_i - r_j + 1$ is available for every paired transfer. Summing the single-count inequalities over the paired transfers gives

$$np\cdot d \le r\cdot d + M,$$

where $M = \sum_{i:d_i>0} d_i$ is the number of transferred counts. Because d has integer components, $|d_i| \ge 1$ on its support, so

$$\sum_i d_i^2 \ge \sum_i |d_i| = 2M.$$

Doubling the displayed inequality and substituting $2M \le \sum_i d_i^2 = \|d\|_2^2$ gives $2np\cdot d \le 2r\cdot d + \|d\|_2^2$. Since $d = s - r$,

$$2np\cdot(s-r) \le 2r\cdot(s-r) + \|s-r\|_2^2 = \|s\|_2^2 - \|r\|_2^2,$$

which is the full projection condition.

It remains to derive the coordinate bounds. Let $m = m(r) = |\{j : r_j > 0\}|$. If $r_i > 0$, summing

$$n(p_u - p_i) \le r_u - r_i + 1$$

over all $u \ne i$ gives

$$1 - p_i \le (k-1)p_i + \frac{n - kr_i + k - 1}{n},$$

so

$$p_i \ge \frac{r_i - 1}{n} + \frac{1}{nk}.$$

For the upper bound, sum

$$n(p_i - p_v) \le r_i - r_v + 1$$

over the $m-1$ positive-report coordinates $v \neq i$; since the remaining coordinates have $r_v = 0$ and therefore contribute $p_v \ge 0$ to the simplex sum, $\sum_{v \neq i, r_v > 0} p_v \le 1 - p_i$, which gives

$$1 - p_i \ge (m-1)p_i + \frac{n - mr_i - m + 1}{n}.$$

This gives

$$p_i \le \frac{r_i + 1}{n} - \frac{1}{nm}.$$

If $r_i = 0$, the lower bound is 0. For the upper bound, sum

$$n(p_i - p_v) \le 1 - r_v$$

over all m positive-report coordinates v:

$$1 - p_i \ge mp_i + \frac{n - m}{n},$$

so

$$p_i \le \frac{m}{n(m+1)}.$$

The bounds are sharp, with an explicit attaining belief for each endpoint. For the lower bound with $r_i > 0$, take $p_i = (r_i - 1)/n + 1/(nk)$ and $p_u = r_u/n + 1/(nk)$ for $u \neq i$. This is a probability vector—it sums to one and every coordinate is at least $1/(nk)$—and $n(p_a - p_b) = r_a - r_b \le r_a - r_b + 1$ for $a, b \neq i$ while $n(p_a - p_i) = r_a - r_i + 1$, so it lies in $P_Q(r)$ and the constraints with $b = i$ bind, attaining the lower endpoint. For the upper bound with $r_i > 0$, take $p_i = (r_i + 1)/n - 1/(nm)$, take $p_v = r_v/n - 1/(nm)$ for the positive-report coordinates $v \neq i$, and set $p_v = 0$ for the zero-report coordinates. This sums to one and is nonnegative, lies in $P_Q(r)$, and makes the constraints $n(p_i - p_v) \le r_i - r_v + 1$ bind for every positive-report $v \neq i$, attaining the upper endpoint. For the upper bound with $r_i = 0$, take $p_i = m/(n(m+1))$, take $p_v = \{(m+1)r_v - 1\}/(n(m+1))$ for the positive-report coordinates v, and set $p_w = 0$ for the other zero-report coordinates. This sums to one and is nonnegative, lies in $P_Q(r)$, and makes the constraints $n(p_i - p_v) \le 1 - r_v$ bind, attaining the upper endpoint. Thus the displayed bounds are sharp. □

A.2 Frequency-Guessing Scoring

Proof of Proposition 2. For any report r,

$$\mathbb{E}_p[S_0(r, \omega)] = a - b \Pr_p(\omega \neq r) = (a - b) + b \Pr_p(\omega = r),$$

so, since $b > 0$, expected-score maximization is equivalent to maximizing the multinomial probability mass $\Pr_p(\omega = r)$.

It remains to characterize the mode region. Compare r with $r + e_i - e_j$, which is feasible when $r_j > 0$. If $p_j = 0$ and $r_j > 0$, then r cannot be a mode unless all probability is zero, which is impossible. For $p_j > 0$,

$$\frac{\Pr_p(r + e_i - e_j)}{\Pr_p(r)} = \frac{r_j}{r_i + 1} \frac{p_i}{p_j}.$$

Thus a necessary condition for r to be a mode is

$$r_j p_i \le (r_i + 1) p_j \quad \forall i, j \text{ with } r_j > 0.$$

Conversely, if these inequalities hold, any feasible s can be reached from r by a sequence of single-count transfers from coordinates where the current vector exceeds s to coordinates where it is below s. At each step the same ratio is at most one, because source counts weakly decrease and destination counts weakly increase along the path. Therefore no feasible s has larger multinomial mass than r.

The coordinate upper bound follows by summing

$$r_j p_i \le (r_i + 1) p_j$$

over $j \neq i$:

$$(n - r_i) p_i \le (r_i + 1)(1 - p_i),$$

so $p_i \le (r_i + 1)/(n + 1)$. The lower bound follows by summing

$$r_i p_j \le (r_j + 1) p_i$$

over $j \neq i$:

$$r_i (1 - p_i) \le (n - r_i + k - 1) p_i,$$

so $p_i \ge r_i/(n + k - 1)$.

The bounds are sharp. The upper endpoint is attained by

$$p_i = \frac{r_i + 1}{n + 1}, \quad p_j = \frac{r_j}{n + 1} \quad (j \neq i).$$

The lower endpoint, when $r_i > 0$, is attained by

$$p_i = \frac{r_i}{n + k - 1}, \quad p_j = \frac{r_j + 1}{n + k - 1} \quad (j \neq i).$$

When $r_i = 0$, the lower endpoint 0 is attained by setting $p_i = 0$ and assigning probabilities proportional to r_j on the reported support. When $r_i = n$, the upper endpoint 1 is attained by $p_i = 1$. □

A.3 Comparing the Two Closed-Form Rules

Proof of Corollary 1. Throughout, $m = m(r)$ is the number of categories with $r_i > 0$, and the average coordinate width is $(1/k) \sum_{i=1}^{k} (\overline{p}_i - \underline{p}_i)$.

Part (1). By Proposition 2 the frequency-guessing coordinate width is

$$\frac{r_i + 1}{n + 1} - \frac{r_i}{n + k - 1} = \frac{r_i (k - 2) + (n + k - 1)}{(n + 1)(n + k - 1)}.$$

Summing over i and using $\sum_i r_i = n$,

$$\sum_{i=1}^{k} \left(\overline{p}_{0,i} - \underline{p}_{0,i} \right) = \frac{(k - 2)n + k(n + k - 1)}{(n + 1)(n + k - 1)} = \frac{(k - 1)(2n + k)}{(n + 1)(n + k - 1)};$$

dividing by k gives $W_0(n,k)$, which does not depend on r.

Part (2). By Proposition 1 a positive-report category has coordinate width $2/n - 1/(nm) - 1/(nk)$ and a zero-report category has width $m/[n(m+1)]$. With m positive and $k-m$ zero categories,

$$\sum_{i=1}^{k}\left(\overline{p}_{Q,i} - \underline{p}_{Q,i}\right) = m\left(\frac{2}{n} - \frac{1}{nm} - \frac{1}{nk}\right) + (k-m)\frac{m}{n(m+1)} = \frac{1}{n}\left[2m - 1 - \frac{m}{k} + \frac{m(k-m)}{m+1}\right];$$

dividing by k gives $W_Q(m;n,k)$, a function of m alone.

Part (3). Evaluating part (2) at $m=k$ gives $W_Q(k;n,k) = 2(k-1)/(nk)$, and

$$W_Q(k;n,k) > W_0(n,k) \iff 2(n+1)(n+k-1) > n(2n+k) \iff k(n+2) > 2,$$

which holds for all $n \geq 1$ and $k \geq 2$. Evaluating at $m=1$ gives $W_Q(1;n,k) = (k-1)(k+2)/(2nk^2)$, and

$$W_Q(1;n,k) < W_0(n,k) \iff (k+2)(n+1)(n+k-1) < 2nk(2n+k).$$

The right side minus the left equals $n^2(3k-2) + nk(k-2) - (k+2)(k-1)$, whose coefficients in n are nonnegative for $k \geq 2$, so it is increasing in n; at $n=2$ it equals $k^2 + 7k - 6 > 0$. Hence $W_Q(1;n,k) < W_0(n,k)$ for all $n \geq 2$ and $k \geq 2$. □

A.4 The Single-Transfer Characterization

Proof of Lemma 1. Write $\Delta\ell(t;q) = \ell(t+1;q) - \ell(t;q)$; discrete convexity of $\ell(\cdot;q)$ means $\Delta\ell(\cdot;q)$ is nondecreasing.

Necessity. For any i,j with $r_j > 0$ the report $r + e_i - e_j$ is feasible, and it changes the objective by $\Delta\ell(r_i;p_i) - \Delta\ell(r_j - 1;p_j)$. If r is optimal this change is nonnegative, which is the displayed inequality.

Sufficiency. Suppose r satisfies every displayed inequality, and let $s \in \Omega$ be any competing report. Since $\sum_i s_i = \sum_i r_i = n$, the surplus units $\sum_i (s_i - r_i)^+$ and the deficit units $\sum_j (r_j - s_j)^+$ are equal in number, say M. Reach s from r by M single-count transfers, each moving one unit from a coordinate still above its target s to one still below. A receiver coordinate then only rises, from r_i toward s_i, and a sender only falls, from r_j toward s_j; so every intermediate report is feasible, and at the step transferring a unit $j \to i$ the running counts satisfy $c_i \geq r_i$ and $c_j \leq r_j$. That step changes the objective by $\Delta\ell(c_i;p_i) - \Delta\ell(c_j - 1;p_j)$. Because $\Delta\ell(\cdot;p)$ is nondecreasing, $c_i \geq r_i$ gives $\Delta\ell(c_i;p_i) \geq \Delta\ell(r_i;p_i)$, and $c_j - 1 \leq r_j - 1$ gives $\Delta\ell(c_j - 1;p_j) \leq \Delta\ell(r_j - 1;p_j)$; so the step's change is at least $\Delta\ell(r_i;p_i) - \Delta\ell(r_j - 1;p_j)$, nonnegative by the displayed inequality for the pair (i,j). Summing over the M steps gives $\sum_i \ell(s_i;p_i) \geq \sum_i \ell(r_i;p_i)$. As s was arbitrary, r minimizes the separable objective.

Necessity and sufficiency together give $r \in R_S(p)$ if and only if the displayed inequalities hold at p, so $P_S(r)$ is exactly the set of beliefs satisfying them. □

A.5 Manhattan-Distance Scoring

Proof of Proposition 3. Linearity of expectation gives

$$\mathbb{E}_p\left[\sum_{i=1}^{k} |r_i - \omega_i|\right] = \sum_{i=1}^{k} \mathbb{E}_p |r_i - \omega_i|.$$

The i-th term depends only on the binomial marginal $W_i \sim \mathrm{Bin}(n, p_i)$. For

$$\ell_i(t; p_i) = \mathbb{E}|t - W_i|,$$

the marginal cost of increasing the reported count from t to $t+1$ is

$$\ell_i(t+1; p_i) - \ell_i(t; p_i) = 2F_i(t; n, p_i) - 1.$$

Moving one reported count from j to i changes expected Manhattan loss by

$$\{2F_i(r_i; n, p_i) - 1\} - \{2F_j(r_j - 1; n, p_j) - 1\}.$$

Hence no feasible single-count transfer improves the report if and only if

$$F_i(r_i; n, p_i) \geq F_j(r_j - 1; n, p_j) \quad \forall i, j \text{ with } r_i < n,\ r_j > 0.$$

The increments $2F_i(t; n, p_i) - 1$ are weakly increasing in t, so each coordinate loss $\ell_i(\cdot; p_i)$ is discrete-convex; Lemma 1 then applies, and the no-transfer condition just derived is sufficient as well as necessary—so $P_M(r)$ is the exact identified set.

The threshold representation follows by taking c between

$$\max_{j: r_j > 0} F_j(r_j - 1; n, p_j) \quad \text{and} \quad \min_{i: r_i < n} F_i(r_i; n, p_i).$$

Inverting the resulting binomial-CDF inequalities coordinate by coordinate gives the threshold-indexed box-simplex description used for Manhattan bounds. □

For $t \in \{0, \ldots, n-1\}$ the binomial CDF $F(t; n, p) = \sum_{j=0}^{t} \binom{n}{j} p^j (1-p)^{n-j}$ is continuous and strictly decreasing in p on $[0, 1]$, with $F(t; n, 0) = 1$ and $F(t; n, 1) = 0$; together with the conventions $F(-1; n, p) = 0$ and $F(n; n, p) = 1$ this lets us define, for each $c \in [0, 1]$ and each coordinate v, the threshold-indexed endpoints

$$p_v^{(L)}(c) = \begin{cases} \text{the unique } p \in [0,1] \text{ with } F(r_v - 1; n, p) = c, & r_v > 0, \\ 0, & r_v = 0, \end{cases}$$

$$p_v^{(U)}(c) = \begin{cases} \text{the unique } p \in [0,1] \text{ with } F(r_v; n, p) = c, & r_v < n, \\ 1, & r_v = n. \end{cases}$$

Both $c \mapsto p_v^{(L)}(c)$ and $c \mapsto p_v^{(U)}(c)$ are continuous and weakly decreasing in c on $[0, 1]$, strictly so on the range where the corresponding case is non-degenerate.

Lemma 2 (Single-equation characterization of Manhattan coordinate bounds)**.** *Fix $r \in \Omega$ and a coordinate i.*

(i) If $r_i < n$, the sharp upper bound $\bar{p}_i(r) = \sup\{p_i : p \in P_M(r)\}$ equals $p_i^{(U)}(c^*)$, where $c^* \in [0, 1]$ is the unique solution of

$$p_i^{(U)}(c) + \sum_{v \neq i} p_v^{(L)}(c) = 1.$$

(ii) If $r_i > 0$, the sharp lower bound $\underline{p}_i(r) = \inf\{p_i : p \in P_M(r)\}$ equals $p_i^{(L)}(c^{**})$, where $c^{**} \in [0,1]$ is the unique solution of

$$p_i^{(L)}(c) + \sum_{v \neq i} p_v^{(U)}(c) = 1.$$

Each defining equation is continuous and strictly monotone in c, so each bound is the solution of a single monotone scalar equation and can be located by bisection on $[0,1]$.

Proof. For (i), Proposition 3 gives $p \in P_M(r)$ iff there is a threshold $c \in [0,1]$ with $p_v^{(L)}(c) \leq p_v \leq p_v^{(U)}(c)$ for all v. At a fixed c, maximizing p_i over the box-simplex slice attains $p_i^{(U)}(c)$ iff the residual mass $1 - p_i^{(U)}(c)$ admits a feasible decomposition over $v \neq i$, i.e. iff $\sum_{v \neq i} p_v^{(L)}(c) \leq 1 - p_i^{(U)}(c) \leq \sum_{v \neq i} p_v^{(U)}(c)$; otherwise the maximum is $1 - \sum_{v \neq i} p_v^{(L)}(c) < p_i^{(U)}(c)$. Hence the sliced maximum is

$$\phi(c) := \min\big(p_i^{(U)}(c),\ 1 - \textstyle\sum_{v \neq i} p_v^{(L)}(c)\big),$$

and $\bar{p}_i(r) = \sup_{c \in [0,1]} \phi(c)$. The first argument $p_i^{(U)}(c)$ is continuous and weakly decreasing in c, while the second argument $1 - \sum_{v \neq i} p_v^{(L)}(c)$ is continuous and weakly increasing in c (since each $p_v^{(L)}$ is weakly decreasing); their unique crossing is the c^* where $p_i^{(U)}(c^*) + \sum_{v \neq i} p_v^{(L)}(c^*) = 1$, and ϕ is unimodal in c with peak at c^*. For $c < c^*$, $\phi(c) = 1 - \sum_{v \neq i} p_v^{(L)}(c) \leq 1 - \sum_{v \neq i} p_v^{(L)}(c^*) = p_i^{(U)}(c^*)$ (strict decrease of $\sum_{v \neq i} p_v^{(L)}$); for $c > c^*$, $\phi(c) = p_i^{(U)}(c) \leq p_i^{(U)}(c^*)$ (strict decrease of $p_i^{(U)}$). So ϕ attains its maximum at c^* with value $p_i^{(U)}(c^*)$, and c^* is the unique solution of the displayed equation since its left-hand side is continuous and strictly decreasing in c (from a value ≥ 1 at $c = 0$ to 0 at $c = 1$). Part (ii) is symmetric, exchanging the roles of $p_v^{(L)}$ and $p_v^{(U)}$. □

A.6 Bounds on Additional Functionals of the Induced Outcome Distribution

The bounds developed in the body have been on individual coordinates of p and on the linear functional $\sum_i x_i p_i$ for a numerical outcome vector x. The same identified sets yield bounds on a broader inferential menu—variance, median, other quantiles—with the available functionals depending on what x represents.

The role of x. Three regimes cover the cases of practical interest.

(i) *No x.* Only coordinate bounds $[\underline{p}_{S,i}(r), \overline{p}_{S,i}(r)]$ are meaningful—probabilities of individual unordered categories.

(ii) *Ordinal x.* Categories carry an ordering $x_1 < x_2 < \cdots < x_k$ but the values are nominal; the menu adds bounds on the median $\mathrm{Med}(Y)$ and other quantiles $q_\tau(Y)$ of the induced random variable $Y = x_J$, $J \sim p$.

(iii) *Numerical x.* Categories have numerical values in $\mathbb{R}^k$; the menu adds bounds on the mean $\mathbb{E}(Y)$, variance $\mathrm{Var}(Y)$, and any polynomial functional of the induced outcome distribution.

General machinery. For any functional $T : \Delta^k \to \mathbb{R}$, the identified bound under rule S is the pair

$$\Big[\inf_{p\in P_S(r)} T(p),\ \sup_{p\in P_S(r)} T(p)\Big].$$

The geometry of $P_S(r)$ determines which optimization machinery applies. Squared-distance and frequency-guessing scoring induce polytope identified sets, so linear T (mean) reduces to a linear program; the variance $\mathrm{Var}(Y) = \sum_i x_i^2 p_i - (\sum_i x_i p_i)^2$ is concave in p, so its upper bound is a convex quadratic program over the polytope and its lower bound is a non-convex minimum attained at a polytope vertex; and a step-function T—the τ-quantile index $j^*_\tau(p) = \min\{j : S_j(p) \geq \tau\}$ with $S_j(p) = \sum_{i\leq j} p_i$, and any monotone transformation of it including the median—reduces to a sequence of linear programs over the polytope, one per candidate value of j^*_τ. Manhattan-distance scoring induces a threshold-indexed slice family rather than a polytope (Proposition 3), but the same operations apply on each slice and the optimum is then taken over the threshold parameter.

The threshold representation gives a particularly clean characterization for cumulative sums, the building blocks of median and quantile bounds.

Lemma 3 (Linear functionals over the Manhattan identified set)**.** *Fix $r \in \Omega$. For any vector $a \in \mathbb{R}^k$,*

$$\sup_{p\in P_M(r)} \sum_i a_i p_i \ =\ \sup_{c\in[0,1]} \sup\Big\{ \textstyle\sum_i a_i p_i \ :\ p_v^{(L)}(c) \leq p_v \leq p_v^{(U)}(c)\ \forall v,\ \sum_v p_v = 1\Big\},$$

and analogously for the infimum, where $p_v^{(L)}(c), p_v^{(U)}(c)$ are the threshold-indexed endpoints from Lemma 2 and the inner supremum is taken with the convention $-\infty$ when the slice is empty. The inner supremum is a linear program over a box-simplex slice. Applied to the indicator vector $a = (\mathbf{1}\{i \leq j\})_{i=1}^k$, this yields the cumulative-sum bounds $\sup_{p\in P_M(r)} S_j(p)$ and $\inf_{p\in P_M(r)} S_j(p)$; the τ-quantile-index bound on the induced Y follows from

$$\min_p j^*_\tau(p) = \min\{j : \sup_p S_j(p) \geq \tau\}, \qquad \max_p j^*_\tau(p) = \max\{j : \inf_p S_{j-1}(p) < \tau\}.$$

Proof. By Proposition 3, $p \in P_M(r)$ iff there is a threshold $c \in [0,1]$ with $p_v^{(L)}(c) \leq p_v \leq p_v^{(U)}(c)$ for all v. Equivalently, $P_M(r) = \bigcup_{c\in[0,1]} \Sigma(c)$, where $\Sigma(c) = \{p \in \Delta^k : p_v^{(L)}(c) \leq p_v \leq p_v^{(U)}(c)\,\forall v\}$. For any $a \in \mathbb{R}^k$,

$$\sup_{p\in P_M(r)} \textstyle\sum_i a_i p_i \ =\ \sup_{c\in[0,1]} \sup_{p\in\Sigma(c)} \sum_i a_i p_i,$$

since the supremum over a union equals the supremum of suprema. The inner problem at fixed c is a linear program over the box-simplex slice $\Sigma(c)$; the analogous identity holds for the infimum. Specialising to $a_i = \mathbf{1}\{i \leq j\}$ gives the cumulative-sum bounds $\sup_{p\in P_M(r)} S_j(p)$ and $\inf_{p\in P_M(r)} S_j(p)$ for each j.

For the quantile-index bounds, since $S_j(p)$ is nondecreasing in j, $j^*_\tau(p) \leq j$ iff $S_j(p) \geq \tau$, and $j^*_\tau(p) \geq j$ iff $S_{j-1}(p) < \tau$ (with $S_0 \equiv 0$). Hence

$$\min_{p\in P_M(r)} j^*_\tau(p) = \min\{j : \sup_{p\in P_M(r)} S_j(p) \geq \tau\}, \quad \max_{p\in P_M(r)} j^*_\tau(p) = \max\{j : \inf_{p\in P_M(r)} S_{j-1}(p) < \tau\}.$$

The one-parameter family of box-simplex linear programs is solved by sweeping c on a fine grid (with the tolerance reported in Appendix B). □

Worked example. Take $n = 10$, $k = 5$, report $r = (2, 2, 2, 2, 2)$, and numerical outcomes $x = (0, 1, 2, 3, 4)$—the symmetric report at which the no-universal-winner conclusion of Section 4 is most visible. Table 3 reports the bounds.

Functional	Frequency-guessing	Squared-distance	Manhattan
Each p_i	[0.143, 0.273]	[0.120, 0.280]	[0.129, 0.280]
Mean $\mathbb{E}(Y)$	[1.75, 2.25]	[1.70, 2.30]	[1.71, 2.29]
Median $\text{Med}(Y)$	$\{1, 2\}$	$\{1, 2, 3\}$	$\{1, 2, 3\}$
Lower quartile $q_{0.25}(Y)$	$\{0, 1\}$	$\{0, 1\}$	$\{0, 1\}$
Upper quartile $q_{0.75}(Y)$	$\{3, 4\}$	$\{3, 4\}$	$\{3, 4\}$

Table 3: Identified bounds for $n = 10$, $k = 5$, $r = (2, 2, 2, 2, 2)$, $x = (0, 1, 2, 3, 4)$. Frequency-guessing gives the tightest mean and median bounds at this symmetric report; the rules coincide on the upper and lower quartiles. Variance bounds follow from a constrained optimization over the same identified sets (a convex quadratic program for the upper bound, a vertex search for the lower) and are omitted to keep the table compact.

Two structural observations. First, the inferential menu—coordinate, mean, variance, median, quantile—is unified through the identified set rather than through rule-specific machinery; each functional is the same optimization problem with a different objective. Second, the no-universal-winner conclusion of Section 4 extends to non-linear functionals: which rule yields the tightest bound depends jointly on the report shape and the chosen functional. The Gneiting (2011) correspondence between absolute-error loss and the median, invoked in the body at the level of optimal reports, is a statement about *elicitation*, not *identified-set tightness*: at the symmetric report displayed in Table 3 the median-eliciting Manhattan rule does not give tighter median bounds than frequency-guessing. Schlag and Tremewan (2021) derive bounds on means, variances, and quantiles for frequency-guessing in their setting; the statement above carries that program across all three rules.

B Design Comparison Details

The design comparison is generated by the script `scripts/design_efficiency.py`. The final grid uses $n \in \{5, 10, 20, 50\}$, $k \in \{2, 3, 5, 10\}$, $\alpha \in \{0.1, 0.3, 1, 3, 10\}$, and 5,000 draws from the symmetric Dirichlet distribution for each cell. For each draw, and for each of the three rules in the comparison—frequency-guessing, squared-distance, and Manhattan—the script computes an optimal report and then bounds for the identified set generated by that report. Hamming and Chebyshev distance are excluded; Section 3.6 explains why their identified sets do not yield tractable sharp bounds.

Frequency-guessing coordinate bounds are closed form and mean bounds are linear programs over the transfer region. Squared-distance coordinate bounds are closed form and mean bounds are linear programs over the single-count transfer polytope. Manhattan coordinate bounds are sharp: the script solves the one-dimensional threshold optimization implied by Appendix A.5 with a monotone root-find. Manhattan mean bounds use the same threshold representation but are computed over a threshold grid with tolerance 10^{-4}.

The generated files are CSV tables under `outputs/design_exercise/`. These outputs are diagnostics for experimental design. They are not simulations of subject behavior and do not replace the mathematical characterization of $R_S(p)$ and $P_S(r)$.

Table 4 substantiates the claim in Section 4 that the number of categories k and the number of trials n are second-order. It reports win shares marginally by k and by n, averaging over the other two design dimensions. Against the variation across belief concentration α in Table 1—where the leading rule's win share runs from near zero to one—the dependence on k and n is slight.

	Average-coordinate			Linear-mean		
	Guessing	Squared	Manhattan	Guessing	Squared	Manhattan
By number of categories k						
2	0.72	0.25	0.03	0.72	0.25	0.03
3	0.59	0.37	0.04	0.34	0.51	0.15
5	0.58	0.35	0.07	0.49	0.42	0.09
10	0.59	0.31	0.10	0.56	0.33	0.11
By number of trials n						
5	0.60	0.34	0.06	0.57	0.33	0.10
10	0.61	0.32	0.07	0.57	0.33	0.11
20	0.62	0.31	0.06	0.53	0.38	0.09
50	0.65	0.30	0.05	0.44	0.47	0.09

Table 4: Win share marginally by number of categories k and number of trials n, averaging over the other two design dimensions (5,000 symmetric-Dirichlet draws per design cell). Relative to belief concentration (Table 1), the dependence is second-order: the coordinate target is nearly flat in both, and the only visible movements are the binary case $k = 2$ and a mild drift toward squared-distance scoring at large n for the mean target.

Robustness. Two implementation choices were checked; neither changes the comparison. First, the linear-functional results are robust to the outcome vector. The same run also evaluates the extreme-category outcome vector $x = (0, \ldots, 0, 1)$ in place of the evenly spaced $x_i = (i-1)/(k-1)$; the ranking is unchanged, with win shares 0.56, 0.34, and 0.10 for frequency-guessing, squared-distance, and Manhattan scoring, against 0.53, 0.38, and 0.10 for the evenly spaced vector. Second, optimal-report ties are a measure-zero event under the continuous Dirichlet draws: the final run recorded a tie rate of 0.0000 for all three rules, so the lexicographic convention used to select a report when $R_S(p)$ is not a singleton has no effect on the reported numbers.

Robustness to asymmetric beliefs. The main design comparison draws beliefs from symmetric Dirichlet distributions, which are exchangeable across categories. A symmetric draw cannot produce a systematic asymmetry—a one-dominant, two-mode, graded, or balanced-but-asymmetric prior on the simplex. Nine asymmetric concentration vectors at $n = 20$, $k = 5$ (5,000 draws each, same seed and bounds machinery as the main grid) check whether the regime story extends. Three are concentrated (at least one $\alpha_i \leq 0.5$); one is a transition vector with smallest component at 0.5; and five are balanced (all $\alpha_i \geq 2$), sorted by the max-to-min ratio of the concentration vector (the "α-spread").

The role of α-spread within the balanced regime. The balanced cells in Table 5 are all balanced in the sense that every $\alpha_i \geq 2$, yet they do not behave identically. On the average-coordinate target the frequency-guessing dominance holds across all five cells, but the margin

	Average-coordinate			Linear-mean		
$\boldsymbol{\alpha}$	Guess	Squared	Manh.	Guess	Squared	Manh.
Concentrated-asymmetric (at least one $\alpha_i \leq 0.5$)						
$(5, 0.5, 0.5, 0.5, 0.5)$	0.06	0.60	0.33	0.04	0.72	0.24
$(3, 3, 0.3, 0.3, 0.3)$	0.03	0.76	0.21	0.04	0.82	0.14
$(5, 2, 1, 0.5, 0.2)$	0.07	0.67	0.26	0.03	0.86	0.11
Transition (smallest $\alpha_i = 0.5$, no large dominant component)						
$(2, 2, 1, 0.5, 0.5)$	0.26	0.54	0.20	0.23	0.63	0.15
Balanced-asymmetric (all $\alpha_i \geq 2$), ordered by α-spread max / min						
$(3, 3, 3, 2, 2)$ (1.5)	0.87	0.08	0.05	0.82	0.14	0.04
$(5, 4, 3, 2, 2)$ (2.5)	0.84	0.10	0.06	0.68	0.26	0.06
$(10, 8, 6, 4, 2)$ (5.0)	0.76	0.17	0.07	0.52	0.41	0.07
$(15, 10, 5, 2, 2)$ (7.5)	0.50	0.33	0.17	0.08	0.77	0.15
$(20, 12, 6, 3, 2)$ (10.0)	0.53	0.32	0.15	0.05	0.82	0.14

Table 5: Win share under nine asymmetric concentration vectors at $n = 20$, $k = 5$ (5,000 draws each); parentheses give the α-spread max / min for the balanced cells. Squared-distance scoring dominates the three concentrated-asymmetric cells and the transition cell. Within the balanced regime the average-coordinate ranking remains in favor of frequency-guessing throughout but the margin shrinks with α-spread; the linear-mean ranking reverses between spread 5 and spread 7.5, with squared-distance taking over at the steepest spreads.

contracts monotonically with the α-spread: from 79 points at the flattest vector $(3, 3, 3, 2, 2)$ to 17 and 21 points at the two steepest. On the linear-mean target the ranking reverses: frequency-guessing wins at spread ≤ 5 (margins 11–68 points), but squared-distance takes over at spread ≥ 7.5 (margins 62–68 points over the runner-up). A plausible explanation is that the induced beliefs under a wide-spread balanced $\boldsymbol{\alpha}$ concentrate tightly around an ordered expected vector, and the mean-target identified set is then sharper under the mean-oriented squared-distance rule than under the mode-oriented frequency-guessing rule. The refinement does not contradict the headline "no universal winner" message, but adds a second dimension to the regime story for asymmetric beliefs: not just concentration versus balance, but, within the balanced regime, how steep the gradient is. A more thorough mapping of α-spread effects is left to future work.

Manhattan stays in the 4–33 % range on both targets across all nine cells, never the outright winner but never far from the leader; its regime-robust character survives intact. Mean regret tells the same story: the leading rule's regret stays at ≤ 0.003 in seven of the nine cells, with the largest leader-regret reached at the transition and balanced-graded cells where the rules are closer in win share.