Online Supplement

Strategic Expert Aggregation: A Hierarchical Bayesian Framework for Robust Decisions in Final Offer Arbitration

This supplement collects material referenced in the main text: the full conditional distributions and sampler used for posterior computation (Section EC.1), detailed specifications of the three comparison methods (Section EC.2), the hyperprior and seed settings of the simulation study together with convergence diagnostics (Section EC.3), additional simulation results including full coverage and the distribution of recommended offers (Section EC.4), prior sensitivity results for the baseline and high disagreement scenarios (Section EC.5), and the outlier removal sensitivity analysis for the application (Section EC.6). Equation and section numbers refer to the main manuscript unless prefixed by EC.

EC.1 Posterior Computation: Full Conditionals and Sampler

Write $z_j = \Phi^{-1}(p_j)$ and, for the current parameter values, the residual

$$r_{k,j} = q_{k,j} - \mu - \sigma z_j - b_k. \tag{EC.1}$$

All full conditionals below follow from the hierarchical model of Section 4.1. Every conditional is conjugate except the one for σ, for which σ enters the mean as a scale on z_j; that parameter is updated by a Metropolis Hastings step on $\log\sigma$ within the Gibbs scan.

Location μ.

$$v_\mu = \Big(\tfrac{1}{v_0^2} + \sum_k \tfrac{J}{\tau_k^2}\Big)^{-1}, \qquad m_\mu = v_\mu\Big(\tfrac{m_0}{v_0^2} + \sum_k \sum_j \tfrac{q_{k,j}-\sigma z_j - b_k}{\tau_k^2}\Big),$$
$$\mu \mid \text{rest} \sim \mathcal{N}(m_\mu,\ v_\mu)\,. \tag{EC.2}$$

Scale σ (Metropolis Hastings on $\log\sigma$). Propose $\log\sigma' = \log\sigma + \xi$ with $\xi \sim \mathcal{N}(0, c^2)$, $c = 0.1$, and accept with probability $\min\{1, \exp(\ell(\sigma') - \ell(\sigma))\}$, where the log target in the $\log\sigma$ parameterization is

$$\ell(\sigma) = -\tfrac{1}{2}\sum_k \sum_j \frac{(q_{k,j}-\mu-\sigma z_j - b_k)^2}{\tau_k^2} - 2a_\sigma \log\sigma - \frac{b_\sigma}{\sigma^2}. \tag{EC.3}$$

The last two terms are the log of the Inv Gamma(a_σ, b_σ) prior on σ^2 expressed in $\log\sigma$ coordinates (including the Jacobian).

Expert bias b_k.

$$v_{b_k} = \Big(\tfrac{1}{\omega^2} + \tfrac{J}{\tau_k^2}\Big)^{-1}, \qquad m_{b_k} = v_{b_k}\sum_j \frac{q_{k,j}-\mu-\sigma z_j}{\tau_k^2},$$
$$b_k \mid \text{rest} \sim \mathcal{N}(m_{b_k},\ v_{b_k})\,. \tag{EC.4}$$

Expert noise variance τ_k^2.

$$\tau_k^2 \mid \text{rest} \sim \text{Inv Gamma}\left(a_\tau + \tfrac{J}{2},\ \ b_\tau + \tfrac{1}{2}\sum_j (q_{k,j}-\mu-\sigma z_j - b_k)^2\right). \tag{EC.5}$$

Algorithm EC.1 Gibbs sampler with embedded Metropolis Hastings step for σ

Initialize μ, σ at the mean of the individual least squares fits; $b_k = 0$, $\tau_k^2 = 0.25$, $\omega^2 = 0.5$.
for $t = 1, \ldots, T$ **do**
  Draw μ from its full conditional (EC.2).
  Propose $\log \sigma'$ and accept or reject by the Metropolis Hastings step (EC.3).
  for $k = 1, \ldots, K$ **do** draw b_k from (EC.4).
  end for
  for $k = 1, \ldots, K$ **do** draw τ_k^2 from (EC.5).
  end for
  Draw ω^2 from (EC.6).
  Record (μ, σ).
end for
return post burn in draws $\{(\mu^{(t)}, \sigma^{(t)})\}$.

Bias variance ω^2.

$$\omega^2 \mid \text{rest} \sim \text{Inv Gamma}\left(a_\omega + \tfrac{K}{2},\ \ b_\omega + \tfrac{1}{2}\sum_k b_k^2\right). \tag{EC.6}$$

The recommended offers and their credible intervals are obtained by mapping the post burn in draws through the closed form equilibrium (Equation (15) in the main text) and summarizing the induced draws of (s_1^*, s_2^*).

EC.2 Comparison Methods: Detailed Specifications

Each comparison method first fits an individual normal distribution to expert k's reported quantiles by least squares, that is, $(\hat{\mu}_k, \hat{\sigma}_k)$ minimizes $\sum_j (q_{k,j} - \mu - \sigma z_j)^2$, which is the ordinary least squares regression of $q_{k,j}$ on $(1, z_j)$. Individual equilibrium offers are then $s_i^*(\hat{\mu}_k, \hat{\sigma}_k)$ from the closed form. The three methods differ only in how they combine experts.

Equal weight linear pooling. The recommendation is the arithmetic mean of the individual equilibrium offers,

$$\hat{s}_i^{\text{LP}} = \frac{1}{K}\sum_{k=1}^{K} s_i^*(\hat{\mu}_k, \hat{\sigma}_k), \qquad i \in \{1, 2\}. \tag{EC.7}$$

Logarithmic pooling. With precisions $\lambda_k = 1/\hat{\sigma}_k^2$, the pooled normal has

$$\bar{\sigma}^2_{\text{LogP}} = \left(\tfrac{1}{K}\textstyle\sum_k \lambda_k\right)^{-1}, \qquad \bar{\mu}_{\text{LogP}} = \bar{\sigma}^2_{\text{LogP}} \cdot \tfrac{1}{K}\textstyle\sum_k \lambda_k \hat{\mu}_k, \tag{EC.8}$$

and the recommendation is $s_i^*(\bar{\mu}_{\text{LogP}}, \bar{\sigma}_{\text{LogP}})$.

Classical performance weighted method. Experts provide quantiles for M seed items with known true values $\{y_m\}_{m=1}^M$. For expert k, the seed quantiles partition the line into $J+1$ bins; let $\hat{\pi}_{k,\ell}$ be the smoothed empirical frequency with which the seed truths fall in bin ℓ, and let π_ℓ be the expected bin probability obtained as the successive differences of $\{0, p_1, \ldots, p_J, 1\}$. The calibration and information scores are

$$\text{cal}_k = \exp\Big(-\sum_\ell \hat{\pi}_{k,\ell} \log \tfrac{\hat{\pi}_{k,\ell}}{\pi_\ell}\Big), \qquad \text{info}_k = \max\Big\{\log \tfrac{\sigma_{\text{bg}}}{\hat{\sigma}_k},\ \epsilon\Big\}, \tag{EC.9}$$

where σ_{bg} is a background scale and ϵ a small floor. Weights are $w_k \propto \text{cal}_k \cdot \text{info}_k$, normalized to sum to one, and the recommendation is $\hat{s}_i = \sum_k w_k \, s_i^*(\hat{\mu}_k, \hat{\sigma}_k)$. Smoothing uses a pseudo count of one half per bin. This is a faithful, simplified implementation of the method of Cooke (1991).

EC.3 Hyperprior, Seed Settings, and Convergence Diagnostics

Hyperpriors. The prior $p(\theta)$ on the arbitrator preference parameters is stated in Section 6.1 of the main text. The remaining components of the hierarchy use $\tau_k^2 \sim \text{Inv Gamma}(a_\tau, b_\tau)$ with $(a_\tau, b_\tau) = (2, 0.5)$ and $\omega^2 \sim \text{Inv Gamma}(a_\omega, b_\omega)$ with $(a_\omega, b_\omega) = (2, 1)$. These are weakly informative and were held fixed across all scenarios.

Seed data for the classical method. Each replication draws $M_{\text{seed}} = 10$ seed items with true values uniform on $[20, 200]$. Each expert's seed quantiles are generated as $y_m + 5\, z_j + b_k + \varepsilon_{k,j}$, using the same expert specific bias b_k and noise τ_k as for the target quantity, with seed scale fixed at 5. The background scale is $\sigma_{\text{bg}} = 50$.

Sampler settings and convergence. The sampler is run for 2,500 iterations with 500 discarded as burn in, leaving 2,000 retained draws per chain. Convergence was assessed on a representative contaminated dataset using four independent chains. The split potential scale reduction factor of Gelman and Rubin (1992) was $\hat{R} = 1.008$ for μ and $\hat{R} = 1.003$ for σ, both comfortably below the conventional 1.01 threshold, and the effective sample sizes were 268 for μ and 680 for σ. The longer run of 8,000 iterations used for the application yields correspondingly larger effective sample sizes.

EC.4 Additional Simulation Results

Full coverage and posterior dispersion. Table EC.1 reports empirical coverage of the 95% credible intervals for both parties' equilibrium offers, together with the mean posterior standard deviations, across the three scenarios. Coverage is at or above nominal in the baseline and contaminated scenarios for both parties, and mildly below nominal in the high disagreement scenario, where the limited information carried by five experts under high inter expert variability widens and slightly miscalibrates the intervals.

Table EC.1: Empirical coverage of the 95% credible intervals and mean posterior standard deviations for the hierarchical Bayesian method, $R = 200$ replications per scenario.

Scenario	Coverage s_1	Coverage s_2	Post. SD s_1	Post. SD s_2
Baseline	96.5%	99.0%	0.422	0.421
High disagreement	86.0%	83.0%	0.980	0.985
Contaminated	95.0%	94.0%	0.676	0.676

Distribution of recommended offers. Figure EC.1 shows the distribution of the recommended team offer s_1 across replications for each method and scenario, with the oracle value marked. In the baseline and high disagreement scenarios the four distributions overlap closely and centre on the oracle. In the contaminated scenario the three comparison methods are displaced away from the oracle, whereas the hierarchical Bayesian distribution remains concentrated near it.

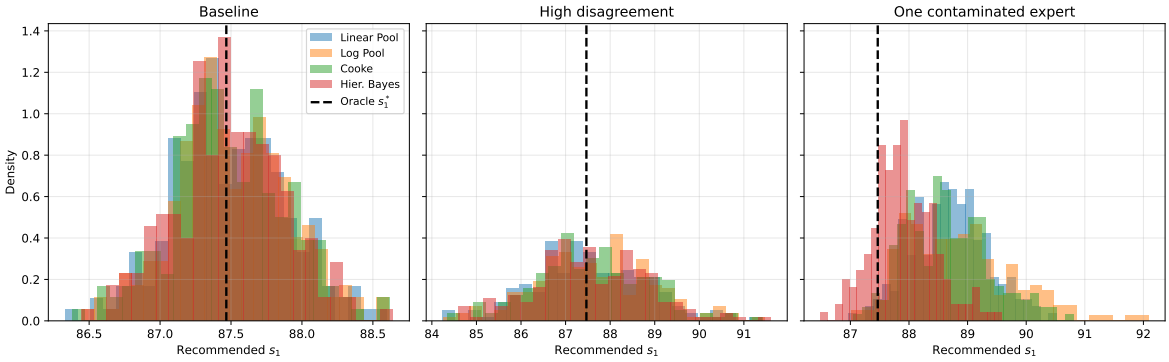

Figure EC.1: Distribution of the recommended team offer s_1 across 200 replications per scenario, by method. The dashed line marks the oracle offer $s_1^\dagger = 87.467$.

EC.5 Prior Sensitivity: Baseline and High Disagreement Scenarios

The main text (Table 3) reports prior sensitivity for the contaminated scenario. Tables EC.2 and EC.3 report the same comparison for the baseline and high disagreement scenarios. The five priors P0 to P4 are as defined in Section 6.3. The qualitative conclusion of the main text holds in every scenario: across the weakly informative, near flat, and well centred priors (P0 to P2) the marginal posterior of θ and the recommended offer are essentially unchanged, and only the strongly and deliberately misspecified priors (P3 and P4) move the recommendation.

Table EC.2: Prior sensitivity, baseline scenario, $R = 200$ replications. $\Delta\bar{s}_1$ is the change in recommended offer relative to P0; the last column is the empirical coverage of the 95% credible interval for s_1.

Prior	Description	$\bar{\mu}$	$\bar{\sigma}$	$\bar{s}_1$	$\Delta\bar{s}_1$	Coverage
P0	Weakly informative (default)	100.00	9.98	87.49	0	96.5%
P1	Near flat	100.01	9.99	87.48	−0.01	96.5%
P2	Informative, well centred	100.01	9.99	87.48	−0.01	96.5%
P3	Informative, location misspecified	100.13	9.98	87.61	+0.12	95.5%
P4	Informative, scale misspecified	100.00	10.86	86.39	−1.10	48.5%

Table EC.3: Prior sensitivity, high disagreement scenario, $R = 200$ replications. Columns as in Table EC.2.

Prior	Description	$\bar{\mu}$	$\bar{\sigma}$	$\bar{s}_1$	$\Delta\bar{s}_1$	Coverage
P0	Weakly informative (default)	100.03	9.96	87.55	0	86.0%
P1	Near flat	100.04	9.98	87.54	−0.02	85.5%
P2	Informative, well centred	100.04	9.98	87.53	−0.02	86.5%
P3	Informative, location misspecified	100.93	9.96	88.45	+0.90	78.5%
P4	Informative, scale misspecified	100.02	13.03	83.69	−3.87	21.0%

A pattern emerges across the three scenarios that is worth noting. Under the scale misspecified

prior P4, the shift in the recommended offer grows as the elicitation becomes less informative: $\Delta\bar{s}_1 = -1.10$ in the baseline, -2.43 in the contaminated scenario (Table 3 of the main text), and -3.87 under high disagreement. This is intuitive. The prior exerts more influence on the posterior precisely when the elicited quantiles are noisier or partly corrupted, so the same misspecification propagates more strongly. It reinforces the practical recommendation in Section 8 to keep $p(\theta)$ diffuse, or matched to the scale of the elicited quantiles, when elicitation quality is uncertain.

EC.6 Application: Outlier Removal Sensitivity

As a complement to the prior sensitivity reported in Table 8 of the main text, we examine how the hierarchical Bayesian recommendation responds to manually removing the discordant expert, Analyst C, from the panel. Table EC.4 compares the full panel recommendation with the recommendation obtained after dropping Analyst C's quantiles.

Table EC.4: Hierarchical Bayesian recommendation with and without Analyst C (in \$M).

	$\bar{\mu}$	$\bar{\sigma}$	Team s_1	Player s_2
Full panel (five experts)	9.61	1.62	7.57	11.64
Analyst C removed (four)	9.21	1.54	7.29	11.14
Change	-0.40	-0.08	-0.29	-0.51

Removing Analyst C moves the recommended team offer by only \$0.29M and the player offer by \$0.51M, confirming the partial pooling behaviour described in the main text: the hierarchical model already discounts the outlier from disagreement alone, so its use of all five experts produces a recommendation close to the one obtained by manually dropping the outlier, without requiring the analyst to identify and remove it by hand.