Online Appendix for "Interpreting TSLS Estimators in Information Provision Experiments"

Vod Vilfort and Whitney Zhang

Table of Contents

A General Setup

As in the main text, we consider beliefs $B \in \mathcal{B} \subseteq \Delta(\Omega)$ and use $\phi \in \mathbb{R}$ to index scalar features. However, we now explicitly allow for the consideration of multiple features $\Phi = (\phi_1, \ldots, \phi_K)' \in \mathbb{R}^K$. If $\Omega \subseteq \mathbb{R}$, then we may have $K = 2$ and $\Phi(B) = (\mu(B), \sigma^2(B))'$. If $\Omega = \times_{\ell=1}^{L} \Omega_\ell \subseteq \mathbb{R}^L$, then we may have $K = L$ and $\Phi(B) = (\mu_1(B), \ldots, \mu_L(B))'$, where $\mu_\ell(B) = \int \omega_\ell dB_\ell(\omega_\ell)$ and $B_\ell \in \Delta(\Omega_\ell)$ is the marginal distribution of B for states $\omega_\ell \in \Omega_\ell$. For instance, the latter with $L = 2$ accommodates Cullen and Perez-Truglia (2022), in which there are two relevant features: expectations of manager wages and expectations of coworker wages.

- We assume that functions of interest are appropriately measurable wherever necessary.

- In some places, we abbreviate partial derivatives in the manner $\partial_x f(x) = \partial f(x)/\partial x$.

Experiment. The information provision experiment proceeds as in the main text, but now we allow more than two groups: $z \in \mathcal{Z}$, where $|\mathcal{Z}| \geq 2$.

- The prior features are $\Phi_{i0} \equiv \Phi(B_{i0}) = (\phi_{ki0})_{k=1}^{K}$, where $\phi_{ki0} \equiv \phi_k(B_{i0})$.
- The posterior features are $\Phi_{i1}^z \equiv \Phi(B_{i1}^z) = (\phi_{ki1}^z)_{k=1}^{K}$, where $\phi_{ki1}^z \equiv \phi_k(B_{i1}^z)$.
- In the realized experiment, we have $\Phi_{i1} \equiv \Phi_{i1}^{Z_i}$ and $\phi_{ki1} \equiv \phi_{ki1}^{Z_i}$.

Instrumental Variables. As in Assumption 1 in the main text, we assume group assignment Z_i is a valid instrument, now stated for the general set of groups $\mathcal{Z}$.

Assumption A1 (Valid Instrument, General Form)**.** *Z_i satisfies*

(i) IV independence: $Z_i \perp\!\!\!\perp (X_i, B_{i0}, \{S_i^z, B_{i1}^z, Y_i^z(B)\}_{z\in\mathcal{Z}, B\in\mathcal{B}})$;

(ii) IV exclusion: $Y_i^z(B) = Y_i(B)$ for all $z \in \mathcal{Z}$ and $B \in \mathcal{B}$.

Belief Exclusion. We assume actions depend on beliefs through a finite number of features $\Phi = (\phi_1, \ldots, \phi_K)'$, similar to Assumption 2 from the main text for the scalar feature case. Let $\Phi(\mathcal{B}) = \{\Phi(B) : B \in \mathcal{B}\}$ denote the set of possible feature values.

Assumption A2 (Belief Exclusion, General Form)**.** *$\Phi(\mathcal{B})$ is convex, and there exist continuously differentiable functions $Y_i(\Phi)$ defined over it satisfying $Y_i(B) = Y_i(\Phi(B))$.*

- We use $Y_i(\cdot)$ to refer to *both* a function that inputs beliefs (i.e., $B \mapsto Y_i(B)$) and to a function that inputs features of beliefs (i.e., $\Phi \mapsto Y_i(\Phi)$).
- Convexity holds when $\mathcal{B}$ is sufficiently rich; see Appendix A.1 below.
- In practice, to recover interpretable causal effects, we will need pairs of groups $\{z, z'\}$ that shift only one feature of interest. This is consistent with Assumption 2 from the main text; see also the active control discussion in Section V.A. We formalize this "ceteris paribus" condition in Appendix A.2.

Partial Effects. Given agent i, features $\Phi \in \mathbb{R}^K$, and feature of interest $\phi_k \in \mathbb{R}$, we formulate the causal effects of beliefs in terms of the partial effects $\partial Y_i(\Phi)/\partial\phi_k$; see Appendix A.3 for alternative formulations.

A.1 Primitive Conditions for Belief Exclusion

To give primitive conditions for Assumption A2, we consider a (net) utility/profit function $u_i(\omega, y)$ that depends on states ω and actions $y \in \mathcal{Y}$. We can motivate the action function $Y_i(B)$ as agent i's utility-maximizing map from beliefs to actions:

$$Y_i(B) = \underset{y \in \mathcal{Y}}{\operatorname{argmax}} \int u_i(\omega, y) dB(\omega), \quad \forall B \in \mathcal{B}. \tag{A1}$$

If agent i has beliefs B, then $\int u_i(\omega, y) dB(\omega)$ is their subjective expected utility from taking action y. In particular, $Y_i \equiv Y_i(B_{i1})$ is the utility-maximizing action taken under i's posterior beliefs B_{i1}. We can use representation (A1) to motivate action functions $Y_i(\Phi)$ that depend on beliefs through a finite number of features $\Phi \in \mathbb{R}^K$.

Restrictions on Preferences. Let $m_i(\mathcal{Y}) = \{m_i(y) : y \in \mathcal{Y}\}$ denote the image of $\mathcal{Y}$ under a strictly monotonic and continuously differentiable function $m_i : \mathcal{Y} \to \mathbb{R}$.

Proposition A1. *Consider functions $\varphi_k : \Omega \to \mathbb{R}$ and features $\phi_k(B) = \int \varphi_k(\omega) dB(\omega)$. Let $\mathcal{B}$ be the set of beliefs such that $\int |\varphi_k(\omega)|^2 dB(\omega) < \infty$ for each k and $B \in \mathcal{B}$. If $Y_i(B)$ satisfies (A1) with squared-error loss function*

$$-u_i(\omega, y) = (\textstyle\sum_{k=1}^K \theta_{ki} \varphi_k(\omega) - m_i(y))^2,$$

and $\sum_{k=1}^K \theta_{ki} \phi_k(B) \in m_i(\mathcal{Y})$ for each $B \in \mathcal{B}$, then $Y_i(B) = m_i^{-1}(\sum_{k=1}^K \theta_{ki} \phi_k(B))$. In particular, Assumption A2 is satisfied for $Y_i(\Phi) = m_i^{-1}(\sum_{k=1}^K \theta_{ki} \phi_k)$.

Proposition A1 considers squared error loss functions $-u_i(\omega, y)$. To give intuition, consider the setting of Jäger et al. (2024), as in Example 1 from the main text. Let $m_i(y) = y$, and suppose there is a latent function $\omega \mapsto f_i(\omega)$ that maps potential wages ω to the optimal rate at which worker i should search for a new job. However, this function is complicated or imperfectly known, and so i uses the approximation $\omega \mapsto \sum_{k=1}^K \theta_{ki} \varphi_k(\omega)$, where $\varphi_k(\omega) = \omega^{k-1}$. In this case, $Y_i(B) = \sum_{k=1}^K \theta_{ki} \phi_k(B)$ is worker i's minimum mean squared error approximation to their optimal rate of job search. $K = 2$ gives a linear approximation, whereas $K = 3$ gives a quadratic approximation. In the linear case, we obtain $Y_i(B) = \theta_{1i} + \theta_{2i} \mu(B)$, and θ_{2i} gives the partial effects of expectations on actions—this specification is considered in Balla-Elliott (2023). In the quadratic case, the agents are allowed to be risk averse, which is relevant for some applications (Kumar, Gorodnichenko, and Coibion 2023b; Coibion et al. 2021).

When $m_i(y) = y$, the partial effects of feature ϕ_k above are homogeneous across $\Phi \in \mathbb{R}^K$:

$$\phi_k \mapsto \frac{\partial Y_i(\Phi)}{\partial \phi_k} = \theta_{ki}.$$

However, some models may allow for heterogeneity across $\Phi \in \mathbb{R}^K$ (Cullen and Perez-Truglia 2022, Section II.A). To accommodate heterogeneity, we can consider nonlinear m_i:

$$\phi_k \mapsto \frac{\partial Y_i(\Phi)}{\partial \phi_k} = \left(\frac{\partial m_i(Y_i(\Phi))}{\partial y} \right)^{-1} \theta_{ki},$$

where the equality follows from the inverse function theorem.

Restrictions on Beliefs. In some cases, $\mathcal{B}$ is a set of parametric belief distributions. A leading class of parametric distributions are exponential families (Lehmann and Casella 2006, Section 3.4). This class includes the set of Normal distributions often considered in models from the information provision literature (Armantier et al. 2016; Cavallo, Cruces, and Perez-Truglia 2017; Armona, Fuster, and Zafar 2019; Cullen and Perez-Truglia 2022; Fuster et al. 2022; Balla-Elliott et al. 2022). There are also classes of parametric distributions that are useful for belief elicitation. For example, Kumar, Gorodnichenko, and Coibion (2023b) and Coibion et al. (2021) consider triangular distributions.

- Formally, consider an open and convex "parameter space" $\Theta \in \mathbb{R}^K$, where $\Phi(\mathcal{B}) \subseteq \Theta$.
- Each $B \in \mathcal{B}$ has density $b(\omega) = dB(\omega)/d\nu$ with respect to some common sigma-finite measure $\nu \in \Delta(\Omega)$.
- $\mathcal{Y}$ is an open and convex subset of $\mathbb{R}$.

Proposition A2. *Consider beliefs $B \in \mathcal{B}$ parameterized by $\Phi \in \Theta$ in the sense that there exists $(\omega, \Phi) \mapsto f(\omega|\Phi) \geq 0$ such that $b(\omega) = f(\omega|\Phi(B))$ for all $B \in \mathcal{B}$. If $\Phi \mapsto f(\omega|\Phi)$ is continuously differentiable over Θ ν-almost surely, $\Phi(\mathcal{B}) = \Theta$, and $Y_i(B)$ satisfies (A1) with $y \mapsto u_i(\omega, y)$ that is strictly concave and twice continuously differentiable ν-almost surely, and such that for each $y \in \mathcal{Y}$ and $\Phi \in \Theta$, there exists $\delta > 0$ for which*

(i) $\int |u_i(\omega, y)| f(\omega|\Phi) d\nu(\omega) < \infty$;

(ii) $\int \sup_{\tilde{y}:|\tilde{y}-y|\leq\delta} |\partial_y^n u_i(\omega, \tilde{y})| f(\omega|\Phi) d\nu(\omega) < \infty$ *for each* $n \in \{1, 2\}$; *and*

(iii) $\int \sup_{\tilde{\phi}_k:|\tilde{\phi}_k-\phi_k|\leq\delta} |\partial_y u_i(\omega, y) \partial_{\phi_k} f(\omega|\tilde{\phi}_k, \phi_{-k})| d\nu(\omega) < \infty$ *for each* k,

then Assumption A2 is satisfied for

$$Y_i(\Phi) = \underset{y \in \mathcal{Y}}{\operatorname{argmax}}\, U_i(y, \Phi), \quad U_i(y, \Phi) = \int u_i(\omega, y) f(\omega|\Phi) d\nu(\omega).$$

Note that, by the implicit function theorem, $Y_i(\Phi)$ has partial effects of the form

$$\phi_k \mapsto \frac{\partial Y_i(\Phi)}{\partial \phi_k} = \frac{\partial_{\phi_k}\partial_y U_i(Y_i(\Phi), \Phi)}{|\partial_y^2 U_i(Y_i(\Phi), \Phi)|}.$$

For intuition, consider agent i's optimal behavior at feature $\Phi \in \Theta$—or equivalently, at the belief corresponding to density $\omega \mapsto f(\omega|\Phi)$. At Φ, the optimal action is $Y_i(\Phi)$. The marginal change in beliefs along feature ϕ_k leads to a marginal change $\partial_{\phi_k} U_i(Y_i(\Phi), \Phi)$ in i's subjective expected utility, so $Y_i(\Phi)$ is no longer optimal. To re-optimize, i can either increase or decrease the intensity of their action. This choice of *direction* depends on the sign of the marginal expected utility $\partial_{\phi_k}\partial_y U_i(Y_i(\Phi), \Phi)$ at $Y_i(\Phi)$. The *magnitude* of the change in actions depends on the curvature $|\partial_y^2 U_i(Y_i(\Phi), \Phi)|$ of the expected utility function at $Y_i(\Phi)$.

A.2 Conditional TSLS and Aggregation

We consider TSLS specifications that condition on pairs of comparison groups $\{z, z'\} \subseteq \mathcal{Z}$. Throughout we maintain Assumptions A1-A2.

Equations. Given feature of interest ϕ_k, consider outcome and first-stage equations

$$Y_i = W_i'\gamma^{z:z'} + \beta^{z:z'}\phi_{ki1} + \varepsilon_i, \tag{A2}$$

$$\phi_{ki1} = W_i'\delta^{z:z'} + \mathbb{1}\{Z_i = z'\}I_i'\pi^{z:z'} + \zeta_i, \tag{A3}$$

where the superscripts mean that we condition to the set of agents with $Z_i \in \{z, z'\}$. Aside from this conditioning, everything else is analogous to Section II.A of the main text.

- W_i is a covariate vector (containing a constant 1) that may include components and/or functions of $(X_i, \Phi_{i0}, (S_i^z)_{z\in\mathcal{Z}})$, and I_i is a scalar component of W_i. As we saw in the main text, valid sign corrections are generally scalars, so we focus on the case where I_i is scalar for ease of exposition: $\pi^{z:z'} \in \mathbb{R}$.

- The conditional TSLS estimand corresponding to equations (A2)-(A3) is the vector of coefficients from the population conditional OLS regression of Y_i on W_i and the fitted values from the population conditional OLS regression of ϕ_{ki1} on W_i and $I_i\mathbb{1}\{Z_i = z'\}$.

- We consider the causal interpretation of $\beta_{\text{TSLS}}^{z:z'}$, which is the component of the conditional TSLS estimand corresponding to the coefficient on ϕ_{ki1} from outcome equation (A2).

- Technically, the above coefficients and estimand depend on k—and so could the choice of covariates—but we suppress these dependencies in the notation.

Proposition A3. *If* $\mathbb{E}[W_i W_i' | Z_i \in \{z, z'\}]$ *is full-rank,* $\mathbb{P}(Z_i = z) > 0, \mathbb{P}(Z_i = z') > 0$*, and* $\mathbb{E}[I_i \phi_{ki1} | Z_i = z'] - \mathbb{E}[I_i \phi_{ki1} | Z_i = z] \neq 0$*, then*

$$\beta_{\text{TSLS}}^{z:z'} = \frac{\mathbb{E}[I_i(Y_i(\Phi_{i1}^{z'}) - Y_i(\Phi_{i1}^{z}))]}{\mathbb{E}[I_i(\phi_{ki1}^{z'} - \phi_{ki1}^{z})]}.$$

Moreover, if the group pair $\{z, z'\}$ *is such that* $\phi_{k'i1}^{z'} = \phi_{k'i1}^{z}$ *for all* $k' \neq k$*, then*

$$\beta_{\text{TSLS}}^{z:z'} = \mathbb{E}[w_i^{I,z:z'} \bar{\beta}_i^{z:z'}], \quad w_i^{I,z:z'} = \frac{I_i(\phi_{ki1}^{z'} - \phi_{ki1}^{z})}{\mathbb{E}[I_i(\phi_{ki1}^{z'} - \phi_{ki1}^{z})]},$$
$$\bar{\beta}_i^{z:z'} = \int \lambda_i^{z:z'}(\phi_k) \frac{\partial Y_i^{z:z'}(\phi_k)}{\partial \phi_k} d\phi_k,$$

where $\lambda_i^{z:z'}(\phi_k)$ *is the density of the uniform distribution on the range of* ϕ_k *between* ϕ_{ki1}^{z} *and* $\phi_{ki1}^{z'}$ *and* $Y_i^{z:z'}(\phi_k) \equiv Y_i(\phi_k, (\phi_{k'i1}^{z})_{k' \neq k})$*.*

Ceteris Paribus Variation. The condition that $\phi_{k'i1}^{z'} = \phi_{k'i1}^{z}$ for all $k' \neq k$ requires the pair of groups $\{z, z'\}$ to generate ceteris paribus variation in ϕ_k (c.f. the active control discussion in Section V.A of the main text). This allows $\beta_{\text{TSLS}}^{z:z'}$ to make a proper normalization in terms of ϕ_{ki1}.

Aggregation. If multiple pairs of comparison groups $\{z, z'\}$ recover positive-weighted APEs for a given feature of interest ϕ_k, then we can aggregate over these APEs. Formally, consider

$$\beta_{\text{APE}}^{z:z'} = \mathbb{E}[w_i^{z:z'} \bar{\beta}_i^{z:z'}], \quad w_i^{z:z'} = \frac{|I_i^{z:z'}(\phi_{ki1}^{z'} - \phi_{ki1}^{z})|}{\mathbb{E}[|I_i^{z:z'}(\phi_{ki1}^{z'} - \phi_{ki1}^{z})|]}.$$

We can view $\beta_{\text{APE}}^{z:z'}$ as the parameter that $\beta_{\text{TSLS}}^{z:z'}$ recovers when using an interaction $I_i^{z:z'}$ that is a valid sign correction for the contrast $\phi_{ki1}^{z'} - \phi_{ki1}^{z}$. For example, suppose $z = C$ is a control group and $z' \neq C$ are various treatment groups designed to shift the same feature ϕ_k, as in Coibion, Gorodnichenko, and Weber (2022). Then, given a choice of weights $\alpha_{z'} \geq 0$ such that $\sum_{z' \neq C} \alpha_{z'} = 1$, we can aggregate to recover

$$\sum_{z' \neq C} \alpha_{z'} \beta_{\text{TSLS}}^{C:z'} = \sum_{z' \neq C} \alpha_{z'} \beta_{\text{APE}}^{C:z'} = \sum_{z' \neq C} \alpha_{z'} \mathbb{E}[w_i^{C:z'} \bar{\beta}_i^{C:z'}].$$

This pairwise approach avoids complications that arise when interpreting TSLS with multiple instruments (Mogstad, Torgovitsky, and Walters 2021).

A.3 Alternative Formulations for the Causal Effects

Let $K = 1$ so that $Y_i(B) = Y_i(\phi(B))$ for a single feature $\phi \in \mathbb{R}$. Moreover, suppose the set of groups is $\mathcal{Z} = \{C, T\}$. We have thus far formulated the causal effects of interest in terms of partial effects $\partial Y_i(\phi)/\partial\phi$. Here we discuss two other potential formulations.

Behavioral Elasticities. If $Y_i(B) \neq 0$ and $\phi(B) \neq 0$, then we can formulate the causal parameters of interest in terms of partial elasticities—these *behavioral elasticities* generate unit-free measures, which facilitates comparisons across applications (Haaland, Roth, and Wohlfart 2023, Section 8). We therefore consider TSLS specifications that replace Y_i and ϕ_{i1} with $\log(Y_i^n)$ and $\log(\phi_{i1}^n)$, for some chosen power $n \in \mathbb{N}$ such that $Y_i(B)^n > 0$ and $\phi(B)^n > 0$. If $Y_i(B) > 0$ and $\phi(B) > 0$, then we can take $n = 1$ and the logarithm will be well-defined. Taking $n = 2$ allows actions and features to take negative values. The analogues of equations (A2)-(A3) and Proposition A3 (or equivalently in this case, equations (1)-(2) and Proposition 1 from the main text) for this setup produce

$$\beta_{\text{TSLS}} = \frac{\mathbb{E}[I_i(\log[Y_i(\phi_{i1}^T)^n] - \log[Y_i(\phi_{i1}^C)^n])]}{\mathbb{E}[I_i(\log[(\phi_{i1}^T)^n] - \log[(\phi_{i1}^C)^n])]}.$$

The fundamental theorem of calculus gives

$$\log[Y_i(\phi_{i1}^T)^n] - \log[Y_i(\phi_{i1}^C)^n] = \psi_i \int \frac{1}{Y_i(v)^n} n Y_i(v)^{n-1} \frac{\partial Y_i(v)}{\partial \phi} \mathbb{1}\{v \in \mathcal{V}_i\} dv$$

$$\log[(\phi_{i1}^T)^n] - \log[(\phi_{i1}^C)^n] = \psi_i \int \frac{1}{v^n} n v^{n-1} \mathbb{1}\{v \in \mathcal{V}_i\} dv,$$

where $\psi_i = \text{sign}(\phi_{i1}^T - \phi_{i1}^C)$ and $\mathcal{V}_i = \{(1-\alpha)\phi_{i1}^C + \alpha\phi_{i1}^T : \alpha \in [0,1]\}$ is the range of feature values v between ϕ_{i1}^C and ϕ_{i1}^T. Therefore, we obtain

$$\beta_{\text{TSLS}} = \mathbb{E}\left[\int \Lambda_i(v)\, \tilde{\beta}_i(v) dv\right], \quad \tilde{\beta}_i(v) = \frac{v}{Y_i(v)} \frac{\partial Y_i(v)}{\partial \phi},$$

$$\Lambda_i(v) = \frac{\psi_i I_i\, v^{-1} \mathbb{1}\{v \in \mathcal{V}_i\}}{\mathbb{E}\left[\int \psi_i I_i\, v^{-1} \mathbb{1}\{v \in \mathcal{V}_i\} dv\right]},$$

which is a weighted average of partial elasticities $\tilde{\beta}_i(v)$ with weights $\Lambda_i(v)$ that integrate to one across v and i; see also Angrist, Graddy, and Imbens (2000, Corollary 1).

Note that a valid sign correction I_i gives $I_i\psi_i = |I_i\psi_i|$, in which case $|I_i\psi_i|\mathbb{1}\{v \in \mathcal{V}_i\} = |I_i||\psi_i|\mathbb{1}\{v \in \mathcal{V}_i\} = |I_i|\mathbb{1}\{v \in \mathcal{V}_i\}$ so that $\Lambda_i(v) \geq 0$ for all i and v.

Discrete Actions. To consider the case of discrete actions, we modify Assumption A2 so that we can take well-defined partial effects. We continue to assume that $\{\phi(B) : B \in \mathcal{B}\}$ is a convex set, but now assume there exist continuously differentiable $\phi \mapsto \bar{Y}_i(\phi)$ defined over it such that $B \mapsto \mathbb{E}[Y_i(B)|R_i] = \bar{Y}_i(\phi(B))$, where R_i is some random vector. This formulation takes actions (which are discrete) and considers "average actions" (which are assumed to be continuous). We assume I_i is a component or function of R_i, so then

$$\begin{aligned}\mathbb{E}[I_i(Y_i(B_{i1}^T) - Y_i(B_{i1}^C))] &= \mathbb{E}[I_i\mathbb{E}[(Y_i(B_{i1}^T) - Y_i(B_{i1}^C))|R_i]] \\ &= \mathbb{E}[I_i(\bar{Y}_i(\phi_{i1}^T) - \bar{Y}_i(\phi_{i1}^C))],\end{aligned}$$

where the first equality follows from iterated expectations. Thus, the analogue of Proposition A3 (or equivalently in this case, Proposition 1 from the main text) for this setup produces

$$\beta_{\text{TSLS}} = \mathbb{E}[w_i^I \bar{\beta}_i], \quad w_i^I = \frac{\pi' I_i(\phi_{i1}^T - \phi_{i1}^C)}{\mathbb{E}[\pi' I_i(\phi_{i1}^T - \phi_{i1}^C)]}, \quad \bar{\beta}_i = \int \lambda_i(\phi) \frac{\partial \bar{Y}_i(\phi)}{\partial \phi} d\phi.$$

We arrive similar expressions to Proposition 1 from the main text, except now the partial effects of features on actions $\partial_\phi Y_i(\phi)$ are replaced by the partial effects of features on (conditional on R_i) average actions $\partial_\phi \bar{Y}_i(\phi)$. The random vector R_i could entirely consist of the observed "pre-determined" variables in the experiment (i.e., prior features ϕ_{i0}, signals S_i^z, and agent characteristics X_i). However, we also allow for latent unobserved components. In the following example, the action is binary $y \in \{0, 1\}$ and the "average action" is the probability of taking action $y = 1$.

Let $\Omega = \mathbb{R}$ and consider binary actions $B \mapsto Y_i(B) \in \{0, 1\}$ that satisfy

$$Y_i(B) = \underset{y \in \{0,1\}}{\text{argmax}} \int y\omega + (1 - y)\xi_i \, dB(\omega), \quad \forall B \in \mathcal{B}.$$

In particular, agents i with beliefs B choose $Y_i(B)$ to maximize their subjective expected utility, as in representation (A1); we can think of ξ_i as some outside option, and ω as the benefit to choosing $y = 1$. The above leads to

$$Y_i(B) = \mathbb{1}\{\mu(B) \geq \xi_i\}.$$

If ξ_i has absolutely continuous CDF conditional on R_i, then we can choose

$$\bar{Y}_i(\mu(B)) = \mathbb{P}(Y_i(B) = 1|R_i) = \mathbb{P}(\xi_i \leq \mu(B)|R_i),$$

and so the marginal effect of expectations on the probability of choosing $y = 1$ is given by the conditional PDF $\partial \bar{Y}_i(\mu)/\partial \mu$. For example, $\xi_i|R_i \sim \mathcal{N}(\theta_i, \Sigma_i)$ for $(\theta_i, \Sigma_i) \equiv (\theta(R_i), \Sigma(R_i))$

gives

$$\frac{\partial \bar{Y}_i(\mu)}{\partial \mu} = \frac{1}{\sqrt{2\pi\Sigma_i}} \exp\left(-\frac{(\mu-\theta_i)^2}{2\Sigma_i}\right).$$

We can think of R_i as a set of determinants for the outside options.

B General Model of Posterior Formation

In Section III of the main text, we derived sign corrections I_i in stylized models of posterior formation that specified (i) what agents assume about the signals and (ii) how agents update their priors. Here we consider a general approach to (i) and (ii). Appendix B.1 covers (i) and Appendix B.2 covers (ii). Appendix B.3 derives the sign corrections of interest, and Appendix B.4 illustrates the general approach in an extension of the stylized Normal-Normal setup.

In what follows, we consider $\omega \in \Omega \subseteq \mathbb{R}$ and suppose beliefs $B \in \mathcal{B} \subseteq \Delta(\Omega)$ have densities $b(\omega) = dB(\omega)/d\nu$ with respect to a common dominating measure $\nu \in \Delta(\Omega)$, such as Lebesgue measure, counting measure, or mixtures of the two.

The experiment provides signals $S_i^z \in \mathcal{S}$ to agents i assigned to group $z \in \mathcal{Z}$, where $\mathcal{S}$ is a set of signals. Agent i assumes that when the true state is ω, their signal was drawn from distribution $Q_i(\cdot|\omega) \in \Delta(\mathcal{S})$. Given i, we suppose each $Q_i(\cdot|\omega)$ has density $q_i(\cdot|\omega)$ with respect to a dominating measure that is common to all ω.

Given signal $s \in \mathcal{S}$, agent i forms posterior beliefs $B_{i1}(\cdot|s)$. In particular, $s \mapsto B_{i1}(\cdot|s)$ is an agent-specific mapping from signals $s \in \mathcal{S}$ to beliefs $B \in \mathcal{B}$ that implicitly depends on i's prior beliefs B_{i0}. In the realized experiment, $B_{i1}^z \equiv B_{i1}(\cdot|S_i^z)$.

B.1 MLR Property

Agents assume the set of signal distributions $\mathcal{Q}_i = \{Q_i(\cdot|\omega) : \omega \in \Omega\} \subseteq \Delta(\mathcal{S})$ is informative for the states. In particular, $\mathcal{Q}_i$ satisfies the following *monotone likelihood ratio* (MLR) property: Agents assume the experiment tends to provide larger signals s under larger realizations of the state $\omega \in \Omega$. In what follows, let $\geq_i$ be a potentially agent-specific partial order on $\mathcal{S}$.

Definition B1 (MLR)**.** $\mathcal{Q}_i \subseteq \Delta(\mathcal{S})$ satisfies the *monotone likelihood ratio* (MLR) property in $(\mathcal{S}, \geq_i)$ if for each $\omega, \omega' \in \Omega$ and $s', s \in \mathcal{S}$ such that $\omega' > \omega$ and $s' >_i s$, we have

$$\frac{q_i(s'|\omega')}{q_i(s'|\omega)} \geq \frac{q_i(s|\omega')}{q_i(s|\omega)}.$$

If signals are quantitative, $\mathcal{S} \subseteq \mathbb{R}$, and $\geq_i$ is the standard order on $\mathbb{R}$, then Definition B1 (c.f. Casella and Berger (2021, Definition 8.3.16)) is satisfied for many signal distributions of

interest, including the set of Normal signal distributions from the main text. More generally, numerous exponential families satisfy Definition B1 in their respective sufficient statistics.

Null Signal. Definition B1 accommodates passive control experiments: If $\mathcal{S} = \mathbb{R} \cup \{s^{\varnothing}\}$, then agent i's ranking of information $s \in \mathbb{R}$ relative to null signal $s^{\varnothing}$ (i.e., no information) may depend on their initial prior beliefs B_{i0} in a manner that is known to the researcher. We use this structure to derive passive control sign corrections in Appendix B.3.

Qualitative Signals. Definition B1 permits qualitative signals, such as educational videos (Alesina, Ferroni, and Stantcheva 2021; Dechezleprêtre et al. 2022). For example, if $\mathcal{S} = \{s^L, s^H\}$ are two videos, then we may have $s^H \geq_i s^L$ in the sense that video s^H conveys more pessimistic information than video s^L, which makes the order $\geq_i$ on $\mathcal{S} = \{s^L, s^H\}$ common across i.

Multiple States. For settings with multi-dimensional states $\Omega = \times_{\ell=1}^{L} \Omega_\ell \subseteq \mathbb{R}^L$, $L > 1$, we can apply the arguments that follow to the marginals $\omega_\ell \in \Omega_\ell$ of interest. For example, Cullen and Perez-Truglia (2022) consider employees i whose effort depends on beliefs $B_1 \in \Delta(\Omega_1)$ over peer wages $\omega_1 \in \Omega_1$ and beliefs $B_2 \in \Delta(\Omega_2)$ over manager wages $\omega_2 \in \Omega_2$. The signal space $\mathcal{S}$ embeds $\mathcal{S}_\ell \subseteq \mathcal{S}$, which gives information pertaining to just ω_ℓ. If we are interested in deriving sign corrections for expectations μ_ℓ over ω_ℓ, then $\geq_i$ may allow agents to compare signals within each $\mathcal{S}_\ell$ (though not necessarily across them). In particular, the MLR structure and the arguments that follow can be argued for at each margin. Thus, for ease of notation, we develop arguments for the case of $\Omega \subseteq \mathbb{R}$.

B.2 Belief Updating Rules

Before we can leverage the above MLR property to derive sign corrections, we must consider agents' belief updating rules in greater detail. Recall that the density of $B \in \mathcal{B}$ is $b(\omega) = dB(\omega)/d\nu$. In particular, the prior and posterior densities are $b_{i1}(\omega|s) = dB_{i1}(\omega|s)/d\nu$ and $b_{i0}(\omega) = dB_{i0}(\omega)/d\nu$, respectively. We first consider Bayesian rules.

Definition B2 (Bayesian)**.** $s \mapsto B_{i1}(\cdot|s)$ is *Bayesian* if

$$b_{i1}(\omega|s) = \frac{q_i(s|\omega)b_{i0}(\omega)}{\int q_i(s|\omega)b_{i0}(\omega)d\nu(\omega)}, \quad \forall s \in \mathcal{S}.$$

These Bayesian rules are prevalent in models from the information provision literature. However, our framework also accommodates non-Bayesian learning.

Definition B3 (Anchored)**.** $s \mapsto B_{i1}(\cdot|s)$ is *anchored* if there exist $\tau_i \in [0,1]$ and $B_i^s \in \mathcal{B}$ such that

$$b_{i1}(\omega|s) = \tau_i b_i^s(\omega) + (1-\tau_i)\frac{q_i(s|\omega)b_{i0}(\omega)}{\int q_i(s|\omega)b_{i0}(\omega)d\nu(\omega)}, \quad \forall s \in \mathcal{S}.$$

Anchored rules distort agents' posteriors away from the Bayesian baseline and towards some anchor belief B_i^s that may depend on s. Anchored rules accommodate numerous behavioral biases from the behavioral economics literature (Gabaix 2019, Section 2.3), including the anchoring-and-adjustment heuristics discussed in Tversky and Kahneman (1974). For instance, choosing the anchors to be the prior beliefs $B_i^s = B_{i0}$ corresponds to *conservatism*, wherein agents insufficiently update their beliefs relative to the Bayesian baseline (Edwards 1968). In the language of Clippel and Zhang (2022), the anchored rules are "affine" distortions of Bayesian updating. Our framework also accommodates "nonlinear" distortions. The following class of nonlinear distortions was developed in Grether (1980).

Definition B4 (Grether)**.** $s \mapsto B_{i1}(\cdot|s)$ is *Grether* if there exist $\theta_{i0}, \theta_{i1} > 0$ such that

$$b_{i1}(\omega|s) = \frac{q_i(s|\omega)^{\theta_{i1}} b_{i0}(\omega)^{\theta_{i0}}}{\int q_i(s|\omega)^{\theta_{i1}} b_{i0}(\omega)^{\theta_{i0}} d\nu(\omega)}, \quad \forall s \in \mathcal{S}.$$

The "inference" and "base-rate" parameters, $\theta_{i1}, \theta_{i0} > 0$, allow for flexible deviations from Bayesian updating. The following discussion closely follows the language in Benjamin (2019, page 103). On the one hand, $\theta_{i1} < 1$ ($\theta_{i1} > 1$) reflects under-inference (over-inference) in that agents update as if the signals $s \in \mathcal{S}$ are less (more) informative for the states $\omega \in \Omega$ than in the Bayesian baseline of $\theta_{i1} = 1$. On the other hand, $\theta_{i0} < 1$ ($\theta_{i0} > 1$) reflects base-rate neglect (over-use) in the sense that agents update as if their priors B_{i0} are less (more) informative for the states $\omega \in \Omega$ than in the Bayesian baseline of $\theta_{i0} = 1$.

Remark B1. If B_{i0} is a conjugate parametric prior belief distribution for the set of signal distributions $\mathcal{Q}_i$, then Bayesian updating implies B_{i1} belongs to the same parametric family as B_{i0}. In this case, if B_{i0} belongs to the same parametric family for each i, then we can take $\mathcal{B}$ to be that parametric family. If this family is parametrized by the feature of interest ϕ, then $Y_i(B) = Y_i(\phi(B))$ for each $B \in \mathcal{B}$, which gives us belief exclusion in ϕ; see Proposition A2. If we allow for richer patterns of heterogeneity, either via non-conjugate belief/signal distributions, or via non-Bayesian learning, then the above argument generally fails. This highlights one sense in which arguments for belief exclusion based on parametric restrictions can be fragile.

Remark B2. Allowing for non-Bayesian learning has practical relevance, since the information provision literature often acknowledges the potential for behavioral biases in belief formation. For example, there is concern that agents may numerically anchor their posterior expectations at values of quantitative signals (Cavallo, Cruces, and Perez-Truglia 2017). This behavior corresponds to anchored rules with B_i^s as the distribution that places probability one on the provided signal s. There is also concern that information provision may induce emotional responses that affect belief updating (Haaland, Roth, and Wohlfart 2023). If such responses are suggestive of motivated reasoning, then we can microfound the anchoring parameters $\tau_i \in [0,1]$ with a model of utility-maximization (Clippel and Zhang 2022, Example 2); similar arguments apply for the inference and base-rate parameters θ_{i1}, θ_{i0} in the Grether case.

B.3 General Sign Corrections

We can derive sign corrections I_i by exploiting the first-order stochastic dominance (FOSD) ordering implied by $(\mathcal{S}, \geq_i)$ and the above updating rules.

Proposition B1. *For each i, suppose the set of signal distributions $\mathcal{Q}_i \subseteq \Delta(\mathcal{S})$ satisfies the MLR property in $(\mathcal{S}, \geq_i)$ and that belief updating rule $s \mapsto B_{i1}(\cdot|s)$ is either Bayesian, anchored (with an anchor B_i^s that increases in the sense of FOSD when $s' >_i s$), or Grether. Then, for any feature $\phi(B) = \int \varphi(\omega) dB(\omega)$ such that $\omega \mapsto \varphi(\omega)$ is increasing, we have that $s' >_i s$ implies*

$$\phi(B_{i1}(\cdot|s')) = \int \varphi(\omega) dB_{i1}(\omega|s') \geq \int \varphi(\omega) dB_{i1}(\omega|s) = \phi(B_{i1}(\cdot|s)). \tag{B1}$$

Proposition B1 exploits the FOSD ordering in posterior beliefs that arises when (i) s', s can be ordered in $(\mathcal{S}, \geq_i)$ and (ii) the belief updating rules respect the orderings—in principle, our framework accommodates any belief updating rule that induces "signal monotonicity" in the sense of condition (B1).

Remark B3. Proposition B1 covers expectations and second moments, which correspond to $\varphi(\omega) = \omega$ and $\varphi(\omega) = \omega^2$, respectively. However, it does not cover the variance, which is the difference of second moments and the squared expectations. That said, it is important to note that this MLR framework only gives a set of *sufficient* conditions for deriving sign corrections. In particular, sign corrections can still be intuited for the variance, as discussed in Section III.C from the main text.

Sign Corrections. Consider an experiment with two groups $\mathcal{Z} = \{C, T\}$, and corresponding signals S_i^C, S_i^T taking values in some signal space $\mathcal{S}$. The feature of interest is the mean μ. If

a researcher knows i's ordering between S_i^C and S_i^T, then the generalized sign interaction

$$I_i^* = \mathbb{1}\{S_i^T >_i S_i^C\} - \mathbb{1}\{S_i^T <_i S_i^C\}$$

is a valid sign correction for $\mu_{i1}^T - \mu_{i1}^C$, given condition (B1). In particular,

$$\begin{aligned} I_i^*(\mu_{i1}^T - \mu_{i1}^C) &= (\mathbb{1}\{S_i^T >_i S_i^C\} - \mathbb{1}\{S_i^T <_i S_i^C\})(\mathbb{1}\{\mu_{i1}^T > \mu_{i1}^C\} - \mathbb{1}\{\mu_{i1}^T < \mu_{i1}^C\})|\mu_{i1}^T - \mu_{i1}^C| \\ &= (\mathbb{1}\{S_i^T >_i S_i^C\}\mathbb{1}\{\mu_{i1}^T > \mu_{i1}^C\} + \mathbb{1}\{S_i^T <_i S_i^C\}\mathbb{1}\{\mu_{i1}^T < \mu_{i1}^C\})|\mu_{i1}^T - \mu_{i1}^C| \\ &= (\mathbb{1}\{S_i^T >_i S_i^C\} + \mathbb{1}\{S_i^T <_i S_i^C\})|\mu_{i1}^T - \mu_{i1}^C| \\ &= \mathbb{1}\{I_i^* \neq 0\}|\mu_{i1}^T - \mu_{i1}^C|, \end{aligned}$$

where the second equality follows from condition (B1), and the third equality follows from

$$\begin{aligned} \mathbb{1}\{S_i^T >_i S_i^C\}\mathbb{1}\{\mu_{i1}^T > \mu_{i1}^C\}|\mu_{i1}^T - \mu_{i1}^C| &= \mathbb{1}\{S_i^T >_i S_i^C\}\mathbb{1}\{\mu_{i1}^T > \mu_{i1}^C\}\mathbb{1}\{\mu_{i1}^T \neq \mu_{i1}^C\}|\mu_{i1}^T - \mu_{i1}^C| \\ &= \mathbb{1}\{S_i^T >_i S_i^C, \mu_{i1}^T \neq \mu_{i1}^C\}\mathbb{1}\{\mu_{i1}^T > \mu_{i1}^C\}|\mu_{i1}^T - \mu_{i1}^C| \\ &= \mathbb{1}\{S_i^T >_i S_i^C, \mu_{i1}^T \neq \mu_{i1}^C\}|\mu_{i1}^T - \mu_{i1}^C| \\ &= \mathbb{1}\{S_i^T >_i S_i^C\}\mathbb{1}\{\mu_{i1}^T \neq \mu_{i1}^C\}|\mu_{i1}^T - \mu_{i1}^C| \\ &= \mathbb{1}\{S_i^T >_i S_i^C\}|\mu_{i1}^T - \mu_{i1}^C|, \end{aligned}$$

and likewise for $\mathbb{1}\{S_i^T <_i S_i^C\}\mathbb{1}\{\mu_{i1}^T < \mu_{i1}^C\}|\mu_{i1}^T - \mu_{i1}^C|$. Thus, TSLS with interaction I_i^* recovers $\beta_{\text{TSLS}} = \mathbb{E}[w_i^I \bar{\beta}_i]$ with weights

$$w_i^* = \frac{\mathbb{1}\{I_i^* \neq 0\}|\mu_{i1}^T - \mu_{i1}^C|}{\mathbb{E}[\mathbb{1}\{I_i^* \neq 0\}|\mu_{i1}^T - \mu_{i1}^C|]} \geq 0.$$

If $\mu_{i1}^T \neq \mu_{i1}^C$ implies $I_i^* \neq 0$, as with I_i^{sign} from the Normal-Normal setup in the main text (or if $I_i^* \neq 0$ for all i), then $\mathbb{1}\{I_i^* \neq 0\}$ can be omitted from the above expression.

Example B1. If $\mathcal{S} = \{s^L, s^H\}$ are two videos, and $(S_i^C, S_i^T) = (s^L, s^H)$ for all i, then we may have $s^H >_i s^L$ in the sense that video s^H conveys more pessimistic information than video s^L, which makes the order $\geq_i$ on $\mathcal{S} = \{s^L, s^H\}$ common across i. In this case, $I_i^* = 1$ for all i. The same logic applies for Example 2 from the main text.

Passive Control. $\mathcal{S}$ contains the null signal $s^\varnothing$ corresponding to no information, and agents in control receive $S_i^C = s^\varnothing$. The generalized sign interaction

$$I_i^* = \mathbb{1}\{S_i^T >_i s^\varnothing\} - \mathbb{1}\{S_i^T <_i s^\varnothing\}$$

is a valid sign correction for $\mu_{i1}^T - \mu_{i1}^C$, given condition (B1), as above.

Alternatively, in the case where $S_i^T \in \mathbb{R}$ and $\geq_i$ respects the ordering on $\mathbb{R}$, one approach is to assume $S_i^T > \mu_{i0} \implies S_i^T >_i s^\varnothing$ and $S_i^T < \mu_{i0} \implies S_i^T <_i s^\varnothing$. In this case,

$$I_i^{sign} = \mathbb{1}\{S_i^T > \mu_{i0}\} - \mathbb{1}\{S_i^T < \mu_{i0}\} = \text{sign}(S_i^T - \mu_{i0})$$

is a valid sign correction, since

$$\begin{aligned}
I_i^{sign}(\mu_{i1}^T - \mu_{i1}^C) &= \mathbb{1}\{S_i^T > \mu_{i0}\}(\mu_{i1}^T - \mu_{i1}^C) - \mathbb{1}\{S_i^T < \mu_{i0}\}(\mu_{i1}^T - \mu_{i1}^C) \\
&= \mathbb{1}\{S_i^T > \mu_{i0}\}\mathbb{1}\{S_i^T >_i s^\varnothing\}(\mu_{i1}^T - \mu_{i1}^C) - \mathbb{1}\{S_i^T < \mu_{i0}\}\mathbb{1}\{S_i^T <_i s^\varnothing\}(\mu_{i1}^T - \mu_{i1}^C) \\
&= \mathbb{1}\{S_i^T > \mu_{i0}\}\mathbb{1}\{S_i^T >_i s^\varnothing\}\mathbb{1}\{\mu_{i1}^T > \mu_{i1}^C\}|\mu_{i1}^T - \mu_{i1}^C| \\
&+ \mathbb{1}\{S_i^T < \mu_{i0}\}\mathbb{1}\{S_i^T <_i s^\varnothing\}\mathbb{1}\{\mu_{i1}^T < \mu_{i1}^C\}|\mu_{i1}^T - \mu_{i1}^C| \\
&= \mathbb{1}\{S_i^T > \mu_{i0}\}\mathbb{1}\{S_i^T >_i s^\varnothing\}\mathbb{1}\{\mu_{i1}^T \neq \mu_{i1}^C\}|\mu_{i1}^T - \mu_{i1}^C| \\
&+ \mathbb{1}\{S_i^T < \mu_{i0}\}\mathbb{1}\{S_i^T <_i s^\varnothing\}\mathbb{1}\{\mu_{i1}^T \neq \mu_{i1}^C\}|\mu_{i1}^T - \mu_{i1}^C| \\
&= \mathbb{1}\{S_i^T > \mu_{i0}\}|\mu_{i1}^T - \mu_{i1}^C| + \mathbb{1}\{S_i^T < \mu_{i0}\}|\mu_{i1}^T - \mu_{i1}^C| \\
&= \mathbb{1}\{S_i^T \neq \mu_{i0}\}|\mu_{i1}^T - \mu_{i1}^C|,
\end{aligned}$$

where the third and fourth equalities follow from condition (B1) and analogous arguments to the above correction for I_i^*. From here, the perception gap $I_i^{gap} = S_i^T - \mu_{i0}$ is also a valid sign correction. Thus, TSLS with interactions I_i^{sign} and I_i^{gap} recover $\beta_{\text{TSLS}} = \mathbb{E}[w_i^I \bar{\beta}_i]$ with weights

$$w_i^{sign} = \frac{\mathbb{1}\{S_i^T \neq \mu_{i0}\}|\mu_{i1}^T - \mu_{i1}^C|}{\mathbb{E}[\mathbb{1}\{S_i^T \neq \mu_{i0}\}|\mu_{i1}^T - \mu_{i1}^C|]} \geq 0, \qquad w_i^{gap} = \frac{|\mu_{i1}^T - \mu_{i1}^C||S_i^T - \mu_{i0}|}{\mathbb{E}[|\mu_{i1}^T - \mu_{i1}^C||S_i^T - \mu_{i0}|]} \geq 0.$$

The above argument assumes signals that are larger than i's prior mean are also larger than the null signal when viewed from the perspective of i's subjective ordering of the signals. In practice, this requires the content of the quantitative information to be comparable to the values of the prior means; Section V.B from the main text provides a related discussion and example.

B.4 Supplementary Results for Normal-Normal Setup

Consider an experiment with two groups $\mathcal{Z} = \{C, T\}$, where each group has positive assignment probability: $\mathbb{P}(Z_i = T), \mathbb{P}(Z_i = C) > 0$.

As in the main text, agents are Bayesian with Normal prior beliefs $B_{i0} = \mathcal{N}(\mu_{i0}, \sigma_{i0}^2)$. If the experiment provides information s, then agent i assumes $s|\omega \sim \mathcal{N}(\omega, \varsigma_i^2)$, where now agents

differ in perceptions of the signal variance ς_i^2. We have $B_{i1}(\cdot|s) = \mathcal{N}(\mu(B_{i1}(\cdot|s)), \sigma^2(B_{i1}(\cdot|s)))$, where

$$\mu(B_{i1}(\cdot|s)) = r_i s + (1 - r_i)\mu_{i0}, \quad \sigma^2(B_{i1}(\cdot|s)) = (1 - r_i)\sigma_{i0}^2, \quad r_i = \frac{\sigma_{i0}^2}{\sigma_{i0}^2 + \varsigma_i^2}. \tag{B2}$$

If agents maintain their priors when the experiment provides the null signal, $B_{i1}(\cdot|s^\varnothing) = B_{i0}$, then the analysis from the main text is unchanged for both passive and active control cases.

TSLS First-Stage Coefficients. In this Normal-Normal setup, the validity of the passive control interactions $I_i^{1,gap} = (1, S_i^T - \mu_{i0})'$ and $I_i^{1,prior} = (1, \mu_{i0})'$ depends on the relationship between prior means μ_{i0} and variances σ_{i0}^2. Let $\Delta_i = S_i^T - \mu_{i0}$ denote the perception gap.

Proposition B2. *Let Assumption A1(i) and rule (B2) be satisfied with* $(S_i^T, \mu_{i0}) \perp\!\!\!\perp (\sigma_{i0}^2, \varsigma_i^2)$ *and* $\mu_{i1}^C = \mu_{i0}$. *Suppose* $\mathbb{E}[W_i W_i']$ *is full-rank,* $\mathbb{P}(Z_i = z) > 0$ *for each* $z \in \{C, T\}$, *and consider the first-stage population OLS regression*

$$\mu_{i1} = W_i'\delta + \mathbb{1}\{Z_i = T\} I_i'\pi + \zeta_i.$$

Given $I_i = I_i^{1,gap}$, *we have*

$$\begin{bmatrix}\pi_1 \\ \pi_2\end{bmatrix} = \begin{bmatrix}1 & \mathbb{E}[\Delta_i] \\ \mathbb{E}[\Delta_i] & \mathbb{E}[\Delta_i^2]\end{bmatrix}^{-1} \begin{bmatrix}\mathbb{E}[r_i\Delta_i] \\ \mathbb{E}[r_i\Delta_i^2]\end{bmatrix}, \qquad \pi' I_i^{1,gap} = \mathbb{E}[r_i]\Delta_i.$$

On the other hand, given $I_i = I_i^{1,prior}$ *and constant* S_i^T, *we have*

$$\begin{bmatrix}\pi_1 \\ \pi_2\end{bmatrix} = \begin{bmatrix}1 & \mathbb{E}[\mu_{i0}] \\ \mathbb{E}[\mu_{i0}] & \mathbb{E}[\mu_{i0}^2]\end{bmatrix}^{-1} \begin{bmatrix}\mathbb{E}[r_i\Delta_i] \\ \mathbb{E}[r_i\Delta_i\mu_{i0}]\end{bmatrix}, \qquad \pi' I_i^{1,prior} = \mathbb{E}[r_i]\Delta_i.$$

Suppose that signals and signal variances are homogeneous: $(S_i^T, \varsigma_i^2) = (S^T, \varsigma^2)$. In this case, Proposition B2 shows that, when prior means are independent of prior variances, we obtain $\pi' I_i = \mathbb{E}[r_i]\Delta_i$ for both $I_i^{1,gap}$ and $I_i^{1,prior}$. Thus, in this special case, these specifications are both valid sign corrections in the Normal-Normal setup; they recover positive-weighted APEs with weights w_i^{gap} that match those of interaction I_i^{gap}. However, outside of such special cases, $I_i^{1,gap}$ and $I_i^{1,prior}$ can generate negative weights; see Section IV.A of the main text for a simulation. By contrast, I_i^{gap} always produces w_i^{gap}.

Exogenous Information Arrival. Now suppose that, after the experiment provides $s \in \mathbb{R} \cup \{s^\varnothing\}$, the agent receives information $\xi_i|\omega \sim \mathcal{N}(\omega, \tau_i^2)$, regardless of group assignment z, and independent of s conditional on ω. In this case, the passive control corrections are generally invalid, while the active control corrections are robust.

Exogenous Information: Passive Control Case. If $S_i^C = s^\varnothing$ and $S_i^T \in \mathbb{R}$, then rule (B2) implies

$$\mu_{i1}^C = \kappa_i^P \xi_i + (1-\kappa_i^P)\mu_{i0}, \qquad \kappa_i^P = \frac{\sigma_{i0}^2}{\sigma_{i0}^2 + \tau_i^2}$$
$$\mu_{i1}^T = \kappa_i^A \xi_i + (1-\kappa_i^A)[r_i S_i^T + (1-r_i)\mu_{i0}], \qquad \kappa_i^A = \frac{(1-r_i)\sigma_{i0}^2}{(1-r_i)\sigma_{i0}^2 + \tau_i^2}.$$

In general, $\text{sign}(\mu_{i1}^T - \mu_{i1}^C) \neq \text{sign}(S_i^T - \mu_{i0})$. Intuitively, agents in the treatment group, who have already received information from the experiment, will update less on the new exogenous information than if they had been assigned to the control group. Therefore, even though the exogenous information is the same in both groups, the updating occurs at different rates κ_i^A, κ_i^P. The situation is less dire when τ_i^2 is very large, which reflects cases where it is difficult to acquire relevant information; see Section V.B for a related discussion.

Exogenous Information: Active Control Case. If $S_i^C, S_i^T \in \mathbb{R}$, then rule (B2) implies

$$\mu_{i1}^C = \kappa_i^A \xi_i + (1-\kappa_i^A)[r_i S_i^C + (1-r_i)\mu_{i0}],$$
$$\mu_{i1}^T = \kappa_i^A \xi_i + (1-\kappa_i^A)[r_i S_i^T + (1-r_i)\mu_{i0}],$$

so then $\mu_{i1}^T - \mu_{i1}^C = (1-\kappa_i^A) r_i (S_i^T - S_i^C)$. Therefore, $\text{sign}(\mu_{i1}^T - \mu_{i1}^C) = \text{sign}(S_i^T - S_i^C)$, so the active control sign correction continues to be valid. Intuitively, since agents in treatment and control both received an initial signal, they respond to the new exogenous information at rates κ_i^A that do not depend on the group. In conclusion, active control designs are robust to exogenous information arrival in the Normal-Normal setup.

C Proofs for Appendix Results

Proof of Proposition A1.
Consider the problem of choosing $a \in m_i(\mathcal{Y})$ to minimize $\int (\sum_{k=1}^K \theta_{ki}\varphi_k(\omega) - a)^2 dB(\omega)$. This is solved by $a_i(B) = \sum_{k=1}^K \theta_{ki}\phi_k(B)$. Note that the inverse of m_i exists by strict monotonicity. Therefore, for each $y \neq m^{-1}(a_i(B))$, we have

$$\int u_i(\omega, m_i^{-1}(a_i(B)))dB(\omega) < \int u_i(\omega, y)dB(\omega), \quad \forall B \in \mathcal{B}.$$

Thus, $Y_i(B) = m_i^{-1}(\sum_{k=1}^K \theta_{ki}\phi_k(B))$. The inverse function theorem implies that $\Phi \mapsto Y_i(\Phi)$ is continuously differentiable over $\mathbb{R}^K$ with derivatives $\partial_{\phi_k} Y_i(\Phi) = (\partial_y m_i(Y_i(\Phi)))^{-1}\theta_{ki}$. Now we show that $\Phi(\mathcal{B})$ is convex. By definition, for each $\Phi, \Phi' \in \Phi(\mathcal{B})$, there exist $B_\Phi, B_{\Phi'} \in \mathcal{B}$

such that $\Phi = (\phi_k(B_\Phi))_{k=1}^K$ and $\Phi' = (\phi_k(B_{\Phi'}))_{k=1}^K$. By definition of $\mathcal{B}$ and $\phi_k(B)$, we have $(1-\alpha)B_\Phi + \alpha B_{\Phi'} \in \mathcal{B}$ for each $\alpha \in [0,1]$. Thus, $(1-\alpha)\Phi + \alpha\Phi' \in \Phi(\mathcal{B})$, and so Assumption A2 is satisfied. □

Proof of Proposition A2.
By assumption, for all $B \in \mathcal{B}$ we have

$$\int u_i(\omega, y)dB(\omega) = \int u_i(\omega, y)b(\omega)d\nu(\omega) = \int u_i(\omega, y)f(\omega|\Phi(B))d\nu(\omega) = U_i(y, \Phi(B)).$$

Conditions (i)-(iii) allow us to exchange differentiation and integrals (i.e., appealing to Leibniz integral rule—and fundamental theorem of calculus for the dominating functions). Since $Y_i(B)$ satisfies representation (A1), we obtain $\partial_y U_i(Y_i(\Phi), \Phi) = \int \partial_y u_i(\omega, Y_i(\Phi))f(\omega|\Phi)d\nu(\omega) = 0$. Strict concavity implies $\partial_y^2 U_i(y, \Phi) = \int \partial_y^2 u_i(\omega, y)f(\omega|\Phi)d\nu(\omega) < 0$ for all y. By the implicit function theorem,

$$\frac{\partial Y_i(\Phi)}{\partial \phi_k} = \frac{\partial_{\phi_k}\partial_y U_i(Y_i(\Phi), \Phi)}{|\partial_y^2 U_i(Y_i(\Phi), \Phi)|},$$

which is continuous over convex $\Theta = \Phi(\mathcal{B})$. Thus, Assumption A2 is satisfied. □

Proof of Proposition A3.
Analogous to Proposition 1 from the main text (see also Blandhol et al. (2022, Proposition 6)), Assumption A1 implies

$$\begin{aligned}
\beta_{\text{TSLS}}^{z:z'} &= \frac{\mathbb{E}[(\mathbb{1}\{Z_i = z'\} - \mathbb{P}(Z_i = z'|Z_i \in \{z, z'\}))I_i Y_i | Z_i \in \{z, z'\}]}{\mathbb{E}[(\mathbb{1}\{Z_i = z'\} - \mathbb{P}(Z_i = z'|Z_i \in \{z, z'\}))I_i \phi_{ki1} | Z_i \in \{z, z'\}]} \\
&= \frac{\mathbb{E}[I_i Y_i | Z_i = z'] - \mathbb{E}[I_i Y_i | Z_i = z]}{\mathbb{E}[I_i \phi_{ki1} | Z_i = z'] - \mathbb{E}[I_i \phi_{ki1} | Z_i = z]} \times \frac{\text{var}(\mathbb{1}\{Z_i = z'\}|Z_i \in \{z, z'\})}{\text{var}(\mathbb{1}\{Z_i = z'\}|Z_i \in \{z, z'\})} \\
&= \frac{\mathbb{E}[I_i(Y_i(B_{i1}^{z'}) - Y_i(B_{i1}^{z}))]}{\mathbb{E}[I_i(\phi_{ki1}^{z'} - \phi_{ki1}^{z})]}.
\end{aligned}$$

Assumption A2 gives $Y_i(B_{i1}^{z'}) - Y_i(B_{i1}^{z}) = Y_i(\Phi_{i1}^{z'}) - Y_i(\Phi_{i1}^{z})$. □

Proof of Proposition B1.
If $\omega' > \omega$, then $q_i(s'|\omega')/q_i(s'|\omega) \geq q_i(s|\omega')/q_i(s|\omega)$ when $s' >_i s$. In particular, if $B_{i1}(\cdot|s)$ is a Bayesian rule, then $s' >_i s$ implies

$$\frac{b_{i1}(\omega'|s')}{b_{i1}(\omega|s')} = \frac{q_i(s'|\omega')}{q_i(s'|\omega)}\frac{b_{i0}(\omega')}{b_{i0}(\omega)} \geq \frac{q_i(s|\omega')}{q_i(s|\omega)}\frac{b_{i0}(\omega')}{b_{i0}(\omega)} = \frac{b_{i1}(\omega'|s)}{b_{i1}(\omega|s)},$$

which implies $B_{i1}(\cdot|s')$ is greater than $B_{i1}(\cdot|s)$ in the sense of first-order stochastic dominance. In particular, since $\omega \mapsto \varphi(\omega)$ is increasing in ω, then $\phi(B_{i1}(\cdot|s')) \geq \phi(B_{i1}(\cdot|s))$.

If $B_{i1}(\cdot|s)$ is instead an anchored rule, then the form of ϕ implies

$$s \mapsto \phi(B_{i1}(\cdot|s)) = \tau_i \phi(B_i^s) + (1-\tau_i) \int \varphi(\omega) \frac{q_i(s|\omega) b_{i0}(\omega)}{\int q_i(s|\omega) b_{i0}(\omega) d\nu(\omega)} d\nu(\omega).$$

The latter is an attenuated value of ϕ at a Bayesian rule, and so is increasing in $s' >_i s$ by the above arguments from the Bayesian case. The additional assumption for the anchor B_i^s implies $\phi(B_i^s)$ is increasing in $s' >_i s$ as well. Altogether, then, we have $\phi(B_{i1}(\cdot|s')) \geq \phi(B_{i1}(\cdot|s))$ for the anchored case.

If $B_{i1}(\omega|s)$ is instead a Grether rule, then $s' >_i s$ implies

$$\frac{b_{i1}(\omega'|s')}{b_{i1}(\omega|s')} = \left(\frac{q_i(s'|\omega')}{q_i(s'|\omega)}\right)^{\theta_{i1}} \left(\frac{b_{i0}(\omega')}{b_{i0}(\omega)}\right)^{\theta_{i0}} \geq \left(\frac{q_i(s|\omega')}{q_i(s|\omega)}\right)^{\theta_{i1}} \left(\frac{b_{i0}(\omega')}{b_{i0}(\omega)}\right)^{\theta_{i0}} = \frac{b_{i1}(\omega'|s)}{b_{i1}(\omega|s)},$$

which implies $B_{i1}(\cdot|s')$ is greater than $B_{i1}(\cdot|s)$ in the sense of first-order stochastic dominance, as in the Bayesian case. Therefore, $\phi(B_{i1}(\cdot|s')) \geq \phi(B_{i1}(\cdot|s))$ for the Grether case. □

Proof of Proposition B2.
The first-stage coefficients are identified and

$$\begin{aligned}\pi &= \mathbb{E}[I_i I_i']^{-1}(\mathbb{E}[I_i \mu_{i1}|Z_i = T] - \mathbb{E}[I_i \mu_{i1}|Z_i = C]) \\ &= \mathbb{E}[I_i I_i']^{-1}\mathbb{E}[I_i r_i \Delta_i],\end{aligned}$$

where the second equality follows from updating rule (B2), $\mu_{i1}^C = \mu_{i0}$, and IV independence.

Therefore, given $I_i = I_i^{1,gap}$ and $(S_i^T, \mu_{i0}) \perp\!\!\!\perp (\sigma_{i0}^2, \varsigma_i^2)$, we have

$$\begin{aligned}\pi' I_i^{1,gap} &= \begin{bmatrix} 1 \\ \Delta_i \end{bmatrix}' \mathrm{var}(\Delta_i)^{-1} \begin{bmatrix} \mathbb{E}[\Delta_i^2] & -\mathbb{E}[\Delta_i] \\ -\mathbb{E}[\Delta_i] & 1 \end{bmatrix} \begin{bmatrix} \mathbb{E}[\Delta_i] \\ \mathbb{E}[\Delta_i^2] \end{bmatrix} \mathbb{E}[r_i] \\ &= \begin{bmatrix} 1 \\ \Delta_i \end{bmatrix}' \mathrm{var}(\Delta_i)^{-1} \begin{bmatrix} 0 \\ \mathrm{var}(\Delta_i) \end{bmatrix} \mathbb{E}[r_i] \\ &= \mathbb{E}[r_i]\Delta_i\end{aligned}$$

On the other hand, given $I_i = I_i^{1,prior}$, $(S_i^T, \mu_{i0}) \perp\!\!\!\perp (\sigma_{i0}^2, \varsigma_i^2)$, and $S_i^T = S^T$ constant, we have

$$\pi' I_i^{1,prior} = \begin{bmatrix} 1 \\ \mu_{i0} \end{bmatrix}' \mathrm{var}(\mu_{i0})^{-1} \begin{bmatrix} \mathbb{E}[\mu_{i0}^2] & -\mathbb{E}[\mu_{i0}] \\ -\mathbb{E}[\mu_{i0}] & 1 \end{bmatrix} \begin{bmatrix} S^T - \mathbb{E}[\mu_{i0}] \\ S^T\mathbb{E}[\mu_{i0}] - \mathbb{E}[\mu_{i0}^2] \end{bmatrix} \mathbb{E}[r_i]$$

$$= \text{var}(\mu_{i0})^{-1} \begin{bmatrix} \mathbb{E}[\mu_{i0}^2] - \mu_{i0}\mathbb{E}[\mu_{i0}] \\ \mu_{i0} - \mathbb{E}[\mu_{i0}] \end{bmatrix}' \begin{bmatrix} S^T - \mathbb{E}[\mu_{i0}] \\ S^T\mathbb{E}[\mu_{i0}] - \mathbb{E}[\mu_{i0}^2] \end{bmatrix} \mathbb{E}[r_i]$$

$$= \mathbb{E}[r_i]\Delta_i.$$

□

D Empirical Applications

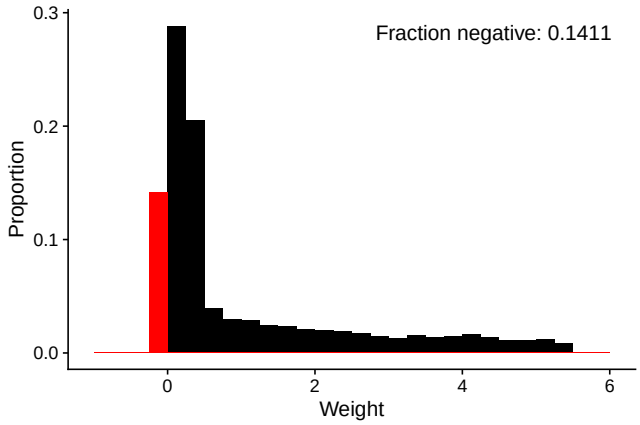
D.1 Interpreting the Simulation Results

In this section, we examine which agents have negative weights, and why negative weights result in an attenuated estimated coefficient. We follow the simulation design from Section IV.A of the main text. Because the partial effects $\partial Y_i(\mu)/\partial\mu = \exp(-3\mu_{i0})$ are homogeneous across μ, TSLS with interaction I_i recovers

$$\beta_{\text{TSLS}} = \mathbb{E}[w_i^I \exp(-3\mu_{i0})], \quad w_i^I = \frac{\pi' I_i(\mu_{i1}^T - \mu_{i1}^C)}{\mathbb{E}[\pi' I_i(\mu_{i1}^T - \mu_{i1}^C)]} = \frac{\pi' I_i r_i(1-\mu_{i0})}{\mathbb{E}[\pi' I_i r_i(1-\mu_{i0})]} \tag{D1}$$

We characterize which agents have negative weights by using formula (D1). First, we can examine the distribution of weights from the contaminated interactions $I_i^{1,gap}$ and $I_i^{1,prior}$ in Figure D1. 14% of agents have negative weight; these weights are close to 0, relative to the positive right tail.

Figure D1: Histogram of weights $w_i^{1,gap}$ or $w_i^{1,prior}$

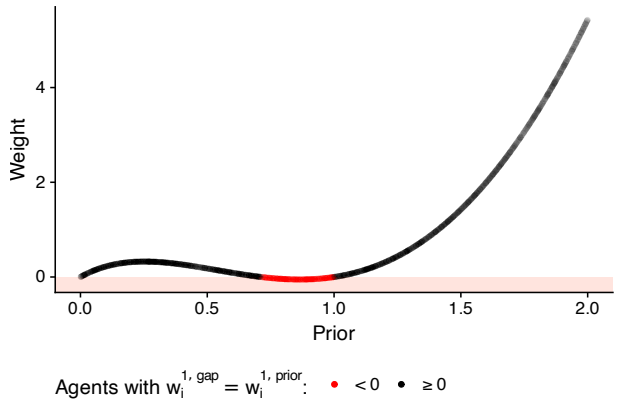


Note: This figure plots the distribution of weights produced by the TSLS specification corresponding to the $I_i^{1,gap}$ or $I_i^{1,prior}$ interactions. Negative weights are highlighted in red. $I_i^{1,gap}$ and $I_i^{1,prior}$ produce the same weights, since the signal is constant. The formula for the weights is in formula (D1).

Which agents have negative weights? Examine the relationship between the prior and

the weights in Figure D2; agents with negative weights are highlighted in red. Agents with negative weights are concentrated among those whose priors are below and near the signal $S_i^T = 1$.

Figure D2: Prior vs. weight $w_i^{1,gap}$

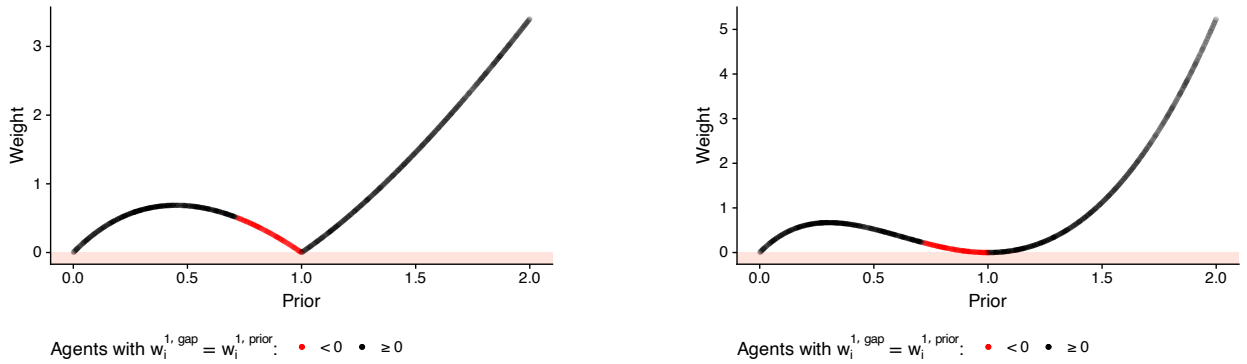


Note: This figure is a scatterplot of the prior against the weights produced by the TSLS specification corresponding to the $I_i^{1,gap}$ or $I_i^{1,prior}$ interactions. Negative weights are highlighted in red. $I_i^{1,gap}$ and $I_i^{1,prior}$ produce the same weights, since the signal is constant. The formula for the weights is in formula (D1).

In contrast, these agents do not have negative weight when using I_i^{sign} and I_i^{gap}. The relationship between the prior and weights in the uncontaminated specifications are shown in Figure D3; agents who had negative weights in the contaminated specifications are highlighted in red. Qualitatively, the relationship between the weights and the priors are similar in I_i^{gap} and $I_i^{1,gap}$. The weights for those with priors less than the signal $S_i^T = 1$ are simply "shifted up" from Figure D2 to Figure D3 such that there are no more negative weights. In contrast, the weights produced by I_i^{sign} in Figure D3 are qualitatively different from those produced by I_i^{gap}. In particular, w_i^{sign} ranges from 0 to 3.5, whereas w_i^{gap} ranges from 0 to 5.5, due to the up-weighting of those with the largest perception gaps. Up-weighting leads the relative contributions of those with large perception gaps to be greater, especially for agents who substantially update their beliefs—here, agents with large priors.

How do these negative weights generate differences in the coefficients? Figure D4 plots each agent's partial effect against their prior. The action function $Y_i(\mu) = \exp(-3\mu_{i0}) \times \mu$ generates positive partial effects for all agents. However, the agents with negative weights enter *negatively* into the weighted average, thus attenuating the weighted average partial effect. Moreover, because this action function generates partial effects that are exponentially decreasing in the prior, the agents with negative weights have relatively large partial effects, compared to those with priors above the signal. Therefore, the attenuation is substantial, generating large differences between the contaminated and uncontaminated estimated weighted average partial effects.

Figure D3: Prior vs. weight in uncontaminated specifications

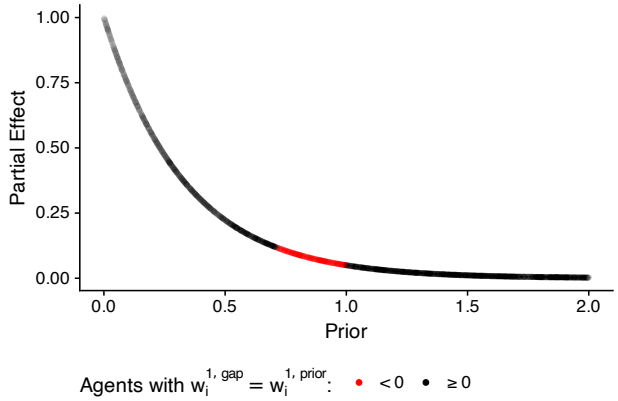


(a) Prior vs. w_i^{sign}

(b) Prior vs. w_i^{gap}

Note: This figure shows scatterplots of the prior against the weights produced by the TSLS specification corresponding to the I_i^{sign} and I_i^{gap} interactions, respectively. Agents who have negative weights when the contaminated $I_i^{1,gap}$ or $I_i^{1,prior}$ specifications are used are highlighted in red. The formula for the weights is in formula (D1).

Figure D4: Prior vs. partial effect

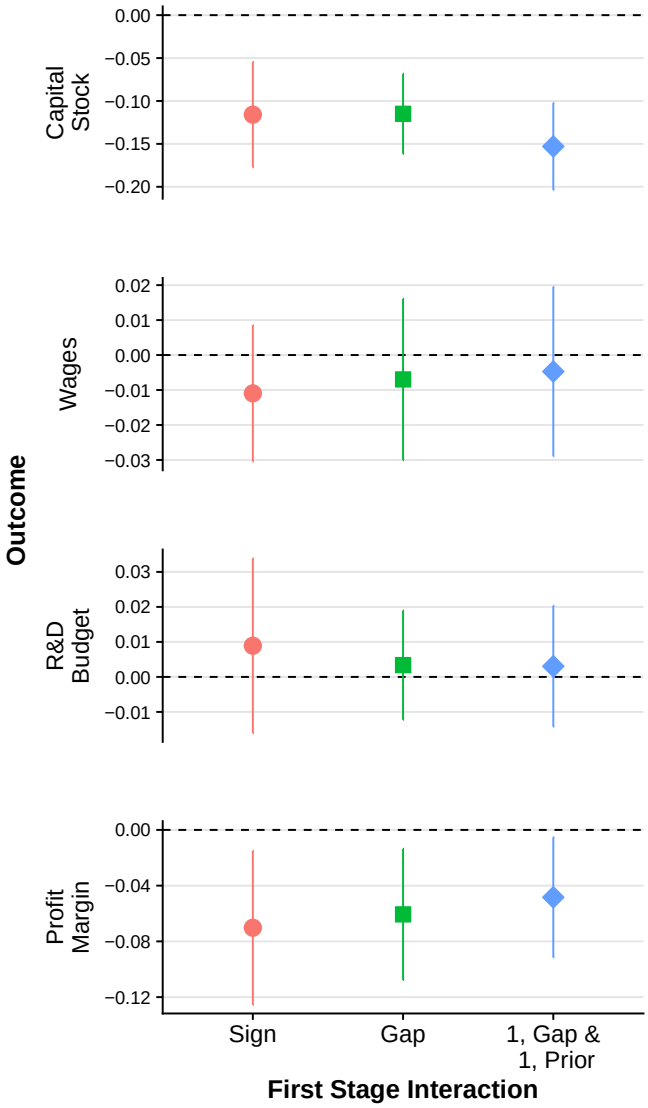


Note: This figure shows a scatterplot of the prior against the partial effect $\bar{\beta}_i = \exp(-3\mu_{i0})$ for each agent. Agents who have negative weights when the contaminated $I_i^{1,gap}$ or $I_i^{1,prior}$ specifications are used are highlighted in red. The formula for the weights is in formula (D1).

D.2 Additional Estimated Coefficients for Kumar, Gorodnichenko, and Coibion (2023b)

Figures D5 displays coefficients from the Kumar, Gorodnichenko, and Coibion (2023b) application that are not shown in the main text.

Figure D5: Estimates of weighted average partial effects using data from Kumar, Gorodnichenko, and Coibion (2023b)



Note: The figure presents point estimates and 95% confidence intervals for the four prominent passive control specifications in the literature using data from two arms in Kumar, Gorodnichenko, and Coibion (2023b). The outcomes are defined as the difference between the planned change and the actualized change in the outcome variable over six months, i.e., the planned change in price versus the actualized changed in price. The outcome variables are the capital stock of the firm, the wages of the firm, the R&D budget of the firm, and the profit margin of the firm. The feature of interest is GDP growth expectations, and the signal is 4% GDP growth. Sign refers to I_i^{sign}, which regresses the outcome on the sign of the perception gap and posterior GDP growth expectations, instrumenting the posterior by the treatment indicator times the sign of the perception gap $S_i^T - \mu_{i0}$. Gap refers to I_i^{gap}, which regresses the outcome on the perception gap and the posterior, instrumenting the posterior by the treatment indicator times the perception gap. "1, Gap" refers to $I_i^{1,gap}$, which regresses the outcome on the perception gap and the posterior, instrumenting the posterior by the treatment indicator and the treatment indicator times the perception gap. "1, Prior" refers to $I_i^{1,prior}$, which regresses the outcome on the prior and the posterior, instrumenting the posterior by the treatment indicator and the treatment indicator times the prior. $I_i^{1,gap}$ and $I_i^{1,prior}$ produce the same estimate because the signal is constant. In all specifications, the coefficient of interest is the coefficient on the posterior expectation.

D.3 Application to Jäger et al. (2024)

Jäger et al. (2024) study how information about workers' outside options—that is, workers' wages if they left their current job to find a new one—affects their labor market decisions. First, Jäger et al. (2024) elicit workers' prior expectations of their outside options. Next, they split their sample into a control group and a single treatment group. The control arm receives no information; those in the treatment group are told the mean wage of workers similar to themselves (based on gender, age, occupation, labor market region, and education level). Then, Jäger et al. (2024) elicit workers' posterior expectations of their outside options.

Figure D6 displays estimated coefficients from all four passive control interactions on the outcomes intended negotiation probability, intended quit probability, intended search probability, intended negotiation magnitude where no negotiation is coded as 0, intended negotiation magnitude where no negotiation is coded as missing, and reservation wage cut. For most outcomes, there are few systematic differences across the specifications.

In general, the difference in the estimated coefficient for I_i^{sign} versus I_i^{gap} tends to be larger than the difference between I_i^{gap} and the contaminated interactions $I_i^{1,gap}$ and $I_i^{1,prior}$. For example, the I_i^{sign} coefficient for the intended negotiation probability outcome is 30-40% larger (though not statistically significantly so) than the coefficients from interactions I_i^{gap}, $I_i^{1,gap}$, and $I_i^{1,prior}$, which are similar to each other. This pattern indicates that in this setting, the up-weighting of workers with larger perception gaps may have a greater impact on estimates than possible negative weights.

We can further investigate effects of up-weighting by characterizing the average weight that TSLS gives to workers for each decile of the perception gap: $\mathbb{E}[w_i^I|(S_i^T - \mu_{i0}) \in [a_j, b_j]]$, where a_j, b_j are the bounds of the deciles and w_i^I denotes the weights on worker i when the interaction is I_i. Given $X_i \perp\!\!\!\perp Z_i$ and Assumption 1 from the main text, we have

$$\mathbb{E}[w_i^I f(X_i)] = \frac{\mathbb{E}[I_i' \pi f(X_i) \mu_{i1} | Z_i = T] - \mathbb{E}[I_i' \pi f(X_i) \mu_{i1} | Z_i = C]}{\mathbb{E}[I_i' \pi \mu_{i1} | Z_i = T] - \mathbb{E}[I_i' \pi \mu_{i1} | Z_i = C]}, \tag{D2}$$

where f is some function of X_i. For example, if $X_i \in [x_{min}, x_{max}]$, then

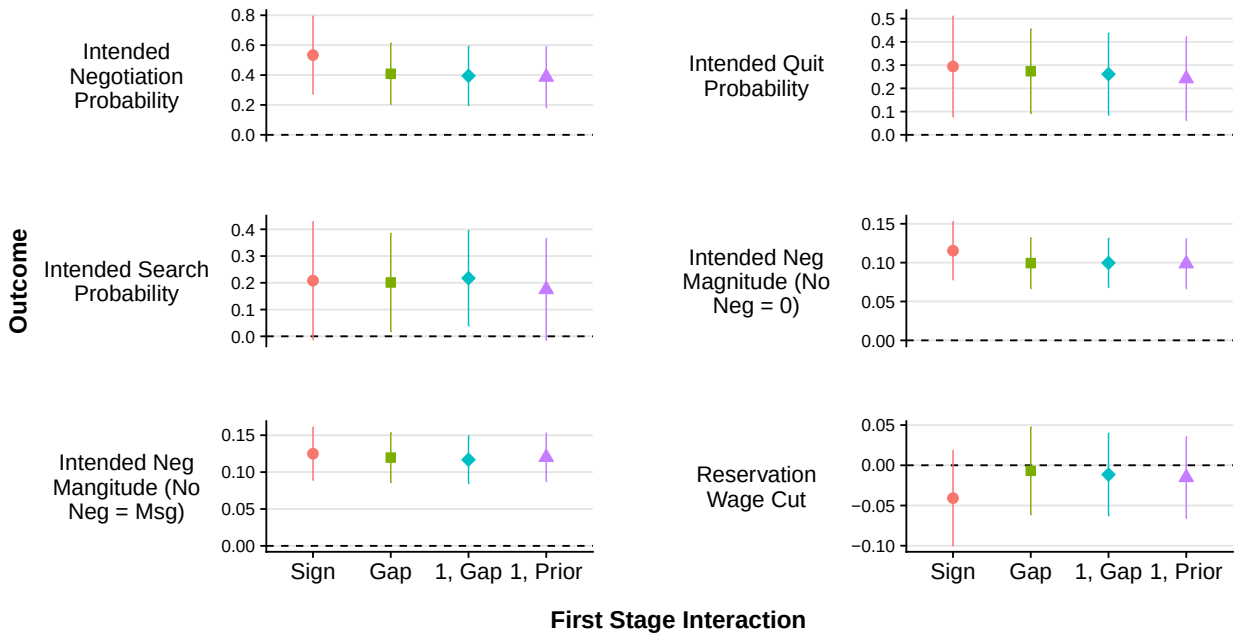
$$f_{a,b}(X_i) := \frac{\mathbb{1}\{X_i \in [a,b]\}}{\mathbb{P}(X_i \in [a,b])}, \quad x_{min} \leq a < b \leq x_{max},$$

gives

$$\mathbb{E}[w_i^I f_{a,b}(X_i)] = \mathbb{E}[w_i^I | X_i \in [a,b]].$$

In this example, if we have a collection of intervals $[a_j, b_j]$ such that $\cup_{j=1}^J [a_j, b_j] = [x_{min}, x_{max}]$, then ranging $\mathbb{E}[w_i^I f_{a_j,b_j}(X_i)]$ over $j = 1, \ldots, J$ allows us to compute the average weight of

Figure D6: Estimates of weighted average partial effects using data from Jäger et al. (2024)



Note: The figure presents point estimates and 95% confidence intervals for the four prominent passive control specifications in the literature using data from Jäger et al. (2024). The outcomes are the worker's intended probability of negotiating with their current employer, intended probability of quitting their current job, intended probability of finding another job, intended negotiation magnitude (no negotiations coded as 0), intended negotiation magnitude (no negotiations coded as missing), and the reservation wage cut as a percent of their current wage. The signal is the mean wage of similar workers, and the feature of interest is workers' expectations of how much their wages would change, as a percentage of their current wage, if they switched jobs. Sign refers to interaction I_i^{sign}, which regresses the outcome on the sign of the perception gap and the posterior, instrumenting the posterior by the treatment indicator times the sign of the perception gap. Gap refers to interaction I_i^{gap}, which regresses the outcome on the perception gap and the posterior, instrumenting the posterior by the treatment indicator times the perception gap. "1, Gap" refers to $I_i^{1,gap}$, which regresses the outcome on the perception gap and the posterior, instrumenting the posterior by the treatment indicator and the treatment indicator times the perception gap. Following Jäger et al. (2024), we normalize agents' perception gaps $S_i^T - \mu_{i0}$ to be a percentage of S_i^T for I_i^{gap} and $I_i^{1,gap}$. "1, Prior" refers to $I_i^{1,prior}$, which regresses the outcome on the prior and the posterior, where the posterior is instrumented by the treatment indicator, the treatment indicator times the prior, and the treatment indicator times the signal. We estimate the version of interaction $I_i^{1,prior}$ that includes both the signal and the prior because signals are personalized to each worker—see the discussion under Interaction 5 of the main text. In all specifications, the coefficient of interest is the coefficient on the posterior expectation.

group $[a_j, b_j]$. We can also consider

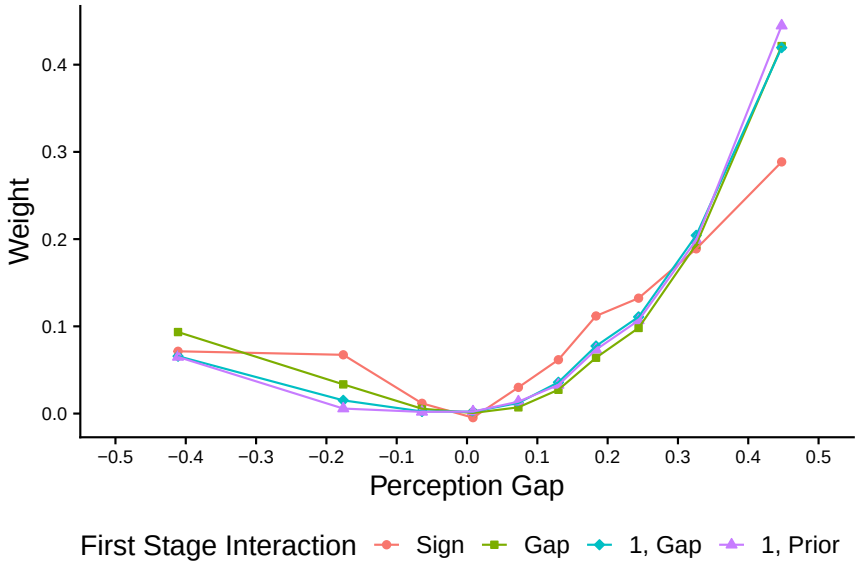
$$\beta_{\text{TSLS}} = \mathbb{E}[w_i^I \bar{\beta}_i] = \sum_{j=1}^{J} \mathbb{P}(X_i \in [a_j, b_j]) \mathbb{E}[w_i^I \bar{\beta}_i | X_i \in [a_j, b_j]],$$

where each $\mathbb{E}[w_i^I \bar{\beta}_i | X_i \in [a_j, b_j]]$ is identified as

$$\mathbb{E}[w_i^I \bar{\beta}_i | X_i \in [a_j, b_j]] = \frac{\mathbb{E}[I_i' \pi f_{a_j,b_j}(X_i) Y_i | Z_i = T] - \mathbb{E}[I_i' \pi f_{a_j,b_j}(X_i) Y_i | Z_i = C]}{\mathbb{E}[I_i' \pi \mu_{i1} | Z_i = T] - \mathbb{E}[I_i' \pi \mu_{i1} | Z_i = C]}.$$

Figure D7 plots the weights against the perception gap. Compared to interactions I_i^{gap}, $I_i^{1,gap}$, and $I_i^{1,prior}$, I_i^{sign} places relatively less weight on those in the lowest decile, and relatively more weight on those in the middle deciles. Therefore, if workers with moderate perception gaps have larger within-agent APEs $\bar{\beta}_i$, then the I_i^{sign} coefficient will be larger than those for interactions I_i^{gap}, $I_i^{1,gap}$, and $I_i^{1,prior}$, which would rationalize the estimates in Figure D6.

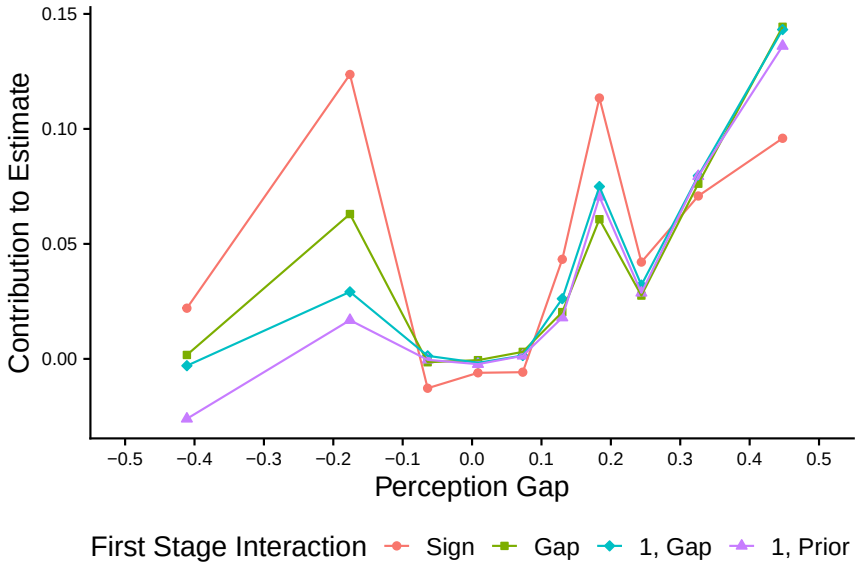
Figure D7: Perception Gap Characterization of Weights—Data from Jäger et al. (2024)



Note: Each point represents the characterized TSLS weight for the corresponding decile bin, using intended negotiation probability as the outcome. For example, the leftmost point corresponds to the characterized weight for those in the lowest decile bin, (-0.572,-0.228]. For a description of the characterization procedure, see the text.

We can also characterize the contribution of each decile of perception gap to the estimated TSLS coefficients; for this exercise, we focus on the intended negotiation probability outcome. Denote the contribution as $\mathbb{E}[w_i^I \bar{\beta}_i | (S_i^T - \mu_{i0}) \in [a_j, b_j]]$. Figure D8 plots these contributions against the perception gaps. Comparing the shape of the plot in Figure D8 against that of Figure D7 shows where the differences in the APEs are. For example, small differences in Figure D7 weight contributions across the interactions for the lowest decile translates to large differences in the analogous Figure D8 TSLS estimate contributions. This is evidence of heterogeneous partial effects—if these effects were constant, then the relative location of each point in Figures D7 and D8 would be identical. Comparing the shape of the plot in

Figure D8: Perception Gap Characterization of TSLS—Data from Jäger et al. (2024)



Note: Each point represents the characterized TSLS estimate for the corresponding decile bin, using intended negotiation probability as the outcome. For example, the leftmost point corresponds to the characterized estimate for those in the lowest decile bin, (-0.572,-0.228]. For a description of the characterization procedure, see the text.

Figure D8 against that of Figure D7 suggests where large APEs are: Across all specifications, the relative contributions of the second and seventh deciles are larger than their relative weight. These deciles have perception gaps of (-0.228, -0.0931] and (0.162, 0.216], respectively. Since the contribution is a weighted average of workers' weights and APEs, the APEs of workers in the second and seventh deciles must be large as well. That is, workers whose priors are 10%-20% off from the signal are the more likely to change their intended negotiation probability following treatment. In contrast, workers in the lowest decile bin of (-0.572, -0.228] have smaller contributions relative to their weight. Therefore, their APEs are relatively smaller than other workers.

One explanation for why workers whose priors are near—but not very close to—the signal are the most likely to change their intended negotiation probability following treatment is endogenous information acquisition, such as in Balla-Elliott (2023). Workers who are more responsive to information are more likely to seek it out, and therefore have more accurate priors. At the same time, workers that are very accurate will not respond much to the information treatment; their actions are already near-optimal. Therefore, the workers that are most responsive are those that have medium-sized perception gaps, and the workers that are least responsive are those that have large or no perception gap. Interactions I_i^{gap}, $I_i^{1,gap}$, and $I_i^{1,prior}$ will up-weight workers with large perception gaps—and thus those who are least responsive—attenuating causal effects towards zero.

References

Alesina, Alberto, Matteo F Ferroni, and Stefanie Stantcheva. 2021. *Perceptions of racial gaps, their causes, and ways to reduce them.* Technical report. National Bureau of Economic Research.

Angrist, Joshua D, Kathryn Graddy, and Guido W Imbens. 2000. "The interpretation of instrumental variables estimators in simultaneous equations models with an application to the demand for fish." *The Review of Economic Studies* 67 (3): 499–527.

Armantier, Olivier, Scott Nelson, Giorgio Topa, Wilbert Van der Klaauw, and Basit Zafar. 2016. "The price is right: Updating inflation expectations in a randomized price information experiment." *Review of Economics and Statistics* 98 (3): 503–523.

Armona, Luis, Andreas Fuster, and Basit Zafar. 2019. "Home price expectations and behaviour: Evidence from a randomized information experiment." *The Review of Economic Studies* 86 (4): 1371–1410.

Balla-Elliott, Dylan. 2023. "Identifying Causal Effects in Information Provision Experiments." *arXiv preprint arXiv:2309.11387.*

Balla-Elliott, Dylan, Zoë B Cullen, Edward L Glaeser, Michael Luca, and Christopher Stanton. 2022. "Determinants of small business reopening decisions after COVID restrictions were lifted." *Journal of Policy Analysis and Management* 41 (1): 278–317.

Benjamin, Daniel J. 2019. "Errors in probabilistic reasoning and judgment biases." *Handbook of Behavioral Economics: Applications and Foundations 1* 2:69–186.

Blandhol, Christine, John Bonney, Magne Mogstad, and Alexander Torgovitsky. 2022. *When is TSLS actually late?* Technical report. National Bureau of Economic Research.

Casella, George, and Roger L Berger. 2021. *Statistical inference.* Cengage Learning.

Cavallo, Alberto, Guillermo Cruces, and Ricardo Perez-Truglia. 2017. "Inflation expectations, learning, and supermarket prices: Evidence from survey experiments." *American Economic Journal: Macroeconomics* 9 (3): 1–35.

Clippel, Geoffroy de, and Xu Zhang. 2022. "Non-bayesian persuasion." *Journal of Political Economy* 130 (10): 2594–2642.

Coibion, Olivier, Dimitris Georgarakos, Yuriy Gorodnichenko, Geoff Kenny, and Michael Weber. 2021. *The effect of macroeconomic uncertainty on household spending.* Technical report. National Bureau of Economic Research.

Coibion, Olivier, Yuriy Gorodnichenko, and Michael Weber. 2022. "Monetary policy communications and their effects on household inflation expectations." *Journal of Political Economy* 130 (6): 1537–1584.

Cullen, Zoë, and Ricardo Perez-Truglia. 2022. "How much does your boss make? The effects of salary comparisons." *Journal of Political Economy* 130 (3): 766–822.

Dechezleprêtre, Antoine, Adrien Fabre, Tobias Kruse, Bluebery Planterose, Ana Sanchez Chico, and Stefanie Stantcheva. 2022. *Fighting climate change: International attitudes toward climate policies.* Technical report. National Bureau of Economic Research.

Edwards, Ward. 1968. "Conservatism in human information processing." *Formal representation of human judgment.*

Fuster, Andreas, Ricardo Perez-Truglia, Mirko Wiederholt, and Basit Zafar. 2022. "Expectations with endogenous information acquisition: An experimental investigation." *Review of Economics and Statistics* 104 (5): 1059–1078.

Gabaix, Xavier. 2019. "Behavioral inattention." In *Handbook of behavioral economics: Applications and foundations 1,* 2:261–343. Elsevier.

Grether, David M. 1980. "Bayes rule as a descriptive model: The representativeness heuristic." *The Quarterly journal of economics* 95 (3): 537–557.

Haaland, Ingar, Christopher Roth, and Johannes Wohlfart. 2023. "Designing information provision experiments." *Journal of economic literature* 61 (1): 3–40.

Jäger, Simon, Christopher Roth, Nina Roussille, and Benjamin Schoefer. 2023. *Replication Data for: "Worker Beliefs about Outside Options".* V1, UNF:6:lDC6uX9GMUFs61mxcLifHg== [fileUNF]. https://doi.org/10.7910/DVN/DCSR0N. https://doi.org/10.7910/DVN/DCSR0N.

———. 2024. "Worker beliefs about outside options." *The Quarterly Journal of Economics* 139 (3): 1505–1556.

Kumar, Saten, Yuriy Gorodnichenko, and Olivier Coibion. 2023a. *Supplement to "The Effect of Macroeconomic Uncertainty on Firm Decisions".* Econometrica. https://www.econometricsociety.org/publications/econometrica/2023/07/01/The-Effect-of-Macroeconomic-Uncertainty-on-Firm-Decisions/supp/21004_Data_and_Programs.zip.

———. 2023b. "The effect of macroeconomic uncertainty on firm decisions." *Econometrica* 91 (4): 1297–1332.

Lehmann, Erich L, and George Casella. 2006. *Theory of point estimation.* Springer Science & Business Media.

Mogstad, Magne, Alexander Torgovitsky, and Christopher R Walters. 2021. "The causal interpretation of two-stage least squares with multiple instrumental variables." *American Economic Review* 111 (11): 3663–3698.

Tversky, Amos, and Daniel Kahneman. 1974. "Judgment under Uncertainty: Heuristics and Biases: Biases in judgments reveal some heuristics of thinking under uncertainty." *science* 185 (4157): 1124–1131.