

Supplementary Information: Leveraging Collective Advice-Taking Behavior to Infer Accuracy and Improve the Wisdom of Crowds

Yunhao Zhang

1 The Advice-taking Model: Derivations, Proofs, and Extensions

1.1 Derivation for $W_{E,0}$ and $W_{L,0}$ in the Main Text

Table 1: Notations of Variables: Actual Distribution vs Internal Representation

True Parameter	Internal Representation by Better-Informed and Worse-Informed Agents
Better-Informed Agents' Initial Estimate ($\hat{\theta}_E$)	$\hat{\theta}_E^E$ and $\hat{\theta}_E^L$
Bias of Better-Informed Agents' Initial Estimate (δ_E)	δ_E^E and δ_E^L
Variance of Better-Informed Agents' Initial Estimate (σ_E^2)	$(\sigma_E^E)^2$ and $(\sigma_E^L)^2$
Worse-Informed Agents' Initial Estimate ($\hat{\theta}_L$)	$\hat{\theta}_L^E$ and $\hat{\theta}_L^L$
Bias of Worse-Informed Agents' Initial Estimate (δ_L)	δ_L^E and δ_L^L
Variance of Worse-Informed Agents' Initial Estimate (σ_L^2)	$(\sigma_L^E)^2$ and $(\sigma_L^L)^2$
Number of Better-Informed Agents (n_E)	n_E^E and n_E^L
Number of Worse-Informed Agents (n_L)	n_L^E and n_L^L
Numerical Advice or Group Mean (y)	y^E and y^L

Notes: Superscripts "E" and "L" denote whose internal perception it is; subscripts "E" and "L" denote which type the quantity refers to. For example, $\hat{\theta}_E^L$ represents what Type L thinks of the initial estimate of Type E. There are $n = n_E + n_L$ agents in total.

This section shows derivations of $W_{E,0}$ and $W_{L,0}$. For notational clarity, superscript which indicates whether a parameter refers to the better-informed (superscript "E") or the worse-informed (superscript "L") is omitted. This does not affect the derivation at all, because the derivation for the two types of agents are separate. For example, when deriving $W_{E,0}$, δ_E refers to δ_E^E in the main text, or the better-informed agents' perceived estimation bias of their own estimate. Omitting the superscript is fine because δ_L^E does not enter the derivation for $W_{E,0}$.

Using the set-up in the main text, $W_{E,0}$ may be computed by solving the following equation:

$$\begin{aligned}
\frac{\partial \mathbb{E}|\tilde{\theta}_{E,0} - \theta|^2}{\partial W_{E,0}} &= \frac{\partial \mathbb{E}|W_{E,0} \cdot y + (1 - W_{E,0}) \cdot \hat{\theta}_{E,0} - \theta|^2}{\partial W_{E,0}} \\
&= \frac{\partial \mathbb{E}|W_{E,0} \cdot \frac{\sum_{i \in E} \hat{\theta}_{E,i} + \sum_{i \in L} \hat{\theta}_{L,i}}{n_E + n_L} + (1 - W_{E,0}) \cdot \hat{\theta}_{E,0} - \theta|^2}{\partial W_{E,0}} \\
&= \frac{\partial \mathbb{E}|W_{E,0} \cdot \frac{\sum_{i \in E} (\theta + \varepsilon_{E,i} + \delta_E) + \sum_{i \in L} (\theta + \varepsilon_{L,i} + \delta_L)}{n_E + n_L} + (1 - W_{E,0}) \cdot (\theta + \varepsilon_{E,0} + \delta_E) - \theta|^2}{\partial W_{E,0}} \\
&= 0
\end{aligned}$$

The numerator is

$$\begin{aligned}
&\mathbb{E} \left\{ \frac{W_{E,0}^2}{(n_E + n_L)^2} \left[\sum (\varepsilon_{E,i} + \delta_E) + \sum (\varepsilon_{L,i} + \delta_L) \right]^2 + (1 - W_{E,0})^2 (\varepsilon_{E,0} + \delta_E)^2 + \right. \\
&\quad \left. \frac{2}{n_E + n_L} W_{E,0} (1 - W_{E,0}) \left[\sum (\varepsilon_{E,i} + \delta_E) + \sum (\varepsilon_{L,i} + \delta_L) \right] (\varepsilon_{E,0} + \delta_E) \right\} \\
&= \frac{W_{E,0}^2}{(n_E + n_L)^2} [n_E^2 \delta_E^2 + n_L^2 \delta_L^2 + n_E \sigma_E^2 + n_L \sigma_L^2 + 2n_E n_L \delta_E \delta_L] + (1 - W_{E,0})^2 (\sigma_E^2 + \delta_E^2) + \\
&\quad \frac{2}{n_E + n_L} W_{E,0} (1 - W_{E,0}) [\sigma_E^2 + n_E \delta_E^2 + n_L \delta_E \delta_L]
\end{aligned}$$

Taking the derivative with respect to $W_{E,0}$, and setting it equal to zero, yields the following equation:

$$\begin{aligned}
&\frac{2W_{E,0}}{(n_E + n_L)^2} [n_E^2 \delta_E^2 + n_L^2 \delta_L^2 + n_E \sigma_E^2 + n_L \sigma_L^2 + 2n_E n_L \delta_E \delta_L] + 2W_{E,0} (\sigma_E^2 + \delta_E^2) - 2(\sigma_E^2 + \delta_E^2) + \\
&\quad \frac{2}{n_E + n_L} [\sigma_E^2 + n_E \delta_E^2 + n_L \delta_E \delta_L] - \frac{4W_{E,0}}{n_E + n_L} [\sigma_E^2 + n_E \delta_E^2 + n_L \delta_E \delta_L] = 0
\end{aligned}$$

Rearranging the equation yields:

$$\begin{aligned}
& W_{E,0} [n_E^2 \delta_E^2 + n_L^2 \delta_L^2 + n_E \sigma_E^2 + n_L \sigma_L^2 + 2n_E n_L \delta_E \delta_L] + W_{E,0} (n_E + n_L)^2 (\sigma_E^2 + \delta_E^2) - \\
& 2W_{E,0} (n_E + n_L) [\sigma_E^2 + n_E \delta_E^2 + n_L \delta_E \delta_L] \\
& = (n_E + n_L)^2 (\sigma_E^2 + \delta_E^2) - (n_E + n_L) [\sigma_E^2 + n_E \delta_E^2 + n_L \delta_E \delta_L]
\end{aligned}$$

Solving for $W_{E,0}$ yields the WOA of a representative better-informed agent:

$$W_{E,0} = \frac{(n_E + n_L)^2 (\sigma_E^2 + \delta_E^2) - (n_E + n_L) (\sigma_E^2 + n_E \delta_E^2 + n_L \delta_E \delta_L)}{n_E^2 \delta_E^2 + n_L^2 \delta_L^2 + n_E \sigma_E^2 + n_L \sigma_L^2 + 2n_E n_L \delta_E \delta_L + (n_E + n_L)^2 (\sigma_E^2 + \delta_E^2) - 2(n_E + n_L) (\sigma_E^2 + n_E \delta_E^2 + n_L \delta_E \delta_L)} \quad (1.1)$$

Solving for $W_{L,0}$ follows the same (symmetric) process:

$$W_{L,0} = \frac{(n_E + n_L)^2 (\sigma_L^2 + \delta_L^2) - (n_E + n_L) (\sigma_L^2 + n_L \delta_L^2 + n_E \delta_E \delta_L)}{n_E^2 \delta_E^2 + n_L^2 \delta_L^2 + n_E \sigma_E^2 + n_L \sigma_L^2 + 2n_E n_L \delta_E \delta_L + (n_E + n_L)^2 (\sigma_L^2 + \delta_L^2) - 2(n_E + n_L) (\sigma_L^2 + n_L \delta_L^2 + n_E \delta_E \delta_L)} \quad (1.2)$$

1.2 An Asymptotic Proof of Proposition 1 (Sign-Independent Case)

Proposition 1. *Fixing all other parameters, when $|\delta_L^E|$ is sufficiently larger than $|\delta_E^E|$, worse-informed agents have larger expected WOA than better-informed agents: $E(\beta_j \cdot W_{L,0} + \zeta_j) > E(\alpha_i \cdot W_{E,0} + \zeta_i)$.*

Proof. $W_{L,0}$ does not depend on δ_L^E (it depends only on worse-informed agents' superscript- L parameters), so $E(\beta_j \cdot W_{L,0} + \zeta_j)$ is a constant in δ_L^E . The limit is therefore determined entirely by the behavior of $W_{E,0}$ as $|\delta_L^E| \rightarrow \infty$:

$$\lim_{|\delta_L^E| \rightarrow \infty} W_{E,0} = 0,$$

which can be verified by inspection of Equation 1.1: as $|\delta_L^E| \rightarrow \infty$, the leading $(\delta_L^E)^2$ term in the denominator dominates while no comparable term grows in the numerator. Therefore,

$$\lim_{|\delta_L^E| \rightarrow \infty} [E(\beta_j \cdot W_{L,0} + \zeta_j) - E(\alpha_i \cdot W_{E,0} + \zeta_i)] = \beta \cdot W_{L,0} > 0,$$

where $W_{L,0} > 0$ holds generically. The same conclusion follows as $\delta_L^E \rightarrow -\infty$: the dominant terms are $(\delta_L^E)^2$ in the denominator, which is sign-independent. $\square$

Intuition. As $|\delta_L^E| \rightarrow \infty$, better-informed agents perceive the group mean (which is heavily influenced by what they take to be very biased worse-informed estimates) as essentially uninformative for belief-updating, and place vanishing weight on it ($W_{E,0} \rightarrow 0$). Worse-informed agents, by contrast, do not perceive their own estimates as similarly biased and continue to place positive weight on the advice.

1.3 Proof of Proposition 1: Explicit Threshold under Same-Sign Configurations

The previous section proves Proposition 1 asymptotically for any sign configuration. This section explicitly solves for the threshold ratio a^* at which $W_{E,0} < W_{L,0}$ under the same-sign assumption $\delta_L^E = a \cdot \delta_E^E$ with $a > 0$. The same-sign case yields a tractable quadratic in a with closed-form roots; an analogous closed-form for the opposite-sign case is less informative because the cross-terms generate a piecewise inequality whose roots depend sensitively on other parameters. As an important note, the opposite-sign case is fully covered by the asymptotic proof in the previous section: a sufficiently large perceived gap in estimation bias produces $W_{E,0} < W_{L,0}$ regardless of sign, though the specific threshold varies depending on the signs of δ_E^E and δ_L^E and other parameters.

$W_{E,0}$ and $W_{L,0}$ may be rewritten as follows:

$$\frac{\left[n^2 \left((\sigma_E^E)^2 + (\delta_E^E)^2 \right) - n \left((\sigma_E^E)^2 + n_E^E (\delta_E^E)^2 + n_L^E a (\delta_E^E)^2 \right) \right]}{\left[(n_E^E \delta_E^E)^2 + (n_L^E a \delta_E^E)^2 + n_E^E (\sigma_E^E)^2 + n_L^E (\sigma_L^E)^2 + 2n_E^E n_L^E a (\delta_E^E)^2 + n^2 \left((\sigma_E^E)^2 + (\delta_E^E)^2 \right) - 2n \left((\sigma_E^E)^2 + n_E^E (\delta_E^E)^2 + n_L^E a (\delta_E^E)^2 \right) \right]} \quad (1.3)$$

$$\frac{\left[n^2 \left((\sigma_L^L)^2 + (\delta_L^L)^2 \right) - n \left((\sigma_L^L)^2 + n_L^L (\delta_L^L)^2 + n_E^L \delta_E^L \delta_L^L \right) \right]}{\left[(n_E^L \delta_E^L)^2 + (n_L^L \delta_L^L)^2 + n_E^L (\sigma_E^L)^2 + n_L^L (\sigma_L^L)^2 + 2n_E^L n_L^L \delta_E^L \delta_L^L + n^2 \left((\sigma_L^L)^2 + (\delta_L^L)^2 \right) - 2n \left((\sigma_L^L)^2 + n_L^L (\delta_L^L)^2 + n_E^L \delta_E^L \delta_L^L \right) \right]} \quad (1.4)$$

Then $W_{E,0} - W_{L,0}$ can be expressed as follows:

First expression: (Numerator of $W_{E,0}$) $\times$ (Denominator of $W_{L,0}$)

$$\underbrace{\left[n^2 \left((\sigma_E^E)^2 + (\delta_E^E)^2 \right) - n \left((\sigma_E^E)^2 + n_E^E (\delta_E^E)^2 \right) \right]}_{\text{(A) independent of } a} - \underbrace{n \left(n_L^E a (\delta_E^E)^2 \right)}_{\text{(B) linear in } a} \underbrace{\left[(n_E^L \delta_E^L)^2 + (n_L^L \delta_L^L)^2 + n_E^L (\sigma_E^L)^2 + n_L^L (\sigma_L^L)^2 + 2n_E^L n_L^L \delta_E^L \delta_L^L + n^2 \left((\sigma_L^L)^2 + (\delta_L^L)^2 \right) - 2n \left((\sigma_L^L)^2 + n_L^L (\delta_L^L)^2 + n_E^L \delta_E^L \delta_L^L \right) \right]}_{\text{(C) independent of } a} \quad (1.5)$$

Second expression: (Numerator of $W_{L,0}$) $\times$ (Denominator of $W_{E,0}$)

$$\underbrace{\left[n^2((\sigma_L^L)^2 + (\delta_L^L)^2) - n((\sigma_L^L)^2 + n_L^L(\delta_L^L)^2 + n_E^L \delta_E^L \delta_L^L) \right]}_{\text{(D) independent of } a} \underbrace{\left[(n_E^E \delta_E^E)^2 + n_E^E (\sigma_E^E)^2 + n_L^E (\sigma_L^E)^2 + n^2((\sigma_E^E)^2 + (\delta_E^E)^2) - 2n((\sigma_E^E)^2 + n_E^E (\delta_E^E)^2) \right]}_{\text{(E) independent of } a} + \underbrace{(n_L^E a \delta_E^E)^2}_{\text{(F) quadratic in } a} + \underbrace{2n_E^E n_L^E a (\delta_E^E)^2}_{\text{(G) linear in } a} - \underbrace{2n(n_L^E a (\delta_E^E)^2)}_{\text{(H) linear in } a} \quad (1.6)$$

Full derivation with letter labels (all base parameters > 0 , and $a > 0$)

Label the two expressions.

$$(A - B)C$$

with

$$\begin{aligned} A &:= n^2((\sigma_E^E)^2 + (\delta_E^E)^2) - n((\sigma_E^E)^2 + n_E^E (\delta_E^E)^2), \\ B &:= n n_L^E a (\delta_E^E)^2, \\ C &:= (n_E^L \delta_E^L)^2 + (n_L^L \delta_L^L)^2 + n_E^L (\sigma_E^L)^2 + n_L^L (\sigma_L^L)^2 + 2n_E^L n_L^L \delta_E^L \delta_L^L \\ &\quad + n^2((\sigma_L^L)^2 + (\delta_L^L)^2) - 2n((\sigma_L^L)^2 + n_L^L (\delta_L^L)^2 + n_E^L \delta_E^L \delta_L^L), \end{aligned}$$

and

$$D(E + F + G - H)$$

with

$$\begin{aligned} D &:= n^2((\sigma_L^L)^2 + (\delta_L^L)^2) - n((\sigma_L^L)^2 + n_L^L (\delta_L^L)^2 + n_E^L \delta_E^L \delta_L^L), \\ E &:= (n_E^E \delta_E^E)^2 + n_E^E (\sigma_E^E)^2 + n_L^E (\sigma_L^E)^2 + n^2((\sigma_E^E)^2 + (\delta_E^E)^2) - 2n((\sigma_E^E)^2 + n_E^E (\delta_E^E)^2), \\ F &:= (n_L^E a \delta_E^E)^2, \quad G := 2n_E^E n_L^E a (\delta_E^E)^2, \quad H := 2n n_L^E a (\delta_E^E)^2. \end{aligned}$$

Factor the a -dependence. Let

$$B = aB_1, \quad B_1 := n n_L^E (\delta_E^E)^2 > 0; \quad F = a^2 F_2, \quad F_2 := (n_L^E \delta_E^E)^2 > 0;$$

$$G = aG_1, \quad G_1 := 2n_E^E n_L^E (\delta_E^E)^2 > 0; \quad H = aH_1, \quad H_1 := 2n n_L^E (\delta_E^E)^2 > 0.$$

Then

$$(A - aB_1)C < D(E + a^2 F_2 + a(G_1 - H_1)),$$

where $G_1 - H_1 = 2n_L^E (\delta_E^E)^2 (n_E^E - n)$.

Organize powers of a .

$$AC - aB_1C < DE + a^2DF_2 + aD(G_1 - H_1).$$

Rearrange

$$a^2 \underbrace{(DF_2)}_p + a \underbrace{(D(G_1 - H_1) + B_1C)}_q > \underbrace{(AC - DE)}_r.$$

Substitute coefficients. Let $k := n_L^E (\delta_E^E)^2 > 0$. Since $F_2 = (n_L^E \delta_E^E)^2 = n_L^E \cdot k$,

$$p = DF_2 = Dn_L^E k, \quad q = k(2D(n_E^E - n) + nC), \quad r = AC - DE.$$

Dividing by $k > 0$:

$$\boxed{Dn_L^E a^2 + (2D(n_E^E - n) + nC)a > \frac{AC - DE}{k}}.$$

Final inequality in a . The comparison

$$[A - ann_L^E (\delta_E^E)^2] C < D(E + a^2(n_L^E \delta_E^E)^2 + a \cdot 2n_L^E (\delta_E^E)^2(n_E^E - n))$$

is equivalent to

$$\boxed{Dn_L^E a^2 + (2D(n_E^E - n) + nC)a > \frac{AC - DE}{k}}, \quad k = n_L^E (\delta_E^E)^2, \quad a > 0.$$

Final Result. Assuming $D \neq 0$, let

$$L := 2D(n_E^E - n) + nC, \quad \Delta := L^2 + 4Dn_L^E \frac{AC - DE}{k},$$

$$\boxed{a_{\pm} = \frac{-L \pm \sqrt{\Delta}}{2Dn_L^E}}.$$

Then:

- If $D > 0$: $a > \frac{-L + \sqrt{\Delta}}{2Dn_L^E}$.
- If $D = 0$: inequality reduces to $a > \frac{AC}{Lk}$.
- If $D < 0$: $a \in (a_-, a_+) \cap (0, \infty)$.

1.4 Alternative Information Structure and Subjective Beliefs

The general structure permits derivations under alternative information structures. The two extensions below recover the classic variance-only advice-taking model and incorporate within-type correlated errors.

1.4.1 Classic Advice-taking Model: No Bias, Variance Only

Assume all agents are unbiased in their estimates or are subjectively unaware of any of the estimation bias terms. In other words, set δ_E^E , δ_L^E , δ_E^L , δ_L^L as 0 in Equation 1.1 and Equation 1.2 above. Thus,

$$W_{E,0} = \frac{(n_E^E + n_L^E)^2 (\sigma_E^E)^2 - (n_E^E + n_L^E) (\sigma_E^E)^2}{n_E^E (\sigma_E^E)^2 + n_L^E (\sigma_L^E)^2 + (n_E^E + n_L^E)^2 (\sigma_E^E)^2 - 2(n_E^E + n_L^E) (\sigma_E^E)^2} \quad (1.7)$$

$$W_{L,0} = \frac{(n_E^L + n_L^L)^2 (\sigma_L^L)^2 - (n_E^L + n_L^L) (\sigma_L^L)^2}{n_E^L (\sigma_E^L)^2 + n_L^L (\sigma_L^L)^2 + (n_E^L + n_L^L)^2 (\sigma_L^L)^2 - 2(n_E^L + n_L^L) (\sigma_L^L)^2} \quad (1.8)$$

This recovers the classic advice-taking results (e.g. Gelman 2013, Enke and Graeber 2019) in which all agents are assumed to be unbiased. In this case, agents are differentiated only by prior variances from their *subjective* standpoint. Therefore, although better-informed agents can objectively be more accurate, whether their WOA will be lower depends on the perceived prior variances as well as the perceived group composition. Comparing average WOA may incorrectly identify worse-informed as better-informed if the better-informed have higher prior uncertainty, which is a boundary condition caused by agents' lack of awareness of their relative heterogeneity in informativeness.

Furthermore, this result more broadly characterizes situations in which only "some" better-informed agents are aware of δ_L while others are unaware. For example, some better-informed agents – despite being more accurate themselves – might not realize that others' estimates can be biased. This scenario is similar to the situation in which the better-informed agents overestimate the proportion of the better-informed agents in the crowd, which is discussed in the main text and in the SI below. In this case, the average WOA of the better-informed is a weighted average of $W_{E,0}$ as derived in Equation 1.7 in the SI and $W_{E,0}$ derived in the main text (assuming the perceived δ_L^E is non-zero). That is, the better-informed who are aware of δ_L^E will have lower WOA than those who are unaware of δ_L , such that the average WOA of the better-informed agents would be higher than the

one derived under the set-up in the main text (assuming the better-informed have some awareness of δ_L), but lower than that derived in SI Equation 1.7 (assuming complete unawareness about δ_L).

Overall, the better-informed unawareness of δ_L is a boundary condition to the algorithm. By a similar logic, if worse-informed agents recognize their systematic inaccuracy, WOA becomes an effective indicator of accuracy. Worse-informed agents who realize their estimates are more biased than the group mean place higher weight on the advice, increasing their WOA.

1.4.2 Model with Correlated Error Terms within Each Agent Type

In addition to the three assumptions used in the main text, the section relaxes the assumption that agents' estimates are independent. This relaxation allows error terms within each agent type to be correlated beyond the common bias terms. As in Section 1.1, the superscript denoting whose internal representation is being modeled is omitted; the derivation applies symmetrically to each agent type.

Additional Assumption: Correlated Error Terms within Each Agent Type. The better-informed agents have correlated error terms with other better-informed agents: $cov(\varepsilon_{E,i}, \varepsilon_{E,j}) = \mathbb{E}(\varepsilon_{E,i}\varepsilon_{E,j}) = \rho_{E,ij}\sigma_E^2 = \sigma_{E,ij}$. The worse-informed agents have correlated error terms with other worse-informed agents: $cov(\varepsilon_{L,i}, \varepsilon_{L,j}) = \mathbb{E}(\varepsilon_{L,i}\varepsilon_{L,j}) = \rho_{L,ij}\sigma_L^2 = \sigma_{L,ij}$. Between the better-informed and the worse-informed, the error terms are uncorrelated: $\mathbb{E}(\varepsilon_{L,i}\varepsilon_{E,j}) = 0$.

$$\begin{aligned} & \mathbb{E} \left\{ \frac{W_{E,0}^2}{(n_E + n_L)^2} \left[\sum (\varepsilon_{E,i} + \delta_E) + \sum (\varepsilon_{L,i} + \delta_L) \right]^2 + (1 - W_{E,0})^2 (\varepsilon_{E,0} + \delta_E)^2 + \right. \\ & \left. \frac{2}{n_E + n_L} W_{E,0}(1 - W_{E,0}) \left[\sum (\varepsilon_{E,i} + \delta_E) + \sum (\varepsilon_{L,i} + \delta_L) \right] (\varepsilon_{E,0} + \delta_E) \right\} \\ & = \frac{W_{E,0}^2}{(n_E + n_L)^2} \left[n_E^2 \delta_E^2 + n_L^2 \delta_L^2 + n_E \sigma_E^2 + \sum_{i \in E} \sum_{j \in E, j \neq i} \sigma_{E,ij} + n_L \sigma_L^2 + \sum_{i \in L} \sum_{j \in L, j \neq i} \sigma_{L,ij} + 2n_E n_L \delta_E \delta_L \right] + \\ & (1 - W_{E,0})^2 (\sigma_E^2 + \delta_E^2) + \\ & \frac{2}{n_E + n_L} W_{E,0}(1 - W_{E,0}) \left[\sigma_E^2 + n_E \delta_E^2 + \sum_{i \in E, i \neq 0} \sigma_{E,i0} + n_L \delta_E \delta_L \right] \end{aligned}$$

Taking the derivative with respect to $W_{E,0}$ and setting it equal to zero results in the following equation:

$$\begin{aligned}
& \frac{2W_{E,0}}{(n_E + n_L)^2} [n_E^2 \delta_E^2 + n_L^2 \delta_L^2 + n_E \sigma_E^2 + \sum_{i \in E} \sum_{j \in E, j \neq i} \sigma_{E,ij} + n_L \sigma_L^2 + \sum_{i \in L} \sum_{j \in L, j \neq i} \sigma_{L,ij} + 2n_E n_L \delta_E \delta_L] + 2W_{E,0}(\sigma_E^2 + \delta_E^2) \\
& - 2(\sigma_E^2 + \delta_E^2) + \frac{2}{n_E + n_L} [\sigma_E^2 + n_E \delta_E^2 + \sum_{i \in E, i \neq 0} \sigma_{E,i0} + n_L \delta_E \delta_L] - \\
& \frac{4W_{E,0}}{n_E + n_L} [\sigma_E^2 + n_E \delta_E^2 + \sum_{i \in E, i \neq 0} \sigma_{E,i0} + n_L \delta_E \delta_L] = 0
\end{aligned}$$

Rearranging the equation results in:

$$\begin{aligned}
& W_{E,0} [n_E^2 \delta_E^2 + n_L^2 \delta_L^2 + n_E \sigma_E^2 + \sum_{i \in E} \sum_{j \in E, j \neq i} \sigma_{E,ij} + n_L \sigma_L^2 + \sum_{i \in L} \sum_{j \in L, j \neq i} \sigma_{L,ij} + 2n_E n_L \delta_E \delta_L] + W_{E,0} (n_E + n_L)^2 (\sigma_E^2 + \delta_E^2) \\
& - 2W_{E,0} (n_E + n_L) [\sigma_E^2 + n_E \delta_E^2 + \sum_{i \in E, i \neq 0} \sigma_{E,i0} + n_L \delta_E \delta_L] \\
& = (n_E + n_L)^2 (\sigma_E^2 + \delta_E^2) - (n_E + n_L) [\sigma_E^2 + n_E \delta_E^2 + \sum_{i \in E, i \neq 0} \sigma_{E,i0} + n_L \delta_E \delta_L]
\end{aligned}$$

Because the expression is very long, the numerator and denominator are described separately. The numerator for $W_{E,0}$ is

$$(n_E + n_L)^2 (\sigma_E^2 + \delta_E^2) - (n_E + n_L) [\sigma_E^2 + n_E \delta_E^2 + \sum_{i \in E, i \neq 0} \sigma_{E,i0} + n_L \delta_E \delta_L]$$

The denominator is

$$\begin{aligned}
& [n_E^2 \delta_E^2 + n_L^2 \delta_L^2 + n_E \sigma_E^2 + \sum_{i \in E} \sum_{j \in E, j \neq i} \sigma_{E,ij} + n_L \sigma_L^2 + \sum_{i \in L} \sum_{j \in L, j \neq i} \sigma_{L,ij} + 2n_E n_L \delta_E \delta_L] + (n_E + n_L)^2 (\sigma_E^2 + \delta_E^2) - 2 \\
& (n_E + n_L) [\sigma_E^2 + n_E \delta_E^2 + \sum_{i \in E, i \neq 0} \sigma_{E,i0} + n_L \delta_E \delta_L]
\end{aligned}$$

Solving for $W_{L,0}$ is straightforward since the derivation is exactly symmetric. Overall, correlated error terms within each agent type increase WOA. Nevertheless, the model's main prediction remains robust because – holding constant the covariance – an increase in δ_L lowers the WOA of the better-informed agents.

2 Mathematical Formulation of the Effect of Misclassification on the Algorithm

2.1 Ashman's D and Cluster Separability

Following the set-up in Section 3 of the main text, cluster separability can be mathematically evaluated by Ashman's D :

$$D = \frac{|\delta_E - \delta_L|}{\sqrt{\frac{(\sigma_E^2 + h^2) + (\sigma_L^2 + h^2)}{2}}} \quad (2.1)$$

where h is the bandwidth, given by the Silverman's Rule (Sheather and Jones 1991, Venables and Ripley 2013).¹ Clean separability (i.e., minimal misclassification) typically requires Ashman's $D > 2$. Intuitively, separability improves with larger bias gaps, smaller variances, and larger sample sizes (reducing bandwidth). Subsequent simulations examine how different parameters affect the algorithm across different values of Ashman's D .

2.2 Misclassification and the WOA of Clusters

The main text conceptually illustrates how misclassification by KDE could affect the algorithm. This section mathematically characterizes the degree of misclassification and how it affects the average WOA in each cluster. Misclassification can happen in reality when the better-informed and the worse-informed agents' initial estimates are not sufficiently distinguishable. Therefore, it is important to understand the impact of misclassification on the algorithm's performance. When agents are classified to the wrong cluster, these misclassified agents still have WOA of their actual type (e.g., some better-informed – despite being more informed – may have a less accurate estimate due to random noise, but their WOA still follows that of a better-informed type.) Consequently, misclassification would reduce the average WOA of the worse-informed cluster or increase the average WOA of the better-informed cluster, creating a potential hazard for the algorithm.

For simplicity, this section assumes the better-informed agents only exhibit conformity bias, or put more weight on advice than they should; and the worse-informed can only exhibit stubbornness, or put less weight on advice than they should. Both are conservative assumptions that reduce the algorithm's performance. Denote c and d the proportion of "conforming" or "over-updating" better-informed agents and the proportion of "stubborn" or "under-updating" worse-informed agents. The

¹The Silverman's Rule bandwidth $h = 0.9 \cdot \min(\hat{\sigma}, 1.34 IQR) \cdot n^{-0.2}$, with $\hat{\sigma}$ the sample standard deviation and IQR the interquartile range.

crowd has the following composition of agent types: $(1 - c) \cdot n_E$ regular better-informed agents, $c \cdot n_E$ conforming better-informed agents, $(1 - d) \cdot n_L$ regular worse-informed, and $d \cdot n_L$ stubborn worse-informed agents. "Regular" refers to the better-informed and the worse-informed whose WOA follows the rational benchmark without behavioral bias in advice-taking such that $\alpha_i = 1$ and $\beta_i = 1$. Furthermore, denote $\delta = \delta_L - \delta_E$ as the gap between δ_E and δ_L . Since the discussion focuses on misclassification of agents, what matters is the relative distance between two clusters representing each type of agents. Without loss of generality, the notations are simplified by centering the better-informed agents' distribution at 0, and the worse-informed agents' distribution at δ .

The theoretical framework quantifies the impact by characterizing the number of correctly and incorrectly classified agents in each cluster identified by KDE (see Table 2).

Agent Type	# of Agents in Better-informed Cluster	# of Agents in Worse-informed Cluster
Regular Better-informed	$(1 - c) \cdot n_E \cdot \Phi_E(T)$	$(1 - c) \cdot n_E \cdot (1 - \Phi_E(T))$
Conforming Better-informed	$c \cdot n_E \cdot \Phi_E(T)$	$c \cdot n_E \cdot (1 - \Phi_E(T))$
Regular Worse-informed	$(1 - d) \cdot n_L \cdot \Phi_L(T)$	$(1 - d) \cdot n_L \cdot (1 - \Phi_L(T))$
Stubborn Worse-informed	$d \cdot n_L \cdot \Phi_L(T)$	$d \cdot n_L \cdot (1 - \Phi_L(T))$

Table 2: $\Phi_E(T)$ is the Cumulative Distribution Function (CDF) of the better-informed agents' prior distribution $(N(0, \sigma_E^2))$, which characterizes the proportion of better-informed with initial answer smaller than T . $\Phi_L(T)$ is the CDF of the worse-informed agents' prior distribution $(N(\delta, \sigma_L^2))$, which characterizes the proportion of worse-informed with initial answer smaller than T .

Because of the misclassification, the "better-informed" cluster and the "worse-informed" cluster here do not necessarily refer to a cluster consisting only of the better-informed or only of the worse-informed. Denote the average or expected WOA of the better-informed cluster $E(W_{CL,E})$ which is the cluster with its peak density at δ_E and the average or expected WOA of the worse-informed cluster $E(W_{CL,L})$ which is the cluster with its peak density at δ_L . Thus, $E(W_{CL,E})$ may be specified as a weighted average of WOA among agents empirically classified into the better-informed cluster, with weights determined by the number of correctly classified regular and conforming better-informed agents and the number of misclassified regular and stubborn worse-informed agents as specified in Table 2. Then the average WOA of the better-informed cluster (i.e., those with an initial estimate less than T) is

$$E(W_{CL,E}) = (1 - c) \cdot \pi_{E,E} W_{E,0} + c \cdot \pi_{E,E} \cdot \alpha \cdot W_{E,0} + (1 - d) \cdot \pi_{L,E} W_{L,0} + d \cdot \pi_{L,E} \cdot \beta \cdot W_{L,0} \quad (2.2)$$

where $\pi_{E,E} = \frac{n_E \cdot \Phi_E(T)}{n_E \cdot \Phi_E(T) + n_L \cdot \Phi_L(T)}$ and $\pi_{L,E} = \frac{n_L \cdot \Phi_L(T)}{n_E \cdot \Phi_E(T) + n_L \cdot \Phi_L(T)}$ are, respectively, the proportions of all better-informed and all worse-informed that fall into the better-informed cluster.

Similarly, the average WOA of the worse-informed cluster (i.e., those with an initial estimate

greater than T) is

$$E(W_{CL,L}) = (1 - c) \cdot \pi_{E,L} W_{E,0} + c \cdot \pi_{E,L} \cdot \alpha \cdot W_{E,0} + (1 - d) \cdot \pi_{L,L} W_{L,0} + d \cdot \pi_{L,L} \cdot \beta \cdot W_{L,0} \quad (2.3)$$

where $\pi_{E,L} = \frac{n_E \cdot (1 - \Phi_E(T))}{n_E \cdot (1 - \Phi_E(T)) + n_L \cdot (1 - \Phi_L(T))}$ and $\pi_{L,L} = \frac{n_L \cdot (1 - \Phi_L(T))}{n_E \cdot (1 - \Phi_E(T)) + n_L \cdot (1 - \Phi_L(T))}$ are, respectively, the proportions of all better-informed and all worse-informed agents in the worse-informed cluster.

A successful identification of the better-informed cluster requires $E(W_{CL,L}) > E(W_{CL,E})$. One sufficient condition is

$$c < \frac{(1 - d) \cdot (\pi_{L,L} - \pi_{L,E}) \cdot W_{L,0} + d \cdot (\pi_{L,L} - \pi_{L,E}) \cdot \beta W_{L,0} - (\pi_{E,E} - \pi_{E,L}) \cdot W_{E,0}}{(\pi_{E,E} - \pi_{E,L}) \cdot (\alpha \cdot W_{E,0} - W_{E,0})} \quad (2.4)$$

The identification of the better-informed cluster becomes more successful as the number of conforming better-informed agents and stubborn worse-informed agents decreases, and when their respective biased advice-taking becomes small (i.e., $\alpha \rightarrow 1$ and $\beta \rightarrow 1$).

2.3 Aggregating the Better-informed Cluster vs Aggregating the Entire Crowd

Assume that the algorithm can correctly identify the better-informed cluster. This section provides a mathematical formulation for the initial cluster mean, updated cluster mean, initial group mean, and updated group mean building on the previous set-up. Denote the mean of the entire group with the subscript "gr" and the better-informed cluster mean with the subscript "cl". The initial mean of the entire group $E(\hat{\theta}_{gr,initial})$ is

$$E(\hat{\theta}_{gr,initial}) = \frac{n_L \cdot \delta}{n_E + n_L} + \frac{n_E \cdot 0}{n_E + n_L} + \delta_E \quad (2.5)$$

The mean of the initial estimates within the better-informed cluster $E(\hat{\theta}_{cl,initial})$ is the conditional mean of better-informed agents and worse-informed agents weighted by their respective numbers within the cluster (see SI Section 2.4 for the derivation). As a reminder, the center of the better-informed agents' distribution is set to 0, while the distance between the centers of the better-informed and worse-informed agents' distributions is maintained (e.g., this can be achieved by subtracting δ_E from every single estimate in the crowd, and subtracting a constant uniformly does not affect misclassification, which only depends on the relative gap between two clusters.) Simply "adding back" δ_E recovers the actual average prediction.

$$E(\hat{\theta}_{cl,initial}) = \frac{n_E \cdot \Phi_E(T) \cdot \mathbb{E}(\hat{\theta}_E|x_e < T) + n_L \cdot \Phi_L(T) \cdot \mathbb{E}(\hat{\theta}_L|x_l < T)}{n_E \cdot \Phi_E(T) + n_L \cdot \Phi_L(T)} + \delta_E \quad (2.6)$$

where $\mathbb{E}(\hat{\theta}_E|x_e < T) = -\sigma_E \cdot \frac{\phi(T/\sigma_E)}{\Phi(T/\sigma_E)}$ and $\mathbb{E}(\hat{\theta}_L|x_l < T) = \delta - \sigma_L \cdot \frac{\phi((T-\delta)/\sigma_L)}{\Phi((T-\delta)/\sigma_L)}$.

Since an agent's updated estimate may be characterized *updated answer = initial answer + WOA * (advice - initial answer)*, the updated mean of the entire group $E(\hat{\theta}_{gr,updated})$ is thus

$$\begin{aligned} E(\hat{\theta}_{gr,updated}) &= \frac{(1-c)n_E}{n_E+n_L} \cdot W_{E,0} \cdot \frac{n_L \delta}{n_E+n_L} + \frac{cn_E}{n_E+n_L} \cdot \alpha W_{E,0} \cdot \frac{n_L \delta}{n_E+n_L} \\ &\quad + \frac{(1-d)n_L}{n_E+n_L} \left(\delta - W_{L,0} \cdot \frac{n_E \delta}{n_E+n_L} \right) \\ &\quad + \frac{dn_L}{n_E+n_L} \left(\delta - \beta W_{L,0} \cdot \frac{n_E \delta}{n_E+n_L} \right) + \delta_E. \end{aligned} \quad (2.7)$$

Similarly, the mean of the updated estimates within the better-informed cluster $E(\hat{\theta}_{cl,updated})$ is

$$\begin{aligned} E(\hat{\theta}_{cl,updated}) &= \frac{1}{n_E \Phi_E(T) + n_L \Phi_L(T)} \left\{ \begin{aligned} &(1-c)n_E \Phi_E(T) \cdot [\mu_E + W_{E,0}(y - \mu_E)] \\ &+ cn_E \Phi_E(T) \cdot [\mu_E + \alpha W_{E,0}(y - \mu_E)] \\ &+ (1-d)n_L \Phi_L(T) \cdot [\mu_L + W_{L,0}(y - \mu_L)] \\ &+ dn_L \Phi_L(T) \cdot [\mu_L + \beta W_{L,0}(y - \mu_L)] \end{aligned} \right\} + \delta_E. \end{aligned} \quad (2.8)$$

2.4 Derivation of Conditional Expectation in the Descriptive Model

This section derives the conditional expectation (the expected proportion of better-informed classified into the better-informed cluster and the proportion of worse-informed classified into the worse-informed cluster). The derivation omits the subscripts for better-informed ("E") and worse-informed ("L") for notational simplicity. To recap the set-up, $\hat{\theta} \sim N(\theta + \delta, \sigma^2)$. Denote Z as the standard normal and $\phi()$ and $\Phi()$ as the density function and CDF of standard normal. Then $\phi(z) = \Phi'(z)$, $\phi'(z) = -z\phi(z)$. Denote T as the threshold determined by Kernel Density Estimation and $t = \frac{T-\theta-\delta}{\sigma}$. Then $P(\hat{\theta} < T) = P(Z < t) = \Phi(t)$.

$$\begin{aligned}
\mathbb{E}(\hat{\theta}|\hat{\theta} < T) &= \mathbb{E}(\sigma Z + \theta + \delta|Z < t) \\
&= \sigma \mathbb{E}(Z|Z < t) + (\theta + \delta) \mathbb{E}(1|Z < t) \\
&= (\sigma \int_{-\infty}^t z \phi(z) dz + (\theta + \delta) \int_{-\infty}^t \phi(z) dz) / P(Z < t) \\
&= (-\sigma \int_{-\infty}^t \phi'(z) dz + (\theta + \delta) \int_{-\infty}^t \Phi'(z) dz) / \Phi(t)
\end{aligned}$$

By the Fundamental Theorem of Calculus, which states $\int_a^b F'(z) dz = F(b) - F(a)$

$$\begin{aligned}
\mathbb{E}(\hat{\theta}|\hat{\theta} < T) &= (\theta + \delta) - \sigma \frac{\phi(t)}{\Phi(t)} \\
&= (\theta + \delta) - \sigma \frac{\phi((T - \theta - \delta)/\sigma)}{\Phi((T - \theta - \delta)/\sigma)}
\end{aligned}$$

where $\frac{\phi((T - \theta - \delta)/\sigma)}{\Phi((T - \theta - \delta)/\sigma)}$ is also known as the Inverse Mills ratio.

3 Beyond Two Clusters: Heterogeneous Estimations Among Worse-informed Agents

In the main text, all worse-informed agents are assumed to have the same estimation bias such that there is only one worse-informed cluster centered at δ_L . When there is only one correct answer, there is plausibly only one better-informed cluster. However, various worse-informed agents may err due to different reasons, resulting in clusters centered around distinct biases in estimation. Modeling heterogeneous estimation bias among worse-informed agents is analytically intractable and does not admit a clean closed-form solution. Nonetheless, the insights from the main text, which assume homogeneous estimation bias among worse-informed agents, remain largely applicable. First, if estimation bias of each worse-informed cluster is large, misclassification is minimal since worse-informed clusters are distinctly separable from the better-informed cluster. Second, when different estimation biases by worse-informed are small, the impact of additional worse-informed clusters on CWOA hinges on the relative positions of these worse-informed clusters. For example, as depicted in Figure 3.1, adding cluster *CL2* (or more clusters to the right of *CL1*)² to the previously discussed *CL0* and *CL1* does not exacerbate misclassification, provided the proportion of stubborn worse-informed in *CL2* remains low. This is because few worse-informed from *CL2* are

²Without loss of generality, denote *CL1* the closest cluster to the better-informed cluster, which introduces the most misclassification.

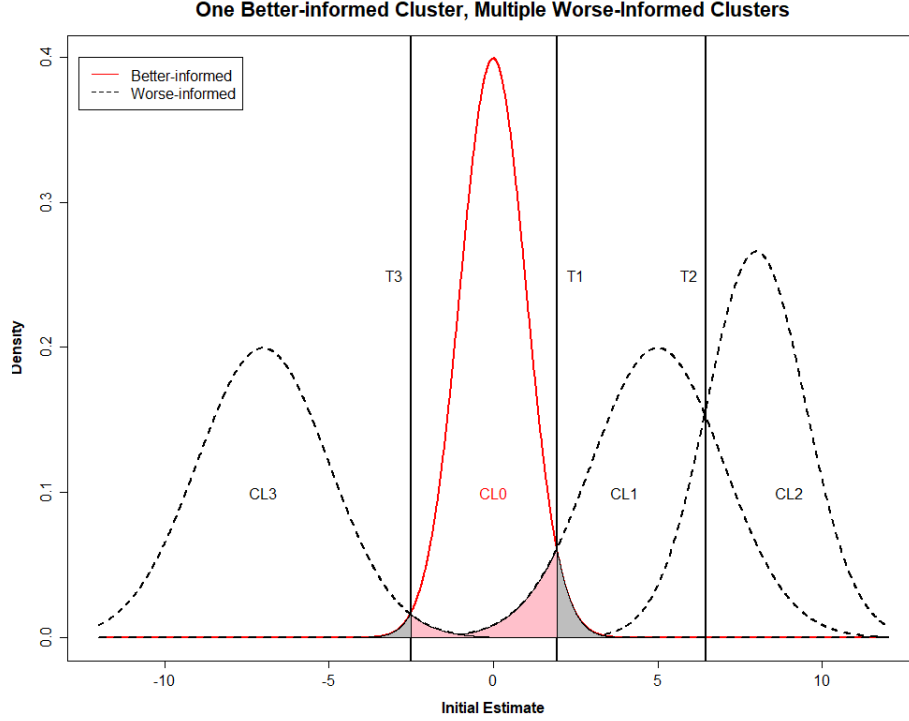


Figure 3.1: The truth is 0.5. The red curve is the distribution of better-informed agents' initial estimates. $T1$, $T2$, and $T3$ refer to the threshold identified by KDE. The black curves are distributions of worse-informed agents' initial estimates centered at $\delta_{L1} = 5.5$, $\delta_{L2} = 8.5$, and $\delta_{L3} = -6.5$. The dashed lines are the clustering threshold (initial estimates between $T3$ and $T1$ are classified into the better-informed cluster "CL0"; initial estimates to the left of $T3$, to the right of $T2$, and between $T1$ and $T2$ are classified into three distinct worse-informed clusters.) The shaded gray regions are the proportion of better-informed incorrectly classified to the worse-informed clusters, and the shaded pink regions represent worse-informed incorrectly classified into the better-informed cluster.

misclassified to the better-informed cluster. Moreover, some regular worse-informed from $CL2$ are classified to $CL1$, which raises the average WOA of the cluster $CL1$. Furthermore, the presence of additional biased clusters like $CL2$ would reduce the accuracy of the average estimate of the entire crowd, making aggregating over the better-informed cluster even more advantageous.

On the other hand, introducing $CL3$, as shown in Figure 3.1, might introduce some additional worse-informed to be misclassified into the better-informed cluster. However, the factors that determine the successful identification of the better-informed cluster remain the same, which has been extensively discussed in the model and via simulations. Perhaps surprisingly, this additional misclassification introduced by $CL3$ may benefit the algorithm – given the algorithm can still identify $CL0$ as the better-informed cluster. For example, when the better-informed cluster is truncated at $T1$, averaging over the better-informed distribution $N(0, 1 \mid x < T1)$ results in a bias to the left of the true value 0. With the introduction of $T3$, the left-skewed bias is mitigated in the new distribution $N(0, 1 \mid T3 < x < T1)$.

Overall, as long as the better-informed cluster's bias is small (i.e., δ_E is close to 0), aggregating over the cluster with the lowest average WOA emerges as the optimal strategy, effectively bypassing the biases of the worse-informed. However, the intuition introduced in SI Section 4 is an important boundary condition (see below). For example, it is possible that various δ_L and δ_E happen to cancel out in the aggregate. In this case, the group mean is already accurate in the first place, leaving little room for improvement.

4 Ecological Assumption: The Better-Informed Cluster Is More Accurate than the Group Mean

CWOA relies on an ecological assumption articulated by Mannes et al. (2014): aggregating over the most accurate subgroup outperforms aggregating over the entire crowd. This assumption can fail when the estimation biases of better-informed and worse-informed agents partially cancel out in the group mean.

To see this, consider a concrete example in which CWOA correctly identifies the better-informed cluster. Suppose objectively $\delta_E = -3$, $\delta_L = 5$, and the two types are equally represented. The group mean has absolute bias $|((-3) + 5)/2| = 1$, while the better-informed cluster mean has absolute bias $|-3| = 3$. Even though CWOA correctly identifies the better-informed cluster, aggregating over it yields a larger absolute bias than aggregating over the entire crowd. By contrast, when $\delta_E = 3$ (rather than -3) and $\delta_L = 5$, the group mean has absolute bias 4 while the better-informed cluster has absolute bias 3, and identifying the better-informed cluster outperforms the group mean.

Opposite-sign case. When δ_E and δ_L have opposite signs such that the truth lies between the two clusters, and the group mean partially benefits from cancellation. Aggregating over the better-informed cluster is therefore not guaranteed to improve over the group mean. However, when $|\delta_E|$ is small relative to $|\delta_L|$, or when better-informed agents are a minority of the crowd, the group mean is dominated by worse-informed agents' more biased estimates, and identifying the better-informed cluster is likely beneficial. In this regime, the updated cluster mean often outperforms the initial cluster mean, since updating toward the group mean partially corrects for the better-informed cluster's small residual bias.

Same-sign case. When δ_E and δ_L have the same sign, worse-informed agents' estimates exacerbate rather than offset bias. In this regime, aggregating only over the better-informed cluster is strongly preferred. Reweighting approaches that incorporate worse-informed agents tend to underperform, and the initial cluster mean often outperforms the updated cluster mean — since the group mean used as advice is more biased than the better-informed agents' initial estimates.

Out-of-scope extensions. Solving for an exact mathematical boundary by comparing Equations 2.5 and 2.6 is not pursued here, since the present simulations and models are intended to illustrate general intuition rather than provide structural prescriptions. A Decision Maker with reliable prior knowledge about δ_E and δ_L could fine-tune CWOA — for example, by opting for the initial cluster

mean over the updated cluster mean in same-sign settings, or by adopting a structural estimation approach. These directions lie beyond the scope of this paper.

Summary. There are cases where aggregating over the better-informed cluster does not improve over the group mean, which is one limitation of CWOA. Without additional contextual knowledge, this limitation is difficult to fully address. One natural extension is to reweight clusters based on average WOA rather than aggregate only over the cluster with the lowest average WOA. However, an effective reweighting scheme would need to specify how to balance WOA, cluster size, and other factors — questions left to future research. Moreover, while reweighting might help in cases where biases offset (e.g., $\delta_E = -3$, $\delta_L = 5$), it could amplify bias when biases compound (e.g., $\delta_E = 3$, $\delta_L = 5$). Accurate contextual knowledge would therefore be essential to refine CWOA in this direction. Overall, the empirical evidence in Section ?? of the main text supports Mannes et al. (2014)’s ecological assumption: CWOA’s consistent improvements over conventional benchmarks suggest that aggregating over a more accurate subgroup outperforms aggregating over the entire crowd in the contexts examined.

5 Simulation Procedure and Results on Determinants of WOA

The simulation procedure for one instance is as follows:

1. Generate n_E better-informed agents' initial estimates from $N(\delta_E, \sigma_E^2)$ and generate n_L worse-informed agents' initial estimates from $N(\delta_L, \sigma_L^2)$. The mean of the simulated initial estimates is the numerical advice.
2. Calculate the WOA of better-informed and worse-informed based on Equation 4.2 and 4.3 in the main text. (Note: This is sufficient for simulation-based analysis that compares WOA.)

(Steps below are for simulations comparing the performance of different methods in Section 6).

3. Calculate the updated estimates based on Equation 4.1 in the main text.
4. Following the algorithm as introduced, apply Kernel Density Estimation to identify the clusters of the simulated initial estimates and then aggregate over the cluster with the lowest average WOA as the algorithm's output.
5. Compute the updated cluster mean (i.e., the mean of the updated estimates within the *identified cluster*), the initial cluster mean (i.e., the mean of the initial estimates within the *identified cluster*), the initial group mean (i.e., the mean of the initial estimates of the *entire crowd*), the updated group mean (i.e., the mean of the updated estimates of the *entire crowd*).
6. Compare the absolute error (i.e., accuracy) of the algorithm's output against that of a benchmark method (e.g., the absolute error of the updated group mean minus the absolute error of the updated cluster mean; a positive (negative) value means a reduction (increase) in absolute error, indicating that the algorithm's output is more (less) accurate than the benchmark).
7. Repeat the above process 15 times³ and plot the average comparison.

All simulations assume that truth is 0.

³The number of simulation runs is set at 15 – neither too small nor too large – for two reasons. First, a much smaller number subjects results to the idiosyncrasies of a single simulated sample. Second, a much larger number can mask variability in the algorithm's performance: when $\delta_E \simeq \delta_L$, KDE may identify only one cluster and CWOA does not proceed; with many redrawn samples, the rare configurations that do trigger the advice-taking step accumulate, and averaging across them yields a *misleading* impression that the algorithm always proceeds beyond Step 2.

5.1 How Each Factor Affects WOA

This section first examines how each factor introduced in the model affect average WOA as a signal of accuracy before examining how they affect the algorithm's performance.

5.1.1 How Perceived Biases Affect WOA

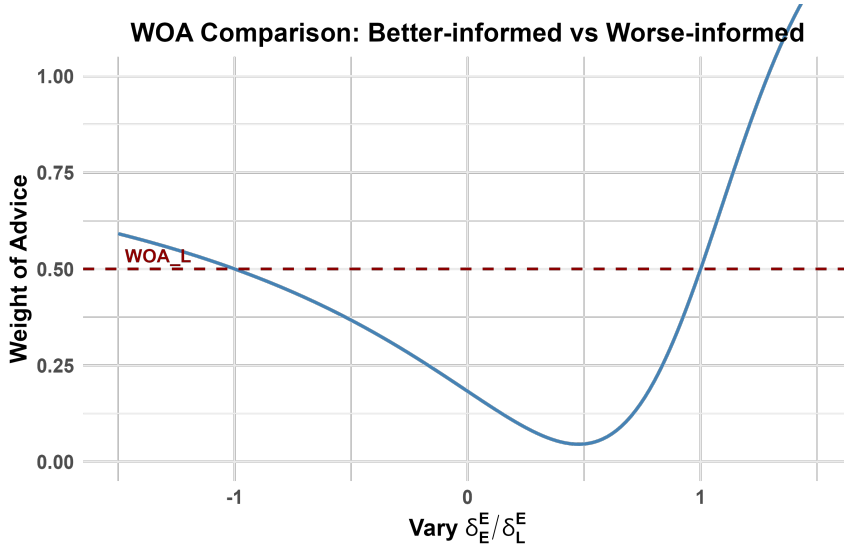


Figure 5.1: The x -axis is how better-informed agents perceive and compare their estimation bias (δ_E^E) with that of the worse-informed agents (δ_L^E). The y -axis is the WOA. Fixing all other parameters, the simulations vary $\frac{\delta_E^E}{\delta_L^E}$ such that $\frac{\delta_E^E}{\delta_L^E} = |1|$ the better-informed perceive the two groups have equal estimation bias. A value above (below) $|1|$ means the better-informed perceive their estimates to be more (less) biased than the worse-informed agents' estimate. The parameter set-up is as follows: equal perceived variance ($\sigma_L^E = \sigma_E^E = 4 = \sigma_L^L = \sigma_E^L$), equal behavioral bias in advice-taking ($\alpha = \beta = 0.5$), equal perceived crowd composition ($n_E^E = n_L^E = n_E^L = n_L^L$).

As shown in Figure 5.1, holding other factors constant, the WOA of better-informed agents increases with the perceived estimation bias of their own estimate relative to their perceived bias in others' estimates for most regions. As a note, WOA depends on the relative rather than the absolute magnitude of perceived bias (i.e., WOA remains the same whether $\sigma_L^E = \sigma_E^E = 4$ or $\sigma_L^E = \sigma_E^L = 40$.) Moreover, because the functional form of WOA is symmetric across groups, the same logic applies to worse-informed agents: their WOA increases with the perceived bias of their own estimate relative to that of the better-informed. Consequently, whether $WOA_E < WOA_L$ depends on how much less biased the better-informed perceive their estimates to be relative to the group mean compared to how the worse-informed perceive theirs. If the worse-informed perceive a larger bias in their own estimates, this loosens the requirement on how much less biased the better-informed must perceive themselves to be.

WOA is not strictly monotonic in $\frac{\delta_E^E}{\delta_L^E}$ (e.g., around $\frac{\delta_E^E}{\delta_L^E} = 0.45$) because of a bias–variance trade-off. When δ_E^E is very small, the group mean’s bias is also small, so placing more weight on advice reduces variance more than it introduces bias — yielding a temporarily higher WOA. Once $\frac{\delta_E^E}{\delta_L^E}$ crosses a threshold where the group mean’s bias dominates the variance reduction, WOA becomes strictly increasing in the ratio.

5.1.2 How Perceived Variances Affect WOA

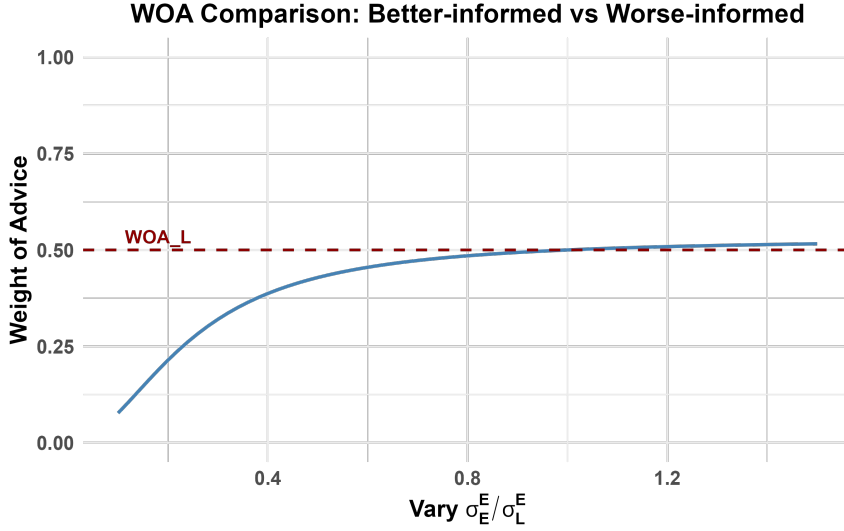


Figure 5.2: The x-axis is how better-informed agents perceive and compare their variance (σ_E^E) with that of the worse-informed agents (σ_L^E). The y-axis is the WOA. Fixing all other parameters, the simulations vary $\frac{\sigma_E^E}{\sigma_L^E}$ such that $\frac{\delta_E^E}{\delta_L^E} = 1$ the better-informed perceive the two groups have equal variances in their initial estimate. A value above (below) 1 means the better-informed perceive their estimates to have higher (lower) variance than the worse-informed agents’ estimate. The parameter set-up is as follows: equal perceived bias ($\delta_L^E = \delta_E^E = 10 = \delta_L^L = \delta_E^L$), equal behavioral bias in advice-taking ($\alpha = \beta = 0.5$), equal perceived crowd composition ($n_E^E = n_L^E = n_L^L = n_E^L$).

As shown in Figure 5.2, holding other factors constant, the WOA of better-informed agents increases with the perceived variance of their own estimate relative to their perceived variance in others’ estimates. As before, whether $WOA_E < WOA_L$ depends on the relative difference in perception of the variances in reference to the group mean between the better-informed and the worse-informed. For example, $WOA_E < WOA_L$ when $\frac{\sigma_E^E}{\sigma_L^E} < 1$, all else equal. This is also an illustration of the “classic” advice-taking result in Section 1.4.1, in which agents perceived to have lower variance in their estimates tend to have lower WOA.

Overall, variance plays a diminishing role, when all agents in the crowd are perceived to have an equal mean. As mentioned in the main text, a small gap in perceived estimation bias is sufficient to

overcome cases wherein the worse-informed comparatively perceive their estimates to be of much lower variance than the better-informed. This is intuitive for two reasons. First, if an estimate is perceived to be, say, heavily biased, even if the estimate has low variance, one should have low WOA, or they risk updating towards a precisely wrong answer. Second, fixing all other parameters, total sample size is another parameter that contributes to variance reduction, which is the foundational argument for Wisdom of Crowds. For example, imagine all agents are perceived to be unbiased, one group has a variance of 10, the other group has a variance of 1, and the perceptions of variances match thereof. When the sample size is sufficiently large, the group mean is accurate. Both groups – despite the differences in perceived variances – will have their WOA converging to 1, because the large sample size can guarantee the low variance of the group mean. This also contributes the diminishing of variance on WOA.

5.1.3 How Perceived Crowd Compositions Affect WOA

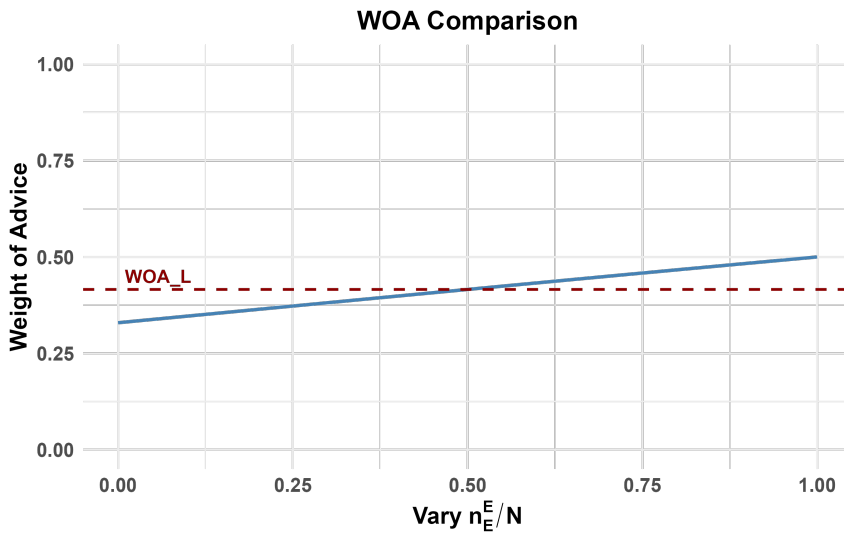


Figure 5.3: The x -axis is the better-informed agents' perceived proportion of better-informed agents in the crowd. The y -axis is the WOA. Fixing all other parameters, the simulations vary $\frac{n_E^E}{N}$ such that $\frac{n_E^E}{N} = 0.5$ means the better-informed perceive the two groups to have equal number in the crowd. A value above (below) 0.5 means the better-informed perceive there are more (less) better-informed agents in the crowd. The parameter set-up is as follows: $\delta_L^E = 10, \delta_E^E = 9.5$; $\delta_L^L = 9.5, \delta_E^L = 10$; equal perceived variance ($\sigma_L^E = \sigma_E^L = 4 = \sigma_L^L = \sigma_E^E$); this set-up allows for some difference between the two types of agents. Otherwise, perceived crowd composition does not matter if all agents' estimates are perceived to be equivalent.

As shown in Figure 5.3, holding other factors constant, the WOA of better-informed agents increases with the perceived proportion of better-informed agents in the crowd. To give one specific example, if better-informed agents believe there are 50% more better-informed agents, whereas worse-informed agents perceive the two types as equally represented, only a $\sim 7\%$ advantage in

perceived estimation bias (e.g., $\delta_E^E = 9.3$ vs $\delta_L^E = 10$) is required for $WOA_E < WOA_L$. This confirms the insight discussed in the main text that a slight advantage in the perceived gap in estimation bias by better-informed agents may be sufficient to overcome a range of adverse factors that increase WOA_E relative to WOA_L .

5.1.4 How Behavioral Biases in Advice-taking Affect WOA

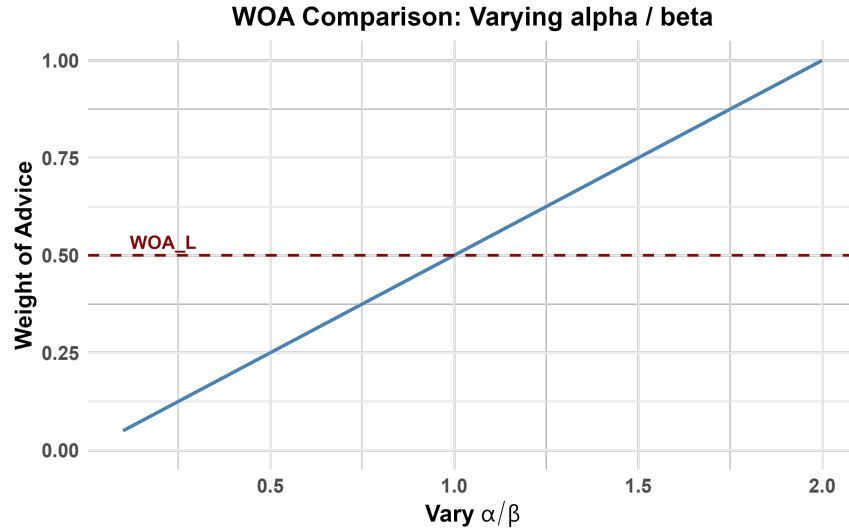


Figure 5.4: The x -axis is the better-informed agents' bias in advice-taking. The y -axis is the WOA. Fixing all other parameters, $\alpha = 1$ means the better-informed agents' WOA follows the rational benchmark. A value above (below) 1 means over-updating (under-updating). The parameter set-up is as follows: equal perceived bias ($\delta_L^E = \delta_E^E = \delta_L^L = \delta_E^L = 10$), equal perceived variance ($\sigma_L^E = \sigma_E^L = 4 = \sigma_L^L = \sigma_E^E$), equal perceived crowd composition ($n_E^E = n_L^E = n_E^L = n_L^L$)

As shown in Figure 5.4, holding other factors constant, the WOA increases with α .

6 How Each Factor Affects the Algorithm's Performance

This section introduces how each factor affects the algorithm's performance. For each factor, the simulations separately discuss scenarios with low (Ashman's $D = \sim 0.5$), medium (Ashman's $D = \sim 1.3$), versus high (Ashman's $D = \sim 2.1$) Ashman's D , small (total $n = 20$) versus large (total $n = 400$), as well as the proportion (i.e., $p = 0\%$, 30% , or 70% where $p = 0\%$ follows the derived benchmark formulations) of agents adopting "choosing strategy" (i.e., instead of WOA following a benchmark, the simulations set randomly p agents' WOA to be either 0 or 1). In sum, as mentioned in the main text, Ashman's D affects cluster separability and misclassification, the algorithm yields larger improvement when Ashman's D is large because (1) there are fewer misclassifications of agents, and (2) more separable clusters imply a larger gap in estimation biases, which means more room for improvement. The total sample size and the proportion of agents adopting the choosing strategy mostly affect the variability of the algorithm's performance. In expectation, the algorithm may remain effective, but a smaller sample size or more people randomly deviating from the benchmark increases the noisiness of the estimation procedure.

Unless otherwise noted, the parameter set-up remains the same as in the previous simulations.

6.1 How Perceived Biases Affect the Algorithm's Performance

6.1.1 Large Ashman's D, Large N

As shown in Figure 6.1, the algorithm improves upon conventional methods when $\frac{\delta_E^E}{\delta_L^E} < |1|$, which is when the better-informed agents have lower WOA than the worse-informed.

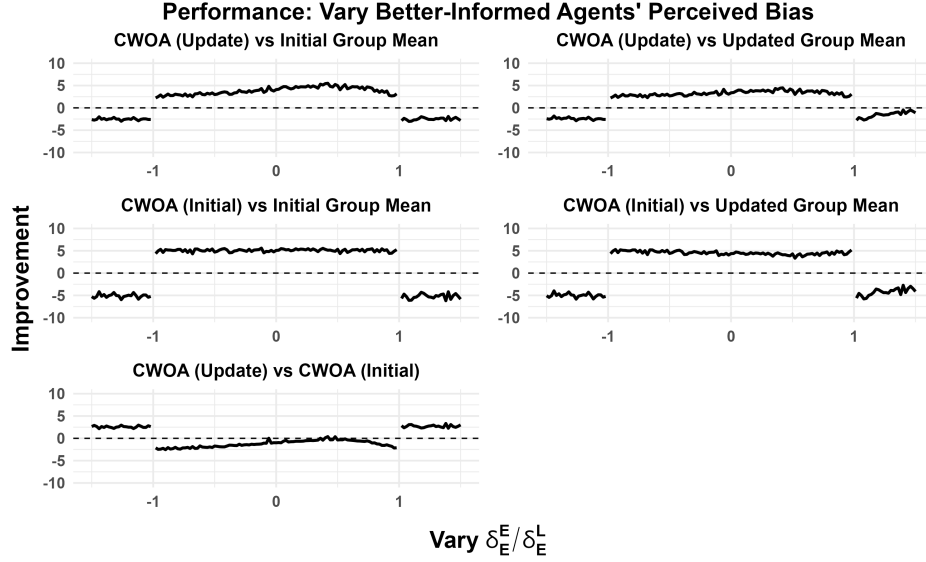


Figure 6.1: The x -axis is $\frac{\delta_E^E}{\delta_E^L}$. The title of each subplot (e.g. "Method A" vs "Method B") specifies the methods being compared. The y -axis is how much "Method A" improves upon "Method B" with the method of comparing accuracy specified in Step 5 of the simulation procedure. Objectively, $\delta_E = 1$, $\delta_L = 10$, $\sigma_E = 4$, $\sigma_L = 4$. Total $n = 200$ with equal number of better-informed and worse-informed agents.

6.1.2 Small Ashman's D , Large N

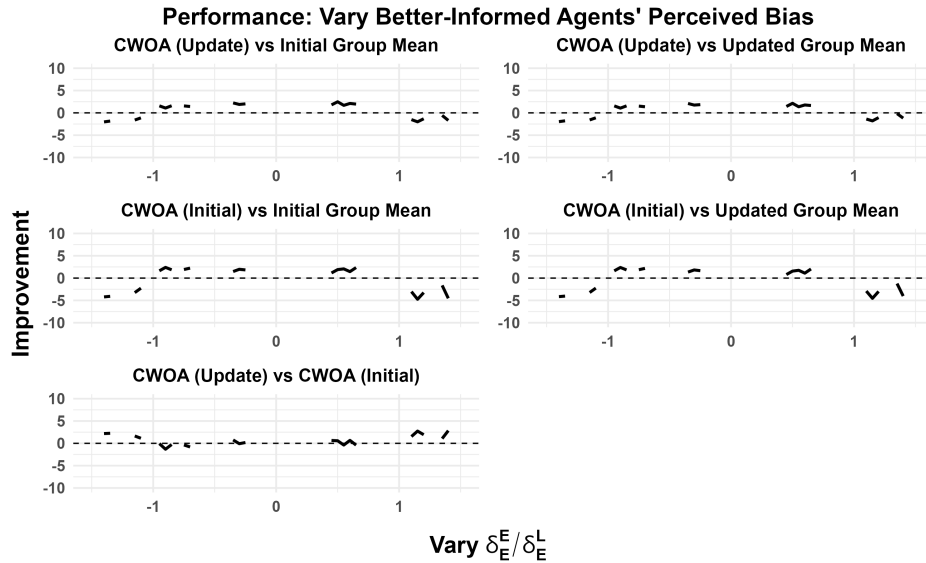


Figure 6.2: Same description as in the caption of Figure 6.1, except the objective set-up is such that Ashman's $D = \sim 0.5$. Total $n = 200$.

As shown in Figure 6.2, when Ashman's D is low. The algorithm often does not identify multiple clusters and therefore does not proceed.

6.1.3 Large Ashman's D, Small N

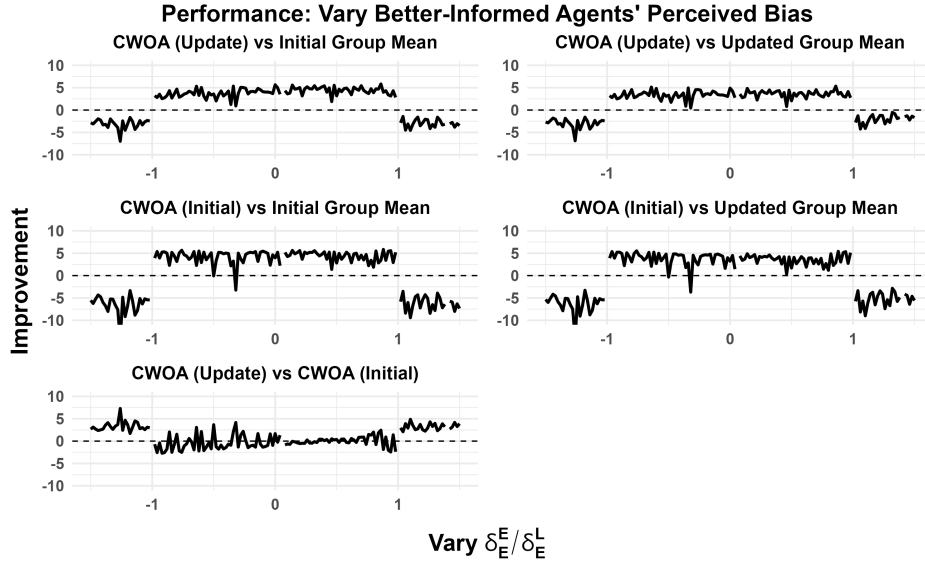


Figure 6.3: Same description as in the caption of Figure 6.1, except total $n = 20$.

As shown in Figure 6.3, when the sample size is small. The algorithm remains effective in expectation, though with greater variability in performance.

6.1.4 Medium Ashman's D, Small N

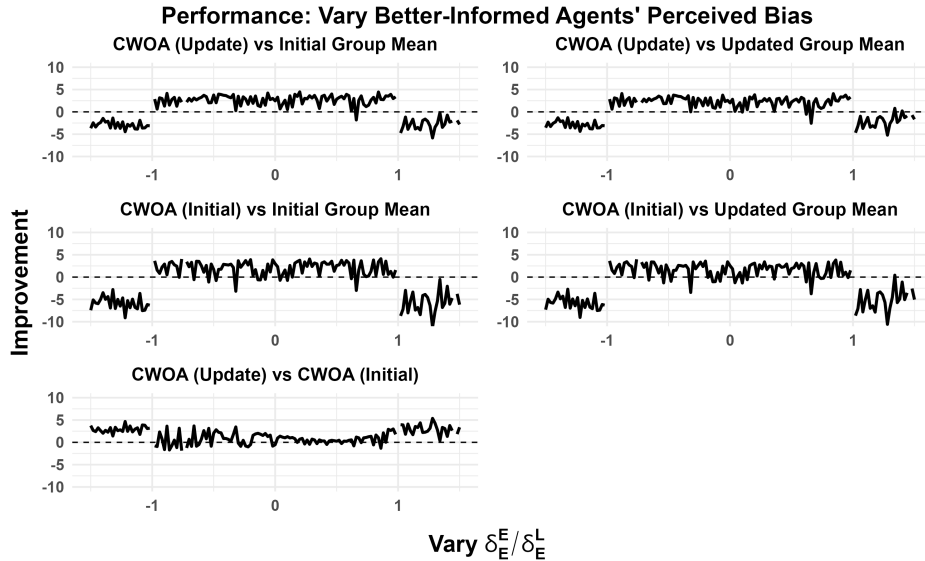


Figure 6.4: Same description as in the caption of Figure 6.1, except total $n = 20$ and Ashman's D is ~ 1.3 .

As shown in Figure 6.4, when the sample size is small and Ashman's D is medium, the algorithm remains effective in expectation except there is larger variability in performance.

6.1.5 Proportion Adopting Choosing Strategy: 30% vs 70%

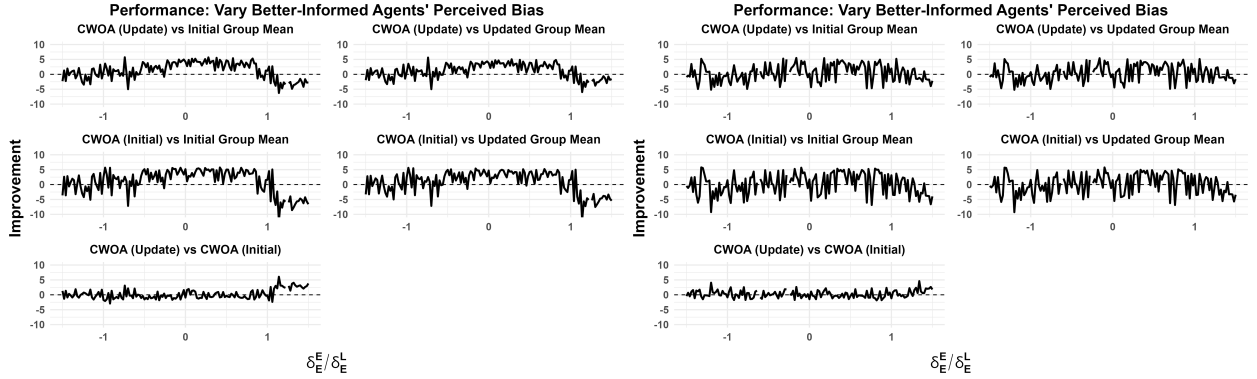


Figure 6.5: Same description as in the caption of Figure 6.1. Total $n = 20$; Ashman's D is large; 30% (left panel) vs. 70% of agents adopting the choosing strategy (i.e., randomly setting 30% or 70% of participants' WOA to be either 0 or 1.)

As shown in Figure 6.5, more agents adopting the choosing strategy increases the noisiness of the algorithm's performance.

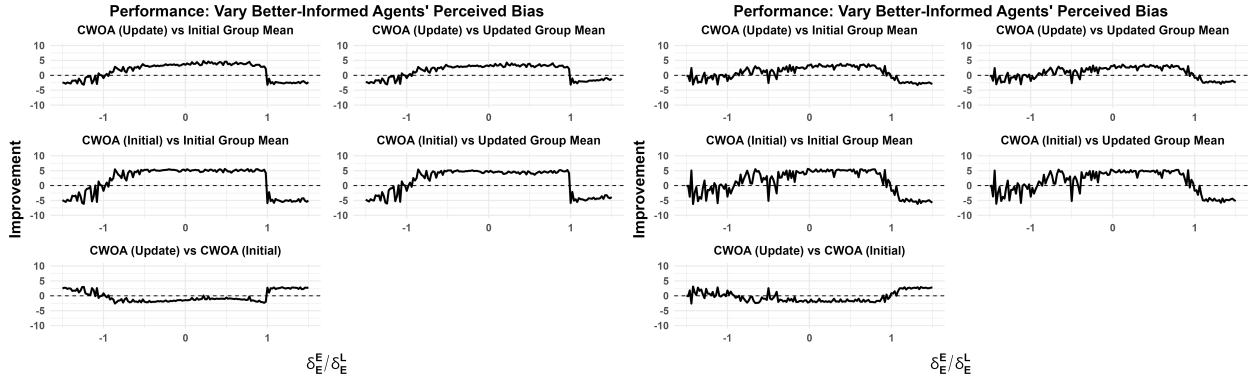


Figure 6.6: Same description as in the caption of Figure 6.5, except total $n = 200$.

Nevertheless, as shown in Figure 6.6, the noisiness from the "choosing strategy" may be offset by a large sample size, which reduces the noisiness of estimation.

6.2 How Perceived Variances Affect the Algorithm's Performance

Overall, the insight is the same as before. Whether CWOA improves upon benchmark depends on whether (1) the better-informed have lower WOA, and (2) whether Ashman's D is large and the overall variance of estimation procedure is low.

6.2.1 Large Ashman's D , Large N

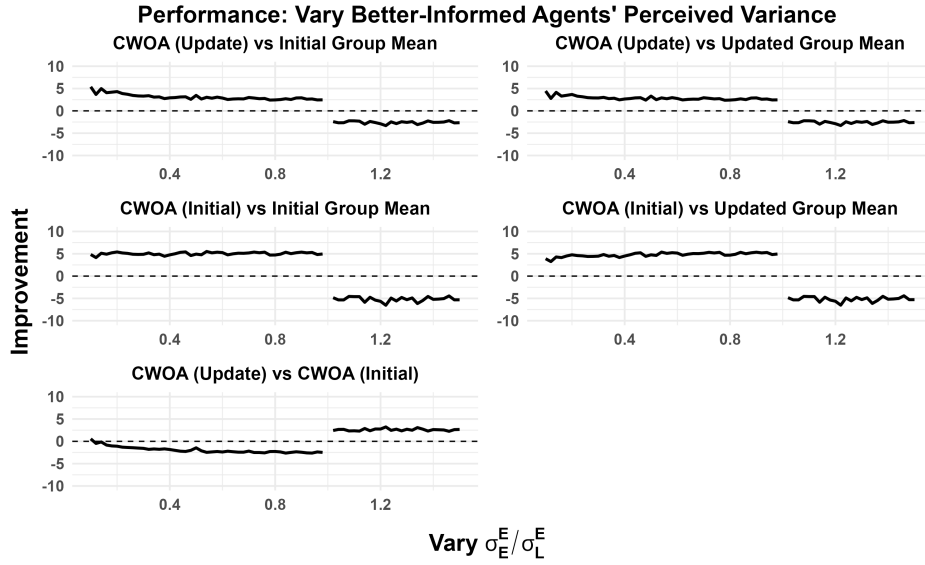


Figure 6.7: The x-axis is $\frac{\sigma_E^E}{\sigma_L^E}$. The title of each subplot (e.g. "Method A" vs "Method B") specifies the methods being compared. The y-axis is how much "Method A" improves upon "Method B" with the method of comparing accuracy specified in Step 5 of the simulation procedure. Objectively, $\delta_E = 1$, $\delta_L = 10$, $\sigma_E = 4$, $\sigma_L = 4$. Total $n = 200$ with equal number of better-informed and worse-informed agents.

As shown in Figure 6.7, the algorithm improves upon conventional methods when $\frac{\sigma_E^E}{\sigma_L^E} < |1|$, which is when the better-informed agents have lower WOA than the worse-informed.

6.2.2 Small Ashman's D , Large N

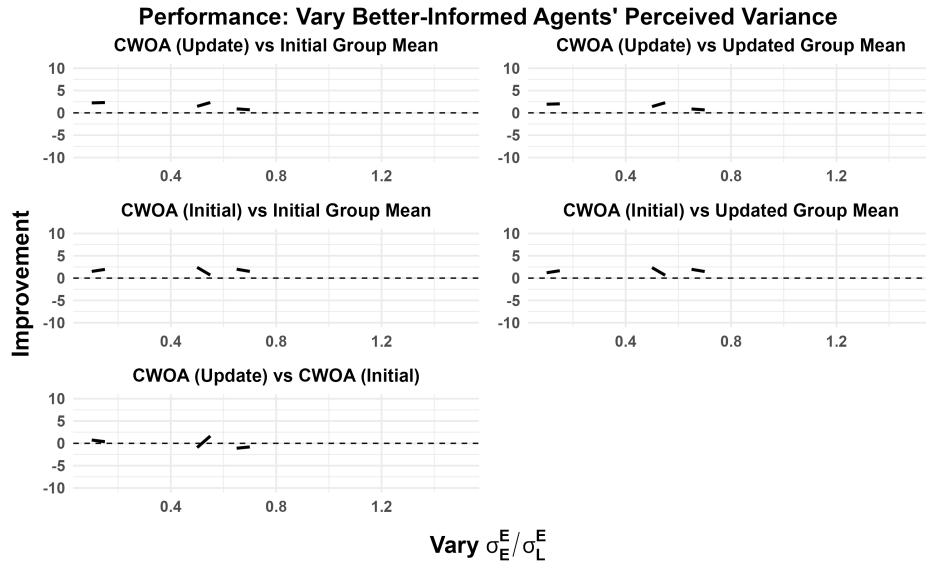


Figure 6.8: Same description as in the caption of Figure 6.7, except the objective set-up is such that Ashman's $D = \sim 0.5$. Total $n = 200$.

As shown in Figure 6.8, when Ashman's D is low, the algorithm often does not identify multiple clusters and does not proceed.

6.2.3 Large Ashman's D, Small N

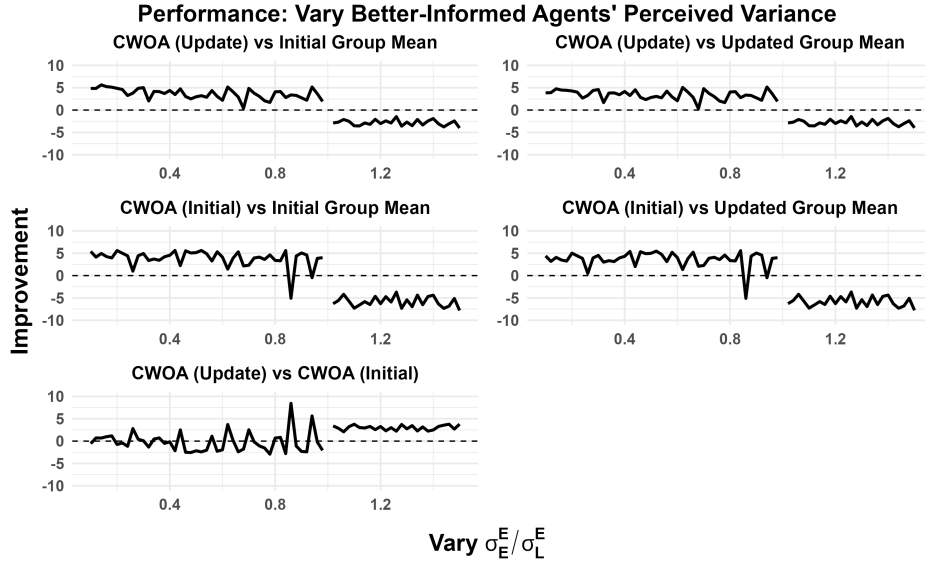


Figure 6.9: Same description as in the caption of Figure 6.7, except total $n = 20$.

As shown in Figure 6.9, when the sample size is small, the algorithm remains effective in expectation except there is larger variability in performance.

6.2.4 Medium Ashman's D, Small N

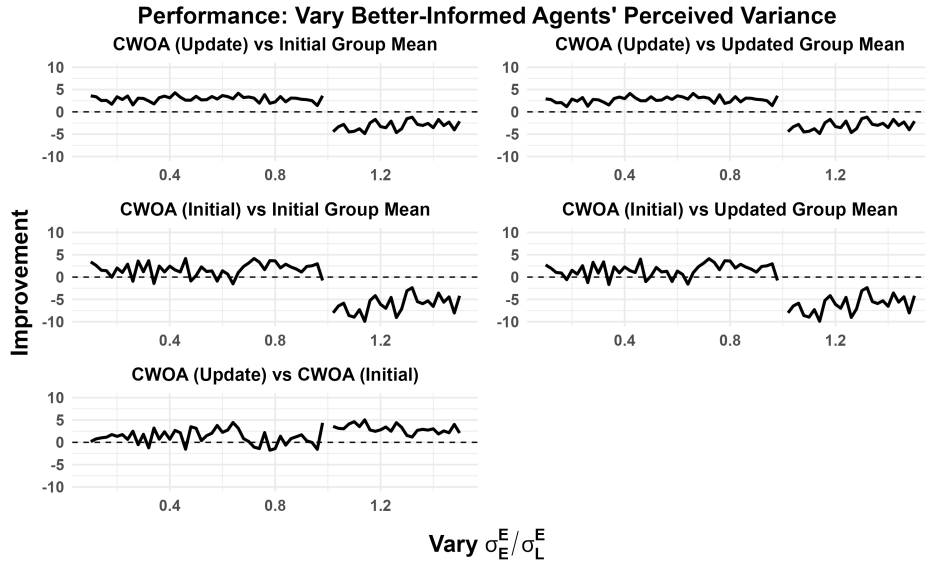


Figure 6.10: Same description as in the caption of Figure 6.7, except total $n = 20$ and Ashman's D is ~ 1.3 .

As shown in Figure 6.10, when the sample size is small and Ashman's D is medium, the algorithm remains effective in expectation except there is larger variability in performance.

6.2.5 Proportion Adopting Choosing Strategy: 30% vs 70%

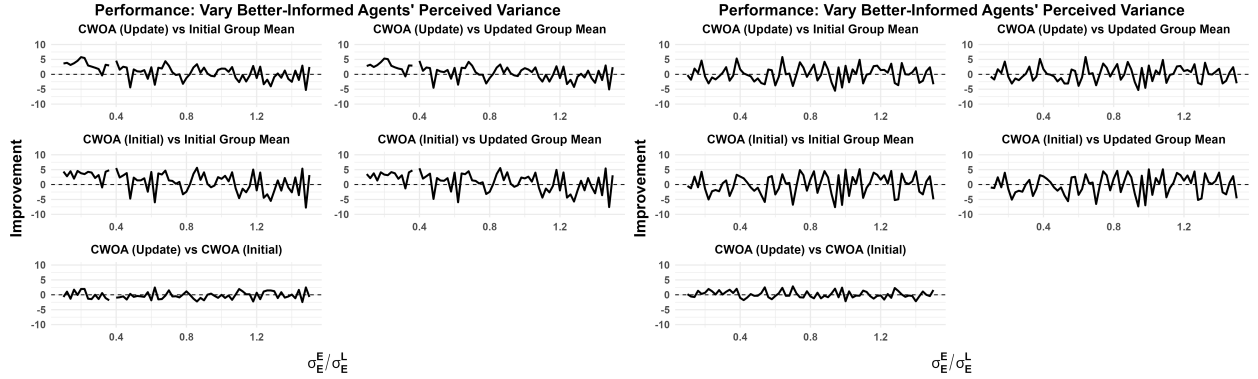


Figure 6.11: Same description as in the caption of Figure 6.7. Total $n = 20$; Ashman's D is large; 30% (left panel) vs. 70% of agents adopting the choosing strategy (i.e., randomly setting 30% or 70% of participants' WOA to be either 0 or 1.)

As shown in Figure 6.11, more agents adopting the choosing strategy increases the noisiness of the algorithm's performance.

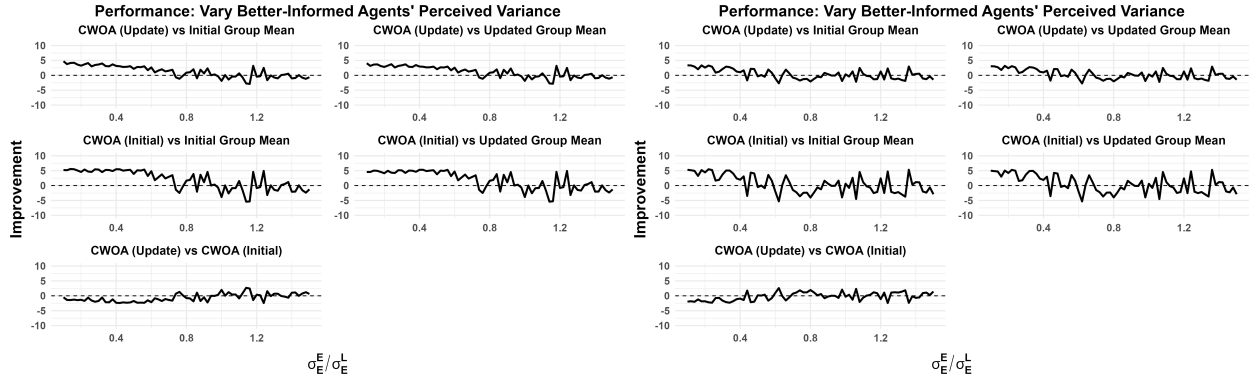


Figure 6.12: Same description as in the caption of Figure 6.11, except total $n = 200$.

Nevertheless, as shown in Figure 6.12, the noisiness from the "choosing strategy" may be offset by a large sample size, which reduces the noisiness of estimation.

6.3 How Perceived Crowd Compositions Affect the Algorithm's Performance

Overall, the insight is the same as before. Whether CWOA algorithm improves upon benchmark depends on whether (1) the better-informed have lower WOA, and (2) whether Ashman's D is large and the overall variance of estimation procedure is low.

6.3.1 Large Ashman's D , Large N

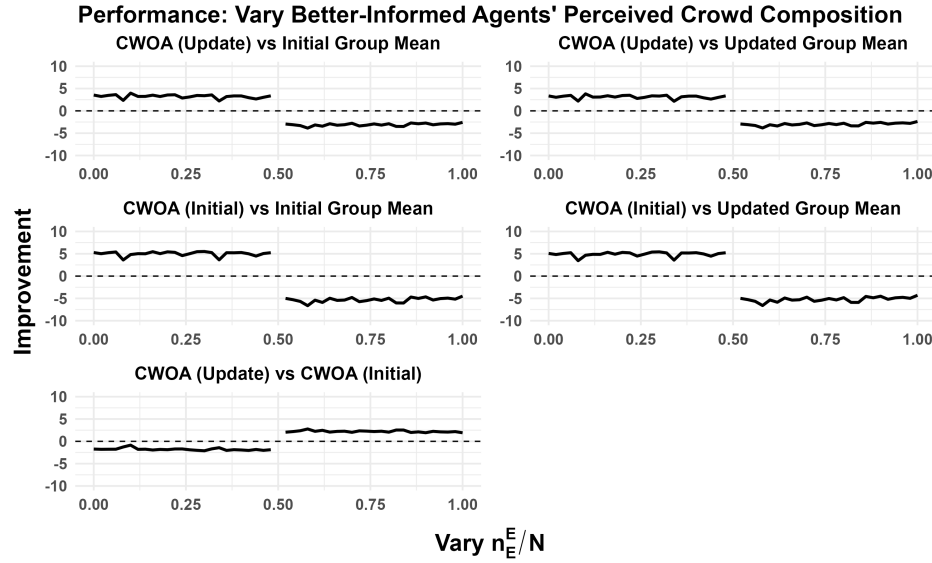


Figure 6.13: The x -axis is $\frac{n_E^E}{N}$ (perceived proportion of better-informed agents in the crowd by the better-informed agents). The title of each subplot (e.g. "Method A" vs "Method B") specifies the methods being compared. The y -axis is how much "Method A" improves upon "Method B" with the method of comparing accuracy specified in Step 5 of the simulation procedure. Objectively, $\delta_E = 1$, $\delta_L = 10$, $\sigma_E = 4$, $\sigma_L = 4$. Total $n = 200$ with equal actual number of better-informed and worse-informed agents.

As shown in Figure 6.13, the algorithm improves upon conventional methods when $\frac{n_E^E}{n_L^E} < |1|$, which is when the better-informed agents have lower WOA than the worse-informed. This is intuitive. When better-informed agents believe there are more better-informed agents in the crowd, they perceive higher quality in the numerical advice, leading to higher WOA. This belief (whether well-calibrated or not) can cause the algorithm to misidentify the wrong cluster as the more accurate one, reducing performance relative to benchmarks. Nevertheless, this claim may be constrained or reversed by the perceived bias-variance tradeoff. Suppose the better-informed subjectively believe that the better-informed agents have low bias but very high variance, such that updating towards the better-informed agents' estimate may not be as beneficial, then perceiving the crowd to be composed of more better-informed agents may actually lower their WOA.

6.3.2 Small Ashman's D , Large N

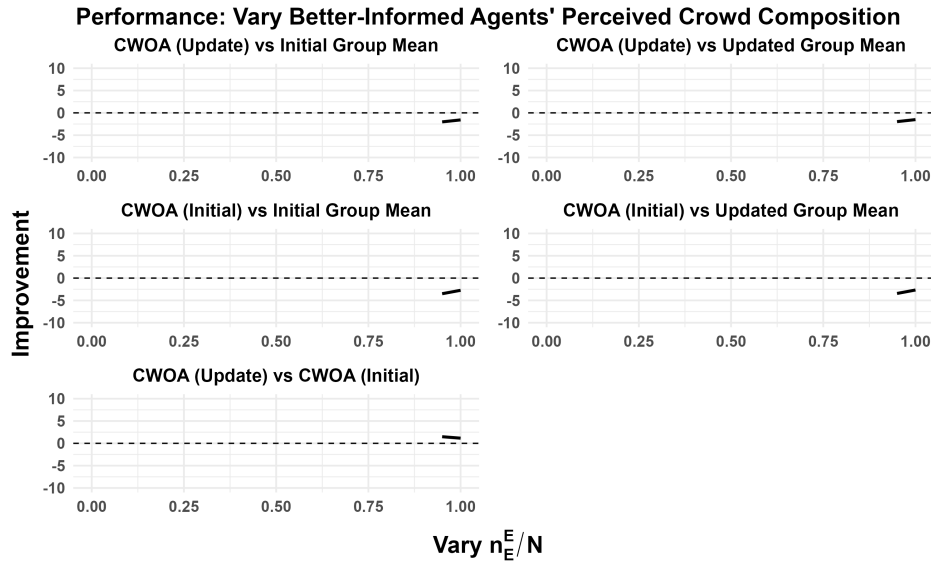


Figure 6.14: Same description as in the caption of Figure 6.13, except the objective set-up is such that Ashman's $D = \sim 0.5$. Total $n = 200$.

As shown in Figure 6.14, when Ashman's D is low, the algorithm often does not identify multiple clusters and the does not proceed.

6.3.3 Large Ashman's D, Small N

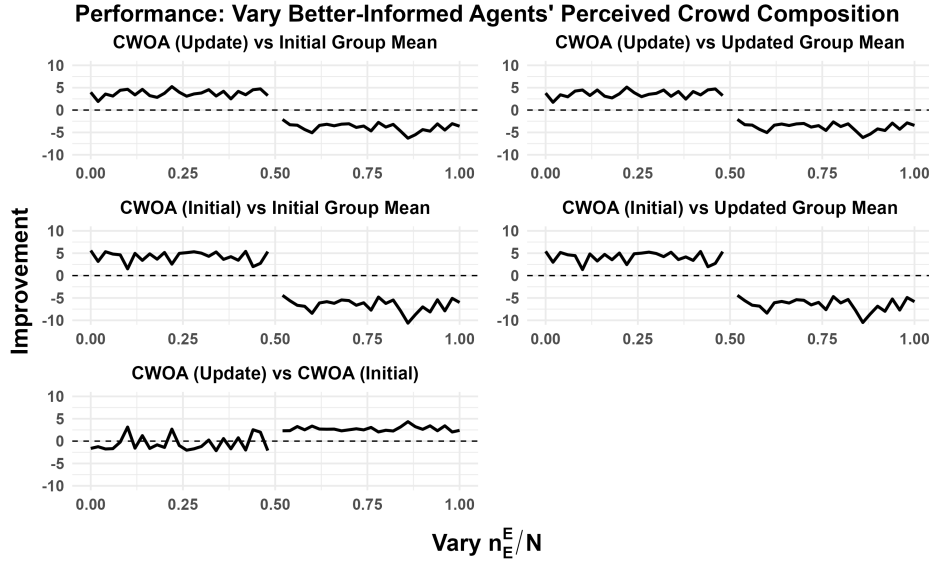


Figure 6.15: Same description as in the caption of Figure 6.13, except total $n = 20$.

As shown in Figure 6.15, when the sample size is small. The algorithm remains effective in expectation except there is larger variability in performance.

6.3.4 Medium Ashman's D, Small N

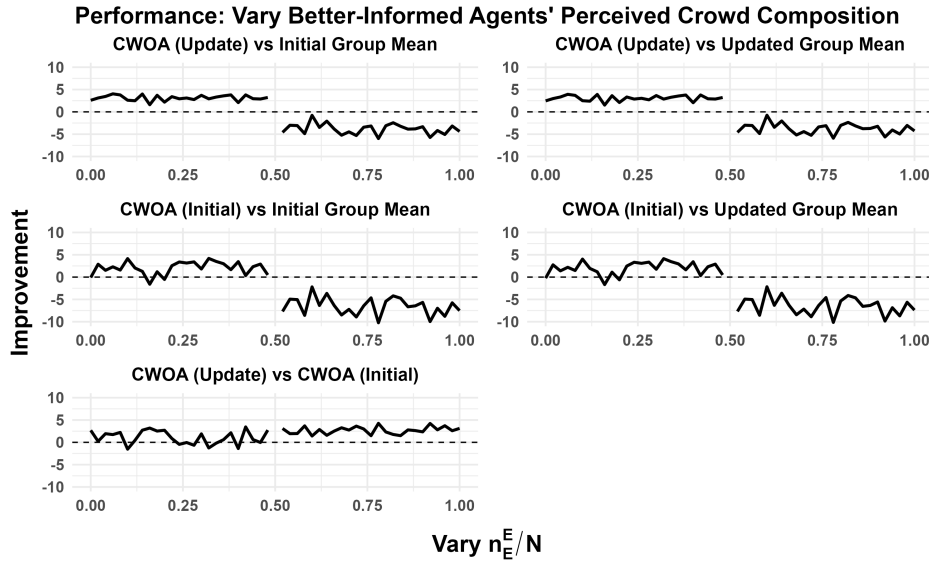


Figure 6.16: Same description as in the caption of Figure 6.13, except total $n = 20$ and Ashman's D is ~ 1.3 .

As shown in Figure 6.16, when the sample size is small and Ashman's D is medium, the algorithm remains effective in expectation except there is larger variability in performance.

6.3.5 Proportion Adopting Choosing Strategy: 30% vs 70%

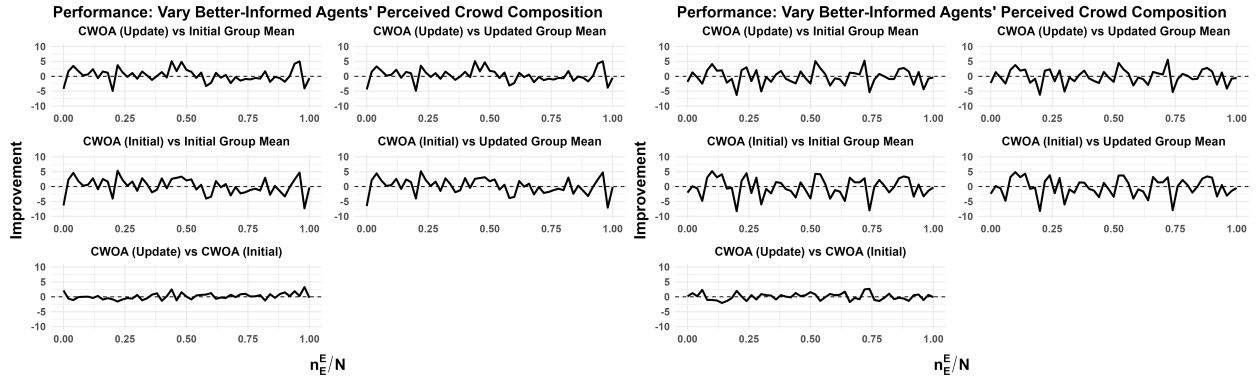


Figure 6.17: Same description as in the caption of Figure 6.13. Total $n = 20$; Ashman's D is large; 30% (left panel) vs. 70% of agents adopting the choosing strategy (i.e., randomly setting 30% or 70% of participants' WOA to be either 0 or 1.)

As shown in Figure 6.17, more agents adopting the choosing strategy increases the noisiness of the algorithm's performance.

6.4 How Biases in Advice-taking Affect the Algorithm's Performance

Overall, the insight is the same as before. Whether CWOA improves upon benchmark depends on whether (1) the better-informed have lower WOA, and (2) whether Ashman's D is large and the overall variance of estimation procedure is low. Over-updating by the better-informed agents makes WOA less indicative of accuracy. However, given that WOA can identify the more accurate group, in cases when the identified cluster may benefit from the numerical advice, over-updating might become helpful (e.g., the better-informed cluster overestimates the ground truth, but the numerical advice underestimates the ground truth). Over-updating may be detrimental in cases where, for instance, both the identified cluster and the numerical advice underestimate the ground truth.

6.4.1 Large Ashman's D , Large N

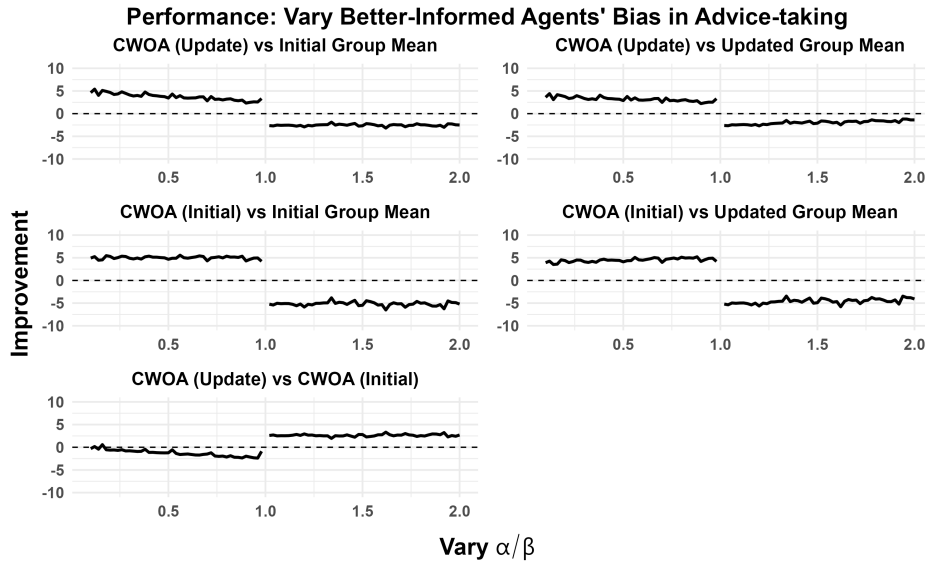


Figure 6.18: The x-axis is $\frac{\alpha}{\beta}$. The title of each subplot (e.g. "Method A" vs "Method B") specifies the methods being compared. The y-axis is how much "Method A" improves upon "Method B" with the method of comparing accuracy specified in Step 5 of the simulation procedure. The subjective perceptions of the two types of agents are set to be equal. Objectively, $\delta_E = 1$, $\delta_L = 10$, $\sigma_E = 4$, $\sigma_L = 4$. Total $n = 200$ with equal number of better-informed and worse-informed agents.

As shown in Figure 6.18, the algorithm improves upon conventional methods when $\frac{\alpha}{\beta} < 1$, which is when the better-informed agents have lower WOA than the worse-informed.

An interesting pattern emerges in the bottom panel of Figure 6.18, which compares the updated cluster mean to the initial cluster mean: as conformity bias increases beyond a certain threshold, the algorithm's performance improves. When α is large, the algorithm may identify the wrong cluster.

However, some better-informed agents are misclassified into the worse-informed cluster. Because of their large conformity bias, these misclassified better-informed agents over-update toward the numerical advice, partially correcting for the bias among the remaining worse-informed agents in the identified cluster. The updated cluster mean therefore becomes accidentally more accurate, despite the misidentification. Consistent with this interpretation, the initial cluster mean does not outperform the initial group mean as α increases. Relatedly, when WOA correctly signals the more accurate group, the update group mean may become much less accurate when α is large due to the better-informed over-updating towards the inaccurate group mean. Nevertheless, this claim is constrained by the relative location of the truth and the clusters. If the truth lies between the clusters, better-informed agents updating toward the group mean may improve both individual and collective accuracy.

6.4.2 Small Ashman's D , Large N

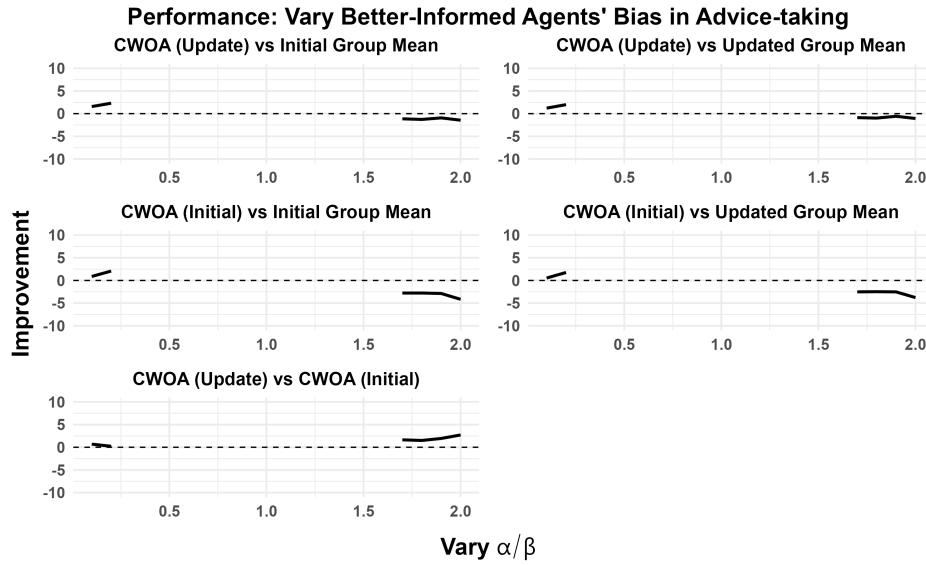


Figure 6.19: Same description as in the caption of Figure 6.18, except the set-up is such that Ashman's $D = \sim 0.5$. Total $n = 200$.

As shown in Figure 6.19, when Ashman's D is low, the algorithm often does not identify multiple clusters and the does not proceed.

6.4.3 Large Ashman's D , Small N

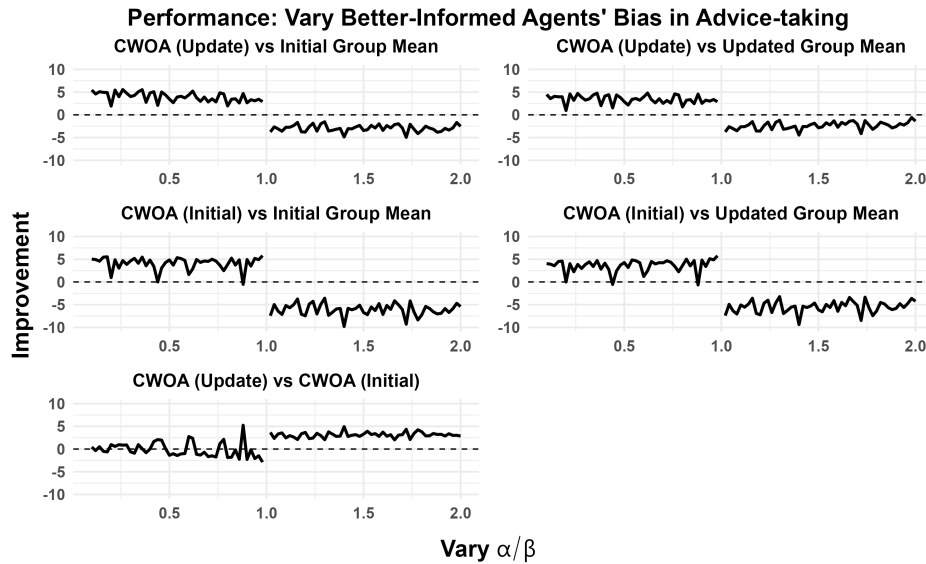


Figure 6.20: Same description as in the caption of Figure 6.18, except total Ashman's $D = 2.1$ and $n = 20$.

As shown in Figure 6.20, when the sample size is small. The algorithm remains effective in expectation except there is larger variability in performance.

6.4.4 Medium Ashman's D , Small N

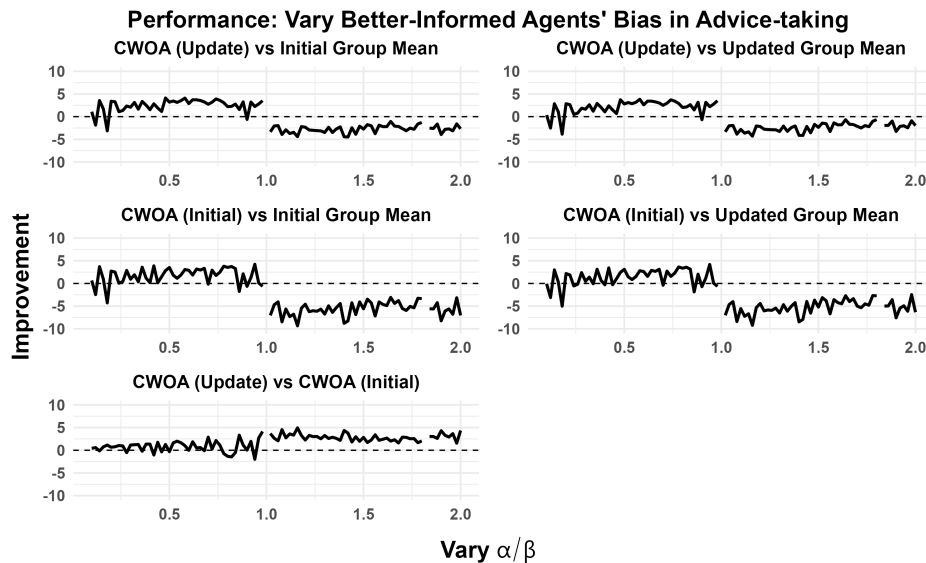


Figure 6.21: Same description as in the caption of Figure 6.18, except total $n = 20$ and Ashman's D is ~ 1.3 .

As shown in Figure 6.21, when the sample size is small and Ashman's D is medium, the algorithm remains effective in expectation except there is larger variability in performance.

6.4.5 Proportion Adopting Choosing Strategy: 30% vs 70%

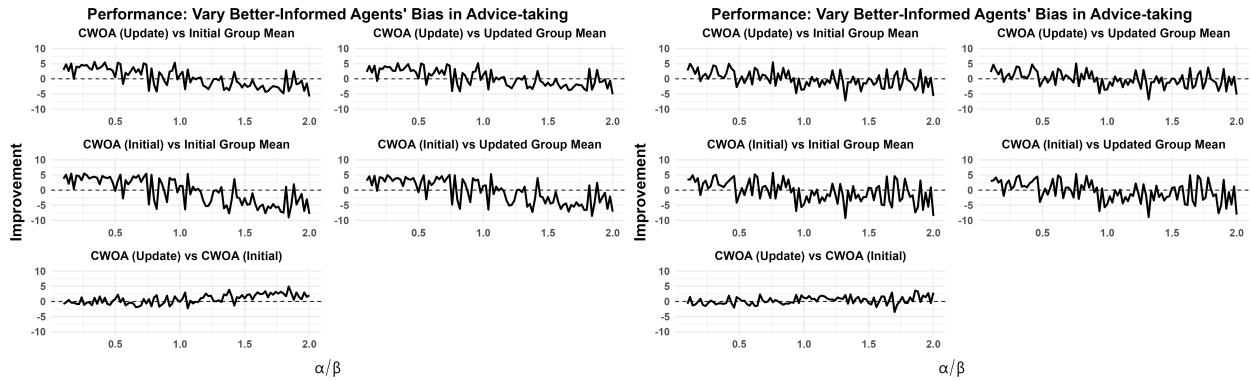


Figure 6.22: Same description as in the caption of Figure 6.18. Total $n = 20$; Ashman's D is large; 30% (left panel) vs. 70% of agents adopting the choosing strategy (i.e., randomly setting 30% or 70% of participants' WOA to be either 0 or 1.)

As shown in Figure 6.22, more agents adopting the choosing strategy increases the noisiness of the algorithm's performance.

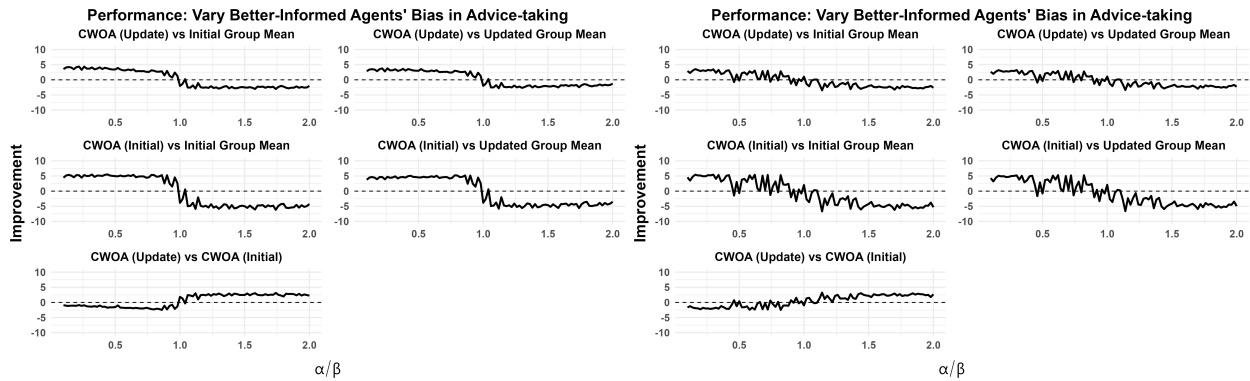


Figure 6.23: Same description as in the caption of Figure 6.22, except total $n = 200$.

Nevertheless, as shown in Figure 6.23, the noisiness from the "choosing strategy" may be offset by a large sample size, which reduces the noisiness of estimation.

6.5 Discussion of the Simulations Results

First, CWOA tolerates a wide range of detrimental factors – including behavioral biases in advice-taking and worse-informed agents being perceived as having lower variance – without losing its core advantage.

Second, whether the updated cluster mean outperforms the initial cluster mean depends on the parameter setup, as discussed in SI Section 4. Broadly speaking, the updated cluster mean is preferable when updating toward the numerical advice makes better-informed agents more accurate; otherwise, the initial cluster mean is preferable. In particular if truth is in-between the two clusters, the updated cluster mean is more likely to be superior as the group mean or the numerical advice is also in-between the two clusters. If both clusters overestimate or underestimate the truth, such that the truth is not bracketed by the two clusters, initial cluster mean tends to be superior as updating towards the group mean makes the better-informed less accurate. The recommendation of using the updated cluster mean as the default is based on empirical observation by past literature that studies whether advice-taking or social influence benefits collective intelligence.

Third, the effect of crowd composition on CWOA’s performance has additional nuance. For example, if the better-informed constitutes the majority of the crowd, the initial group mean would be more accurate to begin with, leaving less room for CWOA to improve upon. In addition, if there are more of a certain type of agents in the crowd, it also makes the perceptions of that type of agent in the crowd more salient in terms of its impact on the algorithm’s performance. (For example, if the better-informed constitute the majority, their perceptions of the bias and variance, which dictate their WOA and belief-updating, would have a larger impact on the final outcome.) Nevertheless, as a clarification, when the number of different types of agents is sufficient to reduce the noise in WOA, whether a group constitutes the majority or minority does not affect the identification of the more accurate cluster, fixing all other perceptions.

Finally, all main conclusions remain qualitatively robust, except the variability in performance becomes larger due to a smaller sample size. This is a common limitation of complex methodologies: as estimation complexity increases, more data is needed to stabilize the procedure. This bias-variance tradeoff and the “no free lunch” principle are fundamental insights in statistical theory (Hastie et al. 2009).

7 Simulation Continued: KDE versus K-means

The section compares K-means with KDE as the statistical clustering method. The key difference between the two methods in 1-dimensional clustering is that K-means requires the specification of the number of clusters *ex ante*, while KDE automatically recognizes the number of clusters based on the default Gaussian Kernel. Past literature in Statistics has well-established that KDE is the better choice of the two methods in 1-dimensional clustering (Sheather and Jones 1991, Venables and Ripley 2013).⁴ The main reason is that when the number of clusters is correctly pre-specified, K-means and KDE have similar performance, but KDE performs much better in case of misspecification in K-means. The results confirm this existing finding in the simulations (see Figure 7.1 through Figure 7.4 for details).

When Ashman's D is large (Figure 7.1 and 7.2), both clustering methods result in similar performance because (1) it is easy to separate the clusters, (2) any sub-clusters of the better-informed cluster identified by K-means tends to have low WOA (so K-Means faces less risk of misidentifying clusters.)

When Ashman's D is small (Figure 7.3 and 7.4, Ashman's $D = \sim 1.1$), KDE performs better as K-means is more likely to result in misidentification due to over-partitioning of clusters. The problem of over-partitioning is that each cluster now suffers from larger variance due to their smaller sample size. In addition, if K-Means accidentally partitions the better-informed cluster into smaller ones, this sub-cluster may be more biased (e.g., say, $N(0, 1)$ is the distribution of the unbiased estimates of the better-informed agents; over-partitioning could make the resulting cluster containing estimates only below 0, which becomes biased). The problem might be exacerbated in cases where there exists only 1 cluster, yet K-means forces the algorithm to identify multiple clusters, increasing the risk of mistaking a worse-informed cluster as the better-informed one.

7.1 Large Ashman's D , Better-informed Constitute the Majority

⁴K-means is more suitable for high-dimensional clustering.

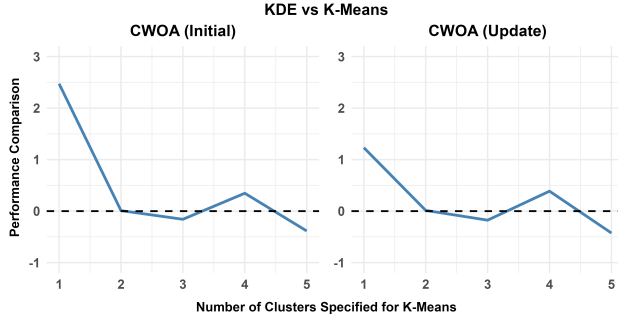


Figure 7.1: The x -axis is the number of clusters specified for K-means. The y -axis is the average improvement of algorithm using KDE outperforms compared to using K-means based on 50 bootstrapped samples under the parameter set-up. The data generating process follows the model introduced in the main text, meaning the correct number of clusters is 2. The parameter set-up is as follows: $N_E = 15 > N_L = 10, \sigma_E = \sigma_L = 1, \delta_E = 1, \delta_L = 6$; for simplicity, the simulations assume an accurate perceptions (subjective perceptions of parameters match objective reality) by both types of agents), $\alpha = \beta = 0.5$

7.2 Large Ashman's D, Better-informed Constitute the Minority

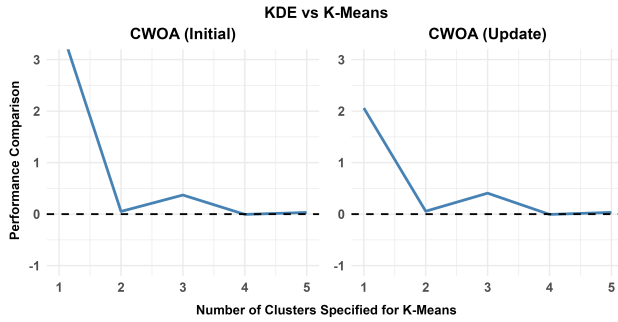


Figure 7.2: Same as the caption of Figure 7.1 except $N_E = 10 < N_L = 15$

7.3 Small Ashman's D, Better-informed Constitute the Majority

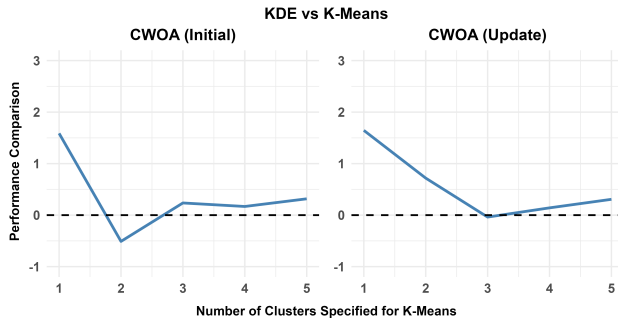


Figure 7.3: $N_E = 15 > N_L = 10, \sigma_E = \sigma_L = 2, \delta_E = 1, \delta_L = 4$

7.4 Small Ashman's D, Worse-informed Constitute the Majority

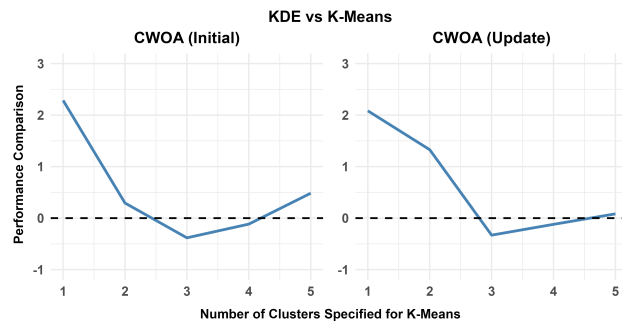


Figure 7.4: Same as the caption of Figure 7.3 except $N_E = 10 < N_L = 15$

8 Two Coin Jar Experiments

The paper contains two coin jar experiments: one presented in the main text (Experiment 1), and one discussed here in the supplement (Experiment 2). The two experiments are similar in their overall concept: manipulating information such that substantial errors may be attributable to estimation bias in lieu of variance. There are two main differences. First, the information structure in Experiment 1 is not nested, but the information structure in Experiment 2 is nested. Experiment 2 further manipulates crowd composition to check whether the algorithm is robust to whether the better-informed constitute the majority versus minority of the crowd. In addition, the wording for the subjective confidence questions is different for two of the three elicitations.

8.1 The First Coin Jar Experiment: Introduced in the Main Text

8.1.1 Additional Study Description

As a clarification, the wording used to elicit confidence in the updated estimate (after receiving numerical advice) is largely identical to the wording used for the initial estimate.

8.1.2 Additional Analysis

As mentioned in the main text, exploratory analysis finds that worse-informed participants (in the WI condition, who do not see the \$100-bill) show a larger decrease in confidence after seeing numerical advice than better-informed participants (in the BI condition). This pattern emerges despite the two groups reporting similar initial confidence. The change in confidence is computed as updated confidence in one's initial estimate minus initial confidence in one's initial estimate. For the Likert scale and probability confidence, a negative (positive) value means a decrease (increase) in confidence in initial estimate after seeing numerical advice. For the estimated error measure, the calculation first takes the absolute value of the measure before doing the subtraction. The interpretation is the opposite: a negative (positive) value means increase (decrease) in confidence in initial estimate after seeing numerical advice.

Here are two-tailed t-test results comparing change in confidence in initial estimate after seeing numerical advice: Likert, -0.522 (BI) vs. -0.151 (WI), $t(181) = -2.31$, $p = 0.022$; likelihood, -19.44% (BI) vs. -13.01% (WI), $t(180) = -1.80$, $p = 0.073$. Nevertheless, the difference in change is directionally consistent, but not statistically significant for the absolute estimated error measure: 6.70 (BI) vs. 9.21 (WI), $t(173) = 0.86$, $p = 0.394$. This happens potentially because the estimation error is an unbounded and thus very noisy measure. A *post-hoc* analysis suggests that

worse-informed participants are more likely to realize their initial estimate is underestimating the true value. Defining 'underestimation' as a signed estimated error less than 0: prior to seeing the numerical advice, the two groups have an equal proportion of participants who believe their initial estimate is underestimating the true value (i.e., estimated error < 0) (t -test comparing proportion with estimated error below 0, 15.1% (BI) vs. 16.7% (WI), $t(180) = -0.297$, $p = 0.767$). After seeing numerical advice, more worse-informed participants believe their initial estimate underestimated the true value (t -test comparing proportion with estimated error below 0, 10.8% (BI) vs. 27.8% (WI), $t(180) = 2.96$, $p = 0.003$).

As an additional note, the results show that participants in the two conditions are equally confident in their *updated* estimate (t -test comparing stated confidence: Likert, 4.05 (BI) vs. 3.96 (WI), $t(179) = 0.44$, $p = 0.66$; likelihood, 53.0% (BI) vs. 53.1% (WI), $t(180) = -0.025$, $p = 0.98$; absolute estimated error: 6.4 (BI) vs. 7.6 (WI), $t(178) = 0.454$, $p = 0.65$).

8.2 The Second Coin Jar Experiment: Omitted in the Main Text

All parameters of the algorithm were pre-registered in detail. Full replication materials including the pre-registration are available on the OSF site: https://osf.io/ga6qu/overview?view_only=78c42f3ccac34e229b7d58071a316ba5

8.2.1 Study Description

I recruited 360 participants from Prolific to estimate the monetary value in a jar. Estimate accuracy was scored with a quadratic scoring rule. Two photos of the jar were taken — one from the front and one from the back (see Figure 8.1).



Figure 8.1: The photo on the left is taken from the front, and the photo on the right is taken from the back. The total monetary value in the jar is \$126.49, and there are 490 coins worth a total of \$26.49.

The study has a between-subjects 2×2 design: {"Uninformed" condition vs. "Informed" condition} $\times$ {"Majority Uninformed" Crowd vs. "Majority Informed" Crowd}.

The worse-informed, or those in the *uninformed* condition, only see the front-view photo, in which only randomly scattered coins are observable. The better-informed, or those in the *informed* condition, see both the front-view photo and the back-view photo, thus they are aware of the \$100-bill. Importantly, participants are not informed of the experimental manipulations (i.e., they are not told some participants would only see the front view and others would see both views), which is a more conservative test of the model compared to if participants are informed of the other conditions. Based on their information, participants provide their responses to the following key questions: (1) the total value of money in the jar; (2) how likely is it that your estimate is close to (within \$5 above or below) the correct value? (probability between 0% and 100%); (3) how confident are you that your estimate is close to (within \$5 above or below) the correct value? from 1 (not at all confident) to 7 (absolutely confident); and (4) the average estimate of other Prolific participants regarding the question "what is the total value of money in the jar?" In addition, they also estimate the total number of coins in the jar and state their uncertainty in similar fashions (within 20 coins above or below the correct number).

The lack of awareness of the \$100-bill should lead the worse-informed to be heavily biased in their estimate of the total monetary value in the jar, which represents the δ_L term in the model. The study also elicits participants' subjective confidence as a proxy for their subjective uncertainty. After answering the questions above, participants receive the following information: *"the average estimate of the total value of money in the jar provided by a large group of previous participants is 'x'."* Based on a pre-test, for participants randomly assigned to the "Majority Uninformed" crowd condition, which contains one-third "informed" participants and two-thirds "uninformed" participants, the numerical advice is \$51. For participants randomly assigned to the "Majority Informed" crowd condition, which contains two-thirds "informed" participants and one-third "uninformed" participants, the numerical advice is \$87. Then participants provide their updated estimate of the total value of money in the jar.

8.2.2 Results: The Algorithm's Performance

As shown in Figure 8.2, the pre-registered output of CWOA (i.e., the mean of the updated estimates in the identified cluster, labeled "CWOA (Update)") outperforms taking the average of the updated or the initial estimates of the entire crowd, the meta-prediction-based methods such as minimal pivoting and knowledge-weighting, and the inverse-WOA-weighting (i.e., placing disproportion-

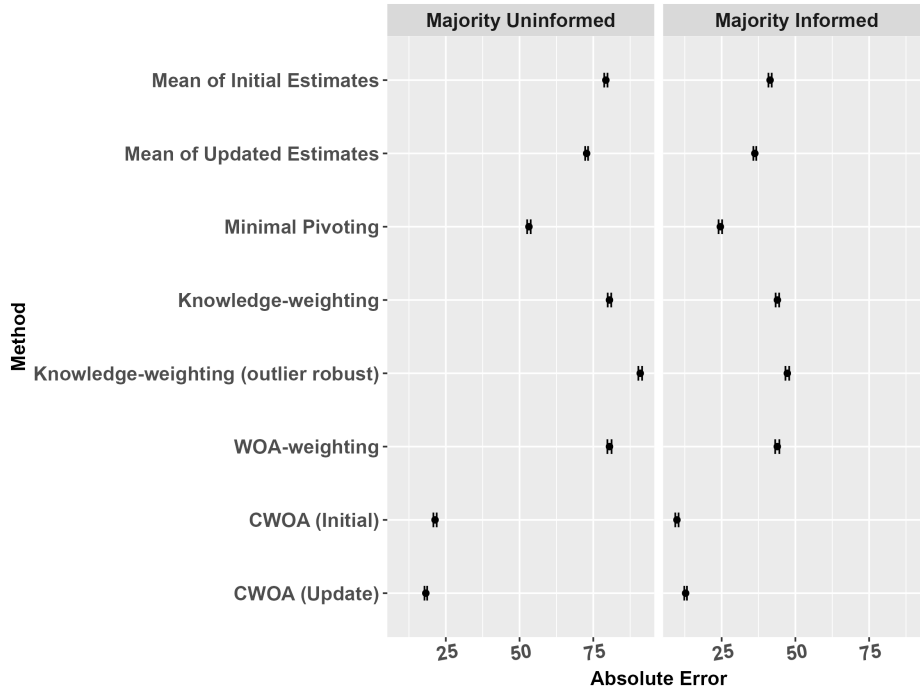


Figure 8.2: The figure reports the absolute error (x -axis) of each method (y -axis). The absolute error is calculated as $\text{method's output} - \text{truth}$. A lower absolute error means better performance. The bars indicate 95% confidence intervals obtained based on 300 bootstrapped samples. The bootstrap procedure is described in the pre-registration.

ately large weight on agents with a WOA of zero⁵) – both when better-informed agents constitute a majority or a minority of the crowd. So does the mean of the initial estimates in the identified cluster (labeled "CWOA (Initial)").

8.2.3 Results: WOA-Accuracy Association, Crowd Estimation Bias, and the Algorithm's Effectiveness

First, as manipulation checks, participants in the *uninformed* condition are less accurate in their estimate of the monetary value in the jar than those in the *informed* condition (two-tailed t -test comparing absolute errors, 116.0 (Uninformed) vs. 20.3 (Informed), $t(236) = 37.7$, $p < 0.0001$). Nevertheless, because the worse-informed are unaware of their bias caused by the \$100-bill, they do not state a lower subjective confidence regarding their estimates than the better-informed (two-tailed t -test, probability: 62.5% (Uninformed) vs. 56.9% (Informed), $t(358) = 2.2$, $p = 0.028$; Likert: 4.54 (Uninformed) vs. 4.12 (Informed), $t(354) = 2.86$, $p = 0.004$) (if anything, the worse-informed are slightly more certain about their estimates than the better-informed). See SI Sections 9.1 and 9.2 for additional analysis.

⁵A WOA of zero means agents' initial and updated estimates are the same.

The key analysis on WOA confirms the algorithm's robustness: regardless of whether the better-informed constitute the majority or the minority of the crowd, the better-informed and more accurate agents – despite expressing lower subjective confidence in their initial estimate – on average have lower WOA than the worse-informed and less accurate agents (t -test comparing WOA between the *informed* and the *uninformed* condition, *Majority Informed* crowd: 0.278 (Uninformed) vs. 0.120 (Informed), $t(166) = 3.46$, $p = 0.0007$; *Majority Uninformed* crowd: t -test, 0.299 (Uninformed) vs. 0.145 (Informed), $t(97) = 2.63$, $p = 0.010$).

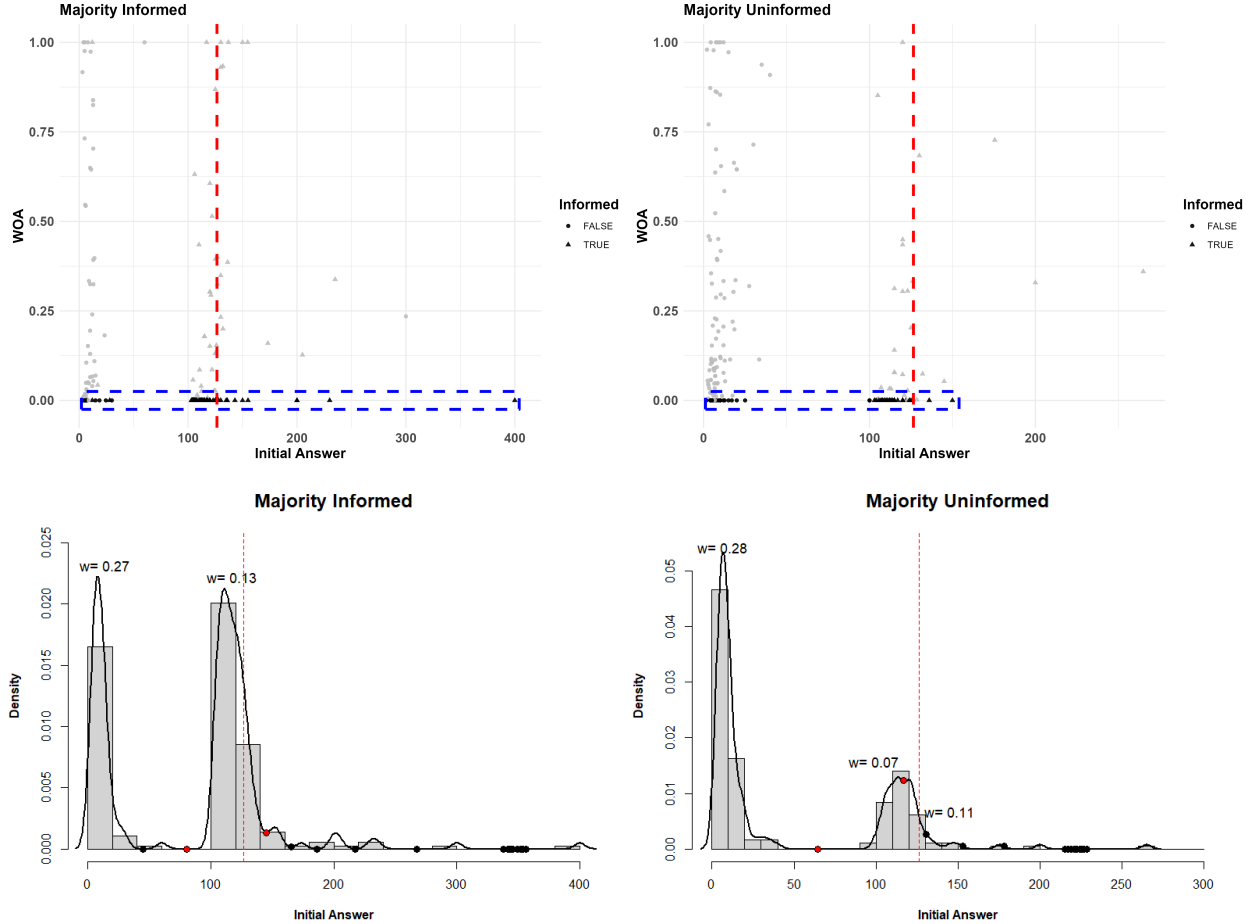


Figure 8.3: The left (right) column describes data from the "Majority Informed" crowd and the "Majority Uninformed" crowd. The top row plots participants' WOA against their initial estimate. WOA is winsorized between 0 and 1 for visualization purpose. The red dashed line indicates the correct value \$126.49. The bottom row contains histograms each with a density curve based on KDE depicting the clusters (between two dots or "pits" is a cluster). The average WOA (w) of each eligible cluster, which contains at least 5% of the subjects, is reported. The average WOA of the identified cluster in the *Majority Informed* crowd and the *Majority Uninformed* crowd is $w = 0.13$ and $w = 0.07$, respectively.

First, because "Uninformed", who are also "worse-informed", agents have systematically larger bias in estimation than the "Informed" agents, who are also "better-informed", the average estimate of the entire group is less accurate. Second, it is not always reliable to equate cluster size

with judgment accuracy, because a Decision Maker cannot always be sure whether a crowd is composed of more "better-informed" or "worse-informed." Third, as highlighted in the top row of Figure 8.3, there are quite a few inaccurate "worse-informed" with a WOA of zero. Thus, simply taking the average among participants with a WOA of zero retains the large bias from those "stubborn but inaccurate" worse-informed agents in this context. However, there are many "non-stubborn worse-informed people" with large WOA who have similar estimates as those "stubborn worse-informed agents." Therefore, as shown in the bottom row of Figure 8.3, since KDE is able to classify better-informed and worse-informed into distinct clusters, the cluster consisting of the better-informed has the lower average WOA than the cluster consisting of worse-informed. This also highlights the importance of using the average WOA of clusters as averaging WOA may reduce the noisiness of WOA at the individual level. In addition, this is true regardless of whether better-informed or worse-informed consist of the majority of the crowd. Therefore, taking the mean of the initial estimates or the updated estimates within the identified cluster outperforms the conventional benchmarks (i.e., the mean of the initial or the updated estimates of the entire group) and the meta-prediction-based minimal pivoting method. All differences in performance comparisons are statistically significant at $\alpha = 0.05$. Another interesting observation from the bottom right panel of Figure 8.3 is that KDE identifies two eligible clusters among the better-informed in the *Majority Uninformed* crowd condition. This is likely an artifact of the small sample size, which produces visible discontinuities in the empirical distribution that KDE picks up as separate clusters. The difference in WOA between the two clusters of better-informed is not statistically significant (t -test, $t(16) = -0.458$, $p = 0.65$), but both are significantly different from the WOA of the worse-informed cluster (t -test, $p = 0.017$ and $p = 0.005$, respectively.)

The results also point out a few potential limitations of the algorithm. First, as shown in the top left panel of Figure 8.3, there is one participant with an initial estimate of "400" and a WOA of zero. KDE classifies this outlier as a distinct cluster. Although Step 2 of the algorithm deems this cluster as ineligible for consideration, it is possible that there are several erroneous outliers that form a cluster and collectively update little in some estimation tasks. More broadly speaking, behavioral biases in advice-taking pose a threat to the algorithm's effectiveness. Second, when the number of better-informed is small, KDE might classify the better-informed into different clusters. This would potentially make the better-informed clusters ineligible or subject the average WOA to larger variance due to small number of observations in each cluster. Third, it is also possible that there is no heterogeneity of informativeness in the crowd. Although it is likely that the algorithm does not proceed beyond Step 1 due to the absence of multiple clusters, it is possible that the empirical distribution has discontinuity as identified KDE. Finally, despite aggregating over either the initial estimates or the updated estimates of the identified cluster could outperform the benchmarks, it is

impossible to conclude which one of the two aggregations is superior using this canonical context. Overall, whether these limitations render CWOA ineffective under the pre-registered parameter set-ups needs to be examined in more externally valid contexts.

8.2.4 Additional Analysis

To provide more insight into the manipulation check, collapsing across the crowd conditions, the two types of agents have similar estimates of the number of coins in the jar (two-tailed t -test comparing raw estimates, 129.7 (Uninformed) vs. 146.4 (Informed), $t(335) = -1.42$, $p = 0.155$) and similar level of stated confidence in the accuracy of their estimated number of coins (two-tailed t -test, probability: 48.6% (Uninformed) vs. 46.3% (Informed), $t(357) = 0.88$, $p = 0.38$; Likert: 3.64 (Uninformed) vs. 3.51 (Informed), $t(358) = 0.89$, $p = 0.375$). The underestimation of the correct number of coins possibly explains why the better-informed underestimate the total monetary value in the jar (t -test comparing better-informed agents' estimates of the total monetary value against the truth, \$120.30 (Informed) vs. \$126.49 (truth), $t(179) = -2.24$, $p = 0.027$). All results here and in the main text are robust to log transformation.

An instrumental variable analysis reveals a causal relationship between estimation accuracy and WOA. I regress WOA on log (initial absolute error) with one's randomly assigned condition as the instrumental variable (*Majority Informed*: $b_{unstandardized} = 0.066$, 95% CI = [0.019, 0.113], $p = 0.007$; *Majority Uninformed*: $b_{unstandardized} = 0.066$, 95% CI = [0.022, 0.109], $p = 0.004$). Both results are robust without log transformation. Therefore, exogenously introduced estimation bias indeed affects WOA (i.e., the less informed thus more biased individuals have larger WOA).

9 Additional Analysis on the Empirical Datasets

9.1 Results: WOA-Accuracy Association

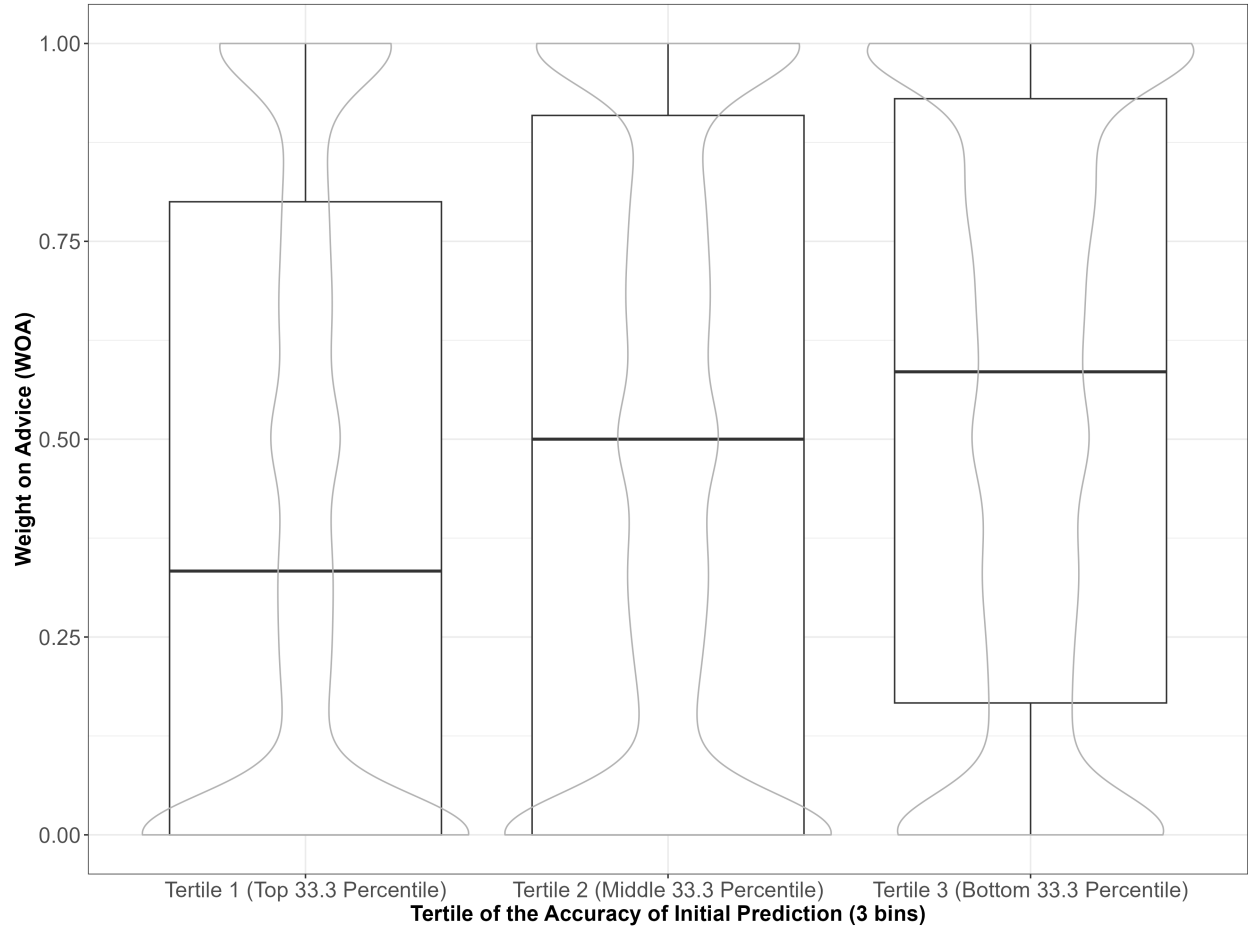


Figure 9.1: The x-axis is the tertile of participant's accuracy. The y-axis is WOA with values greater than 1 winsorized at 1.

The corresponding result in the main text is robust with this alternative binning choice.

9.2 The Error of Initial Group Mean and The Algorithm's Performance

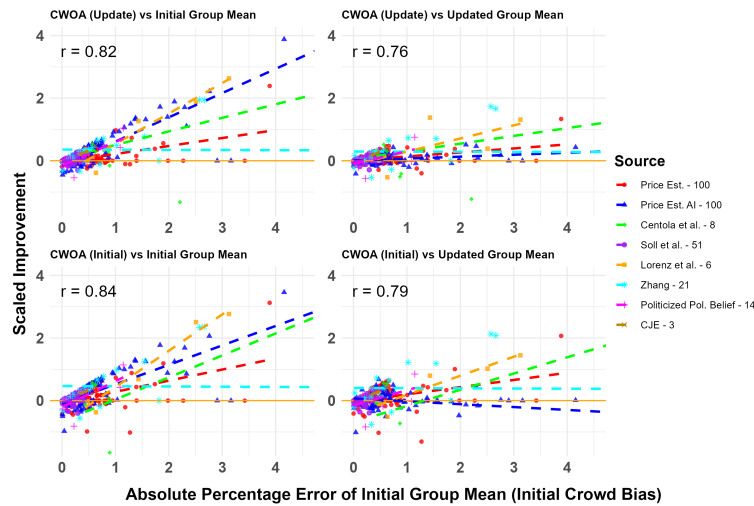


Figure 9.2: The x-axis is the absolute error of initial group mean divided by the truth of each respective question. The y-axis is how much the algorithm's output improves upon conventional wisdom-of-crowds aggregation benchmarks. "Source" refers to the datasets and the number of questions. The colored dashed line is the best linear fitted line within each dataset. The overall correlation pooling all questions in each panel is reported at the top left corner. Five outliers are removed for visualization purposes. The overall correlations remain significantly positive with or without outlier removal.

To facilitate performance comparisons with questions on different scales, the analysis measures a method's performance on a given estimation problem by the absolute error of the method divided by the truth of an estimation problem. Therefore, an absolute error of 1 when the truth is 10 is comparable to an absolute error of 50 when the truth is 500. The improvement is measured by taking the difference between the scaled accuracy of a benchmark with that of the algorithm's output such that a positive (negative) value indicates the algorithm's output is more (less) accurate the benchmark.

Figure 9.2 and results from the controlled experiment support the hypothesis that the algorithm tends to be more effective given a larger initial estimation error of the crowd. As a clarification, the algorithm's improvement is coded as 0 if the algorithm does not identify at least two clusters for a question in Figure 9.2. The correlation becomes even stronger when considering only the 216 out of 303 questions in which the algorithm identified at least two clusters. All results are robust when the y-axis is whether CWOA shows improvement (1 = Yes, 0 = No).

9.2.1 The Error of Initial Group Mean and The Algorithm's Performance: Excluding the "NA" Questions

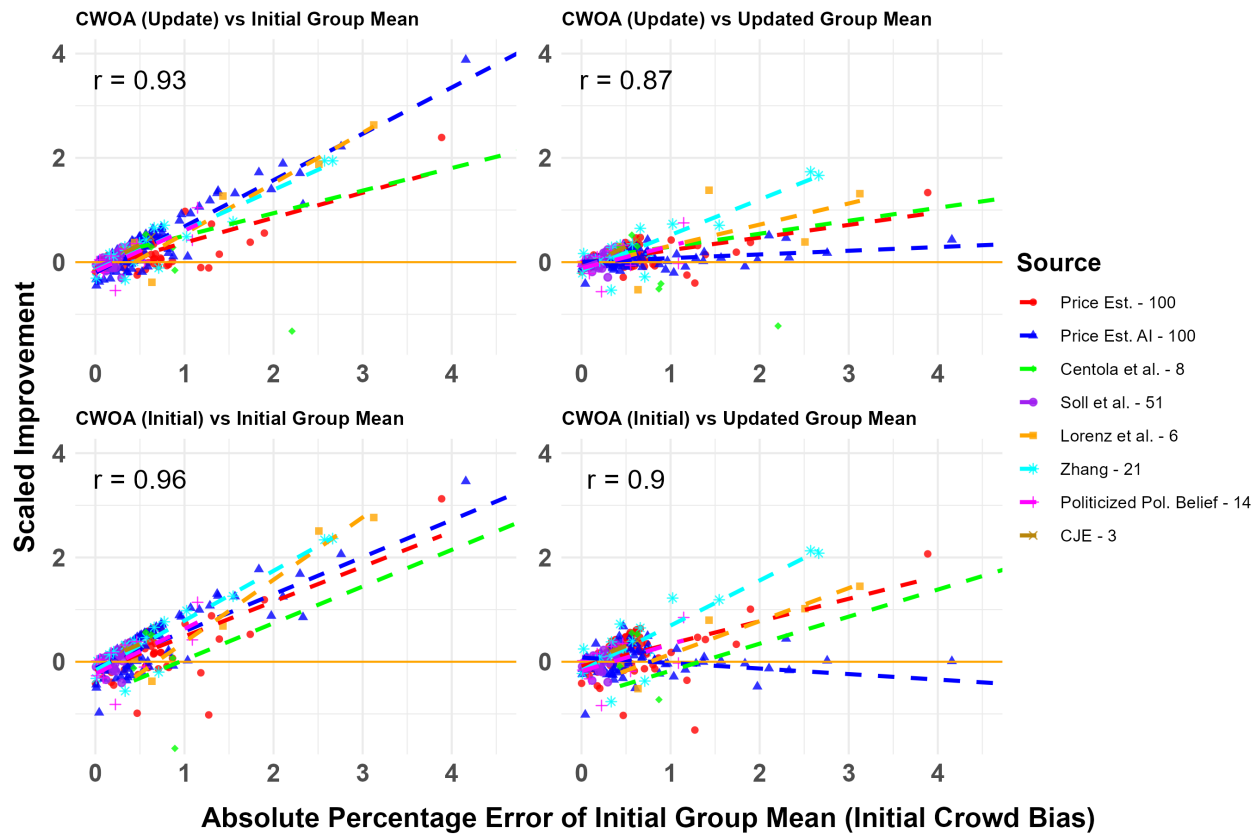


Figure 9.3: The x -axis is the absolute error of initial group mean divided by the truth of each respective question. The y -axis is how much the algorithm's output (the top row: updated cluster mean, the bottom row: initial cluster mean) improves upon conventional wisdom-of-crowds aggregation (the left column: initial group mean, the right column: updated group mean.) "Source" refers to the datasets and the number after hyphen indicates the number of questions in the dataset. The colored dashed line is the best linear fitted line within each dataset. The overall correlation pooling all questions in each panel is reported at the top left corner. Five outliers are removed for visualization purposes. The overall correlations remain significantly positive with or without outlier removal.

In the previous analysis, the algorithm's improvement is coded as 0 if the algorithm does not identify at least two clusters for a question. As shown in Figure 9.3, the correlation becomes even stronger when considering only the 216 out of 303 questions in which the algorithm identified at least two clusters and proceeded beyond Step 2.

Overall, the hypothesized correlation is positive within most datasets. There are two exceptions. In the Zhang (2022) dataset, the correlation is null because there is one question where the group mean is severely inaccurate but the algorithm's improvement is small. Removing this one particular instance makes the correlation significantly positive. In the one hundred price estimation tasks where I used AI's prediction as the numerical advice, the correlation is negative when examining

the improvement of initial cluster mean over the updated group mean.⁶ This is because AI's advice is an external source that happens to be much more accurate. This, in turn, makes the updated group mean much more accurate, and the initial cluster mean does not reliably outperform this particular benchmark. However, the positive correlation is present when examining the improvement of initial cluster mean over initial group mean, which is consistent with the prediction of the model.

⁶The positive correlation is present when examining the improvement of updated cluster mean over the updated group mean: $r = 0.42, p < 0.0001$).

9.2.2 The Error of Initial Group Mean and The Algorithm's Performance: Binary Coding of Performance

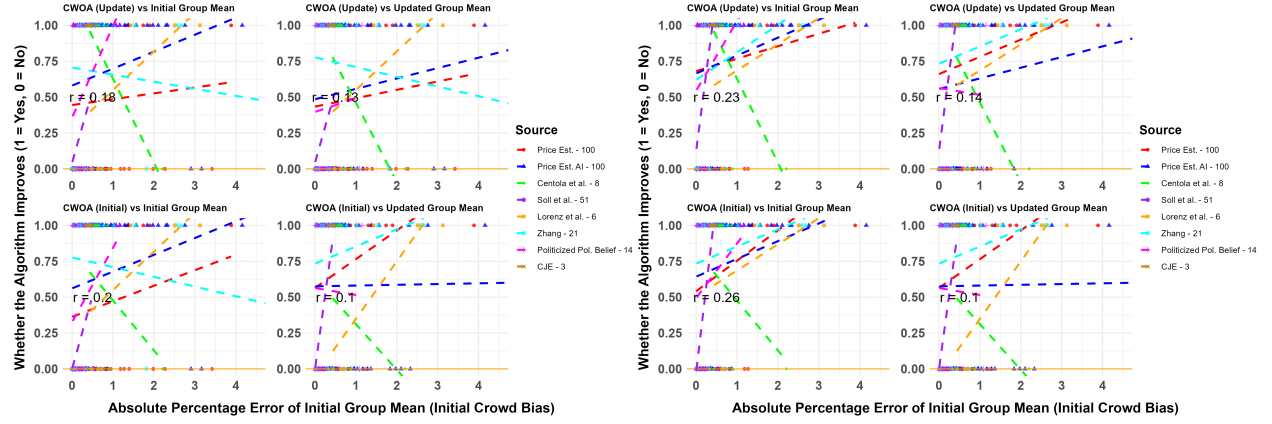


Figure 9.4: The x -axis is the absolute error of initial group mean divided by the truth of each respective question. The y -axis is whether CWOA improves upon a benchmark (1 = Yes, 0 = No). "Source" refers to the datasets and the number after hyphen indicates the number of questions in the dataset. The colored dashed line is the best linear fitted line within each dataset. The left panel includes all 303 questions, and the right panel only considers questions in which CWOA identifies at least two clusters.

The result is robust when CWOA's improvement is coded as a binary outcome (1 = CWOA improves, 0 = otherwise) rather than continuous magnitude. As a clarification, Three outlier questions with scaled initial crowd bias more than 3 SD above the mean are excluded. As an alternative robustness check, the results are robust after a log transformation of the x -axis without outlier removal.

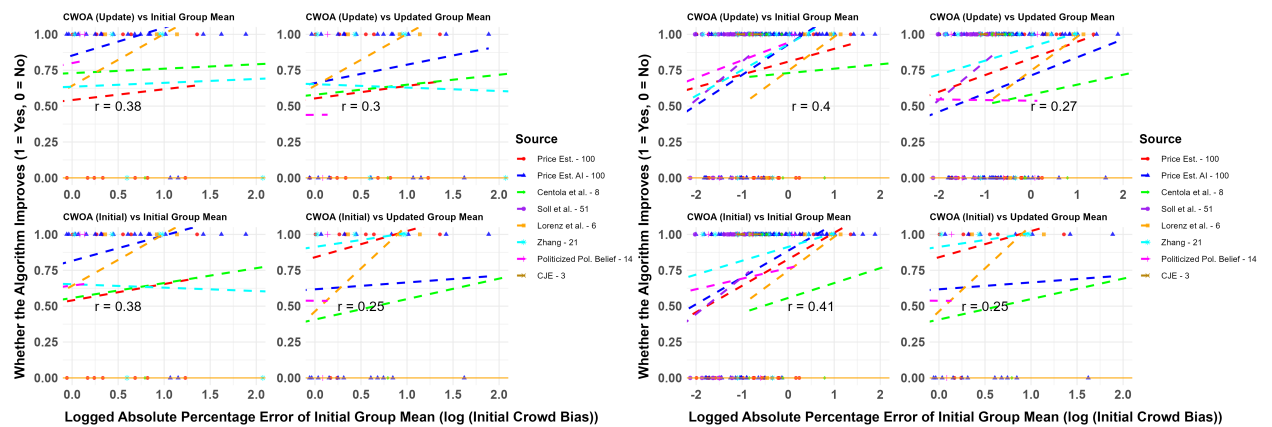


Figure 9.5: The x -axis is the absolute error of initial group mean divided by the truth of each respective question. The y -axis is whether CWOA improves upon a benchmark (1 = Yes, 0 = No). "Source" refers to the datasets and the number after hyphen indicates the number of questions in the dataset. The colored dashed line is the best linear fitted line within each dataset. The left panel includes all 303 questions, and the right panel only considers questions in which CWOA identifies at least two clusters.

9.3 Location of the Identified Cluster and Cluster Location: Suggestive Evidence for Algorithm Benefiting from Bias Reduction

Within each problem where multiple clusters are identified, clusters are labeled in left-to-right order, with the middle cluster receiving the median index. For example, with 5 clusters, the leftmost is Cluster 1, the middle is Cluster 3, and the rightmost is Cluster 5; with 3 clusters, the leftmost is Cluster 1, the middle is Cluster 2, and the rightmost is Cluster 3. The “location extremeness” of a cluster is its absolute distance from the middle cluster index – so the middle cluster has location extremeness 0, and the leftmost/rightmost clusters have the maximum location extremeness. There is a positive correlation between the location extremeness of the identified cluster and the degree of CWOA’s improvement (i.e., scaled measure as introduced in the main text) upon a benchmark. For example, the correlations between how much the initial cluster mean improves upon the initial group mean or the updated group mean are $r = 0.268$ ($p < 0.0001$) and $r = 0.246$ ($p = 0.0003$). The correlations between how much the updated cluster mean improves upon the initial group mean or the updated group mean are $r = 0.260$ ($p = 0.0001$) and $r = 0.224$ ($p = 0.0009$).⁷

Overall, the intuitive findings serve as suggestive evidence that the effectiveness of CWOA comes from bias reduction: if the better-informed cluster is far from the center, it signals that the group mean may be quite inaccurate (e.g., biased). Aggregating over the better-informed cluster therefore achieves greater improvement (e.g., bias reduction) than when the identified cluster is close to the center, whose average is already near the group mean.

9.4 Performance Comparisons using Mean Absolute Percentage Error

⁷The results are qualitatively robust when improvement is coded in a binary fashion (i.e., all correlations are positive but p -values range between 0.098 and 0.419). This is because location extremeness signals the potential gap in estimation bias or potential room for improvement. This granularity is lost when examining improvement in a binary fashion.

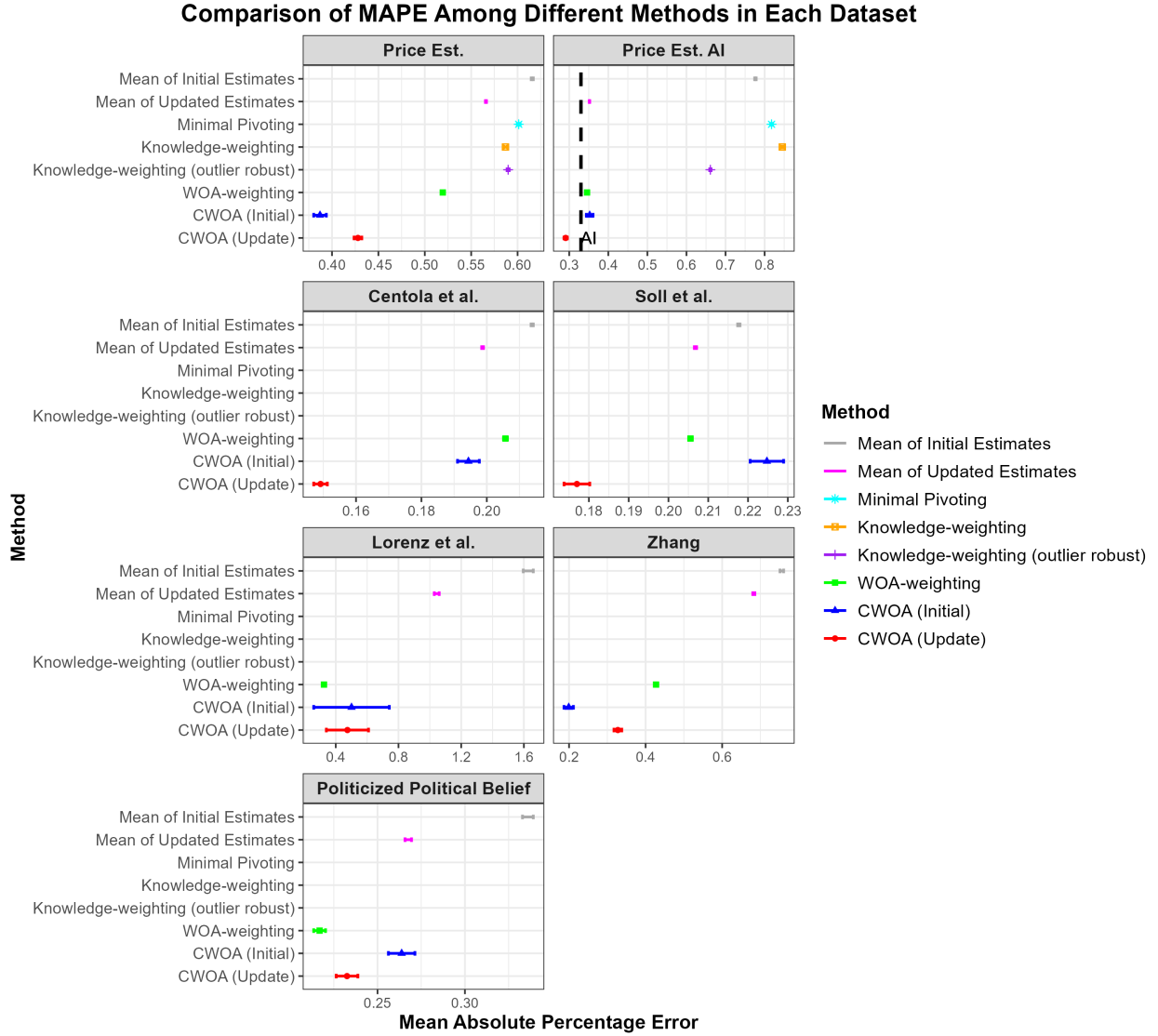


Figure 9.6: The x -axis is each method's mean absolute percentage error (MAPE) across questions in a data source in which CWOA identified at least two clusters. The bars indicate 95% confidence intervals. The y -axis indicates the methods. The analysis does not have performance by meta-prediction-based methods for datasets that do not contain meta-prediction. The black dashed line in the "Price Est. AI" marks the MAPE of AI's predictions.

9.5 The Minimum Cluster Size Threshold (K) and The Algorithm's Performance

Figure 9.7 and Figure 9.8 plot the algorithm's performances, as measured by root mean squared percentage error (RMSPE) and mean absolute percentage error (MAPE) respectively, with varying minimum cluster size K (e.g. $K = 5\%$ means the algorithm only selects from clusters with a number of observations greater than 5% of the total sample size of the entire crowd). The performance remains reasonably stable within a range of values near $K = 5\%$. The general caveat is regarding the

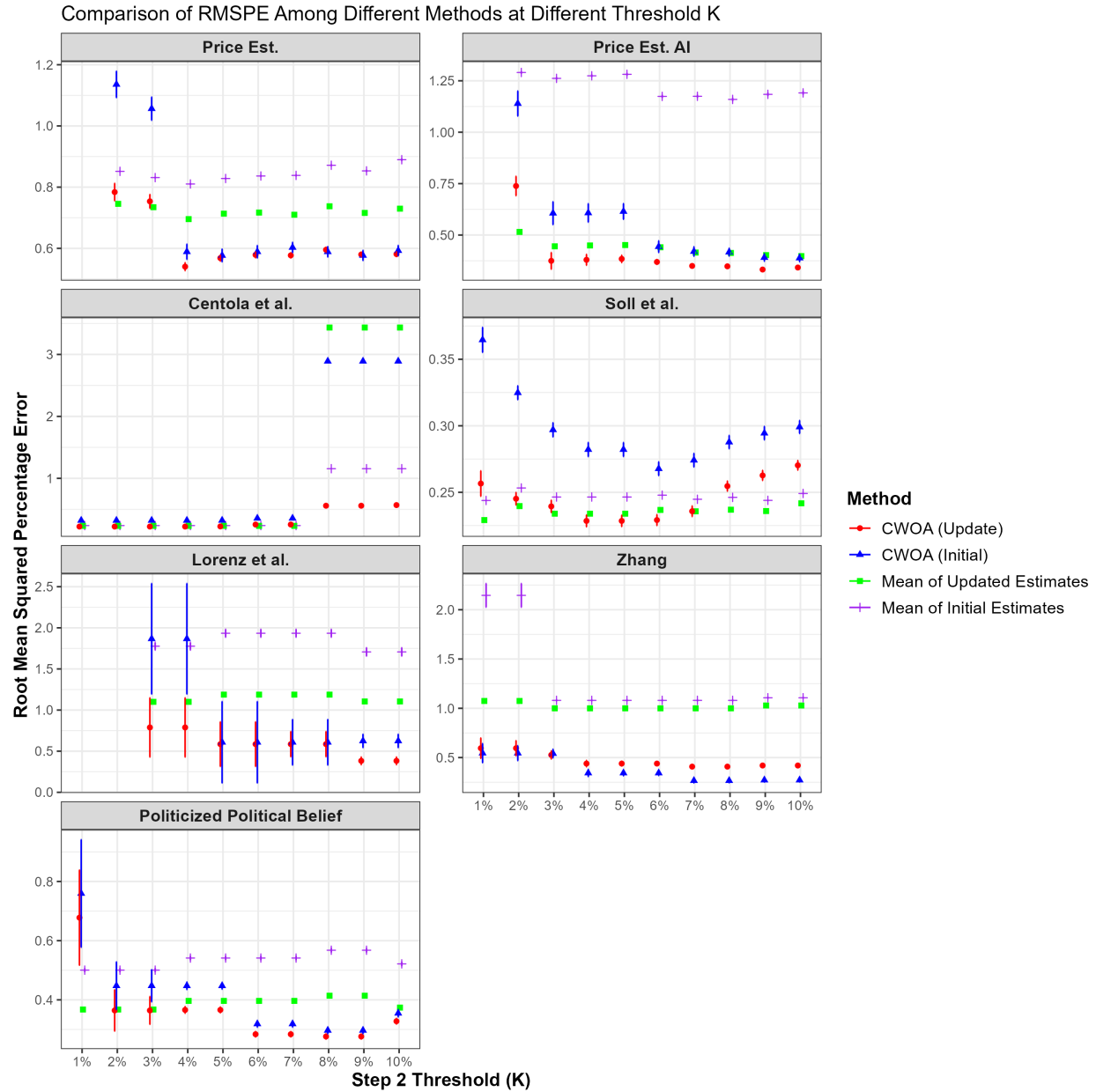


Figure 9.7: The x -axis is the Minimum Cluster Size Threshold (K) – a hyperparameter in Step 2 of the algorithm. The y -axis is each method's root mean squared percentage error (RMSPE) across questions in a data source in which CWOA identified at least two clusters. The bars indicate 95% confidence intervals.

following trade-off: when the threshold K is very small, the algorithm has a chance of identifying better-informed agents if there are very few of them, but the risk is that it may mistake, for example, a "stubborn outlier" as the better-informed cluster. When the threshold K is large, the algorithm is less subject to distinct clusters with few observations which happen to consist mostly of stubborn agents, but the risk is that it accidentally exclude the true better-informed cluster if the number of better-informed agents is not greater than 5% of the total sample size. The choice of $K = 5\%$

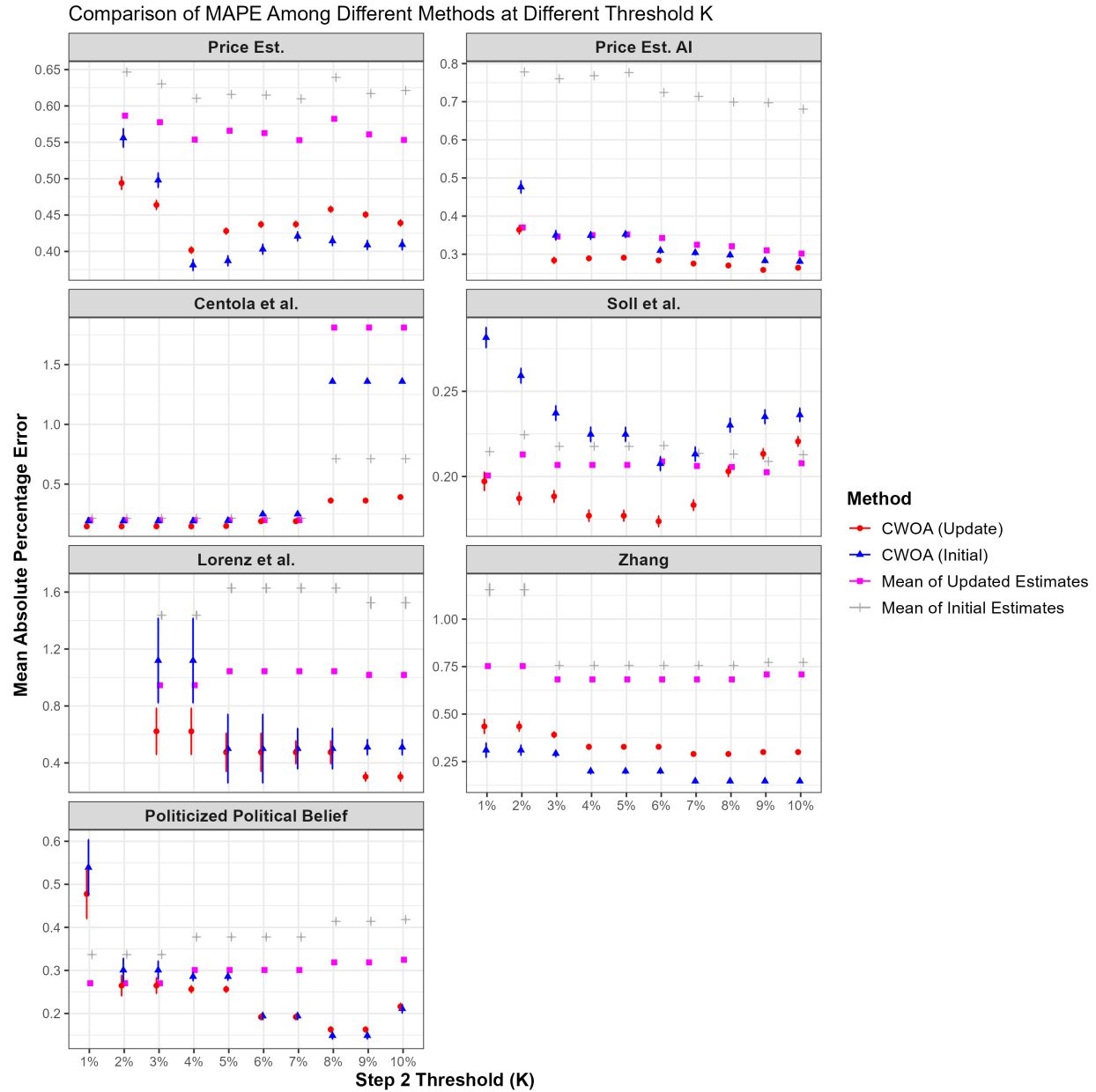


Figure 9.8: The x -axis is each method's mean absolute percentage error across questions in a data source in which CWOA identified at least two clusters. The bars indicate 95% confidence intervals. The y -axis indicates the methods. The analysis does not have performance by meta-prediction-based methods for datasets that do not contain meta-prediction.

aims to balance this trade-off. However, if the Decision Maker has additional knowledge about the question domain, they may fine-tune the 5% to achieve better performance. The next subsection illustrates this idea.

9.6 Varying K on Politicized Political Factual Belief Questions

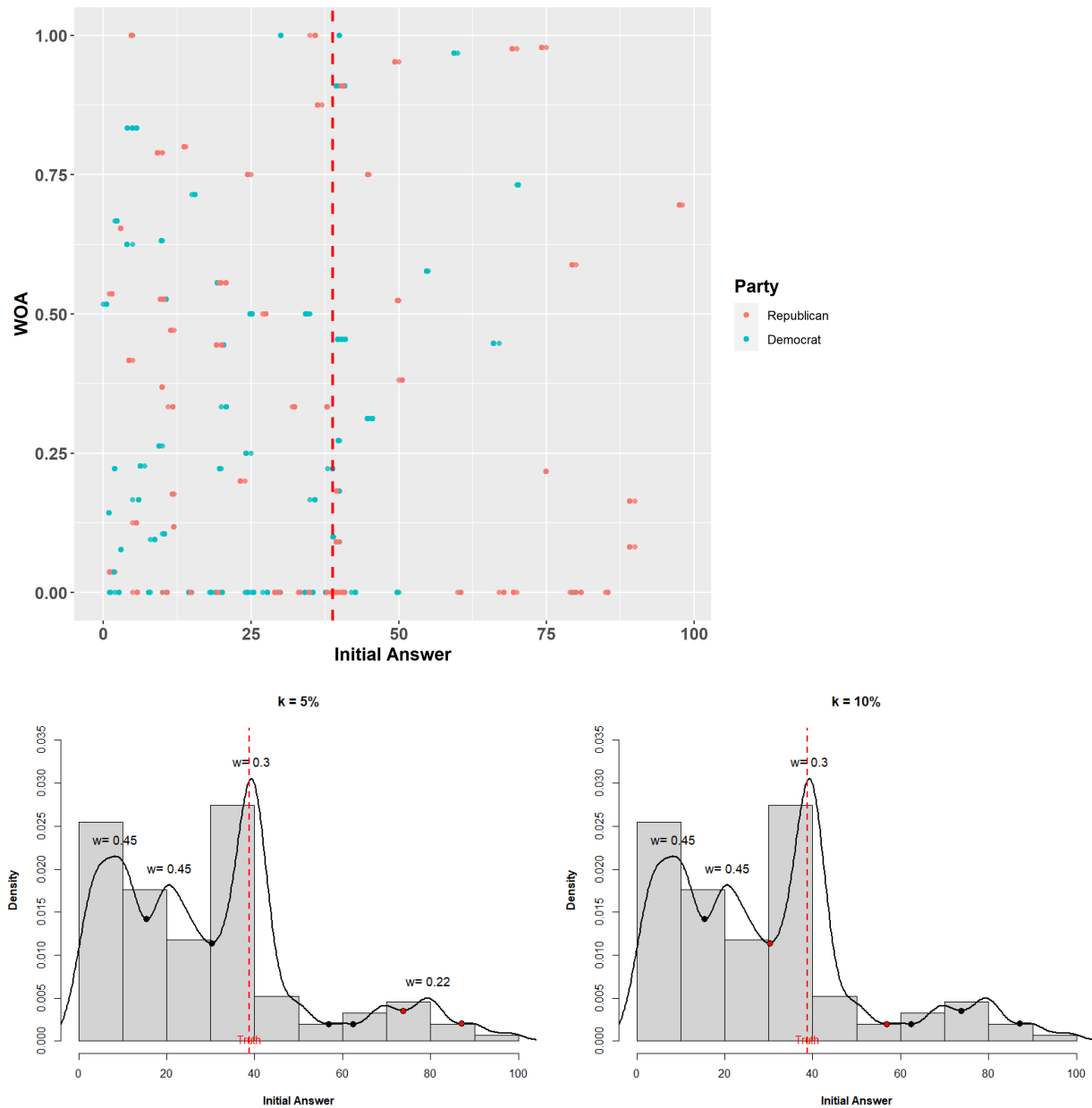


Figure 9.9: The top panel plots the WOA against the corresponding initial estimates. The bottom panels (left, $K = 5\%$; right, $K = 10\%$) plot the clusters emerging from the initial estimates. The cluster with the lowest WOA in each scenario of K lies in between the two cutoff points highlighted in red. The average WOA of clusters that do not meet the minimum cluster size is not reflected in the figure.

The politicized political factual belief questions constitute a context in which there are more likely to exist clusters filled with stubborn agents. For example, participants hold different views on tax policies might be inclined to respond to the question "What percentage of federal income tax revenue is derived from the top 1 percent of earners?" differently and they may refuse to update if the

numerical advice is not an answer that fits their policy views. The upper panel of Figure 9.9 displays the initial answers of Republicans and Democrats to the tax question, as well as their WOA. Evidently, Republicans tend to overestimate the truth. The lower left panel shows that the cluster with the smallest average WOA contains responses ranging from 75 (%) to 85(%). However, this cluster only contains 8 observations, passing the minimum cluster threshold at 5% of the subjects (i.e., there are a total of 153 participants). If K is set to be a higher value (e.g., 6%), as illustrated in the lower right panel, CWOA correctly identifies the better-informed cluster that centers around the truth. Consequently, CWOA's outputs become more accurate.

As showcased in Figure 9.9, persistent Republicans and Democrats respectively overestimate and underestimate the truth to varying degrees. For instance, stubborn Republicans form another minor cluster with responses between 63(%) and 74(%). This is possibly due to different Republicans might exhibit varying degrees of party favoritism, which leads to different and inaccurate initial estimates. Thus, even a modest increase in the required cluster size proves effective in excluding the emergence of stubborn yet inaccurate worse-informed clusters.

9.7 Binary Performance Comparison

This section documents the proportion of questions in which the updated cluster mean (benchmark) outperforms the alternative methods.

9.7.1 Comparison Pooling All Data Sources

Table 3: Binary Comparison Between Updated Cluster Mean (UCM) and Alternative Methods (pooled across all 303 questions). "Questions" is the total number of questions in which both the Updated Cluster Mean and the respective alternative method are available for comparison. "UCM Wins" is the number of questions in which the updated cluster mean is more accurate. "Proportion UCM Wins" is "UCM Wins" divided by "Questions".

Updated Cluster Mean (UCM) vs Method	Questions	UCM Wins	Proportion UCM Wins
Mean of Initial Estimates	219	162	0.740
Mean of Updated Estimates	219	148	0.676
Knowledge-weighting (outlier robust)	155	117	0.755
Minimal Pivoting	155	112	0.723
Knowledge-weighting	155	110	0.710
WOA-weighting	219	137	0.626
Initial Cluster Mean	219	105	0.479

As shown in Table 3, the mean of the updated estimates from the cluster with the lowest average WOA outperforms the meta-prediction-based and conventional methods in 68% to 76% of the problems. Together with the results featuring mean absolute percentage error (MAPE) (SI Section 9.4) and root mean squared percentage error (RMSPE) (main text) norm as the performance metric, the results confirm that CWOA is more likely to produce more accurate outputs. Nevertheless, consistent with the findings from the literature (discussed in SI Section 4 and showcased in simulations), it is hard to conclude *ex ante* whether the updated cluster mean or the initial cluster mean is superior. Although binary comparisons show similar win rates for the updated cluster mean and the initial cluster mean – consistent with mixed findings in the prior literature that the updated cluster mean is more reliable on average.

9.7.2 Separate Comparisons In Each Data Source

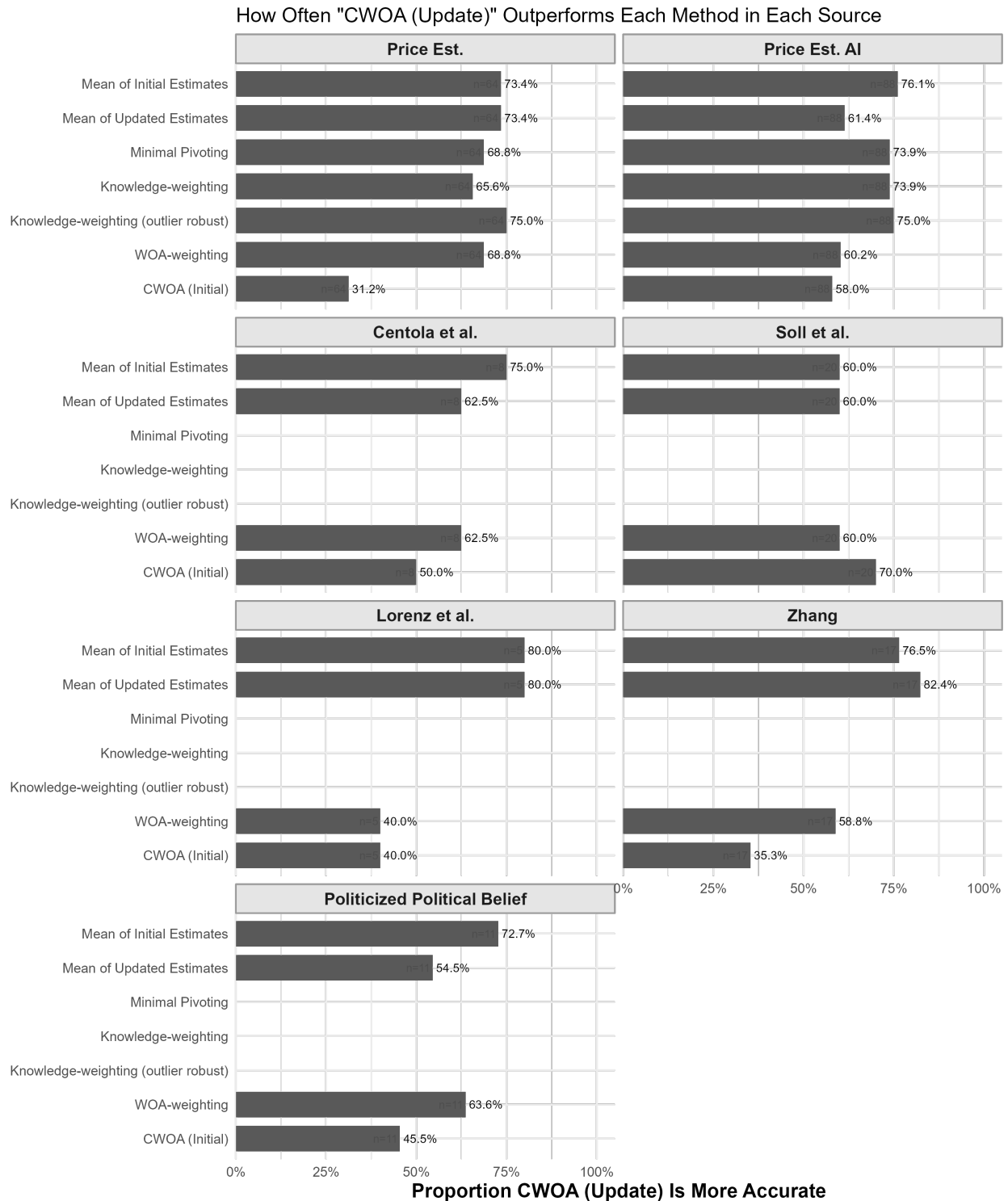


Figure 9.10: The x-axis is the proportion of questions in which the "Updated Cluster Mean" outperforms a benchmark in a data source. The y-axis indicates the methods. The analysis does not have performance by meta-prediction-based methods for datasets that do not contain meta-prediction.

10 Study Materials: Zhang 2022

Because Zhang (2022) is unpublished, I hereby report the detailed materials and methods. The procedures were pre-registered at https://aspredicted.org/FLV_XV1

10.1 Participants.

The subjects were recruited from Prolific. The number of subjects who started the survey was 127, of whom 23 were screened out because they failed a simple arithmetic check or an attention check before starting to answer 21 questions. The 104 remaining subjects ($M_{age} = 39.2$, 70.2% female) provided complete data and were included in the analysis. The median time to complete the study was 22.1 minutes and the average total compensation including bonuses was \$2.95.

10.2 Procedures.

After signing a consent form and passing a simple arithmetic check, subjects were given task instructions and the quadratic scoring rules for determining the bonus payment. After passing an attention check⁸, they were given another quick reminder about the logistics of the tasks before answering the 20 questions in a random order. For each question, subjects provided their initial point estimate, their 90% confidence interval, and the probability that their point estimate was within 10% of the truth. After subjects proceeded to the next page, they were given the advice and were asked to provide their point estimate again. All advice was framed as “*for the previous question, the average answer from ten other participants is about X.*” The advice (group mean) of each question was obtained from a pilot study. Subjects answered the second attention check question⁹ after finishing the 20 questions and then they were asked to estimate the number of dots in a picture. Finally, before finishing the study, they answered some Cognitive Reflection Test questions and some demographic questions.

As a clarification, participants were explicitly permitted to search for answers online in this study. The author used this feature as a way to introduce heterogeneity of informativeness among agents. This permission was unique to this study.

⁸Subjects who failed the arithmetic check or the attention check were not allowed to proceed

⁹No one failed the second attention check question.

10.3 Questions

The following table summarizes the trivia questions and the advice used for this study. In the table, I list the questions under “Question.” Each question’s correct answer is under “Truth.” The advice, or the group mean of each question obtained from a pilot study, is under “Advice.”

Table 4: Questions and Advice

	Question	Truth	Advice
1	What is the median purchase price of a single-family house in San Francisco, CA, in 2018?	1610000	614100
2	What is the price of this Arc'teryx Cerium LT Hoody on Arc'teryx's official website (pre-tax, without promotion or shipping) as of November 2020?	379.00	415
3	What is the price of this Swatch Chequered White watch on Swatch's official US website (pre-tax, without promotion or shipping) as of November 2020?	140.00	582
4	What is the price of this Omega Seamaster Diver 300mco-axial Master Chronometer 42mm watch on Omega's official US website (pre-tax, without promotion or shipping) as of November 2020?	5200.00	2910
5	What is the price of this pair of Nike Blazer Mid '77 Vintage on Nike's official US website (pre-tax, without promotion or shipping) as of November 2019?	100.00	340
6	What is the monthly standard plan subscription fee for Netflix in the U.S. (pre-tax, without promotion) as of November 2020?	13.99	10
7	What is the price of this Martin OMC-X1E acoustic-electric guitar on Martin's official US website (pre-tax, without promotion or shipping) as of November 2020?	549.00	661
8	What is the price of this Fender Vintera 60s Stratocaster electric guitar on Fender's official US website (pre-tax, without promotion or shipping) as of December 2021?	1099	1446
9	What is the price of this Bose Companion 2 Series III multimedia speaker on Bose's official website (pre-tax, without promotion or shipping) as of November 2021?	149.00	483
10	What is the price of this Bose QuietComfort 20 Acoustic Noise Cancelling headphones from Amazon.com (pre-tax, without promotion or shipping) as of November 2021?	249.95	289
11	What is the price of one Wilson NBA Official Game Ball on Wilson's official website (pre-tax, without promotion or shipping) as of December 2021?	199	324
Continued on the next page			

Table 4 – continued from previous page

	Question	Truth	Advice
12	What is the price of this Wilson Blade SWI02 Autograph Tennis Racquet on Wilson's official website (pre-tax, without promotion or shipping) as of November 2020?	249.95	285
13	What is the price of this Omega Constellation Sedna Gold Quartz 28MM watch on Omega's official website (pre-tax, without promotion or shipping) as of November 2020?	5500	2970.00
14	What is the price per share of the most expensive publicly traded stock as of December 2021?	450662	185460
15	According to official data, what is the annual inflation rate in the United States in 2020? (don't include % sign, if you think the inflation is X%, write down X)	1.36	7.4
16	According to official data, what is the unemployment rate in Mexico in 2020? (do not include % sign, if you think the unemployment rate is X%, write down X)	3.65	16.1
17	How many full-time employees does Facebook (the social network company) have globally as of 2020?	58604.00	172708
18	What is the median home sales price in Chicago Metro Area in December 2020?	259000	427000
19	What is the population density (per square mile) of Illinois as of 2018? (population density is the average number of people living in one square mile, e.g., the per square mile population density of the United States is 94.)	229.5	285
20	What is the world record auction price in 2017 (hammer price + fees) of the painting "Salvator Mundi" by Leonardo Da Vinci? Note that the unit of answer is in millions. If you think the price is X million dollars, write down X.	450.30	92
21	How many dots are there in the picture?	144	144

For the product price questions, subjects were also presented with a picture of the product taken from the product's official website.

11 Study Materials: Politicized Political Factual Belief Questions

The procedures were pre-registered at https://aspredicted.org/F9J_Z4P

11.1 Participants

The subjects were recruited from CloudResearch (Turkprime) for a different project in 2022. Turkprime first applied an initial screening, restricting eligibility to Republicans with a moderately right, right, or extremely right political leaning, and to Democrats with a moderately left, left, or extremely left political leaning. In addition, after subjects began, the survey further screened out subjects based on their self-reported political identity.¹⁰ Of 187 subjects who started the survey, 1 was screened out for failing an early attention check, and 33 were screened out because their self-reported political stance did not match the eligibility criteria. The remaining 153 subjects (77 Democrats with left-leaning views, 76 Republicans with right-leaning views; $M_{age} = 44$, 49.7% female) provided complete data and were included in the analysis.

11.2 Procedures

After subjects signed a consent form, they answered demographics questions. I screened out subjects based on their political party and political stance. The remaining subjects were given task instructions and the scoring rule¹¹. Then they answered the 14 questions in random order. For each question, subjects first provided their initial point estimate. After subjects proceeded to the next page, they were given the advice and were asked to provide their point estimate again. All advice was framed as “for the previous question, the average estimate from ten other participants is about X .”¹² The advice (group mean) of each question was obtained from a pilot study.

11.3 14 Political Factual Belief Questions

The following table summarizes the trivia questions and the advice used. In the table, I list the questions under “Question.” Each question’s correct answer is under “Truth.” The advice, or the group mean of each question obtained from a pilot study, is under “Advice.”

¹⁰The two questions are: (1) what is your political position: Republican, Democrat, or other? (2) please indicate your political leaning: Extremely Left, Left, Moderately Left, Neutral, Moderately Right, Right, Extremely Right?

¹¹Initial and updated responses were scored using a quadratic scoring rule, with a maximum bonus of \$3 for accurate responses.

¹²Here, X was rounded to the nearest integer and subjects did not know whether the ten other participants were Republicans or Democrats.

Table 5: Political Factual Belief: Questions and Advice

	Question	Truth	Advice
1	What is the unemployment rate of the U.S. in December 2020 (at the end of Donald Trump's administration)? ?	6.7%	11%
2	What is the unemployment rate of the U.S. in August 2022 (current Joe Biden's administration)?	3.7%	9%
3	In 2018, what percentage of those living in the United States are born outside of the U.S.?	13.7%	18%
4	According PEW Research Center, what share of gun-related deaths are suicides in the U.S. in 2020?	54%	36%
5	What is the percentage of federal income tax revenue that comes from the top 1 percent of earners?	38.8%	29%
6	What is the annual inflation rate in the United States in 2021?	4.7%	8%
7	What share of people in Texas are fully vaccinated (not including booster shots) as of August 31, 2022?	61.6%	57%
8	What share of people in California are fully vaccinated (not including booster shots) as of August 31, 2022?	73.5%	68%
9	How old is Former President Donald Trump as of September 2022?	76	71
10	How old is Vice President Kamala Harris as of September 2022?	57	55
11	According to a national survey conducted by PEW Research Center between Sept. 24-Oct. 7 in 2018, what share of respondents who are Republican or lean-Republican support the notion that "the government should create a federal database to track all gun sales"?	58%	42%
12	According to a national survey conducted by PEW Research Center between Sept. 24-Oct. 7 in 2018, what share of respondents who are Democrat or lean-Democrat support the notion that "the government should create a federal database to track all gun sales"?	88%	72%
13	According to a national survey conducted by PEW Research Center between Sept. 24-Oct. 7 in 2018, what share of respondents who are Democrat or lean-Democrat support the notion that "the government should allow teachers and officials to carry guns in K-12 schools"?	22%	24%
14	According to a national survey conducted by PEW Research Center between Sept. 24-Oct. 7 in 2018, what share of respondents who are Republican or lean-Republican support the notion that "the government should allow teachers and officials to carry guns in K-12 schools"?	69%	58%

12 Dealing with Negative WOA: Absolute Value versus Winsorization at 0

Denote X as the initial estimate, Y as the numerical advice, and Z as the updated estimate. The standard WOA formula in research on advice-taking is $(Z - X)/(Y - X)$, with negative values winsorized at 0 (Bonaccio and Dalal 2006). This section compares the performance of CWOA under this winsorized formulation versus the absolute-value formulation $|Z - X|/|Y - X|$ proposed in the main text.

Empirically, the absolute-value formulation slightly outperforms winsorization at 0 (Table 6). The intuition is as follows. Negative raw WOA arises when an agent updates in the opposite direction of the advice — for example, due to advice aversion or correlated errors among the crowd. Such updating reflects the agent’s behavior toward the advice but does not signal accuracy. Yet winsorizing negative WOA at 0 would treat these contrarian updaters indistinguishably from agents who do not update at all ($\text{WOA} = 0$), artificially lowering the cluster’s average WOA. This can erode the WOA gap between better-informed and worse-informed clusters, weakening CWOA’s signal. The absolute-value formulation mitigates this problem by treating updating-away from the advice as a positive WOA contribution – capturing the underlying behavior more faithfully.

Study	Total # of Questions	Equal	Better	Worse
Price Est.	100	93	4	3
Price Est. AI	100	98	1	1
Soll et al.	51	47	2	2
Centola et al	8	5	2	1
Lorenz et al.	6	4	2	0
Zhang	21	21	0	0
Political	14	13	1	0

Table 6: The "Total # of Questions" is the number of questions in that study. "Equal" / "Better" / "Worse" refer to the number of questions among which using absolute WOA yields the "same" / "better" / "worse" performance compared to using winsorized WOA. Questions wherein only one cluster is identified are coded as "Equal."

13 Robustness Checks: Alternative Ways of Leveraging WOA

13.1 Removing Estimates with $WOA = 0$ When Computing Average Cluster WOA

I conduct a robustness check of CWOA by excluding individuals with $WOA = 0$ when computing the average WOA within each cluster. Specifically, after performing Kernel Density Estimation (KDE) and identifying clusters, I recalculate the average WOA for each cluster using only participants with $WOA > 0$. The cluster with the lowest average WOA is then selected. Importantly, this exclusion affects only the calculation of the average WOA within clusters – it does not influence the KDE procedure, the computation of the mean updated estimate for the identified cluster (since all estimates, including those with $WOA = 0$, remain in that step), or the outcomes of other aggregation methods.

As shown in Figures 13.1 and 13.2, the method that takes the mean of updated estimates from the cluster with the lowest average WOA (CWOA (Update)) (mostly) remains the best performing method, even when individuals with $WOA = 0$ are excluded from the average cluster WOA calculation. This suggests that the lower WOA of the more accurate cluster is not driven merely by the presence of individuals who exhibit zero updating. Rather, it indicates that people's belief updating captures meaningful variation in the granularity of their accuracy or informativeness. However, excluding $WOA = 0$ when computing average cluster WOA tends to make the algorithm perform worse compared to the original method (CWOA (Update) (Original)). It suggests that (1) those with $WOA = 0$ still contribute meaningfully to the algorithm, and (2) the algorithm (or the average cluster WOA) still benefits from a larger sample to help decrease individual-level noisiness of WOA.

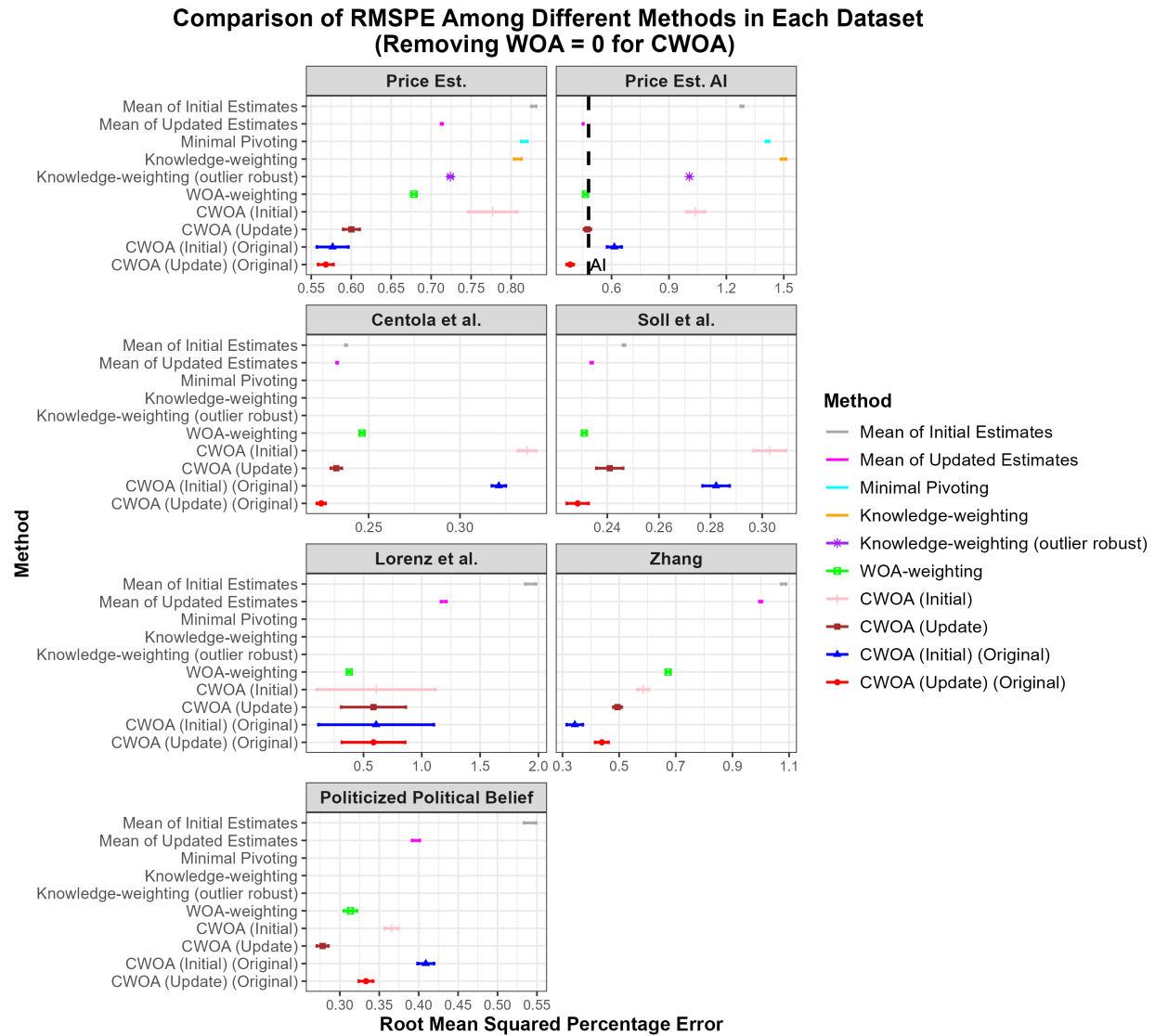


Figure 13.1: The x -axis is each method's root mean squared percentage error across questions in a data source in which CWOA identified at least two clusters. The bars indicate 95% confidence intervals. The y -axis indicates the methods. The analysis does not have performance by meta-prediction-based methods for datasets that do not contain meta-prediction. The black dashed line in the "Price Est. AI" marks the RMSPE of AI's predictions.

(Removing WOA = 0 for CWOA)



Figure 13.2: The x -axis is each method's mean absolute percentage error across questions in a data source in which CWOA identified at least two clusters. The bars indicate 95% confidence intervals. The y -axis indicates the methods. The analysis does not have performance by meta-prediction-based methods for datasets that do not contain meta-prediction. The black dashed line in the "Price Est. AI" marks the MAPE of AI's predictions.

13.2 Using the Proportion of $WOA = 0$ Within Each Cluster as a Signal of Accuracy

Instead of comparing the average WOA of clusters, I conduct a robustness check by comparing the proportion of people who updated (i.e., $WOA \neq 0$) and aggregating over the cluster with the lowest proportion of "updaters" – equivalently, the cluster that contains the most people with $WOA = 0$.

As shown in Figures 13.3 and 13.4, the method that takes the mean of updated estimates from the cluster with the lowest proportion of "Non-Zero WOA" (CWOA (Update)) remains mostly effective but does not consistently deliver the best performance. The approach of "binarizing" WOA to zero versus non-zero loses information conveyed by WOA, which perhaps leads to the inconsistency in performance. Overall, the results suggest that people's belief updating captures meaningful variation in their accuracy and informativeness.

Comparison of RMSPE Among Different Methods in Each Dataset (Lowest Proportion of Non-Zero WOA for CWOA Algorithm)

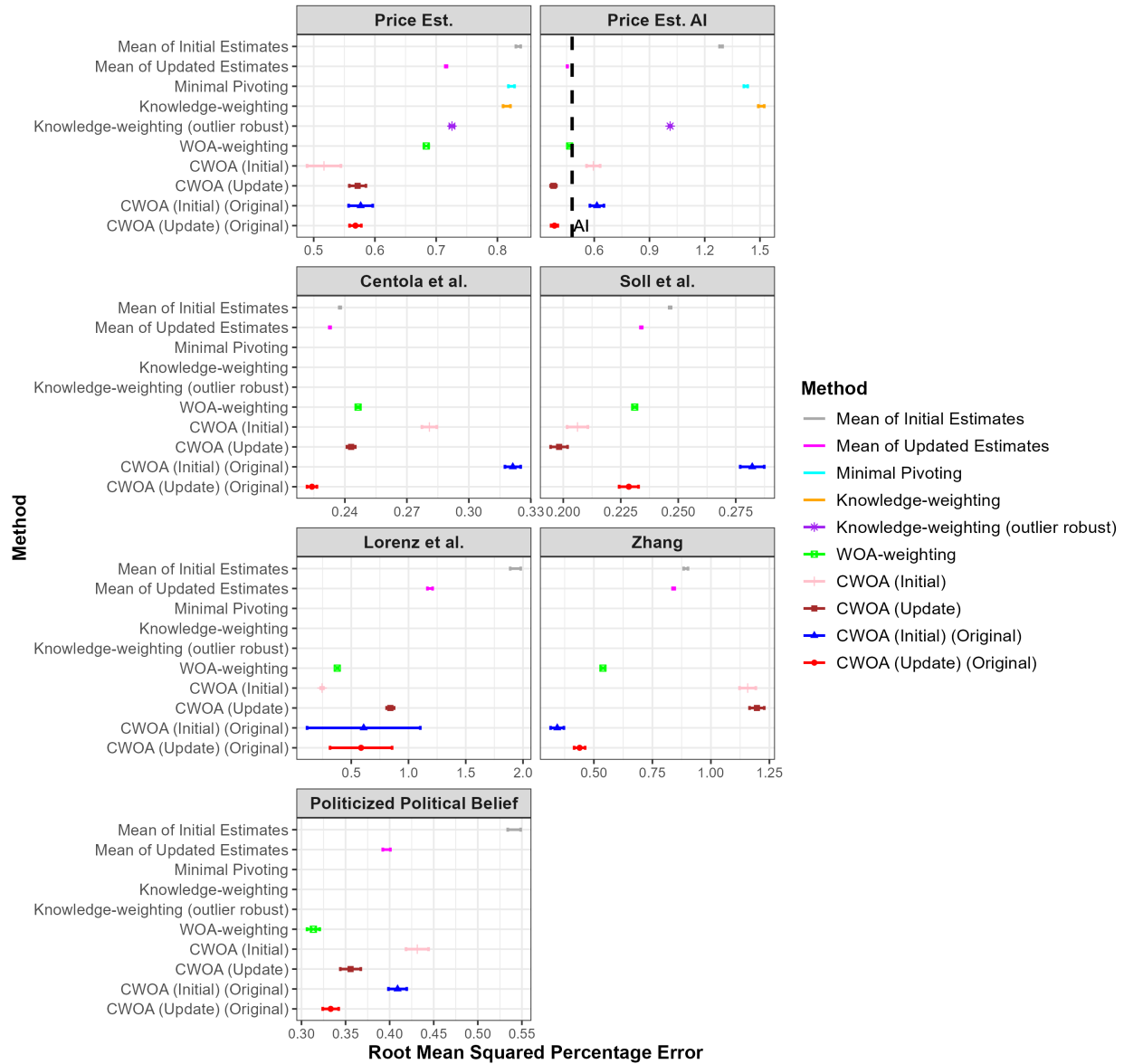


Figure 13.3: The x -axis is each method's root mean squared percentage error across questions in a data source in which CWOA identified at least two clusters. The bars indicate 95% confidence intervals. The y -axis indicates the methods. The analysis does not have performance by meta-prediction-based methods for datasets that do not contain meta-prediction. The black dashed line in the "Price Est. AI" marks the RMSPE of AI's predictions.

**Comparison of MAPE Among Different Methods in Each Dataset
(Lowest Proportion of Non-Zero WOA for CWOA Algorithm)**

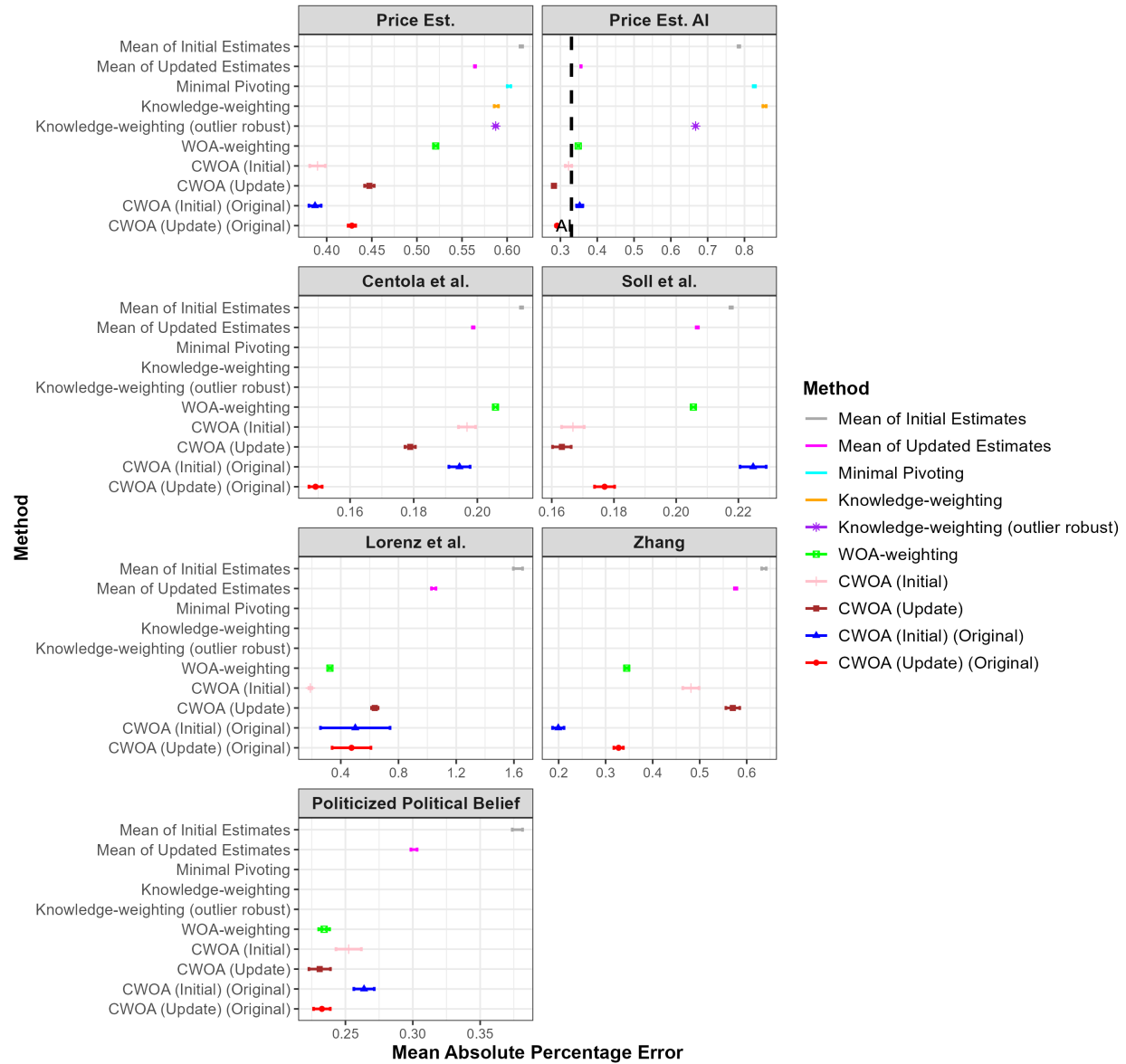


Figure 13.4: The x-axis is each method's mean absolute percentage error across questions in a data source in which CWOA identified at least two clusters. The bars indicate 95% confidence intervals. The y-axis indicates the methods. The analysis does not have performance by meta-prediction-based methods for datasets that do not contain meta-prediction. The black dashed line in the "Price Est. AI" marks the MAPE of AI's predictions.

14 Exploratory Analysis: Including the One-Cluster Cases in Performance Comparison

The main text’s performance analysis focuses on questions where CWOA applies (i.e., KDE identifies at least two clusters). This section examines performance across all questions — including those where CWOA does not apply (single-cluster cases) – by substituting alternative aggregation methods (Knowledge-weighted, mean of updated estimates, mean of initial estimates) in place of CWOA for the single-cluster cases. This allows comparison of overall performance across all 300 questions.¹³

14.1 Eliciting Meta-prediction and Applying the Knowledge-weighting Method Among Questions with One Cluster

CWOA does not apply in problems in which the initial estimates only form one single cluster. However, the algorithm’s two-shot procedure defers further input collection until KDE confirms the existence of multiple clusters. Therefore, a Decision Maker can immediately opt to elicit meta-predictions when there is only a single cluster. Although determining the optimal alternative method for single-cluster cases is beyond this paper’s scope, this analysis applies the outlier-robust knowledge-weighted aggregation method to single-cluster questions, since prior work suggests it is the most effective meta-prediction-based method in single-prediction-problem contexts (Palley and Satopää 2023). As shown in Figure 14.1 and Figure 14.2, the combined approach (i.e., applying the Knowledge-weighted method in questions where KDE only identifies one cluster, and applying CWOA in questions where KDE identifies multiple clusters) results in the best overall performance in the Price Estimation and Price Estimation with AI Advice dataset.

Note: The analysis here includes all 100 estimation problems from the two price estimation datasets. The other datasets do not contain data on meta-prediction.

¹³CWOA applies in the Coin Jar Experiments.

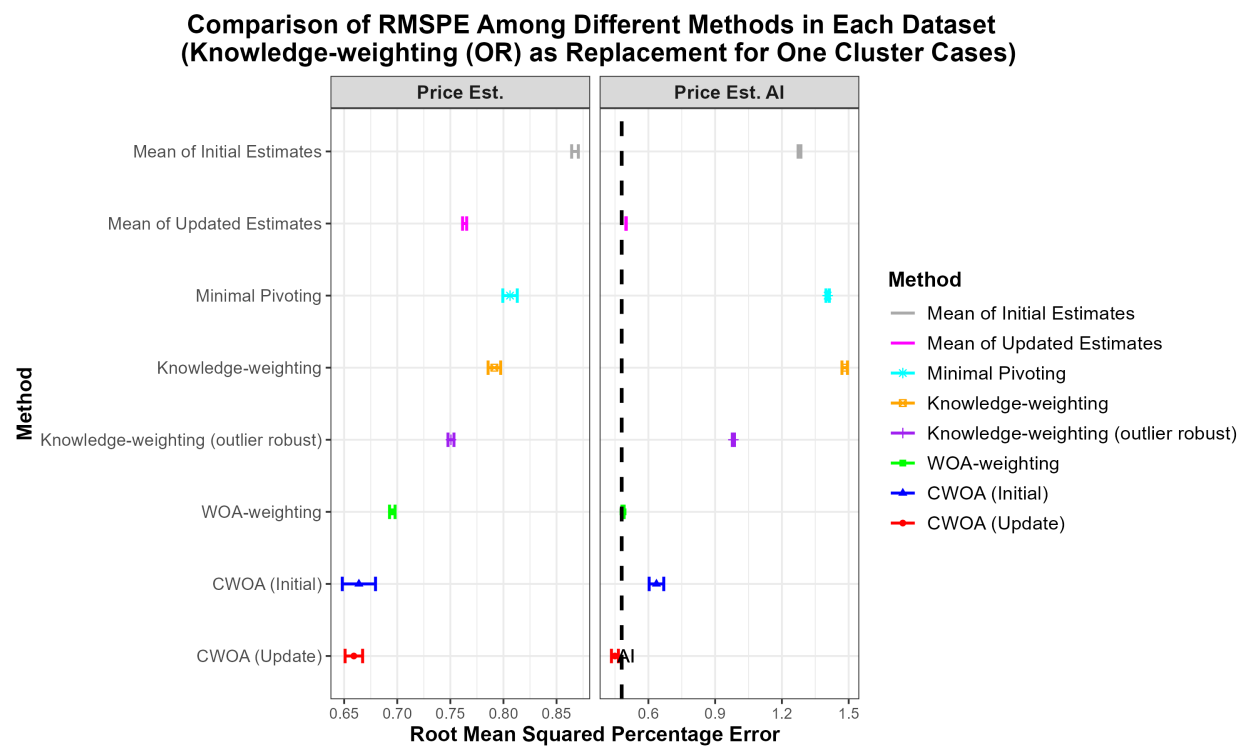


Figure 14.1: The x -axis is each method's root mean squared percentage error across questions in a data source. The bars indicate 95% confidence intervals. The y -axis indicates the methods. The analysis does not have performance by meta-prediction-based methods for datasets that do not contain meta-prediction. The black dashed line in the "Price Est. AI" marks the RMSPE of AI's predictions.

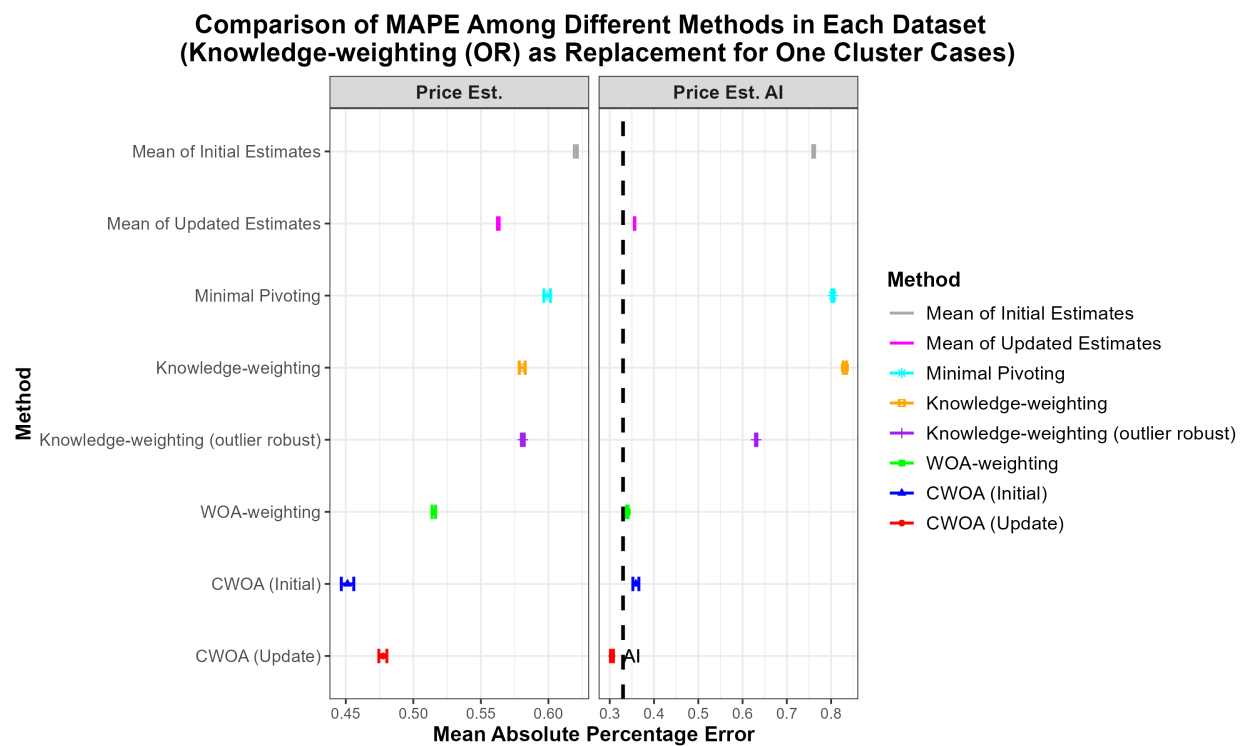


Figure 14.2: The x -axis is each method's mean absolute percentage error across questions in a data source. The bars indicate 95% confidence intervals. The y -axis indicates the methods. The analysis does not have performance by meta-prediction-based methods for datasets that do not contain meta-prediction. The black dashed line in the "Price Est. AI" marks the MAPE of AI's predictions.

14.2 Using "Mean of Updated Estimates" as Replacement for Questions with One Cluster

Past literature has also shown that taking the mean of the updated estimates may improve the Wisdom of Crowds. Suppose the Decision Maker does not wish to elicit meta-predictions but is willing to elicit updated estimates after advice. In single-cluster cases, CWOA can be replaced by the mean of updated estimates from the entire crowd, while CWOA's standard procedure is applied in multi-cluster cases. As shown in Figures 14.3 and 14.4, CWOA remains the best-performing method.

Note: The medical dataset is not shown because CWOA applies in all eight problems.

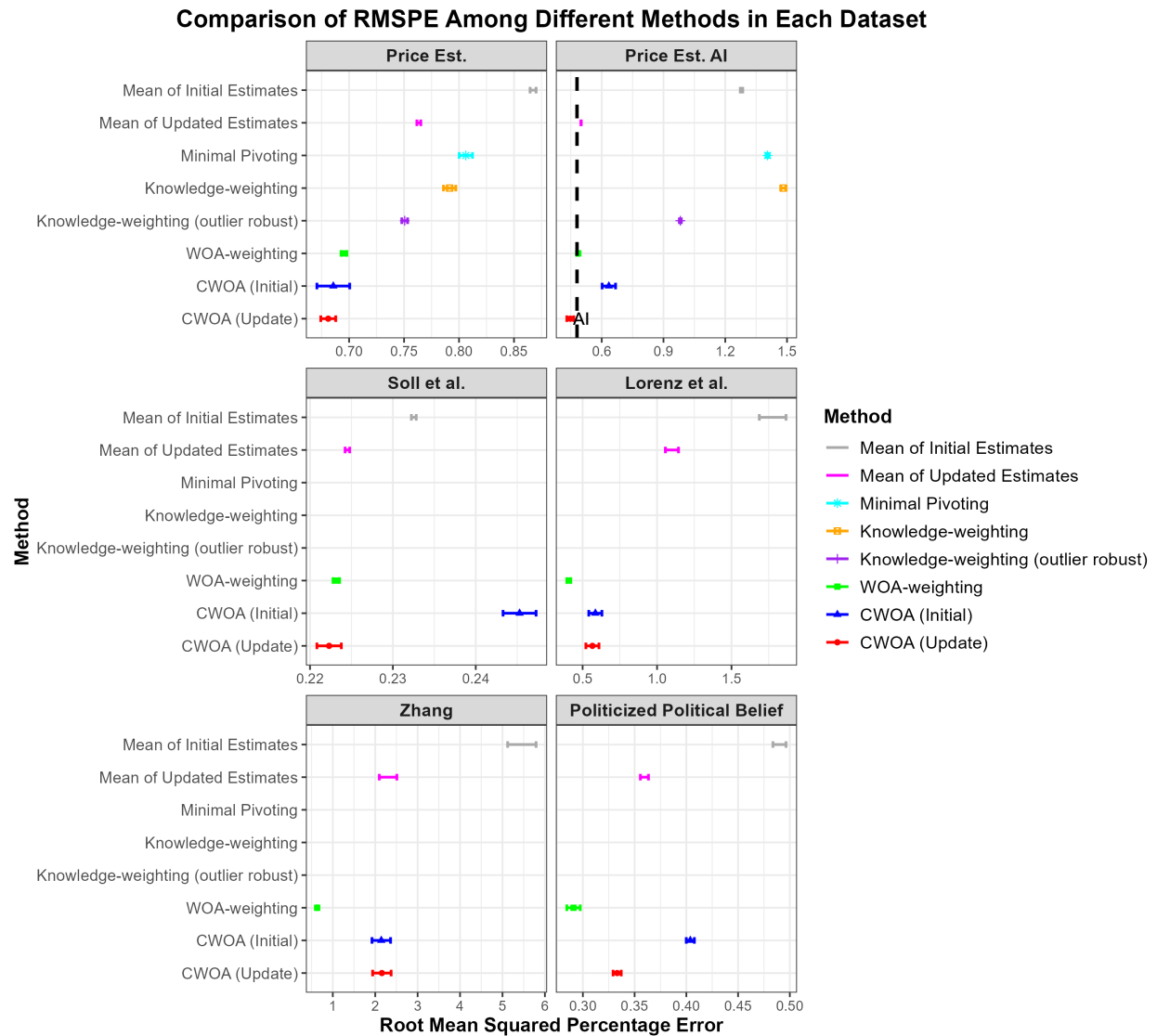


Figure 14.3: The x-axis is each method's root mean squared percentage error across questions. The bars indicate 95% confidence intervals. The y-axis indicates the methods. The black dashed line in the "Price Est. AI" marks the RMSPE of AI's predictions.

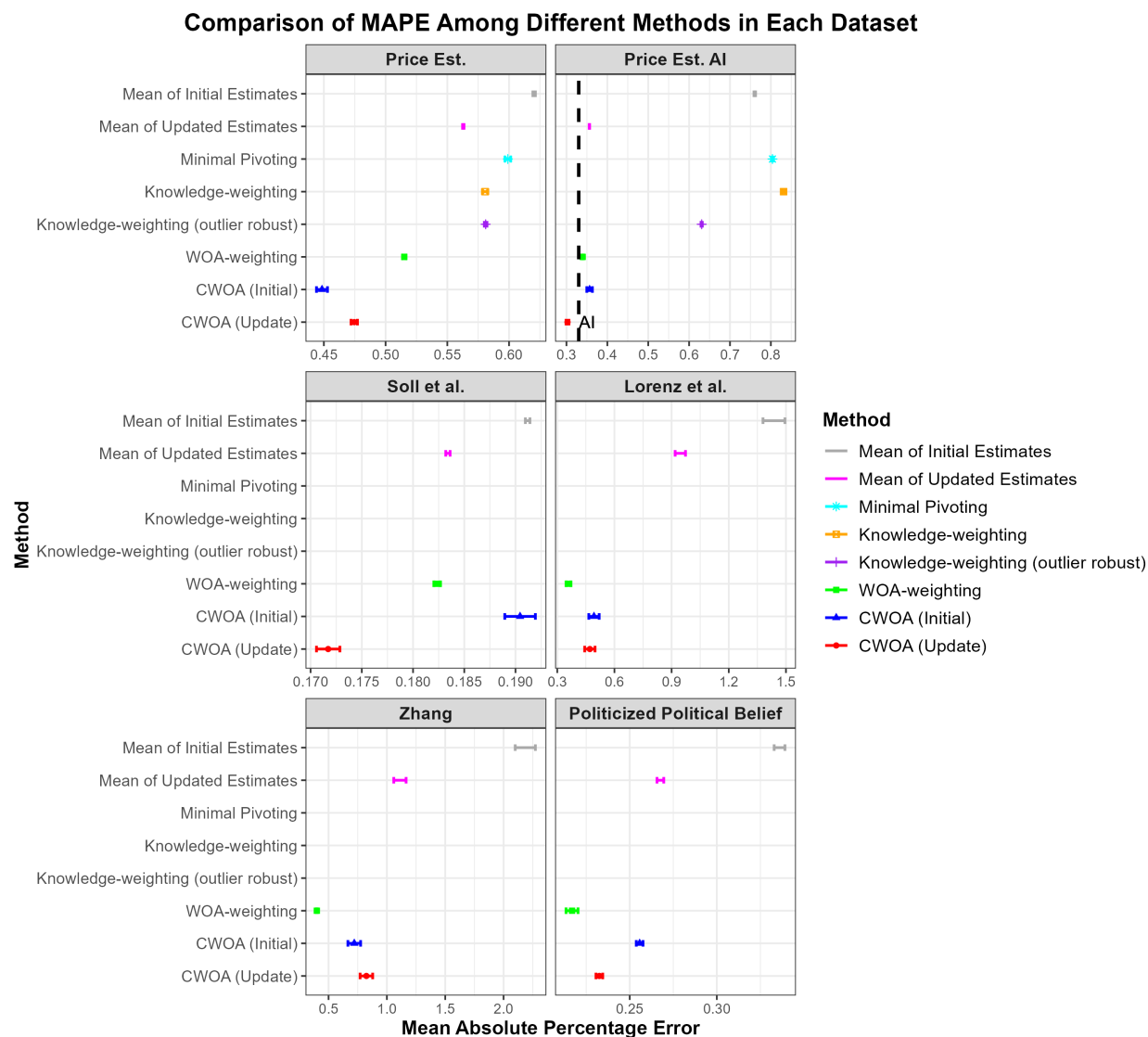


Figure 14.4: The x-axis is each method's mean absolute percentage error across questions in a data source in which CWOA identified at least two clusters. The bars indicate 95% confidence intervals. The y-axis indicates the methods. The black dashed line in the "Price Est. AI" marks the MAPE of AI's predictions.

14.3 Using "Mean of Initial Estimates" as Replacement for Questions with One Cluster

If the Decision Maker does not wish to elicit additional information in single-cluster cases, they may take the mean of initial estimates, while applying CWOA for multi-cluster cases. Importantly, if the Decision Maker collects only initial estimates and does not proceed to advice-taking, alternative methods (mean of updated estimates, meta-prediction-based methods, WOA-weighting) are not available. The analysis below therefore tests CWOA under a minimal-data scenario: when only initial estimates are available for single-cluster cases, CWOA's output for those cases is replaced by the mean of initial estimates. This represents a hypothetical lower bound on CWOA's performance — when richer data is available, the methods discussed in Sections 14.1 and 14.2 are preferred.

As shown in Figures 14.5 and 14.6, CWOA (Update) remains one of the best-performing methods.

Note: The medical dataset is not shown because CWOA algorithm applies in all eight problems.

**Comparison of RMSPE Among Different Methods in Each Dataset
(Mean of Initial Estimates as Replacement for One Cluster Cases)**

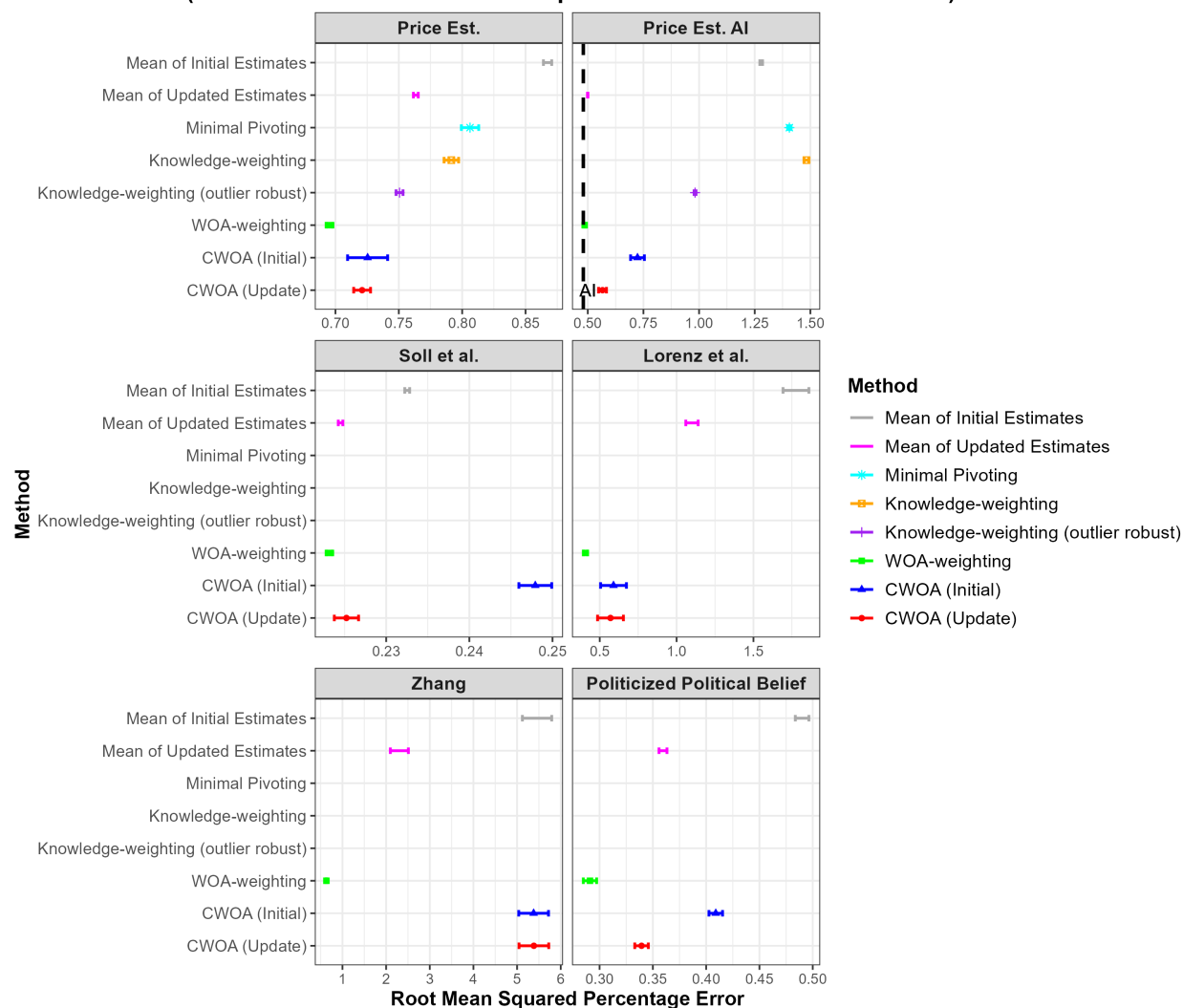


Figure 14.5: The x-axis is each method's root mean squared percentage error across questions. The bars indicate 95% confidence intervals. The y-axis indicates the methods. The black dashed line in the "Price Est. AI" marks the RMSPE of AI's predictions.

**Comparison of MAPE Among Different Methods in Each Dataset
(Mean of Initial Estimates as Replacement for One Cluster Cases)**

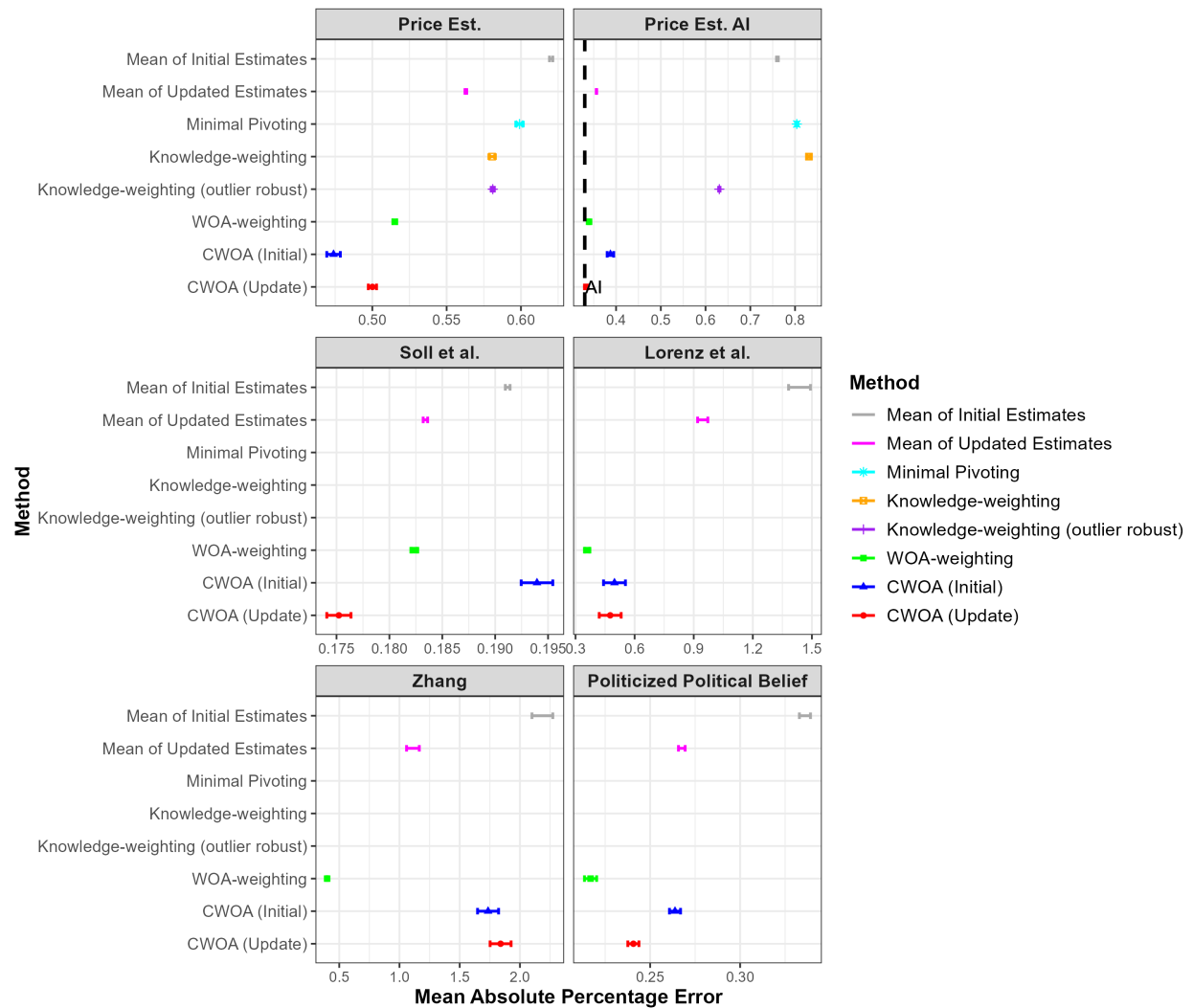


Figure 14.6: The x-axis is each method's mean absolute error across questions in a data source. The bars indicate 95% confidence intervals. The y-axis indicates the methods. The black dashed line in the "Price Est. AI" marks the MAPE of AI's predictions.