Online Appendix:
Reviewing Procedure vs. Judging Substance: How Increasing Bureaucratic Oversight Can Reduce Bureaucratic Accountability

Journal of Political Institutions and Political Economy

Ian R. Turner
Department of Political Science
Yale University
ian.turner@yale.edu

Contents

A Procedural review model

A.1 Equilibrium policy choice

Lemma A.1. *In the procedural review model the agency always matches policy to the state in weakly undominated pure strategies: $x_A^P(\omega) = \omega$.*

Proof of Lemma A.1. At the point in the game at which the agency makes its substantive policy choice, x_A, its effort investment e is a sunk cost. Thus, e and $V_\varepsilon(e)$ are fixed. Additionally, since x_A is not observed by the overseer the overseer's review decision is invariant to the agency's choice. Thus, there are two cases to check: (1) agency will be overturned and (2) agency will be upheld.
Case 1: Agency overturned. The agency's payoff in this case is equivalent regardless of its policy choice since the overseer's review decision is unaffected by the agency's choice of x_A. Thus, the agency has no reason to deviate from setting policy to match the state.
Case 2: Agency upheld. The agency's expected payoff for the proposed strategy is given by,

$$\begin{aligned} EU_A(x_A^P(\omega) = \omega | e, r = 0) &= -(\omega - y)^2 - \kappa e - \pi r, \\ &= -(\omega - (1)(\omega + \varepsilon(e)))^2 - \kappa e, \\ &= -\mathbb{E}[\varepsilon]^2 - V_\varepsilon(e) - \kappa e, \\ &= -V_\varepsilon(e) - \kappa e. \end{aligned}$$

Now suppose the agency deviated by choosing $x_A(\omega) = \omega + 1$ ($x_A(\omega) = \omega - 1$ is similar). Its expected payoff for doing so is given by,

$$\begin{aligned} EU_A(x_A(\omega) = \omega + 1 | e, r = 0) &= -(\omega - (\omega + 1 + \varepsilon(e)))^2 - \kappa e, \\ &= -(\omega - (\omega + 1))^2 - \mathbb{E}[\varepsilon]^2 - V_\varepsilon(e) - \kappa e, \\ &= -1 - V_\varepsilon(e) - \kappa e. \end{aligned}$$

Thus, the net expected utility for deviation is given by,

$$\begin{aligned} \Delta EU_A(x_A(\omega) = \omega + 1 | e, r = 0) &= -1 - V_\varepsilon(e) - \kappa e + V_\varepsilon(e) + \kappa e, \\ &= -1, \end{aligned}$$

implying a net utility loss equal to the policy choice deviation. Thus, the agency is strictly worse off by deviating from the proposed strategy when the overseer will uphold the agency. Taken together these two cases imply that, in weakly undominated pure strategies, the agency will always choose $x_A^P(\omega) = \omega$ in the procedural review model. ■

A.2 Equilibrium oversight

Lemma A.2. *The overseer's optimal review strategy in the procedural review model is,*

$$s_R(e) = \begin{cases} \textit{Uphold: } r=0 & \textit{if } V_\varepsilon(0) - V_\varepsilon(e) \geq p_1(2\beta-1) + p_\theta(2\beta\theta - \theta^2), \\ \textit{Overturn: } r=1 & \textit{otherwise.} \end{cases}$$

Proof of Lemma A.2. Note that from Lemma A.1 we have that $x_A^*(\omega) = \omega$. First, consider the overseer's expected payoff for upholding the agency following a choice of e:

$$\begin{aligned} EU_R(r=0|e,\beta,x_A^*) &= -(\omega - \beta - (x_A^* + \varepsilon(e)))^2, \\ &= -(\omega-\beta-\omega)^2 - \mathbb{E}[\varepsilon]^2 - V_\varepsilon(e), \\ &= -\beta^2 - V_\varepsilon(e). \end{aligned}$$

Now, the overseer's expected payoff for reversing the agency depends on the state ω, which is unknown to the overseer in the procedural review model. For any given ω the overseer's expected payoff for reversal is given by:

$$\begin{aligned} EU_R(r=1|e,\beta,\omega) &= -(\omega-\beta-(0+\varepsilon(0)))^2, \\ &= -(\omega-\beta)^2 - \mathbb{E}[\varepsilon]^2 - V_\varepsilon(0), \\ &= -(\omega-\beta)^2 - V_\varepsilon(0). \end{aligned}$$

Plugging this in for each ω, scaled by the overseer's beliefs about each state having obtained, which is given by $\mathbf{p} = \{p_0, p_1, p_\theta\}$, we have the overseer's overall expected payoff for overturning:

$$\begin{aligned} EU_R(r=1|e,\beta,\mathbf{p}) &= -p_0\left((0-\beta)^2 + V_\varepsilon(0)\right) - p_1\left((1-\beta)^2 + V_\varepsilon(0)\right) - p_\theta\left((\theta-\beta)^2 + V_\varepsilon(0)\right), \\ &= -p_0\beta^2 - p_1(1-\beta)^2 - p_\theta(\theta-\beta)^2 - V_\varepsilon(0). \end{aligned}$$

Incentive compatibility requires that the overseer uphold ($r=0$) if and only if:

$$\begin{aligned} &EU_R(r=0|e,\beta,x_A^*) \geq EU_R(r=1|e,\beta,\mathbf{p}), \\ &\quad -\beta^2 - V_\varepsilon(e) \geq -p_0\beta^2 - p_1(1-\beta)^2 - p_\theta(\theta-\beta)^2 - V_\varepsilon(0), \\ &\quad -\beta^2 - V_\varepsilon(e) \geq -p_0\beta^2 - p_1 + 2\beta p_1 - p_1\beta^2 - p_\theta\beta^2 + 2\beta\theta p_\theta - p_\theta\theta^2 - V_\varepsilon(0), \\ &V_\varepsilon(0) - V_\varepsilon(e) \geq p_1(2\beta-1) + p_\theta(2\beta\theta - \theta^2), \end{aligned}$$

as stated in the result. ■

Now, recall the definitions derived from the overseer's incentive compatibility constraint to uphold. That is, it must be the case that $\beta \in \left(0, \frac{p_1 + p_\theta \theta^2 + V_\varepsilon(0) - V_\varepsilon(e)}{2(p_1 + p_\theta \theta)}\right]$ for the overseer to uphold. We can define two β-thresholds based on whether the agency invested high or low effort: $\beta_1 \equiv \frac{p_1 + p_\theta \theta^2 + V_\varepsilon(0) - V_\varepsilon(1)}{2(p_1 + p_\theta \theta)}$ and $\beta_0 \equiv \frac{p_1 + p_\theta \theta^2}{2(p_1 + p_\theta \theta)}$ where $\beta_0 < \beta_1$ since $V_\varepsilon(1) < V_\varepsilon(0)$.

If $\beta < \beta_1 < \beta_0$ then the overseer always upholds and is *perfectly deferential*. If $\beta_1 < \beta_0 < \beta$ then the overseer always overturns and is *perfectly skeptical*. If $\beta_1 < \beta < \beta_0$ then the overseer upholds if and only if $e = 1$ and is *conditionally deferential*. The next result characterizes how the agency best responds with its effort choices conditional on these oversight regimes.

A.3 Equilibrium effort

Lemma A.3. *Conditional on the overseer's bias β, the agency invests effort as follows:*

1. *If $\beta < \beta_1 < \beta_0$ then the overseer is perfectly deferential and the agency invests high effort if $V_\varepsilon(0) - V_\varepsilon(1) \geq \kappa$.*
2. *If $\beta_1 < \beta_0 < \beta$ then the overseer is perfectly skeptical and the agency never invests high effort.*
3. *If $\beta_1 < \beta < \beta_0$ then the overseer is conditionally deferential and the agency invests high effort if $p_1 + p_\theta \theta^2 + V_\varepsilon(0) - V_\varepsilon(1) + \pi \geq \kappa$.*

Proof of Lemma A.3. I proceed by deriving the agency's incentive compatibility conditions to invest high effort given the type of review it is facing.

Case 1: $\beta < \beta_0 < \beta_1$, perfect deference. In this case the agency knows that it will be upheld regardless of its choice of e. The agency's expected payoff, given it will be upheld for sure, for investing low effort is given by,

$$\begin{aligned} EU_A(e = 0 | r = 0, x_A(\omega) = \omega) &= -(\omega - (\omega + \varepsilon(0)))^2 - \kappa(0) - \pi(0), \\ &= -(\omega - \omega)^2 - \mathbb{E}[\varepsilon]^2 - V_\varepsilon(0), \\ &= -V_\varepsilon(0). \end{aligned}$$

The agency's expected payoff for investing high effort is given by,

$$\begin{aligned} EU_A(e = 1 | r = 0, x_A(\omega) = \omega) &= -(\omega - (\omega + \varepsilon(1))^2 - \kappa - \pi(0), \\ &= -V_\varepsilon(1) - \kappa. \end{aligned}$$

For the agency to invest high effort the following incentive compatibility constraint must be satisfied:

$$\begin{aligned} -V_\varepsilon(1) - \kappa &\geq -V_\varepsilon(0), \\ V_\varepsilon(0) - V_\varepsilon(1) &\geq \kappa. \end{aligned}$$

The precision improvement of investing high effort relative to low effort must outweigh the costs of doing so. This is case 1 in the result.

Case 2: $\beta_0 < \beta_1 < \beta$, perfect skepticism. In this case the agency will be reversed by the overseer with certainty, regardless of its choice of e. The agency will never invest high effort in this case since that would simply lead to a net loss proportional to the cost of that effort. To see why, consider the agency's expected payoff for investing low effort in this case,

$$\begin{aligned} EU_A(e=0|r=1) &= -(\omega-\varepsilon(0))^2-\kappa(0)-\pi, \\ &= -\omega^2-V_\varepsilon(0)-\pi. \end{aligned}$$

The agency's expected payoff for investing high effort is given by,

$$\begin{aligned} EU_A(e=1|r=1) &= -(\omega-\varepsilon(0))^2-\kappa-\pi, \\ &= -\omega^2-V_\varepsilon(0)-\kappa-\pi. \end{aligned}$$

Combining these expected payoffs yields the net expected payoff to the agency for investing high effort given that it will be overturned with certainty,

$$\begin{aligned} \Delta EU_A(e=1|r=1) &= -\omega^2-V_\varepsilon(0)-\kappa-\pi+\omega^2+V_\varepsilon(0)+\pi, \\ &= -\kappa. \end{aligned}$$

Thus, it is never incentive compatible for the agency to invest high effort given that it will overturned by the overseer with certainty. This is case 2 in the result.

Case 3: $\beta_0 < \beta < \beta_1$, conditional-deference. In this case the overseer upholds the agency if and only if the agency invests high effort. The agency's expected payoff for investing high effort, which induces being upheld, is given by,

$$\begin{aligned} EU_A(e=1|r^*(1)=0,x_A^*(\omega)=\omega) &= -(\omega-(\omega+\varepsilon(1)))^2-\kappa(1)-\pi(0), \\ &= -(\omega-\omega)^2-\mathbb{E}[\varepsilon]^2-V_\varepsilon(1)-\kappa, \\ &= -V_\varepsilon(1)-\kappa. \end{aligned}$$

The agency's expected payoff for investing low effort, which induces being overturned, is given by,

$$\begin{aligned} EU_A(e=0|r^*(0)=1) &= -(\omega-\varepsilon(0))^2-\kappa(0)-\pi(1), \\ &= -\mathbb{E}[\omega^2]-\mathbb{E}[\varepsilon]^2-V_\varepsilon(0)-\pi, \\ &= -p_0(0^2)-p_1(1^2)-p_\theta(\theta^2)-V_\varepsilon(0)-\pi, \\ &= -p_1-p_\theta\theta^2-V_\varepsilon(0)-\pi. \end{aligned}$$

Combining and rearranging these expected payoffs yields the agency's incentive compatibility constraint to invest high effort when facing a conditional-deference overseer:

$$\begin{aligned} -V_\varepsilon(1)-\kappa &\geq -p_1-p_\theta\theta^2-V_\varepsilon(0)-\pi, \\ p_1+p_\theta+V_\varepsilon(0)-V_\varepsilon(1)+\pi &\geq \kappa. \end{aligned}$$

This yields case 3 in the result. ■

Proposition 1. *In the equilibrium of the procedural review model the overseer makes review decisions according to $s_R(e)$ (equation 2), the agency always sets substantive policy to match the state and invests effort, conditional on review regime, as follows:*

- When facing a perfectly deferential overseer (i.e., $\beta_0 < \beta_1 < \beta$) the agency invests high effort when $V_\varepsilon(0)-V_\varepsilon(1) \geq \kappa$.
- When facing a perfectly skeptical overseer (i.e., $\beta_0 < \beta_1 < \beta$) the agency never invests high effort.
- When facing a conditional-deference overseer (i.e., $\beta_0 < \beta < \beta_1$) the agency invests high effort if $p_1+p_\theta\theta^2+V_\varepsilon(0)-V_\varepsilon(1)+\pi \geq \kappa$.

Proof of Proposition 1. The result follows from a straightforward combination of Lemma A.1, Lemma A.2, and Lemma A.3. ■

B Substantive review model

B.1 Derivation of overseer best responses to truthful policymaking

In this section I derive the overseer's best response function assuming that the agency always reveals ω by setting policy truthfully, $x_A(\omega)=\omega$. First, consider the overseer's expected utility for

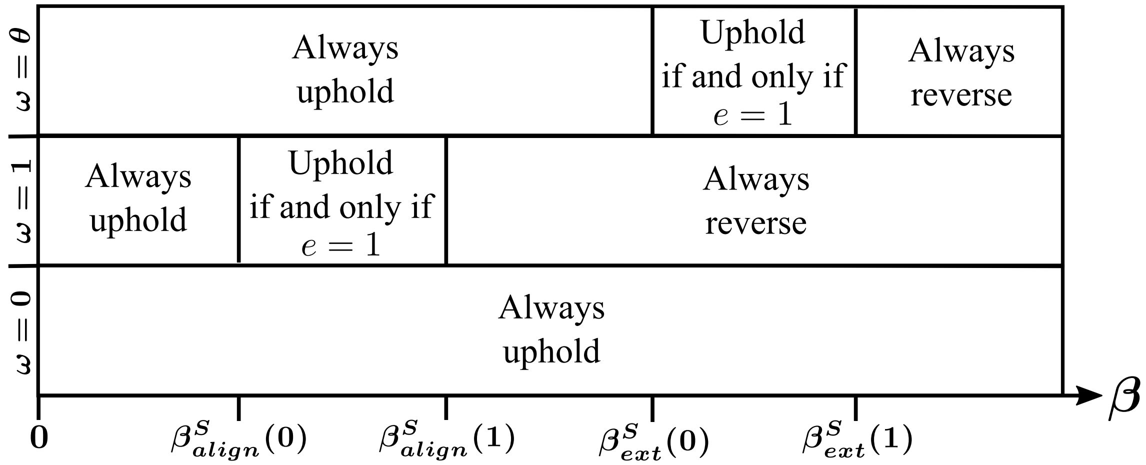

Figure 1: Overseer best responses to $x_A(\omega) = \omega$ given ω, β, and e

Note: Aligned preferences are when $\beta < \beta^S_{align}(0) = 1/2$, conditionally aligned preferences are when $1/2 = \beta^S_{align}(0) < \beta < \beta^S_{align}(1) = \frac{1+V_\varepsilon(0)-V_\varepsilon(1)}{2}$, moderate preferences are when $\frac{1+V_\varepsilon(0)-V_\varepsilon(1)}{1} = \beta^S_{align}(1) < \beta < \beta^S_{ext}(0) = \frac{\theta^2}{2\theta}$, conditionally extreme preferences are when $\frac{\theta^2}{2\theta} = \beta^S_{ext}(0) < \beta < \beta^S_{extreme}(1) = \frac{\theta^2+V_\varepsilon(0)-V_\varepsilon(1)}{2\theta}$, and extreme preferences are when $\beta > \beta^S_{ext}(1)$.

upholding the agency given $x_A(\omega) = \omega$ for any given ω and effort level e:

$$\begin{aligned} EU_R(r=0|x_A(\omega)=\omega,e) &= -(\omega-\beta-(x_A(\omega)+\varepsilon(e)))^2, \\ &= -(\omega-\beta-\omega)^2 - \mathbb{E}[\varepsilon|e]^2 - var[\varepsilon|e], \\ &= -\beta^2 - V_\varepsilon(e). \end{aligned}$$

The analogous expected utilities for reversing the agency depend on ω and are given by:

$$\begin{aligned} EU_R(r=1|\omega=0) &= -(0-\beta-\varepsilon(0))^2 \\ &= -\beta^2 - \mathbb{E}[\varepsilon|0]^2 - var[\varepsilon|0], \\ &= -\beta^2 - V_\varepsilon(0), \\ EU_R(r=1|\omega=1) &= -(1-\beta-\varepsilon(0))^2 \\ &= -(1-\beta)^2 - \mathbb{E}[\varepsilon|0]^2 - var[\varepsilon|0], \\ &= -(1-\beta)^2 - V_\varepsilon(0), \\ EU_R(r=1|\omega=\theta) &= -(\theta-\beta-\varepsilon(0))^2 \\ &= -(\theta-\beta)^2 - \mathbb{E}[\varepsilon|0]^2 - var[\varepsilon|0], \\ &= -(\theta-\beta)^2 - V_\varepsilon(0). \end{aligned}$$

Comparing the overseer's expected utilities for uphold versus reversing the agency in each state, which she knows for sure given the agency's separating strategy, yields the incentive compatibility conditions for the overseer to uphold. These in turn generate the bias thresholds in Figure 1. Consider first the case when $\omega = 0$:

$$\begin{aligned} r = 0 \iff EU_R(r = 0|0, e) &\geq EU_R(r = 1|0), \\ -\beta^2 - V_\varepsilon(e) &\geq -\beta^2 - V_\varepsilon(0), \\ V_\varepsilon(0) - V_\varepsilon(e) &\geq 0, \end{aligned}$$

which is always weakly satisfied since either $e = 0$, in which case the inequality is satisfied with equality, or $e = 1$, in which case the inequality holds strictly. Breaking indifference in the case of $e = 0$ with deference to the agency ($r = 0$ when indifferent) implies that the overseer would always uphold the agency, regardless of e, when the agency truthfully reveals that $\omega = 0$.

Consider now the case when $\omega = 1$:

$$\begin{aligned} r = 0 \iff EU_R(r = 0|1, e) &\geq EU_R(r = 1|1), \\ -\beta^2 - V_\varepsilon(e) &\geq -(1-\beta)^2 - V_\varepsilon(0), \\ \frac{1 + V_\varepsilon(0) - V_\varepsilon(e)}{2} &\geq \beta. \end{aligned}$$

Thus, there are two bias cutoffs depending on agency effort for which the overseer would uphold the agency for truthfully setting $x_A(1) = 1$: (1) $e = 0 \Rightarrow \beta^S_{align}(0) := \frac{1}{2}$ and (2) $e = 1 \Rightarrow \beta^S_{align}(1) := \frac{1+V_\varepsilon(0)-V_\varepsilon(1)}{2}$. Note that $\beta^S_{align}(0) < \beta^S_{align}(1)$ so that the overseer will uphold an agency that sets $x_A(1) = 1$ for a wider range of preference disagreement when the agency exerts high effort. This generates the middle range in Figure 1 for when $\omega = 1$. So long as the agency and overseer do not disagree too much (i.e., $\beta > \beta^S_{align}(1)$) then there is the possibility of the agency being upheld when the agency separates by setting policy truthfully, depending further on the agency's effort.

Finally, consider the case when $\omega = \theta$:

$$\begin{aligned} r = 0 \iff EU_R(r = 0|\theta, e) &\geq EU_R(r = 1|\theta), \\ -\beta^2 - V_\varepsilon(e) &\geq -(\theta-\beta)^2 - V_\varepsilon(0), \\ \frac{\theta^2 + V_\varepsilon(0) - V_\varepsilon(e)}{2\theta} &\geq \beta. \end{aligned}$$

Thus, there is again two bias thresholds that dictate when $r = 0$ following truthful policymaking when the agency plays a separating strategy: (1) $e = 0 \Rightarrow \beta^S_{ext}(0) := \frac{\theta^2}{2\theta}$ and (2) $e = 1 \Rightarrow \beta^S_{ext}(1) := \frac{\theta^2+V_\varepsilon(0)-V_\varepsilon(1)}{2\theta}$. When $\beta > \beta^S_{ext}(e)$ preference divergence is sufficiently extreme, given e and truthful policymaking, to lead the overseer to reverse the agency. Otherwise, the overseer upholds. This captures the top range in Figure 1 when $\omega = \theta$.

Taken together, the overseer's best response function given $x_A(\omega) = \omega$ for all ω is given by:

$$s_R(x_A(\omega) = \omega, e) = \begin{cases} \text{Uphold: } r = 0 & \text{if } x_A = 0, \\ & \text{if } x_A = 1 \text{ and } \beta < \beta^S_{align}(e), \\ & \text{if } x_A = \theta \text{ and } \beta < \beta^S_{ext}(e), \\ \text{Overturn: } r = 1 & \text{otherwise.} \end{cases} \tag{1}$$

B.2 Truthful separating equilibrium and pooling to placate

Proposition 2. *There is a truthful separating equilibrium in which the agency always reveals the state by setting $x^*_A(\omega) = \omega$ if and only if overseer-agency preferences are sufficiently aligned: $\beta < \beta^S_{align}(e) := \frac{1+V_\varepsilon(0)-V_\varepsilon(e)}{2}$, where $\beta^S_{align}(0) < \beta^S_{align}(1)$. Moreover, there is a pooling to placate equilibrium in which the agency sets $x^*_A(\omega) = 0$ for all ω any time preference divergence is sufficiently extreme: $\beta > \beta^S_{ext}(e) := \frac{\theta^2+V_\varepsilon(0)-V_\varepsilon(e)}{2\theta}$, where $\beta^S_{ext}(0) < \beta^S_{ext}(1)$.*

Proof of Proposition 2. Suppose that $\beta < \beta^S_{align}(e)$ so that, given e, the agency is upheld following $x_A(\omega) = \omega, \forall \omega$ (from Lemma A.2). Clearly in this case there is no reason for the agency to deviate from matching policy to the state since doing so will only lead to a net loss equal to the deviation. Suppose instead that $\beta > \beta^S_{align}(e)$, given e, so that the overseer reverses the agency when it truthfully reveals that $\omega = 1$. In that case the agency can avoid reversal by deviating to $x = 0$ given that the overseer will uphold according to $s_R(x_A(\omega) = \omega, e)$ from Lemma A.2, which is beneficial since:

$$\begin{gathered} EU_A(x_A(1) = 1 | r(1,e) = 1, \omega = 1) < EU_A(x_A = 0 | r(0,e) = 0, \omega = 1), \\ -(1-\varepsilon(e))^2 - \kappa e - \pi < -(1-(0+\varepsilon(e)))^2 - \kappa e, \\ -1 - \mathbb{E}[\varepsilon|0]^2 - var[\varepsilon|0] - \kappa e - \pi < -1 - \mathbb{E}[\varepsilon|e]^2 - var[\varepsilon|e] - \kappa e, \\ -1 - V_\varepsilon(0) - \kappa e - \pi < -1 - V_\varepsilon(e) - \kappa e, \\ V_\varepsilon(e) - V_\varepsilon(0) < \pi, \end{gathered}$$

which is satisfied for all $\pi > 0$ given $V_\varepsilon(1) - V_\varepsilon(0) < 0$ and $V_\varepsilon(0) - V_\varepsilon(0) = 0$. Thus, the agency would always deviate from truthful policymaking when $\beta > \beta^S_{align}(e)$ to avoid reversal. Taken together this implies that the agency plays a separating strategy if and only preferences are sufficiently aligned: $\beta < \beta^S_{align}(e)$, as stated in the result.

Now suppose that preference disagreement is extreme: $\beta > \beta^S_{ext}(e)$. The overseer has a strictly dominant strategy of $r(x_A, e) = 1$ for all $x_A \neq 0$ in this case. Given this, iterated elimination of dominated strategies yields $x^*_A(\omega) = 0$ for all ω as the iteratively dominant strategy for the agent. ■

B.3 Semi-pooling obfuscation equilibria

B.3.1 Obfuscating to appease

Proposition 3. *When $\theta > 2$ and preference disagreement is such that $\beta \in \left(\beta^S_{align}(e), \beta^S_{ext}(e)\right)$ there is a semi-pooling equilibrium in which the agency obfuscates to appease by setting $x(\omega) = 0$ for $\omega \in \{0,1\}$ and $x(\theta) = \theta$ and the overseer upholds $x_A \in \{0,\theta\}$ and reverses $x_A = 1$.*

Proof of Proposition 3. If the agent sets $x^*_A(\omega) = 0$ for $\omega \in \{0,1\}$ and $x^*_A(\theta) = \theta$ and overseer beliefs are consistent with this strategy then the overseer will uphold upon observing $x_A = \theta$. Similarly, the overseer will uphold $x_A = 0$ since either (a) she is indifferent since reversal leads to $x = 0$ with a shock $\varepsilon(0)$ or (b) she is better off upholding because the agency chose $e = 1$, which leads to $x = 0$ with $\varepsilon(1)$. Finally, to complete the sequential rationality of the overseer's strategy set the overseer's beliefs following $x_A = 1$ to $Pr[\omega = 1 | x_A = 1] = 1$, which leads the overseer to reverse given β. Verifying the agency's stated strategy as a best response only requires consideration of $x_A(1) = 0$ since $x^*_A(\omega) = \omega$ when $\omega \in \{0,\theta\}$. Setting $x_A(1) = 1$ induces $r(1,e) = 1$ which leads to a net payoff of:

$$\begin{aligned}\Delta EU_A(x_A(1) = 1) &= EU_A(x_A(1) = 1 | r(1,e) = 1) - EU_A(x_A(1) = 0 | r(0,e) = 0)\\ &= -1 - V_\varepsilon(0) - \pi + 1 + V_\varepsilon(e),\\ &= V_\varepsilon(e) - V_\varepsilon(0) - \pi,\end{aligned}$$

which is always negative since $\pi > 0$ and $V_\varepsilon(e) \leq V_\varepsilon(0)$. Thus, the agency never benefits from deviating to $x_A(1) = 1$ from $x^*_A(1) = 0$. Now consider the analogous payoff for $x_A(1) = \theta$, which would induce $r(\theta,e) = 0$:

$$\begin{aligned}\Delta EU_A(x_A(1) = \theta) &= EU_A(x_A(1) = \theta | r(\theta,e) = 0) - EU_A(x_A(1) = 0 | r(0,e) = 0),\\ &= -(1-\theta)^2 - V_\varepsilon(e) + 1 + V_\varepsilon(e),\\ &= 1 - (1-\theta)^2,\end{aligned}$$

which is always negative given that $\theta > 2$ (and therefore $(1-\theta)^2 > 1$). Thus, the agency does not benefit from deviating from $x^*_A(1) = 0$ to $x_A(1) = \theta$. Taken together, the players are best responding and therefore these strategies, along with appropriate and consistent beliefs of the overseer, form a PBE. ■

B.3.2 Obfuscation through exaggeration

Proposition 4. *When $\theta < 2$, preference disagreement is such that $\beta \in \left(\beta^S_{align}(e), \beta^S_{ext}(e)\right)$, and $\omega = \theta$ is sufficiently likely relative to $\omega = 1$:*

$$\frac{p_1}{p_1 + p_\theta} \leq \frac{\theta(\theta - 2\beta) + V_\varepsilon(0) - V_\varepsilon(e)}{2\theta(\theta - 1)}, \tag{2}$$

there is a semi-pooling equilibrium in which the agency obfuscates through exaggeration by setting $x(0) = 0$ and $x(\omega) = \theta$ for $\omega \in \{1, \theta\}$ and the overseer upholds $x_A \in \{0, \theta\}$ and reverses $x_A = 1$.

Proof of Proposition 4. Consider the agency's semi-pooling strategy in which:

$$x^*_A(\omega) = \begin{cases} \theta & \text{if } \omega \in \{1, \theta\} \\ 0 & \text{if } \omega = 0. \end{cases}$$

Given correct beliefs $Pr[\omega = 0|x_A = 0] = 1$ the overseer has no reason to reverse $x_A = 0$. Moreover, set $Pr[\omega = 1|x_A = 1] = 1$ so that the overseer reverses following $x_A = 1$ given β. Finally, consider $r(x_A, e)$ when $x_A = \theta$. Let $b_R(1|\theta) = Pr[\omega = 1|x_A = \theta] = \frac{p_1}{p_1 + p_\theta}$ and $b_R(\theta|\theta) = Pr[\omega = \theta|x_A = \theta] = \frac{p_\theta}{p_1 + p_\theta}$ be the overseer's (correct) beliefs (consistent with $x^*_A(\omega)$) following $x_A = \theta$ that $\omega = 1$ and $\omega = \theta$, respectively. After observing $x_A = \theta$, given $x^*_A(\omega)$, the overseer upholds if and only if:

$$\begin{aligned} -b_R(1|\theta)(1 - \beta - \theta)^2 - b_R(\theta|\theta)\beta^2 - V_\varepsilon(e) &\geq -b_R(1|\theta)(1 - \beta)^2 - b_R(\theta|\theta)(\theta - \beta)^2 - V_\varepsilon(0), \\ -\left(\frac{p_1}{p_1 + p_\theta}(1 - \beta - \theta)^2 + \frac{p_\theta}{p_1 + p_\theta}\beta^2 + V_\varepsilon(e)\right) &\geq -\left(\frac{p_1}{p_1 + p_\theta}(1 - \beta)^2 - \frac{p_\theta}{p_1 + p_\theta}(\theta - \beta)^2 + V_\varepsilon(0)\right), \\ \frac{p_1}{p_1 + p_\theta} &\leq \frac{\theta^2 - 2\beta\theta + V_\varepsilon(0) - V_\varepsilon(e)}{2\theta(\theta - 1)}. \end{aligned}$$

Thus, given the overseer's beliefs consistent with $x^*_A(\omega)$, the overseer's review strategy is a best response to $x^*_A(\omega)$. To verify that $x^*_A(\omega)$ is a best response to the overseer's review strategy we again need only check the case when $\omega = 1$. When $\omega \in \{0, \theta\}$ the agency is matching policy to the state and being upheld so there is clearly no reason to deviate. If the agency were to truthfully set $x_A(1) = 1$, given overseer beliefs fixed above at $Pr[\omega = 1|x_A = 1] = 1$, the overseer will reverse the agency leading to a net negative payoff for deviation:

$$\begin{aligned} \Delta EU_A(x_A(1) = 1) &= EU_A(x_A(1) = 1|r(1, e) = 1) - EU_A(x^*_A(1) = \theta|r(\theta, e) = 0), \\ &= -1 - V_\varepsilon(0) - \kappa e - \pi + (1 - \theta)^2 + V_\varepsilon(e) - \kappa e, \\ &= (1 - \theta)^2 - 1 + V_\varepsilon(e) - V_\varepsilon(0) - \pi, \end{aligned}$$

which is always negative since (i) $(1-\theta)^2 < 1$ because $\theta < 2$, (ii) $V_\varepsilon(e) - V_\varepsilon(1) \leq 0$, and (iii) $\pi > 0$. Thus, the agency is strictly better off under $x_A^*(\omega)$ given $r^*(x_A^*(\omega), e)$ in this case. Finally, consider a deviation to $x_A(1) = 0$, which will also be upheld, instead of $x_A^*(1) = \theta$:

$$\begin{aligned}\Delta EU_A(x_A(1)=0) &= EU_A(x_A(1)=0|r(0,e)=0) - EU_A(x_A^*(1)=\theta|r(\theta,e)=0),\\ &= -1 - V_\varepsilon(e) - \kappa e + (1-\theta)^2 + V_\varepsilon(e) - \kappa e,\\ &= (1-\theta)^2 - 1,\end{aligned}$$

which is always negative since $\theta < 2$ (implying $(1-\theta)^2 < 1$). Thus, $x_A^*(\omega)$ is a best response to the overseer's review strategy. Taken together, these strategies, along with consistent overseer beliefs, form a PBE. ■

Proposition B.1. *When $\theta < 2$, preference disagreement is such that $\beta \in \left(\beta^S_{align}(e), \beta^s_{ext}(e)\right)$, and equation* (2) *does not hold:*

$$\frac{p_1}{p_1 + p_\theta} > \frac{\theta(\theta - 2\beta) + V_\varepsilon(0) - V_\varepsilon(e)}{2\theta(\theta-1)}, \tag{3}$$

there is a semi-pooling equilibrium in which the agency mixes between $x_A = \theta$ and $x_A = 0$ when $\omega = 1$ and chooses $x_A = \omega$ for $\omega \in \{0, \theta\}$:

$$x_A^{mix}(\omega) = \begin{cases} \theta & \textit{with probability } \frac{p_\theta(\theta(\theta-2\beta)+V_\varepsilon(0)-V_\varepsilon(e))}{p_1(\theta(\theta-2+2\beta)+V_\varepsilon(e)-V_\varepsilon(0))} \textit{ if } \omega = 1,\\ 0 & \textit{with probability } 1 - \frac{p_\theta(\theta(\theta-2\beta)+V_\varepsilon(0)-V_\varepsilon(e))}{p_1(\theta(\theta-2+2\beta)+V_\varepsilon(e)-V_\varepsilon(0))} \textit{ if } \omega = 1,\\ \omega & \textit{if } \omega \in \{0, \theta\},\end{cases}$$

and the overseer reverses x_A according the following strategy:

$$s_R^*(x_A, e) = \begin{cases} \textit{Uphold: } 0 & \textit{if } x_A = 0,\\ \textit{Reverse: } 1 & \textit{if } x_A = 1,\\ \textit{Uphold: } 0 & \textit{with probability } 1 - \frac{\theta(\theta-2)}{\theta(\theta-2)+V_\varepsilon(e)-V_\varepsilon(0)-\pi} \textit{ if } x_A = \theta,\\ \textit{Reverse: } 1 & \textit{with probability } \frac{\theta(\theta-2)}{\theta(\theta-2)+V_\varepsilon(e)-V_\varepsilon(0)-\pi} \textit{ if } x_A = \theta.\end{cases}$$

Proof of Proposition B.1. In this environment it is straightforward to show that, given the overseer's strategy $s_R^*(x_A, e)$, the agency's best response is $x_A = 0$ when $\omega = 0$ and, supposing that the agency is mixing between $x_A = 0$ and $x_A = \theta$ when $\omega = 1$, that $x_A = \theta$ when $\omega = \theta$ is also a best response. The overseer upholding $x_A = 0$ is also clearly a best response since, given the preference environment, the overseer prefers both (i) the agency to set $x_A = 0$ when $\omega = 0$ and (ii) the agency to set $x_A = 0$ when $\omega = 1$. So in all instances in which $x_A = 0$ would be observed, given $x_A^{mix}(\omega)$, the agency is setting policy in line with overseer preferences. Note also that the preference environment implies that overturning $x_A = 1$ is a best response. Thus, what remains to be shown is that (a) the agency is best responding when $\omega = 1$ and (b) the overseer is best responding when $x_A = \theta$. This involves deriving when the two players are indifferent between the two actions prescribed by their respective strategies ($x_A^{mix}(\omega)$ and $s_R^*(x_A, e)$) in those settings.

First, note that the agency's payoff for setting $x_A(1) = 1$ given $s_R^*(x_A, e)$ is $EU_A(x_A = 1|\omega = 1, r(1,e) = 1) = -1 - V_\varepsilon(0) - \kappa e - \pi$ while the payoff for setting $x_A(1) = 0$ given $s_R^*(x_A, e)$ is $EU_A(x_A = 0|\omega = 1, r(1,e) = 0) = -1 - V_\varepsilon(e) - \kappa e$. Clearly $EU_A(x_A = 0|\omega = 1, r(1,e) = 0) > EU_A(x_A = 1|\omega = 1, r(1,e) = 1)$ since $V_\varepsilon(0) - V_\varepsilon(e) \geq 0$ and $\pi > 0$ so agency mixing when $\omega = 1$ is between $x_A = 0$ and $x_A = \theta$. Letting $\phi := Pr[x_A = \theta|\omega = 1]$ the overseer can mix between $r(\theta, e) = 0$ and $r(\theta, e) = 1$ when $x_A = \theta$ only if:

$$\frac{p_1\phi}{p_1\phi + p_\theta} = \frac{\theta(\theta - 2\beta) + V_\varepsilon(0) - V_\varepsilon(e)}{2\theta(\theta - 1)}, \tag{4}$$

which implies that:

$$\phi = \frac{p_\theta\left(\theta(\theta - 2\beta) + V_\varepsilon(0) - V_\varepsilon(e)\right)}{p_1\left(\theta(\theta - 2 + 2\beta) + V_\varepsilon(e) - V_\varepsilon(0)\right)} \tag{5}$$

The restrictions in the environment ensure that ϕ as defined in equation (5) is positive and less than one (i.e., $\phi \in (0,1)$).

Now, letting $\rho := Pr[r(x_A, e) = 1|x_A = \theta]$ denote the probability the overseer reverses following $x_A = \theta$ the agency's expected payoff from setting $x_A(1) = \theta$ is given by:

$$EU_A(x_A = \theta|\omega = 1, \rho) = \rho\left[-1 - V_\varepsilon(0) - \kappa e - \pi\right] + (1 - \rho)\left[-(1-\theta)^2 - V_\varepsilon(e) - \kappa e\right],$$

and the agency's analogous expected payoff from setting $x_A(1) = 0$ (which induces $r(1,e) = 0$) is given by:

$$EU_A(x_A = 0|\omega = 1, r(1,e) = 0) = -1 - V_\varepsilon(e) - \kappa e.$$

In order to choose $\phi \in (0,1)$ it must be that:

$$EU_A(x_A = \theta|\omega = 1, \rho) = EU_A(x_A = 0|\omega = 1, r(1,e) = 0),$$
$$\rho = \frac{\theta(\theta - 2)}{\theta(\theta - 2) + V_\varepsilon(e) - V_\varepsilon(0) - \pi}.$$

Note that $\theta \in (\underline{\theta}, 2)$ and $\pi > 0$ ensure that $\rho \in (0,1)$. So long as $r(\theta, e) = \rho$ the agent is indifferent between choosing $x_A = \theta$ and $x_A = 0$ when $\omega = 1$ and can set $x_A = \theta$ with probability ϕ and $x_A = 0$ with probability $1-\phi$. Doing so and choosing $x_A = \omega$ when $\omega \in \{0, \theta\}$ with probability one, along with consistency of overseer beliefs, implies that the overseer is indifferent between $r(\theta, e) = 1$ and $r(\theta, e) = 0$ so that she can choose the former with probability ρ and the latter with probability $1-\rho$. To complete the equilibrium, set the overseer's off-path beliefs after observing $x_A = 1$ to $Pr[\theta = 1 | x_A = 1] = 1$. ■

B.4 Equilibrium effort

Lemma B.1. *Define $\beta^S_{align}(e) := \frac{1+V_\varepsilon(0)-V_\varepsilon(e)}{2}$ and $\beta^S_{ext}(e) := \frac{\theta^2+V_\varepsilon(0)-V_\varepsilon(e)}{2\theta}$. Note that $0 < \beta^S_{align}(0) < \beta^S_{align}(1) < \beta^S_{ext}(0) < \beta^S_{ext}(1)$ since $V_\varepsilon(0) > V_\varepsilon(1)$. Conditional on the overseer's bias β, which dictates the type of signaling outlined in Figure 3, the agency invests effort as follows:*

1. *If $\beta < \beta^S_{align}(0)$ then there is a truthful separating equilibrium in which the agency invests high effort if and only if $V_\varepsilon(0) - V_\varepsilon(1) \geq \kappa$.*

2. *If $\beta^S_{align}(0) < \beta < \beta^S_{align}(1)$ then there is a truthful separating equilibrium following high effort and a semi-pooling obfuscation equilibrium after low effort, which depends further on θ.*

   (a) When $\theta < 2$ the agency invests high effort and obfuscates through exaggeration if and only if $p_1(1-\theta)^2 + V_\varepsilon(0) - V_\varepsilon(1) \geq \kappa$.

   (b) When $\theta > 2$ the agency invests high effort and obfuscates to appease if and only if $p_1 + V_\varepsilon(0) - V_\varepsilon(1) \geq \kappa$.

3. *If $\beta^S_{align}(1) < \beta < \beta^S_{ext}(0)$ then there is a semi-pooling obfuscation equilibrium following both high and low effort, obfuscation through exaggeration or obfuscation to appease depending on θ, and the agency invests high effort if and only if $V_\varepsilon(0) - V_\varepsilon(1) \geq \kappa$.*

4. *If $\beta^S_{ext}(0) < \beta < \beta^S_{ext}(1)$ then there is a semi-pooling obfuscation equilibrium following high effort, which depends further on θ, and a pooling to placate equilibrium following low effort.*

   (a) When $\theta < 2$ the agency invests high effort and obfuscates through exaggeration if and only if $p_1(1-(1-\theta)^2) + p_\theta \theta^2 + V_\varepsilon(0) - V_\varepsilon(1) \geq \kappa$.

   (b) When $\theta > 2$ the agency invests high effort and obfuscates to appease if and only if $p_\theta \theta^2 + V_\varepsilon(0) - V_\varepsilon(1) \geq \kappa$.

5. *If $\beta > \beta^S_{ext}(1)$ then there is a pooling to placate equilibrium following both high and low effort. The agency invests high effort if and only if $V_\varepsilon(0) - V_\varepsilon(1) \geq \kappa$.*

Proof of Lemma B.1. Each case requires deriving the agency's incentive compatibility condition for high effort given the type of policymaking strategy (and review strategy) that will follow.

Case 1: $\beta < \beta^S_{align}(0)$**.** In this case there is a truthful separating equilibrium following both low and high effort. Thus, $r^*(x_A(\omega) = \omega, e) = 0$ for all e and the agency's incentive compatibility condition for high effort is given by:

$$\begin{aligned} EU_A(e = 1|\beta \leq \beta^S_{align}(0)) &\geq EU_A(e = 0|\beta \leq \beta^S_{align}(0)), \\ -p_0V_\varepsilon(1) - p_1V_\varepsilon(1) - p_\theta V_\varepsilon(1) - \kappa &\geq -p_0V_\varepsilon(0) - p_1V_\varepsilon(0) - p_\theta, \\ -V_\varepsilon(1) - \kappa &\geq -V_\varepsilon(0), \\ V_\varepsilon(0) - V_\varepsilon(1) &\geq \kappa, \end{aligned}$$

as stated in the result.

Case 2: $\beta^S_{align}(0) < \beta \leq \beta^S_{align}(1)$**.** In this case there is a truthful separating equilibrium after $e = 1$ and a semi-pooling obfuscation equilibrium after $e = 0$. If $\theta > 2$ then the agency obfuscates to appease and if $\theta < 2$ then the agency obfuscates through exaggeration. Consider first the case in which $\theta > 2$:

$$\begin{aligned} EU_A(e = 1|\beta^S_{align}(0) < \beta \leq \beta^S_{align}(1), \theta > 2) &\geq EU_A(e = 0|\beta^S_{align}(0) < \beta \leq \beta^S_{align}(1), \theta > 2), \\ -V_\varepsilon(1) - \kappa &\geq -p_0V_\varepsilon(0) - p_1(1 + V_\varepsilon(0)) - p_\theta V_\varepsilon(0), \\ -V_\varepsilon(1) - \kappa &\geq -p_1 - V_\varepsilon(0), \\ p_1 + V_\varepsilon(0) - V_\varepsilon(1) &\geq \kappa, \end{aligned}$$

as stated in the result. Now suppose that $\theta < 2$. The associated incentive compatibility condition for high effort is:

$$\begin{aligned} EU_A(e = 1|\beta^S_{align}(0) < \beta \leq \beta^S_{align}(1), \theta < 2) &\geq EU_A(e = 0|\beta^S_{align}(0) < \beta \leq \beta^S_{align}(1), \theta < 2), \\ -V_\varepsilon(1) - \kappa &\geq -p_0V_\varepsilon(0) - p_1((1 - \theta)^2 + V_\varepsilon(0)) - p_\theta V_\varepsilon(0), \\ -V_\varepsilon(1) - \kappa &\geq -p_1(1 - \theta)^2 - V_\varepsilon(0), \\ p_1(1 - \theta)^2 + V_\varepsilon(0) - V_\varepsilon(1) &\geq \kappa, \end{aligned}$$

as stated in the result.

Case 3: $\beta^S_{align}(1) < \beta < \beta^S_{ext}(0)$**.** In this case the agency semi-pools following both $e = 0$ and $e = 1$. Since the substance of policy will be the same regardless of e it follows that the incentive compatibility condition for high effort is simply $V_\varepsilon(0) - V_\varepsilon(1) \geq \kappa$ as in the first case in which there is a truthful separating equilibrium regardless of e. So long as the spatial location of x does not change based on e the only relevant consideration is whether the precision improvement from high effort, $V_\varepsilon(0) - V_\varepsilon(1)$, outweighs the cost of that improvement, κ.

Case 4: $\beta^S_{ext}(0) < \beta < \beta^S_{ext}(1)$**.** In this case the agency obfuscates through semi-pooling after $e = 1$ and pools to placate after $e = 0$. Accordingly, there are again two cases depending on θ. Suppose first that $\theta > 2$. The agency invests high effort if and only if:

$$\begin{aligned}
EU_A(e = 1|\beta^S_{align}(1) < \beta < \beta^S_{ext}(0), \theta > 2) &\geq EU_A(e = 0|\beta^S_{align}(1) < \beta < \beta^S_{ext}(0), \theta > 2),\\
-p_1 - V_\varepsilon(1) - \kappa &\geq -p_1 - p_\theta\theta^2 - V_\varepsilon(0),\\
p_\theta\theta^2 + V_\varepsilon(0) - V_\varepsilon(1) &\geq \kappa,
\end{aligned}$$

as stated in the result. The analogous condition when $\theta < 2$ is given by:

$$\begin{aligned}
EU_A(e = 1|\beta^S_{align}(1) < \beta < \beta^S_{ext}(0), \theta < 2) &\geq EU_A(e = 0|\beta^S_{align}(1) < \beta < \beta^S_{ext}(0), \theta < 2),\\
-p_1(1-\theta)^2 - V_\varepsilon(1) - \kappa &\geq -p_1 - p_\theta\theta^2 - V_\varepsilon(0),\\
p_1 - p_1(1-\theta)^2 + p_\theta\theta^2 + V_\varepsilon(0) - V_\varepsilon(1) &\geq \kappa,
\end{aligned}$$

as stated in the result.

Case 5: $\beta > \beta^S_{ext}(1)$**.** In this case the agency pools to placate by setting $x_A = 0$ regardless of ω and e. Similar to cases 1 and 3 above, this implies that the relevant incentive compatibility condition for high effort is $V_\varepsilon(0) - V_\varepsilon(1) \geq \kappa$ since the spatial location of policy is zero regardless of e. Thus, the only relevant consideration is whether the induced precision improvements from high effort are sufficient to outweigh the cost for that improvement. ■

Corollary B.1. *The agency's discretion to match policy to the state is higher when the agency invests high effort.*

Proof of Corollary B.1. This follows from comparison of Case 2 to Case 1 and Case 4 to Case 3 in the proof of Lemma B.1. Specifically, note that, regardless of θ, in both cases the agency is able to set $x_A = \omega$ for more instances of ω. In the first comparison (case 2 and case 1), the agency is able to match $x_A(\omega) = \omega$ for all ω after high effort whereas it can only match $x_A(\omega) = \omega$ for $\omega \in \{0, \theta\}$ after low effort. This implies that the agency is able to match policy to the state strictly more for a higher level of preference disagreement than following low effort. In the second comparison (case 4 to case 3), the agency is able to match $x_A(\omega) = \omega$ for $\omega \in \{0, \theta\}$ after high effort and only matches $x_A(\omega) = \omega$ for $\omega = 0$ after low effort. Again, this implies that the agency is better able to match policy to the state for higher levels of preference disagreement following high effort. Thus, overall, high effort allows the agency to better match policy to the state for higher levels of bias. ■

C Reviewing procedure vs. judging substance

Proposition 5. *Suppose $\beta \in \left(\beta^P(0), \beta^P(1)\right)$, $\beta \in \left(\beta^S_{align}(1), \beta^S_{ext}(0)\right)$, and equation* (3) *is satisfied so that procedural review is conditionally-deferential and substantive review always leads to a semi-pooling obfuscation equilibrium. If $\theta > 2$ then the agency obfuscates to appease and overseer welfare is higher under procedural review if and only if effort costs are intermediate, $\kappa^S < \kappa < \kappa^P$, and $V_\varepsilon(0) - V_\varepsilon(1) > p_1(2\beta - 1)$. If $\theta < 2$ then the agency obfuscates through exaggeration and overseer welfare is always higher under procedural review.*

Proof of Proposition 5. Consider an environment where $\beta \in \left(\beta^P_0, \beta^P_1\right)$ and $\beta \in \left(\beta^S_{align}(1), \beta^S_{ext}(0)\right)$ so that under procedural review we are in a conditional-deference environment and, from Lemma A.3, the agency exerts high effort if and only if $p_1 + p_\theta \theta^2 + V_\varepsilon(0) - V_\varepsilon(1) + \pi \geq \kappa$, and under substantive review we are in a semi-pooling obfuscation equilibrium. If $\theta < 2$ then the agency obfuscates through exaggeration (i.e., $x^*_A(0) = 0$ and $x^*_A(\omega) = \theta$ for $\omega \in \{1, \theta\}$) and if $\theta > 2$ then the agency obfuscates to appease (i.e., $x^*_A(\omega) = 0$ for $\omega \in \{0, 1\}$ and $x^*_A(\theta) = \theta$). In both cases the agency exerts high effort if and only if $V_\varepsilon(0) - V_\varepsilon(1) \geq \kappa$ (from Lemma B.1). Denote the upper bounds on κ to support high effort under procedural review and substantive review, respectively, as $\kappa^P := p_1 + p_\theta \theta^2 + V_\varepsilon(0) - V_\varepsilon(1) + \pi$ and $\kappa^S := V_\varepsilon(0) - V_\varepsilon(1)$. Note that $\kappa^P > \kappa^S$ so that high effort is more likely in the sense of set inclusion under procedural review. Thus there are three cases of effort across regimes: (1) if $\kappa > \kappa^P$ then $e = 0$ under both procedural and substantive review; (2) if $\kappa \in \left(\kappa^S, \kappa^P\right]$ then $e = 0$ under substantive review and $e = 1$ under procedural review; (3) if $\kappa \leq \kappa^S$ then $e = 1$ under both procedural and substantive review.

With these possibilities in mind we can compute the overseer's ex ante welfare for each:

$$\begin{aligned}
W^P(e = 1|\beta \text{ moderate}) &= -p_0(\beta^2 + V_\varepsilon(1)) - p_1(\beta^2 + V_\varepsilon(1)) - p_\theta(\beta^2 + V_\varepsilon(1)), \\
&= -\beta^2 - V_\varepsilon(1), \\
W^P(e = 0|\beta \text{ moderate}) &= -p_0(\beta^2 + V_\varepsilon(0)) - p_1((1-\beta)^2 + V_\varepsilon(0)) - p_\theta((\theta - \beta)^2 + V_\varepsilon(0)), \\
&= -p_0\beta^2 - p_1(1-\beta)^2 - p_\theta(\theta - \beta)^2 - V_\varepsilon(0), \\
W^S(e|\beta \text{ moderate}, \theta > 2) &= -p_0\beta^2 - p_1(1-\beta)^2 - p_\theta\beta^2 - V_\varepsilon(e), \\
W^S(e|\beta \text{ moderate}, \theta < 2) &= -p_0\beta^2 - p_1(1-\beta-\theta)^2 - p_\theta\beta^2 - V_\varepsilon(e).
\end{aligned}$$

Obfuscation to appease. Consider the case in which $\theta > 2$ so the agency obfuscates to appease under substantive review. We can evaluate each case that depends on κ. First, let $\kappa > \kappa^P > \kappa^S$ so that $e = 0$ regardless of review type:

$$\begin{aligned} W^P(e=0|\beta \text{ moderate}, \theta > 2) &< W^S(e=0|\beta \text{ moderate}, \theta > 2), \\ -p_0\beta^2 - p_1(1-\beta)^2 - p_\theta(\theta-\beta)^2 - V_\varepsilon(0) &< -p_0\beta^2 - p_1(1-\beta)^2 - p_\theta\beta^2 - V_\varepsilon(0), \\ p_\theta\beta^2 &< p_\theta(\theta-\beta)^2, \\ 0 &< p_\theta\theta(\theta-2\beta), \end{aligned}$$

which is always satisfied for $\theta > 2$ and the range of β in this environment. Thus, the overseer is better off under substantive review.

Now consider the case in which $\kappa \in \left(\kappa^S, \kappa^P\right]$ so that $e = 0$ under substantive review and $e = 1$ under procedural review:

$$\begin{aligned} W^P(e=1|\beta \text{ moderate}, \theta > 2) &< W^S(e=0|\beta \text{ moderate}, \theta > 2), \\ -\beta^2 - V_\varepsilon(1) &< -p_0\beta^2 - p_1(1-\beta)^2 - p_\theta\beta^2 - V_\varepsilon(0), \\ V_\varepsilon(0) - V_\varepsilon(1) &< p_1(2\beta-1), \end{aligned}$$

which yields the condition for substantive review to be preferred to procedural review. Thus, $W^P > W^S$ when the precision improvement from high effort outweighs the overseer's ability to induce $x_A = 0$ when $\omega = 1$: $V_\varepsilon(0) - V_\varepsilon(1) > p_1(2\beta-1)$. Otherwise, substantive review produces higher welfare for the overseer.

Finally, suppose that $\kappa < \kappa^S < \kappa^P$ so that $e = 1$ under both procedural and substantive review:

$$\begin{aligned} W^P(e=1|\beta \text{ moderate}, \theta > 2) &< W^S(e=1|\beta \text{ moderate}, \theta > 2), \\ -\beta^2 - V_\varepsilon(1) &< -p_0\beta^2 - p_1(1-\beta)^2 - p_\theta\beta^2 - V_\varepsilon(1), \\ 0 &< p_1(2\beta-1), \end{aligned}$$

which is always satisfied since $\beta > \frac{1}{2}$ in this setting. Thus, the overseer always benefits from substantive review in this setting. Taken together this implies that when $\theta > 2$ the only time the overseer's welfare is higher under procedural review is when (a) $\kappa \in \left(\kappa^S, \kappa^P\right]$ and (b) $V_\varepsilon(0) - V_\varepsilon(1) > p_1(2\beta-1)$.

Obfuscation through exaggeration. Let $\theta < 2$ so the agency obfuscates through exaggeration under substantive review. Suppose that $\kappa > \kappa^P > \kappa^S$ so that $e = 0$ under both procedural and substantive review:

$$
\begin{aligned}
W^P(e=0|\beta \text{ moderate}, \theta<2) &> W^S(e=0|\beta \text{ moderate}, \theta<2),\\
-p_0\beta^2 - p_1(1-\beta)^2 - p_\theta(\theta-\beta)^2 - V_\varepsilon(0) &> -p_0\beta^2 - p_1(1-\beta-\theta)^2 - p_\theta\beta^2 - V_\varepsilon(0),\\
0 &> p_1(1-\beta)^2 - p_1(1-\beta-\theta)^2 + p_\theta(\theta-\beta)^2 - p_\theta\beta^2.
\end{aligned}
$$

This inequality is always satisfied in this setting, which implies that the overseer always benefits from procedural review in this case.

Now suppose that $\kappa \in \left(\kappa^S, \kappa^P\right]$ so $e=0$ under substantive review and $e=1$ under procedural review:

$$
\begin{aligned}
W^P(e=1|\beta \text{ moderate}, \theta<2) &> W^S(e=0|\beta \text{ moderate}, \theta<2),\\
-\beta^2 - V_\varepsilon(1) &> -p_0\beta^2 - p_1(1-\beta-\theta)^2 - p_\theta\beta^2 - V_\varepsilon(0),\\
V_\varepsilon(0) - V_\varepsilon(1) + p_1 + 2p_1\beta\theta + p_1\theta^2 - 2p_1\beta - 2p_1\theta &> 0,\\
V_\varepsilon(0) - V_\varepsilon(1) + p_1(1+\theta(2\beta+\theta) - 2(\beta+\theta)) &> 0.
\end{aligned}
$$

Again, this inequality is always satisfied in this setting, which implies that the overseer once again benefits from procedural review.

Finally, suppose that $\kappa < \kappa^S < \kappa^P$ so that $e=1$ under both procedural and substantive review:

$$
\begin{aligned}
W^P(e=1|\beta \text{ moderate}, \theta<2) &> W^S(e=1|\beta \text{ moderate}, \theta<2),\\
-\beta^2 - V_\varepsilon(1) &> -p_0\beta^2 - p_1(1-\beta-\theta)^2 - p_\theta\beta^2 - V_\varepsilon(1),\\
p_1(1+\theta(2\beta+\theta) - 2(\beta+\theta)) &> 0.
\end{aligned}
$$

Once again this inequality is always satisfied and the overseer benefits from procedural review. Taken together this implies that when $\theta < 2$ the overseer's welfare is always higher under pure procedural review. ■