Online Supplemental Appendix:
Reviewing Procedure vs. Judging Substance: The Scope of Review and Bureaucratic Policymaking

Ian R. Turner*

Contents

*Assistant Professor, Department of Political Science, Yale University, Email: ian.turner@yale.edu.

A Procedural review

A.1 Agency substantive policy choice

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

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) the agency will be upheld and (2) the agency will be overturned.

Agency upheld. The agency's expected payoff for the proposed strategy is given by,[1]

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

Now suppose the agency deviated by choosing $x_A(\omega) = \omega + \delta$, where $\delta > 0$ denotes the deviation. Its expected payoff for doing so is given by,

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

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

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

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.

Agency overturned. The agency's payoff is equivalent in this case since the agency's choice of x_A will not change whether it is overturned and by this point in the game the only oversight-relevant

[1] Line 3 follows from the mean-variance property of quadratic utility in the presence of uncertainty (see, e.g., p. 649 in Callander, Steven. 2011. "Searching for Good Policies." *American Political Science Review* 105(4): 622–643). I will use this notation throughout.

choice, e, has been chosen. Thus, there is no incentive for the agency to deviate from setting policy so that $x_A(\omega) = \omega$. Taken together these two cases imply that, in weakly undominated strategies, the agency will always choose $x_A(\omega) = \omega$ in the procedural review model. ■

A.2 Optimal procedural oversight

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

$$s_R(e) = \begin{cases} \textit{Uphold: } r = 0 & \textit{if } V_{SQ} - V_\varepsilon(e) \geq p_1(2\beta - 1) + p_2(4\beta - 4), \\ \textit{Overturn: } r = 1 & \textit{otherwise.} \end{cases}$$

Proof of Lemma A.2. First, consider the overseer's expected payoff for upholding the agency following a choice of e:

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

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. The overseer's expected payoff for overturning given p_ω for each ω, is given by,

$$\begin{aligned} EU_R(r = 1|e, \beta, p_\omega) &= \mathbb{E}[-(\omega - \beta - (1-r)x)^2|e, p_\omega] - V_{SQ}, \\ &= p_0(-(0 - \beta - (0)x)^2 - V_{SQ}) + p_1(-(1 - \beta - (0)x)^2 - V_{SQ}) + p_2(-(2 - \beta - (0)x)^2 - V_{SQ}), \\ &= -p_0(\beta^2) - p_1((1-\beta)^2) - p_2((2-\beta)^2) - V_{SQ}. \end{aligned}$$

Combining and rearranging these two expected payoffs yields the incentive compatibility constraint that must be satisfied in order for the overseer to uphold:

$$\begin{aligned} -\beta^2 - V_\varepsilon(e) &\geq -p_0\beta^2 - p_1(1-\beta)^2 - p_2(2-\beta)^2 - V_{SQ}, \\ V_{SQ} - V_\varepsilon(e) &\geq p_1(2\beta - 1) + p_2(4\beta - 4), \end{aligned}$$

as stated in the lemma. ■

We can rearrange the condition to uphold in terms of overseer bias to yield an upper bound for upholding on β: $\beta \in \left(0, \frac{p_1 + 4p_2 + V_{SQ} - V_\varepsilon(e)}{2p_1 + 4p_2}\right]$. We can further define two β-thresholds based on whether the agency invested high or low effort: Let $\beta_1 := \frac{p_1 + 4p_2 + V_{SQ} - V_\varepsilon(1)}{2p_1 + 4p_2}$ and $\beta_0 := \frac{p_1 + 4p_2 + V_{SQ} - V_\varepsilon(0)}{2p_1 + 4p_2}$ 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 defer-*

ential. 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 section characterizes how the agency best responds with its effort choices conditional on these oversight regimes.

A.3 Agency effort investment

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 $V_{SQ} - V_\varepsilon(1) + p_1 + 4p_2 + \pi \geq \kappa$.*

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.

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,

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

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

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

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

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

That is, 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.

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. Policy outcomes are the same regardless of the agency's effort choice ($x = 0$) so any high effort investment simply produces a net cost κ. 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.

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,

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

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

$$\begin{aligned} EU_A(e=0|r^*(0)=1) &= \mathbb{E}[-(\omega-(1-1)x)^2|p_\omega] - V_{SQ} - \kappa(0) - \pi(1), \\ &= -\mathbb{E}[\omega^2|p_\omega] - V_{SQ} - \pi, \\ &= -p_0(0^2) - p_1(1^2) - p_2(2^2) - V_{SQ} - \pi, \\ &= -p_1 - 4p_2 - V_{SQ} - \pi. \end{aligned}$$

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

$$\begin{aligned} -V_\varepsilon(1) - \kappa &\geq -p_1 - 4p_2 - V_{SQ} - \pi, \\ V_{SQ} - V_\varepsilon(1) + p_1 + 4p_2 + \pi &\geq \kappa. \end{aligned}$$

This is case 3 in the result. Taken together the analysis above completes the proof. ■

Proposition 1. *In the equilibrium of the procedural review model the overseer makes review decisions according to $s_R(e)$, 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 $V_{SQ} - V_\varepsilon(1) + p_1 + 4p_2 + \pi \geq \kappa$.*

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

B.1 Truthful equilibrium

Lemma B.1. *When the agency sets substantive policy truthfully* $\left(i.e.,\ x_A^{truth}(\omega)\right)$ *the overseer's optimal review strategy, given effort investment e, is given by,*

$$s_R^*(x_A^{truth}(\omega), e) = \begin{cases} \textit{Uphold: } r=0 & \textit{if } \omega = 0 \textit{ and } V_{SQ} \geq V_\varepsilon(e), \\ & \textit{or } \omega = 1 \textit{ and } \beta < \frac{1+V_{SQ}-V_\varepsilon(e)}{2}, \\ & \textit{or } \omega = 2 \textit{ and } \beta \leq \frac{4+V_{SQ}-V_\varepsilon(e)}{4}, \\ \textit{Overturn: } r=1 & \textit{otherwise.} \end{cases}$$

Proof of Lemma B.1. There are three cases to check, assuming that the agency always matches policy to the state, $x_A(\omega) = \omega$: when $\omega = 0$, $\omega = 1$, and $\omega = 2$. Before analyzing each possibility, first note that the overseer's payoff is constant for all values of ω should she uphold the agency:

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

With this expected payoff for $r(e) = 0$ we can now proceed to the cases.

Case 1: $\omega = 0$**.** The overseer's expected payoff for reversing the agency when $\omega = 0$ and $x_A(0) = 0$, fixing e, is given by,

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

Incentive compatibility requires that the following condition hold for the overseer to uphold the agency when $\omega = 0$,

$$\begin{aligned} -\beta^2 - V_\varepsilon(e) &\geq -\beta^2 - V_{SQ}, \\ V_{SQ} - V_\varepsilon(e) &\geq 0. \end{aligned}$$

Thus, the overseer upholds the agency when $x_A^{\text{truth}}(0) = 0$ so long as $V_\varepsilon(e) \leq V_{SQ}$.

Case 2: $\omega = 1$. The overseer's expected payoff for reversing the agency when $\omega = 1$ and $x_A(1) = 1$, for a given e, is given by,

$$\begin{aligned} EU_R(r = 1 | x_A(1) = 1, e) &= \mathbb{E}[-(\omega - \beta - (1-r)x)^2 | x_A^{\text{truth}}, e] - (1)V_{SQ}, \\ &= -(1-\beta)^2 - V_{SQ}, \\ &= 2\beta - \beta^2 - 1 - V_{SQ}. \end{aligned}$$

For the overseer to uphold incentive compatibility requires that,

$$\begin{aligned} -\beta^2 - V_\varepsilon(e) &\geq 2\beta - \beta^2 - 1 - V_{SQ}, \\ V_{SQ} - V_\varepsilon(e) &\geq 2\beta - 1, \\ \frac{1 + V_{SQ} - V_\varepsilon(e)}{2} &\geq \beta. \end{aligned}$$

Case 3: $\omega = 2$. The overseer's expected payoff for reversing when $\omega = 2$ is given by,

$$\begin{aligned} EU_R(r = 1 | x_A(2) = 2, e) &= \mathbb{E}[-(\omega - \beta - (1-r)x)^2 | x_A^{\text{truth}}, e] - (1)V_{SQ}, \\ &= -(2-\beta)^2 - V_{SQ}, \\ &= 4\beta - \beta^2 - 4 - V_{SQ}. \end{aligned}$$

This yields the following incentive compatibility constraint to uphold:

$$\begin{aligned} -\beta^2 - V_\varepsilon(e) &\geq 4\beta - \beta^2 - 4 - V_{SQ}, \\ V_{SQ} - V_\varepsilon(e) &\geq 4\beta - 4, \\ \frac{4 + V_{SQ} - V_\varepsilon(e)}{4} &\geq \beta. \end{aligned}$$

Combining the cases analyzed above yields the result. ■

The oversight rule derived above leads to five cases based on the level of effort the agency invests earlier in the game. The cases, along with the technical conditions on β, are displayed in Table 1, which corresponds to Figure 1 in the main body.

With the overseer's review strategy in hand, I now turn to analysis of when the agency will truthfully set policy, and the accompanying effort investments in those cases. First, the following lemma is useful for the rest of the analysis.

Lemma B.2. *When $\omega = 0$ the agency always sets $x_A = 0$.*

ω	**Aligned Preferences:** $\beta \in \left[0, \frac{1+V_{SQ}-V_\varepsilon(0)}{2}\right)$	**Conditionally Aligned Preferences:** $\beta \in \left[\frac{1+V_{SQ}-V_\varepsilon(0)}{2}, \frac{1+V_{SQ}-V_\varepsilon(1)}{2}\right)$	**Moderate Preferences:** $\beta \in \left[\frac{1+V_{SQ}-V_\varepsilon(1)}{2}, \frac{4+V_{SQ}-V_\varepsilon(0)}{4}\right]$	**Conditionally Extreme Preferences:** $\beta \in \left(\frac{4+V_{SQ}-V_\varepsilon(0)}{4}, \frac{4+V_{SQ}-V_\varepsilon(1)}{4}\right]$	**Extreme Preferences:** $\beta > \frac{4+V_{SQ}-V_\varepsilon(1)}{4}$
0	$r(e)=0$ if $V_{SQ} \geq V_\varepsilon(e)$	$r(e)=0$ if $V_{SQ} \geq V_\varepsilon(e)$	$r(e)=0$ if $V_{SQ} \geq V_\varepsilon(e)$	$r(e)=0$ if $V_{SQ} \geq V_\varepsilon(e)$	$r(e)=0$ if $V_{SQ} \geq V_\varepsilon(e)$
1	$r(e)=0, \forall e$	$r(0)=1, r(1)=0$	$r(e)=1, \forall e$	$r(e)=1, \forall e$	$r(e)=1, \forall e$
2	$r(e)=0, \forall e$	$r(e)=0, \forall e$	$r(e)=0, \forall e$	$r(0)=1, r(1)=0$	$r(e)=1, \forall e$

Table 1: Overseer best responses given truthful policymaking, conditional on ω, e, and β.

Proof of Lemma B.2. First, note that there is no reason for the agency to deviate from $x_A = 0$ when $\omega = 0$ if it will be upheld by the overseer. Thus, we need only check whether the agency would benefit by deviating from $x_A = 0$ when it will be overturned. First, suppose that $r(0,e) = 1$ and $r(1,e) = 0$ so that deviating to $x_A = 1$ would lead to being upheld. The agency's expected utilities from $x_A = 0$ and $x_A = 1$ in this case are given by,

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

Incentive compatibility requires that the following inequality be satisfied for the agency to stick with $x_A(0) = 0$ even though $r(0,e) = 1$:

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

The LHS is positive since $\pi \in (0,1)$ and the RHS is negative since $r(0,e) = 1$ requires $V_{SQ} < V_\varepsilon(e)$. Thus, the condition is always satisfied, implying that the agency never benefits from deviating from $x_A(0) = 0$ even if it will be overturned. An analogous argument also rules out the possibility that the agency could benefit from deviating to $x_A = 2$ to be upheld. ■

Proposition 2. *There is a truthful separating equilibrium in which, for all ranges of preference disagreement, the agency always matches policy to the state if and only if reversal costs are not too punitive:* $V_\varepsilon(e) - V_{SQ} \geq \pi$.

Proof of Proposition 2. Lemma B.2 shows that the agency is always truthful when $\omega = 0$ so we derive the incentive compatibility conditions for the agency to truthfully reveal $\omega \in \{1,2\}$. First, consider the case in which $\omega = 1$. If the agency is upheld when it truthfully sets $x_A(1) = 1$ then there

is no incentive to deviate. Similarly, if the agency is overturned when $x_A(1) = 1$ and also overturned whenever $x_A = 0$ and $x_A = 2$ then there is no reason to deviate. Thus, we need only check whether the agency would deviate when $x_A(1) = 1$ is overturned but either $x_A = 0$ or $x_A = 2$ would be upheld. In either case the agency deviates spatially by one so expected utility is equivalent for deviating to $x_A = 0$ and $x_A = 2$ when $\omega = 1$ so we only show the derivations for deviating to $x_A = 0$ noting that the calculations are equivalent when $x_A = 2$ is the deviation. The agency's expected utilities for setting $x_A = 1$ truthfully and deviating by one to be upheld are given by,

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

To stick with truthfully setting $x_A(1) = 1$ incentive compatibility requires that the following inequality holds,

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

Now consider the case in which $\omega = 2$. Again, when $x_A(2) = 2$ is upheld there is no reason for the agency to deviate. When $x_A(2) = 2$ is overturned it must be the case that $x_A = 1$ is also overturned given that the preference divergence that leads to overturning $x_A(2) = 2$ is strictly larger than overturning $x_A = 1$. Thus, the only opportunity for the agency to deviate to be upheld is when $x_A = 0$, which only happens when $V_{SQ} \geq V_\varepsilon(e)$. The agency's expected utilities for sticking to $x_A(2) = 2$ when it will be overturned and deviating to $x_A = 0$ (assuming that will be upheld) are:

$$\begin{aligned}
EU_A(x_A = 2|\omega = 2, r(2,e) = 1) &= -(2-(1-1)x)^2 - (1)V_{SQ} - \kappa e - \pi(1), \\
&= -4 - V_{SQ} - \kappa e - \pi, \\
EU_A(x_A = 0|\omega = 2, r(0,e) = 0) &= -\mathbb{E}[(2-(1-0)(0+\varepsilon))^2|e] - (0)V_{SQ} - \kappa e - \pi(0), \\
&= -4 - V[\varepsilon|e] - \kappa e, \\
&= -4 - V_\varepsilon(e) - \kappa e.
\end{aligned}$$

For the agency to optimally set $x_A(2) = 2$ incentive compatibility requires that,

$$\begin{aligned} -4 - V_{SQ} - \kappa e - \pi &\geq -4 - V_\varepsilon(e) - \kappa e, \\ V_\varepsilon(e) - V_{SQ} &\geq \pi. \end{aligned}$$

Taken together, the agency never deviates from $x_A = 0$ when $\omega = 0$ even if it leads to being overturned and when $\omega \in \{1,2\}$ the agency will remain truthful (and separate) if and only if $V_\varepsilon(e) - V_{SQ} \geq \pi$, as stated in the result. The statement regarding the sufficient condition for high effort in the result follows from Lemma B.3. ■

The next result characterizes agency effort decisions assuming that it will subsequently set policy truthfully so that $x_A(\omega) = \omega$. Note that this requires that $V_\varepsilon(e) - V_{SQ} \geq \pi$ for all e, which requires that $V_{SQ} > V_\varepsilon(0) > V_\varepsilon(1)$.

Lemma B.3. *Assume $V_\varepsilon(e) - V_{SQ} \geq \pi$ for all e, which requires that $V_\varepsilon(0) > V_\varepsilon(1) > V_{SQ}$, so that the agency always sets policy truthfully. Conditional on $s_R(x_A, e)$ the agency makes effort investments as follows. The agency invests high effort when the overseer is aligned if $(p_1 + p_2)(V_\varepsilon(0) - V_\varepsilon(1)) \geq \kappa$, when the overseer is conditionally aligned if $p_1(1 + V_{SQ} - V_\varepsilon(1) + \pi) + p_2(V_\varepsilon(0) - V_\varepsilon(1)) \geq \kappa$, when the overseer is moderately biased if $p_2(V_\varepsilon(0) - V_\varepsilon(1)) \geq \kappa$, when the overseer is conditionally extreme if $p_2(4 + V_{SQ} - V_\varepsilon(1) + \pi) \geq \kappa$, and never invests high effort when the overseer is extremely biased.*

Proof of Lemma B.3. First, note that if $V_{SQ} > V_\varepsilon(0) > V_\varepsilon(1)$ then there cannot be a truthful separating equilibrium since for any e, $V_\varepsilon(e) - V_{SQ} < \pi$. Similarly, we set aside the case in which $V_\varepsilon(0) > V_{SQ} > V_\varepsilon(1)$ since at the moment we are interested in effort incentives assuming that the agency will subsequently set policy truthfully, which cannot hold for all ranges of overseer biases under this ordering following $e = 1$. Thus, the only case we need to characterize is when $V_{SQ} < V_\varepsilon(1) < V_\varepsilon(0)$. I will derive the condition for high effort in each of the agency-overseer preference environments assuming this ordering.

Consider the agency's generic expected utilities for $e = 1$ and $e = 0$, conditional on $r(x_A, e)$, given ω and $x_A^{\text{truth}}(\omega) = \omega$:

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

I will plug these general expected utilities into the relevant incentive compatibility conditions to

analyze each case given $r(x_A, e)$ from Lemma B.1.

Consider an aligned overseer: $\beta \in \left[0, \frac{1+V_{SQ}-V_\varepsilon(0)}{2}\right)$. In this case $r(0,e) = 1$ and $r(x,e) = 0$ for $x \in \{1,2\}$. The agency's expected utilities for investing high and low effort, respectively, are:

$$\begin{aligned} EU_A(e=1|x_A^{\text{truth}}(\omega), s_R(x_A,e), p) &= -p_0(V_{SQ}+\kappa+\pi) - p_1(V_\varepsilon(1)+\kappa) - p_2(V_\varepsilon(1)+\kappa), \\ EU_A(e=0|x_A^{\text{truth}}(\omega), s_R(x_A,e), p) &= -p_0(V_{SQ}+\pi) - p_1(V_\varepsilon(0)) - p_2(V_\varepsilon(0)). \end{aligned}$$

Incentive compatibility dictates that $e = 1$ if and only if,

$$\begin{aligned} -p_0(V_{SQ}+\kappa+\pi) - p_1(V_\varepsilon(1)+\kappa) - p_2(V_\varepsilon(1)+\kappa) &\geq -p_0(V_{SQ}+\pi) - p_1(V_\varepsilon(0)) - p_2(V_\varepsilon(0)), \\ (p_1+p_2)(V_\varepsilon(0) - V_\varepsilon(1)) &\geq \kappa. \end{aligned}$$

Now consider a conditionally aligned overseer: $\beta \in \left[\frac{1+V_{SQ}-V_\varepsilon(0)}{2}, \frac{1+V_{SQ}-V_\varepsilon(1)}{2}\right)$. In this case, $r(0,e) = 1$, $r(1,1) = 0$, $r(1,0) = 1$, and $r(2,e) = 0$. The agency's expected utilities for high and low effort are given by,

$$\begin{aligned} EU_A(e=1|x_A^{\text{truth}}(\omega), s_R(x_A,e), p) &= -p_0(V_{SQ}+\kappa+\pi) - p_1(V_\varepsilon(1)+\kappa) - p_2(V_\varepsilon(1)+\kappa), \\ EU_A(e=0|x_A^{\text{truth}}(\omega), s_R(x_A,e), p) &= -p_0(V_{SQ}+\pi) - p_1(1+V_{SQ}+\pi) - p_2(V_\varepsilon(0)). \end{aligned}$$

Incentive compatibility requires that the following inequality holds to support $e = 1$,

$$\begin{aligned} -p_0(V_{SQ}+\kappa+\pi) - p_1(V_\varepsilon(1)+\kappa) - p_2(V_\varepsilon(1)+\kappa) &\geq -p_0(V_{SQ}+\pi) - p_1(1+V_{SQ}+\pi) - p_2(V_\varepsilon(0)), \\ p_1(1+V_{SQ}-V_\varepsilon(1)+\pi) + p_2(V_\varepsilon(0) - V_\varepsilon(0)) &\geq \kappa. \end{aligned}$$

Now consider a moderately biased overseer: $\beta \in \left[\frac{1+V_{SQ}-V_\varepsilon(1)}{2}, \frac{4+V_{SQ}-V_\varepsilon(0)}{4}\right]$. In this case $r(0,e) = 1$, $r(1,e) = 1$, and $r(2,e) = 0$. The agency's expected utilities for high and low effort are given by,

$$\begin{aligned} EU_A(e=1|x_A^{\text{truth}}(\omega), s_R(x_A,e), p) &= -p_0(V_{SQ}+\kappa+\pi) - p_1(1+V_{SQ}+\kappa+\pi) - p_2(V_\varepsilon(1)+\kappa), \\ EU_A(e=0|x_A^{\text{truth}}(\omega), s_R(x_A,e), p) &= -p_0(V_{SQ}+\pi) - p_1(1+V_{SQ}+\pi) - p_2(V_\varepsilon(0)). \end{aligned}$$

Incentive compatibility dictates the $e = 1$ if and only if,

$$\begin{aligned} -p_0(V_{SQ}+\kappa+\pi) - p_1(1+V_{SQ}+\kappa+\pi) - p_2(V_\varepsilon(1)+\kappa) &\geq -p_0(V_{SQ}+\pi) - p_1(1+V_{SQ}+\pi) - p_2(V_\varepsilon(0)), \\ p_2(V_\varepsilon(0) - V_\varepsilon(1)) &\geq \kappa. \end{aligned}$$

Now consider a conditionally extreme overseer: $\beta \in \left(\frac{4+V_{SQ}-V_\varepsilon(0)}{4}, \frac{4+V_{SQ}-V_\varepsilon(1)}{4}\right]$. In this case $r(0,e) = 1$, $r(1,e) = 1$, $r(2,1) = 0$, and $r(2,0) = 1$. The agency's expected utilities in this case are

given by,

$$\begin{aligned} EU_A(e=1|x_A^{\text{truth}}(\omega), s_R(x_A,e), p) &= -p_0(V_{SQ}+\kappa+\pi) - p_1(1+V_{SQ}+\kappa+\pi) - p_2(V_\varepsilon(1)+\kappa), \\ EU_A(e=0|x_A^{\text{truth}}(\omega), s_R(x_A,e), p) &= -p_0(V_{SQ}+\pi) - p_1(1+V_{SQ}+\pi) - p_2(4+V_{SQ}+\pi). \end{aligned}$$

Incentive compatibility dictates that $e=1$ if and only if the following inequality holds,

$$\begin{aligned} -p_0(V_{SQ}+\kappa+\pi) - p_1(1+V_{SQ}+\kappa+\pi) - p_2(V_\varepsilon(1)+\kappa) &\geq -p_0(V_{SQ}+\pi) - p_1(1+V_{SQ}+\pi) - p_2(4+V_{SQ}+\pi), \\ p_2(4+V_{SQ}-V_\varepsilon(1)+\pi) &\geq \kappa. \end{aligned}$$

Finally, consider an extreme overseer: $\beta > \frac{4+V_{SQ}-V_\varepsilon(1)}{4}$. In this case the agency is reversed for any e given $x_A^{\text{truth}}(\omega) = \omega$. Accordingly, it is easy to show that investing $e=1$ is never optimal since outcomes never change, but the agency has to pay κ. Thus, when the agency will always be overturned it never invests high effort. ■

The following corollary states one of the main insights in the article: there is a trade-off between information and effort when oversight is substantive.

Corollary 1. *The incentive for the agency to obfuscate with its substantive policy choice is stronger when the agency invests high effort.*

Proof of Corollary 1. This follows from the fact that when $e=1$, relative to $e=0$, there is a larger set of π such that truthful policymaking does not hold. That is, $\{\pi : V_\varepsilon(e) - V_{SQ} < \pi | e=0\} \subset \{\pi : V_\varepsilon(e) - V_{SQ} < \pi | e=1\}$ for a fixed V_{SQ} since $V_\varepsilon(1) < V_\varepsilon(0)$. ■

Proposition B.1. *Suppose preferences are aligned: $\beta < \frac{1+V_{SQ}-V_\varepsilon(0)}{2}$. Then the agency sets policy truthfully and the overseer upholds the agency following $x_A = 0$ if $V_{SQ} \geq V_\varepsilon(e)$ and $x_A \in \{1,2\}$ for any e. Furthermore, when $V_{SQ} > V_\varepsilon(0) > V_\varepsilon(1)$ the agency will invest high effort if $V_\varepsilon(0) - V_\varepsilon(1) \geq \kappa$; when $V_\varepsilon(0) > V_{SQ} > V_\varepsilon(1)$ the agency invests high effort if $p_0(V_{SQ} - V_\varepsilon(1) + \pi) + (p_1 + p_2)(V_\varepsilon(0) - V_\varepsilon(1)) \geq \kappa$; and when $V_\varepsilon(0) > V_\varepsilon(1) > V_{SQ}$ the agency invests high effort if $(p_1 + p_2)(V_\varepsilon(0) - V_\varepsilon(1)) \geq \kappa$.*

Proof of Proposition B.1. From Lemma B.2 it follows that $x_A(0) = 0$ regardless of whether the agency will be upheld or not. Moreover, we know from Lemma B.1 that the overseer will uphold a truthful choice of $x_A = 0$ if and only if $V_{SQ} \geq V_\varepsilon(e)$ and any truthful policy change when $\beta < \frac{1+V_{SQ}-V_\varepsilon(0)}{2}$ even if $e=0$. Note that this holds for any ordering of agency and reversion variances: $V_{SQ} > V_\varepsilon(0) > V_\varepsilon(1)$, $V_\varepsilon(0) > V_{SQ} > V_\varepsilon(1)$, and $V_\varepsilon(0) > V_\varepsilon(1) > V_{SQ}$.

In terms of the sufficient condition for effort, consider first the case in which $V_{SQ} > V_\varepsilon(0) > V_\varepsilon(1)$ so that $r(0,e) = 0$, $r(1,e) = 0$, and $r(2,e) = 0$ given truthful policymaking (Lemma B.1). The

agency's expected utilities for high and low effort are given by,

$$\begin{aligned}
EU_A(e|x_A^{\text{truth}}(\omega), r(x_A,e)) &= \mathbb{E}[-(\omega-(1)(x_A^{\text{truth}}+\varepsilon))^2-(0)V_{SQ}-(0)\pi|e,r(x_A,e)],\\
EU_A(1|x_A^{\text{truth}}(\omega), r(x_A,e)) &= -p_0(V_\varepsilon(1)+\kappa)-p_1(V_\varepsilon(1)+\kappa)-p_2(V_\varepsilon(1)+\kappa),\\
EU_A(0|x_A^{\text{truth}}(\omega), r(x_A,e)) &= -p_0V_\varepsilon(0)-p_1V_\varepsilon(0)-p_2V_\varepsilon(0).
\end{aligned}$$

Combining and rearranging we get the incentive compatibility condition for high effort given that the agency will always be upheld, regardless of e, following truthful policymaking:

$$\begin{aligned}
-p_0V_\varepsilon(1)-p_1V_\varepsilon(1)-p_2V_\varepsilon(1)-\kappa &\geq -p_0V_\varepsilon(0)-p_1V_\varepsilon(0)-p_2V_\varepsilon(0),\\
V_\varepsilon(0)-V_\varepsilon(1) &\geq \kappa.
\end{aligned}$$

Now consider the case where $V_\varepsilon(0) > V_{SQ} > V_\varepsilon(1)$ so that $r(0,1)=0$, $r(0,0)=1$, $r(1,e)=0$, and $r(2,e)=0$. The agency's expected utility for high and low effort in this case are:

$$\begin{aligned}
EU_A(e|x_A^{\text{truth}}(\omega), r(x_A,e)) &= \mathbb{E}[-(\omega-(1)(x_A^{\text{truth}}+\varepsilon))^2-(0)V_{SQ}-(0)\pi|e,r(x_A,e)],\\
EU_A(1|x_A^{\text{truth}}(\omega), r(x_A,e)) &= -p_0(V_\varepsilon(1)+\kappa)-p_1(V_\varepsilon(1)+\kappa)-p_2(V_\varepsilon(1)+\kappa),\\
EU_A(0|x_A^{\text{truth}}(\omega), r(x_A,e)) &= -p_0(V_{SQ}+\pi)-p_1V_\varepsilon(0)-p_2V_\varepsilon(0).
\end{aligned}$$

This yields the following incentive compatibility condition for high effort when $x_A(0)=0$ is upheld only when $e=1$,

$$\begin{aligned}
-p_0V_\varepsilon(1)-p_1V_\varepsilon(1)-p_2V_\varepsilon(1)-\kappa &\geq -p_0(V_{SQ}+\pi)-p_1V_\varepsilon(0)-p_2V_\varepsilon(0),\\
p_0(V_{SQ}-V_\varepsilon(1)+\pi)+(p_1+p_2)(V_\varepsilon(0)-V_\varepsilon(1)) &\geq \kappa.
\end{aligned}$$

Finally, consider the case in which $V_\varepsilon(0) > V_\varepsilon(1) > V_{SQ}$ so that the agency is never upheld following $x_A(0)=0$: $r(0,e)=1$, $r(1,e)=0$, $r(2,e)=0$. The agency's expected utilities for high and low effort in this case are,

$$\begin{aligned}
EU_A(e|x_A^{\text{truth}}(\omega), r(x_A,e)) &= \mathbb{E}[-(\omega-(1)(x_A^{\text{truth}}+\varepsilon))^2-(0)V_{SQ}-(0)\pi|e,r(x_A,e)],\\
EU_A(1|x_A^{\text{truth}}(\omega), r(x_A,e)) &= -p_0(V_{SQ}+\kappa+\pi)-p_1(V_\varepsilon(1)+\kappa)-p_2(V_\varepsilon(1)+\kappa),\\
EU_A(0|x_A^{\text{truth}}(\omega), r(x_A,e)) &= -p_0(V_{SQ}+\pi)-p_1V_\varepsilon(0)-p_2V_\varepsilon(0).
\end{aligned}$$

Combining and rearranging yields the incentive compatibility for high effort in this environment,

$$\begin{aligned} -p_0(V_{SQ}+\pi)-p_1V_\varepsilon(1)-p_2V_\varepsilon(1)-\kappa &\geq -p_0(V_{SQ}+\pi)-p_1V_\varepsilon(0)-p_2V_\varepsilon(0), \\ (p_1+p_2)(V_\varepsilon(0)-V_\varepsilon(1)) &\geq \kappa. \end{aligned}$$

$(p_1+p_2)(V_\varepsilon(0)-V_\varepsilon(1)) \geq \kappa$ being sufficient to ensure the agency always invests high effort in this environment follows from inspection of the three incentive compatibility conditions. This is the most demanding condition in the sense that if it is satisfied then the other two conditions are necessarily satisfied. ■

B.2 Obfuscation equilibria

First, I establish two results that are useful in constructing obfuscation equilibria: (1) Following lemma B.2 the only possible pooling equilibrium involves the agency setting $x_A(\omega)=0$ for all ω, and (2) Regardless of the agency's policy strategy (e.g., pooling, semi-pooling, etc.) the overseer upholds any observation of $x_A=0$ if and only if $V_{SQ}\geq V_\varepsilon(e)$.

Corollary B.1. *If there is a pooling equilibrium then it involves the agency choosing $x_A(\omega)=0$ for all ω.*

Proof of Corollary B.1. This follows straightforwardly from Lemma B.2. If the agency will never deviate from $x_A(0)=0$, even when doing so would avoid reversal, then every agency policymaking strategy must involve $x_A(0)=0$ implying that if there is a pooling equilibrium then it involves $x(\omega)=0,\forall\omega$. ■

Lemma B.4. *For any agency policymaking strategy, the overseer upholds $x_A=0$, given e, if and only if $V_{SQ}\geq V_\varepsilon(e)$.*

Proof of Lemma B.4. Consider any (possibly mixed) agency policymaking strategy that involves setting $x_A=0$ with positive probability. (Note that Lemma B.2 ensures that so long as $p_0>0$ any agency strategy involves setting $x_A=0$ with positive probability.) Let $q_0 \coloneqq Pr(\omega=0|x_A=0)$, $q_1 \coloneqq Pr(\omega=1|x_A=0)$, and $q_2 \coloneqq Pr(\omega=2|x_A=0)$ represent the overseer's posterior beliefs over ω given observation of $x_A=0$. For instance, if the agency's equilibrium strategy is $x_A(\omega)=\omega$ then $q_0=1$ and $q_1=q_2=0$ or if $x_A(\omega)=\omega$ for $\omega\in\{0,1\}$ and $x_A(2)=2$ (in equilibrium) then $q_0=\frac{p_0}{p_0+p_1}$, $q_1=\frac{p_1}{p_0+p_1}$, and $q_2=0$. The overseer's expected utilities for upholding and overturning

following $x_A = 0$ are given by,

$$\begin{aligned}
EU_R(r=0|x_A=0) &= q_0(\mathbb{E}[-(0-\beta-(1-0)(0+\varepsilon))^2|e]-(0)V_{SQ})+q_1(\mathbb{E}[-(1-\beta-(1-0)(0+\varepsilon))^2|e]-(0)V_{SQ}) \\
&+ q_2(\mathbb{E}[-(2-\beta-(1-0)(0+\varepsilon))^2|e]-(0)V_{SQ}), \\
&= -q_0(\beta^2+V_\varepsilon(e))-q_1((1-\beta)^2+V_\varepsilon(e))-q_2((2-\beta)^2+V_\varepsilon(e)), \\
&= -q_0\beta^2-q_1(1-\beta)^2-q_2(2-\beta)^2-V_\varepsilon(e), \\
EU_R(r=1|x_A=0) &= q_0(-(0-\beta-(0)x)^2-V_{SQ})+q_1((1-\beta-(0)x)^2-V_{SQ})+q_2((2-\beta-(0)x)^2-V_{SQ}), \\
&= -q_0\beta^2-q_1(1-\beta)^2-q_2(2-\beta)^2-V_{SQ}.
\end{aligned}$$

Incentive compatibility requires that the following inequality hold for the overseer to uphold following $x_A = 0$,

$$\begin{aligned}
EU_R(r=0|x_A=0) &\geq EU_R(r=1|x_A=0), \\
-q_0\beta^2-q_1(1-\beta)^2-q_2(2-\beta)^2-V_\varepsilon(e) &\geq -q_0\beta^2-q_1(1-\beta)^2-q_2(2-\beta)^2-V_{SQ}, \\
V_{SQ} &\geq V_\varepsilon(e).
\end{aligned}$$

Thus, any time the agency sets $x_A = 0$, given e, the overseer upholds if and only if $V_{SQ} \geq V_\varepsilon(e)$. ■

B.2.1 Substantive policy choices

Pooling. The next result establishes the fact that if the agency is being overturned, given e, for either $x_A = 1$ or $x_A = 2$, or both, but would be upheld for instead setting $x_A = 0$ that it will do so in equilibrium.

Proposition B.2. *Suppose preference disagreement is such that any agency choice to change policy, $x_A \in \{1,2\}$, is overturned given e. Then there is a pooling equilibrium in which the agency sets $x_A(\omega) = 0$ for all ω and the overseer upholds $x_A = 0$ if $V_{SQ} \geq V_\varepsilon(e)$.*

Proof of Proposition B.2. First, Lemma B.4 ensures that the overseer is best responding to the agency's pooling policymaking strategy by upholding only when $V_{SQ} \geq V_\varepsilon(e)$.

Lemma B.2 ensures that the agency always sets $x_A(0) = 0$. Now assume $x_A = 1$ leads to being overturned but $x_A = 0$ leads to being upheld, implying that $V_{SQ} \geq V_\varepsilon(e)$. The agency's expected utilities in that case are given by,

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

The agency benefits from deviating to $x_A = 0$, given incentive compatibility, if and only if,

$$\begin{aligned} -\omega^2 - V_\varepsilon(e) - \kappa e &\geq -\omega^2 - V_{SQ} - \kappa e - \pi, \\ \pi &\geq V_\varepsilon(e) - V_{SQ}. \end{aligned}$$

This is always satisfied since $r(0,e) = 0$ if and only if $V_{SQ} \geq V_\varepsilon(e)$ and $\pi \in (0,1)$ (i.e., the LHS is positive and the RHS is negative given that $r(0,e) = 0$). An analogous argument shows that the agency also benefits by setting $x_A(2) = 0$ rather than $x_A(2) \in \{1,2\}$ when $r(0,e) = 0$, $r(1,e) = 1$, and $r(2,e) = 1$.

Now assume $V_{SQ} < V_\varepsilon(e)$ so that the overseer overturns $x_A = 0$. Since the overseer also overturns $x_A = 1$ and $x_A = 2$ due to being extremely biased there is no benefit to the agency for deviating from always setting $x_A = 0$. Finally, note that upholding $x_A = 0$ if and only $V_{SQ} \geq V_\varepsilon(e)$ given that the agency is pooling on $x_A = 0$ follows from Lemma B.4. ■

Semi-pooling. The next results characterize two types of semi-pooling equilibria.

Proposition 3. *Assume the overseer is moderately biased, $\beta \in \left[\frac{1+V_{SQ}-V_\varepsilon(e)}{2}, \frac{4+V_{SQ}-V_\varepsilon(e)}{4}\right]$ and $\pi > V_\varepsilon(e) - V_{SQ}$ so that a truthful separating equilibrium does not exist. If $V_{SQ} \geq V_\varepsilon(e)$ then there is a pure strategy semi-pooling equilibrium in which the agency sets $x_A(\omega) = 0$ for $\omega \in \{0,1\}$ and $x_A(2) = 2$, and the overseer upholds $x_A = 0$, overturns $x_A = 1$, and upholds $x_A = 2$.*

Proof of Proposition 3. Suppose the agency sets $x_A(\omega) = 0$ for $\omega \in \{0,1\}$ and $x_A(2) = 2$. The derivation of overseer best responses to truthful policymaking in Lemma B.1 and the assumption that the overseer is moderately biased ensure that the overseer is best responding by upholding $x_A(2) = 2$. The equilibrium also follows from overseer off-path beliefs such that observation of $x_A = 1$ induces belief $b_R^*(\omega = 1 | x_A = 1) = Pr(\omega = 1 | x_A = 1) = 1$, which is consistent with PBE and leads the overseer to overturn $x_A = 1$ due to being moderately biased (per the overseer best responses derived in Lemma B.1). The overseer's posterior beliefs following observation of $x_A = 0$ given the agency's strategy are given by,

$$\begin{aligned} b_R^*(\omega = 0 | x_A = 0) &= \frac{p_0}{p_0 + p_1}, \\ b_R^*(\omega = 1 | x_A = 0) &= \frac{p_1}{p_0 + p_1}. \end{aligned}$$

By Lemma B.4 the overseer upholds $x_A = 0$ if and only if $V_{SQ} \geq V_\varepsilon(e)$, as stated in the result.

Since the agency is upheld when $x_A(0) = 0$ and $x_A(2) = 2$ there is no reason to deviate from the stated equilibrium strategy when $\omega \in \{0,2\}$. The assumption that $\pi > V_\varepsilon(e) - V_{SQ}$ ensures, per Proposition 2, that $x_A(1) = 0$ is also a best response when $r(0,e) = 0$, which requires that $V_{SQ} \geq V_\varepsilon(e)$ as stated in the result. ■

Proposition 4. *Assume the overseer is moderately biased,* $\beta \in \left[\frac{1+V_{SQ}-V_\varepsilon(e)}{2}, \frac{4+V_{SQ}-V_\varepsilon(e)}{4}\right]$ *and* $\pi > V_\varepsilon(e) - V_{SQ}$ *so that a truthful separating equilibrium does not exist. If* $\omega = 2$ *is sufficiently likely relative to* $\omega = 1$*:* $\frac{p_1}{p_1+p_2} \leq \frac{1}{4}(4-4\beta+V_{SQ}-V_\varepsilon(e))$ *then there is a pure strategy semi-pooling equilibrium in which the agency sets* $x_A(\omega) = \omega$ *for* $\omega \in \{0,2\}$ *and* $x_A = 2$ *when* $\omega = 1$*, and the overseer upholds* $x_A = 0$ *if* $V_{SQ} \geq V_\varepsilon(e)$*, overturns* $x_A = 1$*, and upholds* $x_A = 2$*.*

Proof of Proposition 4. Suppose that the agency sets policy according to the following strategy:

$$x_A(\omega) = \begin{cases} 0 & \text{if } \omega = 0, \\ 2 & \text{if } \omega \in \{1,2\}. \end{cases} \tag{1}$$

From Lemma B.4, the overseer upholds $x_A(0) = 0$ if and only if $V_{SQ} \geq V_\varepsilon(e)$ and, from Lemma B.1, does not uphold $x_A(1) = 1$ for any e given moderately biased preferences. In this equilibrium the overseer's off-path beliefs are such that observation of $x_A = 1$ induces the belief $b_R^*(\omega = 1|x_A = 1) = Pr(\omega = 1|x_A = 1) = 1$, which is consistent with PBE. Thus, we need only check whether $r(x_A = 2, e) = 0$ is incentive compatible. First, note that the overseer's equilibrium posterior beliefs about ω following $x_A = 2$ are given by,

$$\begin{aligned} b_R^*(\omega = 1|x_A = 2) &= \frac{p_1}{p_1+p_2}, \\ b_R^*(\omega = 2|x_A = 2) &= \frac{p_2}{p_1+p_2} \end{aligned}$$

Let $b_R^*(\omega = 1|x_A = 2) \coloneqq q$ and $b_R^*(\omega = 2|x_A = 2) \coloneqq (1-q)$. Given these posterior beliefs the overseer's expected utilities for upholding and overturning $x_A = 2$ in this case are given by,

$$\begin{aligned} EU_R(r=0|x_A=2,b_R^*) &= -q((1-\beta-(1-0)(2+\varepsilon))^2+(0)V_{SQ})-(1-q)((2-\beta-(1-0)(2+\varepsilon))^2+(0)V_{SQ}), \\ &= -q((1-\beta-2)^2+V_\varepsilon(e))-(1-q)((2-\beta-2)^2+V_\varepsilon(e)), \\ &= -q(\beta+1)^2-(1-q)\beta^2-V_\varepsilon(e), \\ EU_R(r=1|x_A=2,b_R^*) &= -q((1-\beta-(0)x)^2+(1)V_{SQ})-(1-q)((2-\beta-(0)x)^2+(1)V_{SQ}), \\ &= -q((1-\beta)^2)-(1-q)((2-\beta)^2)-V_{SQ}. \end{aligned}$$

Combining and re-arranging provides the incentive compatibility condition that must be met in order for the overseer to uphold $x_A = 2$:

$$\begin{aligned} EU_R(r=0|x_A=2,b_R^*) &\geq EU_R(r=1|x_A=2,b_R^*), \\ -q(\beta+1)^2-(1-q)\beta^2-V_\varepsilon(e) &\geq -q((1-\beta)^2)-(1-q)((2-\beta)^2)-V_{SQ}, \\ \frac{1}{4}(4-4\beta+V_{SQ}-V_\varepsilon(e)) &\geq q\left(\coloneqq \frac{p_1}{p_1+p_2}\right), \\ \frac{1}{4}(4-4\beta+V_{SQ}-V_\varepsilon(e)) &\geq \frac{p_1}{p_1+p_2}, \end{aligned}$$

as stated in the result.

The agency is best responding when $x_A = 0$ since Lemma B.2 shows that the agency never benefits by deviating from $x_A = 0$ when $\omega = 0$. Moreover, since $\pi > V_\varepsilon(e) - V_{SQ}$ truthful separating is not a best response when $\omega = 1$ if there is a deviation that will lead to being upheld. Thus, the deviation from $x_A(1) = 1$ to $x_A(1) = 2$ is optimal in this case so long as $r(2,e) = 0$ which requires that $\frac{p_1}{p_1+p_2} \leq \frac{1}{4}(4 - 4\beta + V_{SQ} - V_\varepsilon(e))$. Finally, the agency clearly has no reason to deviate from $x(2) = 2$ since the overseer is upholding $x_A = 2$ in equilibrium. ■

B.2.2 Effort choices

Before going through each preference environment it is useful to derive some general expected utility expressions that are subsequently plugged into the specific cases below. When the agency chooses its effort investment it does not yet know what state will obtain. Thus, the expected utilities below are scaled by p_ω and can be plugged into overall expected utility expressions by scaling each possibility by p_ω given the subsequent substantive policy strategy the agency will pursue in each case.

First, consider the agency's expected utility for investing effort e given state ω (that is realized with probability p_ω) when it can subsequently set policy truthfully and be upheld:

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

Now consider the analogous expected utility expression when the agency will set policy truthfully, but will be overturned. This requires that there be no profitable deviations to avoid reversal (including cases in which $V_\varepsilon(e) - V_{SQ} > \pi$) and is given by,

$$\begin{aligned} EU_A(e|p_\omega, x_A^{\text{truth}}, r = 1) &= p_\omega(-(\omega - (1-1)x)^2 - (1)V_{SQ} - \kappa e - \pi(1)), \\ &= p_\omega(-\omega^2 - V_{SQ} - \kappa e - \pi), \\ &= -p_\omega(\omega^2 + V_{SQ} + \kappa e + \pi). \end{aligned}$$

Finally, the other possibility is that the agency obfuscates by setting $x_A \neq \omega$ in order to avoid reversal. The expression for this case is given by,

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

These expressions are used in the proofs below to derive overall incentive compatibility conditions for each preference environment conditional on the type of policymaking strategy the agency will subsequently play after it learns ω (pooling, semi-pooling, etc.). In a sense the agency's effort choice dictates the type of policymaking strategy available once the agency learns ω, which in turn informs the agency's effort choices conditional on the likelihood of each state being realized (and the substantive policy strategy that is optimal in that state conditional on overseer bias). Before turning to the specific environments it is useful to derive an example to illustrate how these general expressions will be used in the proofs below.

Consider an environment in which $\beta \in \left[\frac{1+V_{SQ}-V_{\varepsilon}(1)}{2}, \frac{4+V_{SQ}-V_{\varepsilon}(0)}{4}\right]$ (moderate preference disagreement), the policy environment is moderately volatile, $V_{\varepsilon}(0) > V_{SQ} > V_{\varepsilon}(1)$, and reversal costs are moderately punitive, $V_{\varepsilon}(0) - V_{SQ} > \pi > V_{\varepsilon}(1) - V_{SQ}$. Using the general expression above for each possible state given that when $e = 1$ the agency is upheld for setting policy truthfully if $\omega = 0$ or $\omega = 2$ and will obfuscate by setting $x_A = 0$ or $x_A = 2$ (if the conditions for Proposition 4 hold) when $\omega = 1$ while when $e = 0$ the agency is overturned when $\omega = 0$ because it truthfully sets $x_A = 0$ and preferences are moderate, does not deviate when $\omega = 1$ to be upheld and instead sets policy truthfully and accepts being overturned since reversal costs are only moderately punitive, and is upheld for truthfully setting $x_A = 2$ when $\omega = 2$. These cases dictate which expression from above applies to each possible state conditional on the agency's effort investment. This leads to the following expected utility expressions for $e = 1$ and $e = 0$:

$$\begin{aligned} EU_A(e=1) &= -p_0(V_{\varepsilon}(1)+\kappa) - p_1(1+V_{\varepsilon}(1)+\kappa) - p_2(V_{\varepsilon}(1)+\kappa), \\ EU_A(e=0) &= -p_0(V_{SQ}+\pi) - p_1(1+V_{SQ}+\pi) - p_2(V_{\varepsilon}(0)). \end{aligned}$$

Now, combining and rearranging these expressions yields the incentive compatibility condition that must be met for the agency to invest high effort in this environment:

$$\begin{aligned} -p_0V_{\varepsilon}(1) - p_1(1+V_{\varepsilon}(1)) - p_2V_{\varepsilon}(1) - \kappa &\geq -p_0(V_{SQ}+\pi) - p_1(1+V_{SQ}+\pi) - p_2(V_{\varepsilon}(0)), \\ (p_0+p_1)(V_{SQ}-V_{\varepsilon}(1)+\pi) + p_2(V_{\varepsilon}(0)-V_{\varepsilon}(1)) &\geq \kappa. \end{aligned}$$

That is, in this environment the agency will be reversed in states zero and one if it invests low effort, but it invests high effort it will be upheld for truthful policymaking in state zero and will obfuscate to be upheld in state one. In state two the agency will be upheld for truthful policymaking for both low and high effort. Thus, conditional on states zero or one obtaining (probability $p_0 + p_1$) the agency invests high effort if the implementation improvement from doing so ($V_{SQ} - V_{\varepsilon}(1)$) and the benefit of avoiding reversal (π) are high enough, and conditional on state two obtaining (probability p_2) the agency wants to invest high effort if the implementation improvement between high and low effort

investment are high enough ($V_\varepsilon(0) - V_\varepsilon(1)$). As long as those collective potential benefits are higher than the cost of investing high effort then the agency will do so. Equivalent derivations produce analogous incentive compatibility conditions for each preference environment conditional on the volatility of the policy environment (ordering of $V_{SQ}, V_\varepsilon(0)$, and $V_\varepsilon(1)$) and how punitive reversal is (ordering of π and $V_\varepsilon(e) - V_{SQ}$ for both e), captured in the next set of results.

Lemma B.5. *Assume preferences are aligned: $\beta < \frac{1+V_{SQ}-V_\varepsilon(0)}{2}$. If reversal costs are highly punitive so that the agency always obfuscates to be upheld when possible, $\pi > V_\varepsilon(0) - V_{SQ} > V_\varepsilon(1) - V_{SQ}$, it invests high effort as follows:*

- *When $V_{SQ} > V_\varepsilon(0) > V_\varepsilon(1)$ the agency invests high effort if $V_\varepsilon(0) - V_\varepsilon(1) \geq \kappa$.*
- *When $V_\varepsilon(0) > V_{SQ} > V_\varepsilon(1)$ the agency invests high effort if $p_0(V_{SQ} - V_\varepsilon(1) + \pi) + (p_1 + p_2)(V_\varepsilon(0) - V_\varepsilon(1)) \geq \kappa$.*
- *When $V_\varepsilon(0) > V_\varepsilon(1) > V_{SQ}$ the agency invests high effort if $(p_1 + p_2)(V_\varepsilon(0) - V_\varepsilon(1)) \geq \kappa$*

If instead reversal costs are moderately punitive so that the agency obfuscates when possible to avoid reversal only after $e = 1$, $V_\varepsilon(0) - V_{SQ} > \pi > V_\varepsilon(1) - V_{SQ}$, it invests high effort as follows:

- *When $V_\varepsilon(0) > V_{SQ} > V_\varepsilon(1)$ the agency invests high effort if $p_0(V_{SQ} - V_\varepsilon(1) + \pi) + (p_1 + p_2)(V_\varepsilon(0) - V_\varepsilon(1)) \geq \kappa$.*
- *When $V_\varepsilon(0) > V_\varepsilon(1) > V_{SQ}$ the agency invests high effort if $(p_1 + p_2)(V_\varepsilon(0) - V_\varepsilon(1)) \geq \kappa$.*

Proof of Lemma B.5. The result follows from plugging in the relevant general expression from above that correspond to the appropriate potential state and the subsequent equilibrium oversight and policymaking behavior.

Assume first that $\pi > V_\varepsilon(e) - V_{SQ}$ for all e. When $V_{SQ} > V_\varepsilon(0) > V_\varepsilon(1)$ the agency will set policy truthfully in all states, regardless of effort, and be upheld in all cases due to preferences being aligned. Plugging in the relevant expressions yields the agency's expected utility for high and low effort:

$$\begin{aligned}
EU_A(e=1) &= -p_0(V_\varepsilon(1)+\kappa) - p_1(V_\varepsilon(1)+\kappa) - p_2(V_\varepsilon(1)+\kappa), \\
&= -V_\varepsilon(1) - \kappa, \\
EU_A(e=0) &= -p_0 V_\varepsilon(0) - p_1 V_\varepsilon(0) - p_2 V_\varepsilon(0), \\
&= -V_\varepsilon(0).
\end{aligned}$$

Incentive compatibility requires that the following inequality hold for the agency to invest high effort in equilibrium,

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

When $V_\varepsilon(0) > V_{SQ} > V_\varepsilon(1)$ the agency will set policy truthfully and be upheld in all states if $e=1$ and will set policy truthfully if $e=0$ but will be overturned when $\omega = 0$ and upheld when $\omega \in \{1,2\}$, yielding the following expected utility expressions:

$$\begin{aligned} EU_A(e=1) &= -p_0(V_\varepsilon(1)+\kappa)-p_1(V_\varepsilon(1)+\kappa)-p_2(V_\varepsilon(1)+\kappa), \\ &= -V_\varepsilon(1)-\kappa, \\ EU_A(e=0) &= -p_0(V_{SQ}+\pi)-p_1V_\varepsilon(0)-p_2V_\varepsilon(0), \\ &= -V_\varepsilon(0). \end{aligned}$$

Incentive compatibility requires that the following inequality hold for the agency to invest high effort in equilibrium,

$$\begin{aligned} -V_\varepsilon(1)-\kappa &\geq -p_0(V_{SQ}+\pi)-p_1V_\varepsilon(0)-p_2V_\varepsilon(0), \\ p_0(V_{SQ}-V_\varepsilon(1)+\pi)+(p_1+p_2)(V_\varepsilon(0)-V_\varepsilon(1)) &\geq \kappa. \end{aligned}$$

When $V_\varepsilon(0) > V_\varepsilon(1) > V_{SQ}$ the agency always sets policy truthfully regardless of effort but is overturned for $x_A = 0$ and upheld for $x_A \in \{1,2\}$, yielding the following expected utilities for high and low effort:

$$\begin{aligned} EU_A(e=1) &= -p_0(V_{SQ}+\kappa+\pi)-p_1(V_\varepsilon(1)+\kappa)-p_2(V_\varepsilon(1)+\kappa), \\ &= -p_0(V_{SQ}+\pi)-p_1V_\varepsilon(1)-p_2V_\varepsilon(1)-\kappa, \\ EU_A(e=0) &= -p_0(V_{SQ}+\pi)-p_1V_\varepsilon(0)-p_2V_\varepsilon(0). \end{aligned}$$

Combining and rearranging yields the agency's incentive compatible condition for high effort,

$$\begin{aligned} -p_0(V_{SQ}+\pi)-p_1V_\varepsilon(1)-p_2V_\varepsilon(1)-\kappa &\geq -p_0(V_{SQ}+\pi)-p_1V_\varepsilon(0)-p_2V_\varepsilon(0), \\ (p_1+p_2)(V_\varepsilon(0)-V_\varepsilon(1)) &\geq \kappa. \end{aligned}$$

This set of conditions constitutes the first part of the result in which $\pi > V_\varepsilon(e) - V_{SQ}$ for all e.

Now let $V_\varepsilon(0) - V_{SQ} > \pi > V_\varepsilon(1) - V_{SQ}$. Note that $V_{SQ} > V_\varepsilon(0) > V_\varepsilon(1)$ can not be true in this

setting. Because preference are aligned agency-overseer equilibrium behavior is the same: There is no reason to obfuscate since the agency is upheld for setting policy truthfully when $\omega \in \{1,2\}$ and lemma B.2 implies the agency also never deviates when $\omega = 0$. Thus, the incentive compatibility conditions for high effort are equivalent to those derived for the $\pi > V_\varepsilon(e) - V_{SQ}$ for all e cases above, as stated in the result. ■

Lemma B.6. *Assume preferences are conditionally aligned:* $\beta \in \left[\frac{1+V_{SQ}-V_\varepsilon(0)}{2}, \frac{1+V_{SQ}-V_\varepsilon(1)}{2}\right)$. *If reversal costs are highly punitive so that the agency will always obfuscate to be upheld when possible,* $\pi > V_\varepsilon(0) - V_{SQ} > V_\varepsilon(1) - V_{SQ}$, *it invests high effort as follows:*

- *When* $V_{SQ} > V_\varepsilon(0) > V_\varepsilon(1)$ *the agency invests high effort if* $p_1 + V_\varepsilon(0) - V_\varepsilon(1) \geq \kappa$.
- *When* $V_\varepsilon(0) > V_{SQ} > V_\varepsilon(1)$ *the agency invests high effort if* $p_0(V_{SQ} - V_\varepsilon(1) + \pi) + p_1 + (p_1 + p_2)(V_\varepsilon(0) - V_\varepsilon(1)) \geq \kappa$ *assuming proposition 4 conditions hold so it can obfuscate with* $x_A(1) = 2$ *to be upheld and invests high effort if* $(p_0 + p_1)(V_{SQ} - V_\varepsilon(1) + \pi) + p_1 + p_2(V_\varepsilon(0) - V_\varepsilon(1)) \geq \kappa$ *when proposition 4 conditions do not hold.*
- *When* $V_\varepsilon(0) > V_\varepsilon(1) > V_{SQ}$ *the agency invests high effort if* $p_1 + (p_1 + p_2)(V_\varepsilon(0) - V_\varepsilon(1)) \geq \kappa$ *assuming proposition 4 conditions hold so it can obfuscate with* $x_A(1) = 2$ *to be upheld and invests high effort if* $p_1(1 + V_{SQ} - V_\varepsilon(1) + \pi) + p_2(V_\varepsilon(0) - V_\varepsilon(1)) \geq \kappa$ *when proposition 4 conditions do not hold.*

If instead reversal costs are moderately punitive so that the agency obfuscates when possible to avoid reversal only after $e = 1$, $V_\varepsilon(0) - V_{SQ} > \pi > V_\varepsilon(1) - V_{SQ}$, *it invests high effort as follows:*

- *When* $V_\varepsilon(0) > V_{SQ} > V_\varepsilon(1)$ *the agency invests high effort if* $(p_0 + p_1)(V_{SQ} - V_\varepsilon(1) + \pi) + p_1 + p_2(V_\varepsilon(0) - V_\varepsilon(1)) \geq \kappa$.
- *When* $V_\varepsilon(0) > V_\varepsilon(1) > V_{SQ}$ *the agency invests high effort if* $p_1(1 + V_{SQ} - V_\varepsilon(1) + \pi) + p_2(V_\varepsilon(0) - V_\varepsilon(1)) \geq \kappa$.

Proof of Lemma B.6. To derive the results I plug in the relevant general expressions from above that correspond to the appropriate potential state and the subsequent equilibrium oversight and policy-making behavior.

First consider $\pi > V_\varepsilon(e) - V_{SQ}$ for all e. When $V_{SQ} > V_\varepsilon(0) > V_\varepsilon(1)$ the agency will set policy truthfully in all states if it invested high effort and will be upheld in all cases. If instead the agency invests low effort it will set policy truthfully when $\omega = 0$ and be upheld, obfuscate by setting either $x_A = 0$ or $x_A = 2$ when $\omega = 1$ and be upheld, and set policy truthfully and be upheld when $\omega = 2$. Plugging in the relevant expressions yields the agency's expected utility for high and low effort in

this setting:

$$\begin{aligned} EU_A(e=1) &= -p_0(V_\varepsilon(1)+\kappa)-p_1(V_\varepsilon(1)+\kappa)-p_2(V_\varepsilon(1)+\kappa), \\ &= -V_\varepsilon(1)-\kappa, \\ EU_A(e=0) &= -p_0V_\varepsilon(0)-p_1(1+V_\varepsilon(0))-p_2V_\varepsilon(0), \\ &= -(p_0+p_2)(V_\varepsilon(0))-p_1(1+V_\varepsilon(0)). \end{aligned}$$

Incentive compatibility requires that the following inequality hold for the agency to invest high effort in equilibrium,

$$\begin{aligned} -V_\varepsilon(1)-\kappa &\geq -(p_0+p_2)(V_\varepsilon(0))-p_1(1+V_\varepsilon(0)), \\ p_1+V_\varepsilon(0)-V_\varepsilon(1) &\geq \kappa. \end{aligned}$$

When $V_\varepsilon(0) > V_{SQ} > V_\varepsilon(1)$ and the agency invests high effort it can again set policy truthfully and be upheld in all states, while if $e=0$ it will be overturned for truthfully setting policy when $\omega=0$, obfuscate to $x_A=2$ when $\omega=1$ if the conditions in proposition 4 are satisfied, truthfully set policy and be reversed if $\omega=1$ and the conditions in proposition 4 are not satisfied, and set policy truthfully and be upheld when $\omega=2$. This yields the following expected utility expressions:

$$\begin{aligned} EU_A(e=1) &= -p_0(V_\varepsilon(1)+\kappa)-p_1(V_\varepsilon(1)+\kappa)-p_2(V_\varepsilon(1)+\kappa), \\ &= -V_\varepsilon(1)-\kappa, \\ EU_A(e=0|\text{obfuscate}) &= -p_0(V_{SQ}+\pi)-p_1(1+V_\varepsilon(0))-p_2V_\varepsilon(0), \\ EU_A(e=0|\text{not obfuscate}) &= -p_0(V_{SQ}+\pi)-p_1(1+V_{SQ}+\pi)-p_2(V_\varepsilon(0)). \end{aligned}$$

If the agency can obfuscate then its incentive compatibility condition to invest high effort is,

$$\begin{aligned} -V_\varepsilon(1)-\kappa &\geq -p_0(V_{SQ}+\pi)-p_1(1+V_\varepsilon(0))-p_2V_\varepsilon(0), \\ p_0(V_{SQ}+\pi)+p_1+(p_1+p_2)(V_\varepsilon(0)-V_\varepsilon(1)) &\geq \kappa. \end{aligned}$$

If instead the agency cannot obfuscate then its incentive compatibility condition is,

$$\begin{aligned} -V_\varepsilon(1)-\kappa &\geq -p_0(V_{SQ}+\pi)-p_1(1+V_{SQ}+\pi)-p_2(V_\varepsilon(0)), \\ (p_0+p_1)(V_{SQ}-V_\varepsilon(1)+\pi)+p_1+p_2(V_\varepsilon(0)-V_\varepsilon(1)) &\geq \kappa. \end{aligned}$$

When $V_\varepsilon(0) > V_\varepsilon(1) > V_{SQ}$ and the agency invests high effort it sets policy truthfully and is overturned when $\omega=0$ since $V_{SQ} < V_\varepsilon(1)$, and sets policy truthfully and is upheld when $\omega=1$ or

$\omega = 2$. If it invests low effort then it sets policy truthfully when $\omega = 0$ and is overturned, obfuscates by setting $x_A = 2$ when $\omega = 1$ and is upheld when the proposition 4 conditions are satisfied and sets policy truthfully when $\omega = 1$ and is reversed when those conditions are not satisfied, and sets policy truthfully and is upheld when $\omega = 2$. Together this yields the following expected utility expressions:

$$\begin{aligned}
EU_A(e=1) &= -p_0(V_{SQ}+\kappa+\pi) - p_1(V_\varepsilon(1)+\kappa) - p_2(V_\varepsilon(1)+\kappa), \\
&= -p_0(V_{SQ}+\pi) - p_1 V_\varepsilon(1) - p_2 V_\varepsilon(1) - \kappa, \\
EU_A(e=0|\text{obfuscate}) &= -p_0(V_{SQ}+\pi) - p_1(1+V_\varepsilon(0)) - p_2 V_\varepsilon(0), \\
EU_A(e=0|\text{not obfuscate}) &= -p_0(V_{SQ}+\pi) - p_1(1+V_{SQ}+\pi) - p_2 V_\varepsilon(0).
\end{aligned}$$

Combining and rearranging yields the agency's incentive compatible condition for high effort when it can obfuscate,

$$\begin{aligned}
-p_0(V_{SQ}+\pi) - p_1 V_\varepsilon(1) - p_2 V_\varepsilon(1) - \kappa &\geq -p_0(V_{SQ}+\pi) - p_1(1+V_\varepsilon(0)) - p_2 V_\varepsilon(0), \\
p_1(1+V_\varepsilon(0)-V_\varepsilon(1)) + p_2(V_\varepsilon(0)-V_\varepsilon(1)) &\geq \kappa.
\end{aligned}$$

When the agency cannot obfuscate its incentive compatible condition for high effort is,

$$\begin{aligned}
-p_0(V_{SQ}+\pi) - p_1 V_\varepsilon(1) - p_2 V_\varepsilon(1) - \kappa &\geq -p_0(V_{SQ}+\pi) - p_1(1+V_{SQ}+\pi) - p_2 V_\varepsilon(0), \\
p_1(1+V_{SQ}-V_\varepsilon(1)+\pi) + p_2(V_\varepsilon(0)-V_\varepsilon(1)) &\geq \kappa.
\end{aligned}$$

This set of conditions constitutes the first part of the result in which $\pi > V_\varepsilon(e) - V_{SQ}$ for all e.

Now let $V_\varepsilon(0) - V_{SQ} > \pi > V_\varepsilon(1) - V_{SQ}$. The incentive compatibility conditions are the same in this case as those above when the agency cannot obfuscate. The difference here in equilibrium is that the agency would never obfuscate, so it's not a matter of whether the relevant conditions are satisfied (as in proposition 4). This yields all the conditions as stated in the result. ■

Lemma B.7. *Assume preferences are moderately divergent:* $\beta \in \left[\frac{1+V_{SQ}-V_\varepsilon(1)}{2}, \frac{4+V_{SQ}-V_\varepsilon(0)}{4}\right]$. *If reversal costs are highly punitive so that the agency will always obfuscate to be upheld when possible,* $\pi > V_\varepsilon(0) - V_{SQ} > V_\varepsilon(1) - V_{SQ}$, *it invests high effort as follows:*

- *When* $V_{SQ} > V_\varepsilon(0) > V_\varepsilon(1)$ *the agency invests high effort if* $V_\varepsilon(0) - V_\varepsilon(1) \geq \kappa$.
- *When* $V_\varepsilon(0) > V_{SQ} > V_\varepsilon(1)$ *the agency invests high effort if* $p_0(V_{SQ} - V_\varepsilon(1) + \pi) + (p_1 + p_2)(V_\varepsilon(0) - V_\varepsilon(1)) \geq \kappa$ *assuming proposition 4 conditions hold so it can obfuscate with* $x_A(1) = 2$ *to be upheld and invests high effort if* $(p_0 + p_1)(V_{SQ} - V_\varepsilon(1) + \pi) + p_2(V_\varepsilon(0) - V_\varepsilon(1)) \geq \kappa$ *when proposition 4 conditions do not hold.*

- *When $V_\varepsilon(0) > V_\varepsilon(1) > V_{SQ}$ the agency invests high effort if $(p_1+p_2)(V_\varepsilon(0)-V_\varepsilon(1)) \geq \kappa$ assuming proposition 4 conditions hold so it can obfuscate with $x_A(1)=2$ to be upheld and invests high effort if $p_2(V_\varepsilon(0)-V_\varepsilon(1)) \geq \kappa$ when proposition 4 conditions do not hold.*

If instead reversal costs are moderately punitive so that the agency obfuscates when possible to avoid reversal only after $e=1$, $V_\varepsilon(0)-V_{SQ} > \pi > V_\varepsilon(1)-V_{SQ}$, it invests high effort as follows:

- *When $V_\varepsilon(0) > V_{SQ} > V_\varepsilon(1)$ the agency invests high effort if $(p_0+p_1)(V_{SQ}-V_\varepsilon(1)+\pi)+p_2(V_\varepsilon(0)-V_\varepsilon(1)) \geq \kappa$.*

- *When $V_\varepsilon(0) > V_\varepsilon(1) > V_{SQ}$ the agency invests high effort if $p_1(V_{SQ}-V_\varepsilon(1)+\pi)+p_2(V_\varepsilon(0)-V_\varepsilon(1)) \geq \kappa$ assuming proposition 4 conditions hold so it can obfuscate with $x_A(1)=2$ to be upheld and invests high effort if $p_2(V_\varepsilon(0)-V_\varepsilon(1)) \geq \kappa$ when proposition 4 conditions do not hold.*

Proof of Lemma B.7. I derive the conditions for each case stated in the result.

First, let $\pi > V_\varepsilon(e)-V_{SQ}$ for all e. When $V_{SQ} > V_\varepsilon(0) > V_\varepsilon(1)$ and the agency invests high effort it sets policy truthfully and is upheld when $\omega=0$ or $\omega=2$ and obfuscates by setting $x_A=0$ or $x_A=2$ (if possible) when $\omega=1$ which is also upheld. If it invests low effort then it sets policy truthfully and is upheld when $\omega \in \{0,2\}$ and again obfuscates when $\omega=1$ and is upheld. Plugging in the relevant payoffs for each state yields the following expected utilities,

$$\begin{aligned} EU_A(e=1) &= -p_0(V_\varepsilon(1)+\kappa)-p_1(1+V_\varepsilon(1)+\kappa)-p_2(V_\varepsilon(1)+\kappa), \\ &= -p_0V_\varepsilon(1)-p_1(1+V_\varepsilon(1))+p_2V_\varepsilon(1)-\kappa, \\ EU_A(e=0) &= -p_0V_\varepsilon(0)-p_1(1+V_\varepsilon(0))-p_2V_\varepsilon(0). \end{aligned}$$

Thus, the agency will invest high effort when,

$$\begin{aligned} -p_0V_\varepsilon(1)-p_1(1+V_\varepsilon(1))+p_2V_\varepsilon(1)-\kappa &\geq -p_0V_\varepsilon(0)-p_1(1+V_\varepsilon(0))-p_2V_\varepsilon(0), \\ V_\varepsilon(0)-V_\varepsilon(1) &\geq \kappa. \end{aligned}$$

When $V_\varepsilon(0) > V_{SQ} > V_\varepsilon(1)$ and the agency invests high effort it can set policy truthfully and be upheld if $\omega \in \{0,2\}$ and will obfuscate to either $x_A=0$ or $x_A=2$ (if possible) when $\omega=1$ and be upheld. If $e=0$ it will be overturned for truthfully setting policy when $\omega=0$, obfuscate to $x_A=2$ when $\omega=1$ if the conditions in proposition 4 are satisfied, truthfully set policy and be reversed if $\omega=1$ and the conditions in proposition 4 are not satisfied, and set policy truthfully and be upheld

when $\omega = 2$. This yields the following expected utility expressions:

$$\begin{aligned} EU_A(e=1) &= -p_0(V_\varepsilon(1)+\kappa) - p_1(1+V_\varepsilon(1)+\kappa) - p_2(V_\varepsilon(1)+\kappa), \\ EU_A(e=0|\text{obfuscate}) &= -p_0(V_{SQ}+\pi) - p_1(1+V_\varepsilon(0)) - p_2 V_\varepsilon(0), \\ EU_A(e=0|\text{not obfuscate}) &= -p_0(V_{SQ}+\pi) - p_1(1+V_{SQ}+\pi) - p_2(V_\varepsilon(0)). \end{aligned}$$

If the agency can obfuscate then its incentive compatibility condition to invest high effort is,

$$\begin{aligned} -p_0(V_\varepsilon(1)+\kappa) - p_1(1+V_\varepsilon(1)+\kappa) - p_2(V_\varepsilon(1)+\kappa) &\geq -p_0(V_{SQ}+\pi) - p_1(1+V_\varepsilon(0)) - p_2 V_\varepsilon(0), \\ p_0(V_{SQ} - V_\varepsilon(1)+\pi) + (p_1+p_2)(V_\varepsilon(0) - V_\varepsilon(1)) &\geq \kappa. \end{aligned}$$

If instead the agency cannot obfuscate then its incentive compatibility condition is,

$$\begin{aligned} -p_0(V_\varepsilon(1)+\kappa) - p_1(1+V_\varepsilon(1)+\kappa) - p_2(V_\varepsilon(1)+\kappa) &\geq -p_0(V_{SQ}+\pi) - p_1(1+V_{SQ}+\pi) - p_2(V_\varepsilon(0)), \\ (p_0+p_1)(V_{SQ} - V_\varepsilon(1)+\pi) + p_2(V_\varepsilon(0) - V_\varepsilon(1)) &\geq \kappa. \end{aligned}$$

When $V_\varepsilon(0) > V_\varepsilon(1) > V_{SQ}$ the agency sets policy truthfully and is overturned when $\omega = 0$ since $V_{SQ} < V_\varepsilon(1)$, obfuscates to $x_A = 2$ when $\omega = 1$ to be upheld when proposition 4 conditions are satisfied and truthfully sets policy and is overturned when they are not, and truthfully sets policy and is upheld when $\omega = 2$ for both effort levels. If it invests low effort then policymaking choices and review choices are the same as when $e = 1$. Together this yields the following expected utility expressions:

$$\begin{aligned} EU_A(e=1|\text{obfuscate}) &= -p_0(V_{SQ}+\pi) - p_1(1+V_\varepsilon(1)) - p_2(V_\varepsilon(1)) - \kappa, \\ EU_A(e=1|\text{not obfuscate}) &= -p_0(V_{SQ}+\pi) - p_1(1+V_{SQ}+\pi) - p_2 V_\varepsilon(1) - \kappa, \\ EU_A(e=0|\text{obfuscate}) &= -p_0(V_{SQ}+\pi) - p_1(1+V_\varepsilon(0)) - p_2 V_\varepsilon(0), \\ EU_A(e=0|\text{not obfuscate}) &= -p_0(V_{SQ}+\pi) - p_1(1+V_{SQ}+\pi) - p_2 V_\varepsilon(0). \end{aligned}$$

Combining and rearranging yields the agency's incentive compatible condition for high effort when it can obfuscate,

$$\begin{aligned} -p_0(V_{SQ}+\pi) - p_1(1+V_\varepsilon(1)) - p_2(V_\varepsilon(1)) - \kappa &\geq -p_0(V_{SQ}+\pi) - p_1(1+V_\varepsilon(0)) - p_2 V_\varepsilon(0), \\ (p_1+p_2)(V_\varepsilon(0) - V_\varepsilon(1)) &\geq \kappa. \end{aligned}$$

When the agency cannot obfuscate its incentive compatible condition for high effort is,

$$\begin{aligned} -p_0(V_{SQ}+\pi) - p_1(1+V_{SQ}+\pi) - p_2 V_\varepsilon(1) - \kappa &\geq -p_0(V_{SQ}+\pi) - p_1(1+V_{SQ}+\pi) - p_2 V_\varepsilon(0), \\ p_2(V_\varepsilon(0) - V_\varepsilon(1)) &\geq \kappa. \end{aligned}$$

This set of conditions constitutes the first part of the result in which $\pi > V_\varepsilon(e) - V_{SQ}$ for all e.

When $V_\varepsilon(0) - V_{SQ} > \pi > V_\varepsilon(1) - V_{SQ}$ the conditions are the same as above when $V_\varepsilon(0) > V_{SQ} > V_\varepsilon(1)$ and the agency cannot obfuscate following $e = 0$. They are also the same when $V_\varepsilon(0) > V_\varepsilon(1) > V_{SQ}$ when the agency cannot obfuscate following $e = 1$, but differ when it can given $e = 1$ because in this case the agency will never obfuscate following $e = 0$ whereas in above it will. The relevant expected utilities in that case are,

$$\begin{aligned} EU_A(e=1) &= -p_0(V_{SQ}+\pi) - p_1(1+V_\varepsilon(1)) - p_2(V_\varepsilon(1)) - \kappa, \\ EU_A(e=0) &= -p_0(V_{SQ}+\pi) - p_1(1+V_{SQ}+\pi) - p_2 V_\varepsilon(0), \end{aligned}$$

which yields the following incentive compatibility condition for high effort,

$$\begin{aligned} -p_0(V_{SQ}+\pi) - p_1(1+V_\varepsilon(1)) - p_2(V_\varepsilon(1)) - \kappa &\geq -p_0(V_{SQ}+\pi) - p_1(1+V_{SQ}+\pi) - p_2 V_\varepsilon(0), \\ p_1(V_{SQ} - V_\varepsilon(1) + \pi) + p_2(V_\varepsilon(0) - V_\varepsilon(1)) &\geq \kappa. \end{aligned}$$

This derives all the results in the lemma. ■

Lemma B.8. *Assume preferences are conditionally extreme:* $\beta \in \left[\frac{4+V_{SQ}-V_\varepsilon(0)}{4}, \frac{4+V_{SQ}-V_\varepsilon(1)}{4}\right]$. *If reversal costs are highly punitive so that the agency will always obfuscate to be upheld when possible,* $\pi > V_\varepsilon(0) - V_{SQ} > V_\varepsilon(1) - V_{SQ}$, *it invests high effort as follows:*

- *When* $V_{SQ} > V_\varepsilon(0) > V_\varepsilon(1)$ *the agency invests high effort if* $4p_2 + V_\varepsilon(0) - V_\varepsilon(1) \geq \kappa$.
- *When* $V_\varepsilon(0) > V_{SQ} > V_\varepsilon(1)$ *the agency invests high effort if* $4p_2 + V_{SQ} - V_\varepsilon(1) + \pi \geq \kappa$.
- *When* $V_\varepsilon(0) > V_\varepsilon(1) > V_{SQ}$ *the agency invests high effort if* $4p_2 + (p_1+p_2)(V_{SQ} - V_\varepsilon(1) + \pi) \geq \kappa$ *assuming proposition 4 conditions hold so it can obfuscate with* $x_A(1) = 2$ *to be upheld and invests high effort if* $p_2(4 + V_{SQ} - V_\varepsilon(1) + \pi) \geq \kappa$ *when proposition 4 conditions do not hold.*

If instead reversal costs are moderately punitive so that the agency obfuscates when possible to avoid reversal only after $e = 1$, $V_\varepsilon(0) - V_{SQ} > \pi > V_\varepsilon(1) - V_{SQ}$, *it invests high effort as follows:*

- *When* $V_\varepsilon(0) > V_{SQ} > V_\varepsilon(1)$ *the agency invests high effort if* $4p_2 + V_{SQ} - V_\varepsilon(1) + \pi \geq \kappa$.
- *When* $V_\varepsilon(0) > V_\varepsilon(1) > V_{SQ}$ *the agency invests high effort if* $4p_2 + (p_1+p_2)(V_{SQ} - V_\varepsilon(1) + \pi) \geq \kappa$ *assuming proposition 4 conditions hold so it can obfuscate with* $x_A(1) = 2$ *to be upheld and invests high effort if* $p_2(4 + V_{SQ} - V_\varepsilon(1) + \pi) \geq \kappa$ *when proposition 4 conditions do not hold.*

Proof of Lemma B.8. First, let $\pi > V_\varepsilon(e) - V_{SQ}$ for all e. When $V_{SQ} > V_\varepsilon(0) > V_\varepsilon(1)$ the agency will match policy to the state and be upheld when $\omega \in \{0, 2\}$ and obfuscate to $x_A = 0$ or $x_A = 2$ (if

possible) and be upheld when $\omega = 1$. After $e = 0$ the agency will pool by setting policy at $x_A = 0$ for all ω and be upheld. This yields the following expected utilities for high and low effort,

$$\begin{aligned} EU_A(e=1) &= -p_0 V_\varepsilon(1) - p_1(1 + V_\varepsilon(1) - p_2 V_\varepsilon(1) - \kappa, \\ EU_A(e=0) &= -p_0 V_\varepsilon(0) - p_1(1 + V_\varepsilon(0)) - p_2(4 + V_\varepsilon(0)), \end{aligned}$$

which further yields the following incentive compatibility condition for high effort,

$$\begin{aligned} -p_0 V_\varepsilon(1) - p_1(1 + V_\varepsilon(1) - p_2 V_\varepsilon(1) - \kappa &\geq -p_0 V_\varepsilon(0) - p_1(1 + V_\varepsilon(0)) - p_2(4 + V_\varepsilon(0)), \\ (p_0 + p_1)(V_\varepsilon(0) - V_\varepsilon(1)) + p_2(4 + V_\varepsilon(0) - V_\varepsilon(1)) &\geq \kappa. \end{aligned}$$

When $V_\varepsilon(0) > V_{SQ} > V_\varepsilon(1)$ the agency will set policy truthfully when $\omega \in \{0, 2\}$ and obfuscate to be upheld when $\omega = 1$ if it invests high effort. If it invests low effort then it cannot obfuscate at all because both $x_A = 0$ and $x_A = 2$ are now being overturned so it sets policy truthfully and accepts reversal. This yields the following expected utilities,

$$\begin{aligned} EU_A(e=1) &= -p_0 V_\varepsilon(1) - p_1(1 + V_\varepsilon(1)) - p_2 V_\varepsilon(1) - \kappa, \\ EU_A(e=0) &= -p_0 V_{SQ} - p_1(1 + V_{SQ}) - p_2(4 + V_{SQ}) - \pi, \end{aligned}$$

which, combining and rearranging, yields the incentive compatibility condition for high effort,

$$\begin{aligned} -p_0 V_\varepsilon(1) - p_1(1 + V_\varepsilon(1)) - p_2 V_\varepsilon(1) - \kappa &\geq -p_0 V_{SQ} - p_1(1 + V_{SQ}) - p_2(4 + V_{SQ}) - \pi, \\ 4p_2 + V_{SQ} - V_\varepsilon(1) + \pi &\geq \kappa. \end{aligned}$$

When $V_\varepsilon(0) > V_\varepsilon(1) > V_{SQ}$ and the agency invests high effort then it sets policy truthfully when $\omega \in \{0, 2\}$ and is overturned when $x_A = 0$ and upheld when $x_A = 2$, and it obfuscates when $\omega = 1$ by setting $x_A = 2$ if it can (proposition 4 conditions hold) and otherwise sets policy truthfully and accepts reversal. If it invests low effort then it can never obfuscate and sets policy truthfully and is reversed. This yields the following expected utilities,

$$\begin{aligned} EU_A(e=1|\text{obfuscate}) &= -p_0(V_{SQ} + \pi) - p_1(1 + V_\varepsilon(1)) - p_2(V_\varepsilon(1)) - \kappa, \\ EU_A(e=1|\text{not obfuscate}) &= -p_0 V_{SQ} - p_1(1 + V_{SQ}) - p_2(4 + V_{SQ}) - \pi, \\ EU_A(e=0|\text{obfuscate}) &= -p_0 V_{SQ} - p_1(1 + V_{SQ}) - p_2(4 + V_{SQ}) - \pi, \\ EU_A(e=0|\text{not obfuscate}) &= -p_0 V_{SQ} - p_1(1 + V_{SQ}) - p_2(4 + V_{SQ}) - \pi. \end{aligned}$$

When the agency can obfuscate if $\omega = 1$ incentive compatibility requires the following holds to

invest high effort,

$$\begin{aligned} -p_0(V_{SQ}+\pi)-p_1(1+V_\varepsilon(1))-p_2(V_\varepsilon(1))-\kappa &\geq -p_0V_{SQ}-p_1(1+V_{SQ})-p_2(4+V_{SQ})-\pi, \\ 4p_2+(p_1+p_2)(V_{SQ}-V_\varepsilon(1)+\pi) &\geq \kappa. \end{aligned}$$

When the agency cannot obfuscate to avoid reversal it never invests high effort as this simply leads to a net loss equal to κ (since the agency is always overturned regardless of e).

Now let $V_\varepsilon(0)-V_{SQ}>\pi>V_\varepsilon(1)-V_{SQ}$. When $V_\varepsilon(0)>V_{SQ}>V_\varepsilon(1)$ the incentive compatibility conditions are the same as above since policymaking and review behavior is equivalent in both settings. When $V_\varepsilon(0)>V_\varepsilon(1)>V_{SQ}$ the incentive compatibility conditions in this case are again equivalent to the conditions when $\pi>V_\varepsilon(e)-V_{SQ}$ for all e since policymaking and review behavior is the same. ■

Lemma B.9. *Assume preferences are extreme: $\beta>\frac{4+V_{SQ}-V_\varepsilon(1)}{4}$. If reversal costs are highly punitive so that the agency will always obfuscate to be upheld when possible, $\pi>V_\varepsilon(0)-V_{SQ}>V_\varepsilon(1)-V_{SQ}$, it invests high effort as follows:*

- *When $V_{SQ}>V_\varepsilon(0)>V_\varepsilon(1)$ the agency invests high effort if $V_\varepsilon(0)-V_\varepsilon(1)\geq\kappa$.*
- *When $V_\varepsilon(0)>V_{SQ}>V_\varepsilon(1)$ the agency invests high effort if $V_{SQ}-V_\varepsilon(1)+\pi\geq\kappa$.*
- *When $V_\varepsilon(0)>V_\varepsilon(1)>V_{SQ}$ the agency never invests high effort because there are no deviations that avoid reversal.*

If instead reversal costs are moderately punitive so that the agency obfuscates when possible to avoid reversal only after $e=1$, $V_\varepsilon(0)-V_{SQ}>\pi>V_\varepsilon(1)-V_{SQ}$, it invests high effort as follows:

- *When $V_\varepsilon(0)>V_{SQ}>V_\varepsilon(1)$ the agency invests high effort if $V_{SQ}-V_\varepsilon(1)+\pi\geq\kappa$.*
- *When $V_\varepsilon(0)>V_\varepsilon(1)>V_{SQ}$ the agency never invests high effort because there are no deviations that avoid reversal.*

Proof of Lemma B.9. Let $\pi>V_\varepsilon(e)-V_{SQ}$ for all e so that the agency always obfuscates when possible to avoid reversal. When $V_{SQ}>V_\varepsilon(0)>V_\varepsilon(1)$ the agency pools on $x_A=0$ for all ω to avoid reversal when $\omega\in\{1,2\}$ for both effort levels (i.e., it obfuscates by setting $x_A=0$ when $\omega\in\{1,2\}$ and is subsequently upheld since $V_{SQ}>V_\varepsilon(e)$ for all e). This yields the following expected utilities:

$$\begin{aligned} EU_A(e=1) &= -p_0V_\varepsilon(1)-p_1(1+V_\varepsilon(1))-p_2(4+V_\varepsilon(1))-\kappa, \\ EU_A(e=0) &= -p_0V_\varepsilon(0)-p_1(1+V_\varepsilon(0))-p_2(4+V_\varepsilon(0)), \end{aligned}$$

which yields the following incentive compatibility condition to invest high effort,

$$\begin{aligned} -p_0V_\varepsilon(1)-p_1(1+V_\varepsilon(1))-p_2(4+V_\varepsilon(1))-\kappa &\geq -p_0V_\varepsilon(0)-p_1(1+V_\varepsilon(0))-p_2(4+V_\varepsilon(0)), \\ V_\varepsilon(0)-V_\varepsilon(1) &\geq \kappa. \end{aligned}$$

When $V_\varepsilon(0) > V_{SQ} > V_\varepsilon(1)$ and it invests high effort then it again pools on $x_A = 0$ for all ω to avoid reversal when $\omega \in \{1,2\}$, but if it invests low effort $x_A = 0$ is reversed so it is always overturned given its policy choices. This yields the following expected utilities,

$$\begin{aligned} EU_A(e=1) &= -p_0V_\varepsilon(1)-p_1(1+V_\varepsilon(1))-p_2(4+V_\varepsilon(1))-\kappa, \\ EU_A(e=0) &= -p_0(V_{SQ}+\pi)-p_1(1+V_{SQ}+\pi)-p_2(4+V_{SQ}+\pi). \end{aligned}$$

Combining and rearranging yields the incentive compatibility condition for high effort,

$$\begin{aligned} -p_0V_\varepsilon(1)-p_1(1+V_\varepsilon(1))-p_2(4+V_\varepsilon(1))-\kappa &\geq -p_0(V_{SQ})-p_1(1+V_{SQ})-p_2(4+V_{SQ})-\pi, \\ V_{SQ}-V_\varepsilon(1)+\pi &\geq \kappa. \end{aligned}$$

When $V_\varepsilon(0) > V_\varepsilon(1) > V_{SQ}$ the agency is never upheld following $x_A = 0$ and is also never upheld for $x_A \in \{1,2\}$ since preferences are extreme. Thus, outcomes do not vary according to effort choice, implying that the agency never invests in high effort because doing so would produce a net utility loss equal to κ.

Now let $V_\varepsilon(0) - V_{SQ} > \pi > V_\varepsilon(1) - V_{SQ}$ so the agency would only obfuscate (when possible) if $e = 1$. In this case the incentive compatibility conditions are exactly the same as above for both variance orderings. In both cases there is no obfuscation following $e = 0$ and the same type of obfuscation following $e = 1$ when $V_\varepsilon(0) > V_{SQ} > V_\varepsilon(1)$, which implies that effort investment behavior (and the conditions to support it) are the same as when $\pi > V_\varepsilon(e) - V_{SQ}$ for all e. This completes the derivations underpinning the conditions in the result. ■

B.2.3 In-text examples

Example 1. *(High effort to tell the truth)* Let $\beta \in \left[\frac{1+V_{SQ}-V_\varepsilon(0)}{2}, \frac{1+V_{SQ}-V_\varepsilon(1)}{2}\right)$ so that preferences are *conditionally aligned*. Further, fix the following parameter values: $\beta = 7/16$, $p_0 = 1/4$, $p_1 = 1/4$, $p_2 = 1/2$, $V_{SQ} = 1/4$, $V_\varepsilon(0) = 1/2$, $V_\varepsilon(1) = 1/8$, $\pi = 1/2$. These parameter values further imply that $\frac{p_1}{p_1+p_2} \leq \frac{1}{4}(4-4\beta+V_{SQ}-V_\varepsilon(0))$ holds so that the conditions for semi-pooling characterized in Proposition 4 are satisfied, $V_\varepsilon(0) > V_{SQ} > V_\varepsilon(1)$ so that the policy environment is *moderately volatile*, and that $\pi > V_\varepsilon(0) - V_{SQ} > V_\varepsilon(1) - V_{SQ}$ so that reversal costs are *highly punitive*.

These environmental features imply that when the agency invests high effort it will be upheld for truthfully matching policy to the state for all ω. If instead the agency invests low effort it will be

overturned for truthfully setting $x_A = 0$ when $\omega = 0$ but will nonetheless do so,[2] it will obfuscate by setting $x_A = 2$ when $\omega = 1$ to avoid reversal since $\pi > V_\varepsilon(0) - V_{SQ}$ (the semi-pooling equilibrium in Proposition 4), and can set policy truthfully when $\omega = 2$ and be upheld since preferences are conditionally aligned. Thus, whether the agency invests high effort in equilibrium depends on whether the net benefits of being able to always set policy truthfully with high effort implementation are large enough to offset the costs to obtain those benefits.

Given the equilibrium dynamics described above the agency's expected utility expressions for high and low effort investment are given by,

$$\begin{aligned} EU_A(e=1) &= -\underbrace{p_0 V_\varepsilon(1)}_{\substack{\text{payoff if } \omega=0,\, e=1 \\ \text{since } x_A^{\text{truth}}(0) \text{ upheld}}} - \underbrace{p_1 V_\varepsilon(1)}_{\substack{\text{payoff if } \omega=1,\, e=1 \\ \text{since } x_A^{\text{truth}}(1) \text{ upheld}}} - \underbrace{p_2 V_\varepsilon(1)}_{\substack{\text{payoff if } \omega=2,\, e=1 \\ \text{since } x_A^{\text{truth}}(2) \text{ upheld}}} - \underbrace{\kappa,}_{\text{effort cost}} \\ EU_A(e=0) &= -\underbrace{p_0(V_{SQ}+\pi)}_{\substack{\text{payoff if } \omega=0,\, e=0 \\ \text{since } x_A^{\text{truth}}(0) \text{ reversed}}} - \underbrace{p_1(1+V_\varepsilon(0))}_{\substack{\text{payoff if } \omega=1,\, e=0 \\ \text{and obfuscate to be upheld}}} - \underbrace{p_2 V_\varepsilon(0),}_{\substack{\text{payoff if } \omega=2,\, e=1 \\ \text{since } x_A^{\text{truth}}(2) \text{ upheld}}} \end{aligned}$$

Combining and rearranging these expressions yields the condition that must be satisfied in this environment for the agency to invest high effort,

$$\underbrace{p_0(V_{SQ} - V_\varepsilon(1) + \pi)}_{\substack{\text{net benefit from high effort} \\ \text{to avoid reversal when } \omega=0}} + \underbrace{p_1(1 + V_\varepsilon(0) - V_\varepsilon(1))}_{\substack{\text{net benefit from high effort} \\ \text{and not obfuscating when } \omega=1}} + \underbrace{p_2(V_\varepsilon(0) - V_\varepsilon(1))}_{\substack{\text{net benefit from high effort} \\ \text{and always upheld when } \omega=2}} \geq \underbrace{\kappa.}_{\text{effort cost}}$$

Plugging in the parameter values specified at the beginning of the example reduces this condition to,

$$\begin{aligned} \left(\frac{1}{4}\right)\left(\frac{1}{4} - \frac{1}{8} + \frac{1}{2}\right) + \left(\frac{1}{4}\right)\left(1 + \frac{1}{2} - \frac{1}{8}\right) + \left(\frac{1}{2}\right)\left(\frac{1}{2} - \frac{1}{8}\right) &\geq \kappa, \\ \frac{11}{16} \approx 0.69 &\geq \kappa. \end{aligned}$$

If $\kappa < 11/16$ then the agency will invest high effort, which will allow it to subsequently set policy truthfully and be upheld. If instead $\kappa > 11/16$ then when $\omega = 1$ the agency will obfuscate by setting $x_A = 2$ to avoid reversal. ■

Example 2 (*High effort to obfuscate*)**.** Let $\beta \in \left(\frac{1+V_{SQ}-V_\varepsilon(1)}{2}, \frac{4+V_{SQ}-V_\varepsilon(0)}{4}\right]$ so that preferences are *moderate*. Further, fix the following parameter values: $p_0 = 1/4$, $p_1 = 1/4$, $p_2 = 1/2$, $V_{SQ} = 1/2$, $V_\varepsilon(0) = 3/4$, $V_\varepsilon(1) = 1/4$, and $\pi = 1/8$. These parameter values further imply that $V_\varepsilon(0) > V_{SQ} > V_\varepsilon(1)$ so that the policy environment is *moderately volatile*, and that $V_\varepsilon(0) - V_{SQ} > \pi > V_\varepsilon(1) - V_{SQ}$ so that reversal costs are *moderately punitive*.

[2] This follows from Lemma B.2 in the online appendix.

This environment is one in which when the agency invests high effort it is upheld for truthfully setting policy when either $\omega = 0$ or $\omega = 2$, and it obfuscates by setting either $x_A = 0$ or $x_A = 2$ (when the conditions for Proposition 4 are satisfied) when $\omega = 1$ to avoid reversal. When the agency invests low effort it accepts being overturned for setting policy truthfully when $\omega \in \{0,1\}$ since $\pi < V_\varepsilon(0) - V_{SQ}$ implies that it is never incentive compatible to deviate from truthful policymaking (from Proposition 2) and is upheld for truthfully setting policy when $x_A = 2$. Thus, whether the agency invests high effort involves whether the net benefits from doing so – being upheld when $\omega = 0$, obfuscating to be upheld when $\omega = 1$, and being upheld when $\omega = 2$ with lower implementation variance – outweigh the costs of those benefits (κ).

Given the equilibrium dynamics in this environment the agency's expected utility expressions for high and low effort investment are given by,

$$\begin{aligned} EU_A(e=1) &= -\underbrace{p_0 V_\varepsilon(1)}_{\substack{\text{payoff if } \omega=0,\, e=1 \\ \text{since } x_A^{\text{truth}}(0) \text{ upheld}}} - \underbrace{p_1(1+V_\varepsilon(1))}_{\substack{\text{payoff if } \omega=1,\, e=1 \\ \text{and obfuscate to be upheld}}} - \underbrace{p_2 V_\varepsilon(1)}_{\substack{\text{payoff if } \omega=2,\, e=1 \\ \text{since } x_A^{\text{truth}}(2) \text{ upheld}}} - \underbrace{\kappa,}_{\text{effort cost}} \\ EU_A(e=0) &= -\underbrace{p_0(V_{SQ}+\pi)}_{\substack{\text{payoff if } \omega=0,\, e=0 \\ \text{since } x_A^{\text{truth}}(0) \text{ reversed}}} - \underbrace{p_1(1+V_{SQ}+\pi)}_{\substack{\text{payoff if } \omega=1,\, e=0 \\ \text{since obfuscation not IC}}} - \underbrace{p_2 V_\varepsilon(0),}_{\substack{\text{payoff if } \omega=2,\, e=1 \\ \text{since } x_A^{\text{truth}}(2) \text{ upheld}}} \end{aligned}$$

Combining and rearranging these expressions yields the agency's incentive compatibility condition to invest high effort in this setting,

$$\underbrace{p_0(V_{SQ} - V_\varepsilon(1) + \pi)}_{\substack{\text{net benefit from high effort} \\ \text{to avoid reversal when } \omega=0}} + \underbrace{p_1(V_{SQ} - V_\varepsilon(1) + \pi)}_{\substack{\text{net benefit from high effort} \\ \text{and obfuscation when } \omega=1}} + \underbrace{p_2(V_\varepsilon(0) - V_\varepsilon(1))}_{\substack{\text{net benefit from high effort} \\ \text{and always upheld when } \omega=2}} \geq \underbrace{\kappa.}_{\text{effort cost}}$$

Plugging in the parameter values specified at the beginning of the example reduces this condition to,

$$\begin{aligned} \left(\frac{1}{4}\right)\left(\frac{1}{2} - \frac{1}{4} + \frac{1}{8}\right) + \left(\frac{1}{4}\right)\left(\frac{1}{2} - \frac{1}{4} + \frac{1}{8}\right) + \left(\frac{1}{2}\right)\left(\frac{3}{4} - \frac{1}{4}\right) &\geq \kappa, \\ \frac{7}{16} &\geq \kappa, \end{aligned}$$

If $\kappa < 7/16$ then the agency will invest high effort, which will allow it to subsequently set policy truthfully and be upheld when $\omega = 0$, obfuscate to be upheld when $\omega = 1$, and improve implementation precision when $\omega = 2$. If instead $\kappa > 7/16$ then the agency accepts being overturned whenever $\omega = 0$ and $\omega = 1$, following truthful policymaking, and is upheld with lower implementation precision when $\omega = 2$. ■

C Comparing Review Institutions

C.1 Aligned and extreme preferences

Proposition C.1. *When preferences are sufficiently aligned so that the agency is always upheld following procedural review and under substantive review truthful policymaking leads the overseer to uphold $x_A = 0$ only if $V_{SQ} \geq V_\varepsilon(e)$ and uphold $x_A \in \{1,2\}$ for any e the overseer weakly prefers substantive review from an ex ante perspective.*

Proof of Proposition C.1. Consider the overseer's ex ante utility under procedural review when the agency is always upheld (for any e),

$$\begin{aligned} EU_R^P(r(e)=0) &= -p_0((0-\beta-(1)(0+\varepsilon))^2) - p_1((1-\beta-(1)(1+\varepsilon))^2) - p_2((2-\beta-(1)(2+\varepsilon))^2), \\ &= -p_0\beta^2 - p_1\beta^2 - p_2\beta^2 - V_\varepsilon(e), \\ &= -\beta^2 - V_\varepsilon(e). \end{aligned}$$

Now, supposing that $r(0,e) = 0$ for a given e following substantive review the overseer's ex ante utility is given by,

$$\begin{aligned} EU_R^S(r(0,e)=0) &= -p_0\beta^2 - p_1\beta^2 - p_2\beta^2 - V_\varepsilon(e), \\ &= -\beta^2 - V_\varepsilon(e). \end{aligned}$$

In this case $EU_R^S(r(0,e)=0)$ and $EU_R^P(r(e)=0)$ are equivalent so the type of review is inconsequential to the overseer from an ex ante welfare perspective. Suppose instead that $r(0,e) = 1$ for a given e under substantive review. Then the overseer's ex ante utility is given by,

$$\begin{aligned} EU_R^S(r(0,e)=1) &= -p_0((0-\beta-(0)x)^2 + V_{SQ}) - p_1((1-\beta-(1)(1+\varepsilon))^2) - p_2((2-\beta-(1)(2+\varepsilon))^2), \\ &= -p_0(\beta^2 + V_{SQ}) - p_1(\beta^2 + V_\varepsilon(e)) - p_2(\beta^2 + V_\varepsilon(e)), \\ &= -\beta^2 - p_0 V_{SQ} - (p_1+p_2)V_\varepsilon(e). \end{aligned}$$

For procedural review to be preferred in this case it must be that,

$$\begin{aligned} EU_R^P(r(e)=0) &> EU_R^S(r(0,e)=1), \\ -\beta^2 - V_\varepsilon(e) &\geq -\beta^2 - p_0 V_{SQ} - (p_1+p_2)V_\varepsilon(e), \\ p_0(V_{SQ} - V_\varepsilon(e)) &\geq 0, \end{aligned}$$

which can never be satisfied since $V_{SQ} < V_\varepsilon(e)$ is required to ensure $r(0,e) = 1$. Thus, in this case the overseer benefits from substantive review due to increased control over the agency when $\omega = 0$.

Overall, then, the overseer is either indifferent between procedural review and substantive review or strictly benefits from substantive review when preferences are sufficiently aligned. ■

Proposition C.2. *When preferences are so extreme that the agency is always overturned following procedural review and $x_A = 0$ is upheld only if $V_{SQ} \geq V_\varepsilon(e)$ and $x_A \in \{1,2\}$ are both overturned for all e following substantive review procedural review is weakly preferred by the overseer in terms of ex ante utility.*

Proof of Proposition C.2. Consider first the overseer's ex ante utility for procedural review when the agency is never upheld regardless of e:

$$EU_R^P(r(e)=1) \quad = \quad -p_0\beta^2 - p_1(1-\beta)^2 - p_2(2-\beta)^2 - V_{SQ}.$$

Now suppose that, given e, the agency is upheld when $x_A = 0$ and it pays the pooling strategy where $x_A(\omega) = 0, \forall\omega$. The overseer's payoff in that case is given by,

$$EU_R^S(r(0,e)=0) \quad = \quad -p_0\beta^2 - p_1(1-\beta)^2 - p_2(2-\beta)^2 - V_\varepsilon(e).$$

Finally, if the overseer overturns $x_A = 0$ given e then her ex ante utility is equivalent to the procedural review case since regardless of what the agency chooses it is overturned:

$$EU_R^S(r(0,e)=1) \quad = \quad -p_0\beta^2 - p_1(1-\beta)^2 - p_2(2-\beta)^2 - V_{SQ}.$$

Obviously when $r(0,e) = 1$ under substantive review and the agency is always overturned the overseer is ex ante indifferent between procedural and substantive oversight – both yield the same payoff in expectation. However, when $r(0,e) = 0$ the overseer (weakly) benefits from substantive review, relative to procedural review, since,

$$\begin{aligned} EU_R^S(r(0,e)=0) &\geq EU_R^P(r(e)=1), \\ -p_0\beta^2 - p_1(1-\beta)^2 - p_2(2-\beta)^2 - V_\varepsilon(e) &\geq -p_0\beta^2 - p_1(1-\beta)^2 - p_2(2-\beta)^2 - V_{SQ}, \\ V_{SQ} &\geq V_\varepsilon(e), \end{aligned}$$

which is exactly the condition that must hold in order for the agency to be upheld following pooling on $x_A = 0$: $r(0,e) = 0$. Thus, when the preference environment is such that the agency will never be upheld when the overseer reviews procedure and will never be upheld for changing policy under substantive review (i.e., $x_A \in \{1,2\} \Rightarrow r(x_A,e) = 1, \forall e$) the overseer weakly benefits from substantive oversight in expectation. ■

C.2 Moderate preference examples

The following result is useful for defining the potential effort incentives in the environment analyzed in this section.

Lemma C.1. *When the overseer is conditionally-deferential under procedural review and the environment is as defined in example 2 under substantive review the threshold on κ to support high effort investment is higher under procedural review:*

$$p_1 + 4p_2 + V_{SQ} - V_\varepsilon(1) + \pi > (p_0 + p_1)(V_{SQ} - V_\varepsilon(1) + \pi) + p_2(V_\varepsilon(0) - V_\varepsilon(1)).$$

Proof of Lemma C.1. The relevant inequality is derived from the incentive compatibility constraints for each relevant environment (lemma A.3 for procedural review and lemma B.7 for substantive review). The fact that the inequality in the result is always satisfied is straightforward given the assumptions of the model:

$$\begin{aligned} p_1 + 4p_2 + V_{SQ} - V_\varepsilon(1) + \pi &> (p_0 + p_1)(V_{SQ} - V_\varepsilon(1) + \pi) + p_2(V_\varepsilon(0) - V_\varepsilon(1)), \\ p_2(4 + V_{SQ} - V_\varepsilon(0) + \pi) + p_1 &> 0. \end{aligned}$$

This is always trivially satisfied so long as either $p_1 > 0$ or $p_2 > 0$ since $V_{SQ}, V_\varepsilon(0), \pi \in (0,1)$. ■

Thus, the only possibilities in the setting analyzed below are that the agency invests high effort under both review systems (high κ), the agency invests high effort under procedural review and low effort under substantive review (intermediate κ), and invests low effort under both types of review (low κ). I now consider each possibility in turn.

Proposition C.3. *Assume under procedural review the overseer is conditionally deferential so she upholds if and only if the agency invests high effort, and under substantive review the environment is as in example 2 so that when the agency invests high effort it obfuscates by setting $x_A(1) = 2$ and the overseer upholds that choice. Further, assume that $p_1 + 4p_2 + V_{SQ} - V_\varepsilon(1) + \pi > (p_0 + p_1)(V_{SQ} - V_\varepsilon(1) + \pi) + p_2(V_\varepsilon(0) - V_\varepsilon(1)) \geq \kappa$ so that the agency invests high effort in equilibrium under both procedural and substantive review. Then procedural review is always preferred to substantive review.*

Proof of Proposition C.3. Under procedural review the agency invests high effort and is upheld. The agency matches policy to the state and implementation uncertainty is given by $V_\varepsilon(1)$ for all ω. Under substantive review the agency truthfully sets $x_A(1) = 1$ and is upheld, obfuscates when $\omega = 1$ and sets $x_A(1) = 2$ and is upheld, and truthfully sets $x_A(2) = 2$ and is upheld. In all cases, implementation uncertainty is given by $V_\varepsilon(1)$. This yields the following ex ante expected utility expressions for the

overseer conditional on scope of review:

$$\begin{aligned} EU_R^P &= -p_0(\beta^2+V_\varepsilon(1))-p_1(\beta^2+V_\varepsilon(1))-p_2(\beta^2+V_\varepsilon(1)), \\ EU_R^S &= -p_0(\beta^2+V_\varepsilon(1))-p_1((\beta+1)^2+V_\varepsilon(1))-p_2(\beta^2+V_\varepsilon(1)). \end{aligned}$$

Now, for procedural review to be beneficial relative to substantive review the following inequality must be satisfied:

$$\begin{aligned} EU_R^P &> EU_R^S, \\ -p_0(\beta^2)-p_1(\beta^2)-p_2(\beta^2)-V_\varepsilon(1) &> -p_0(\beta^2)-p_1((\beta+1)^2)-p_2(\beta^2)-V_\varepsilon(1), \\ 2\beta+1 &> 0, \end{aligned}$$

which is always satisfied since β is non-negative. This implies that procedural review is always preferred to substantive review in this environment, as stated in the result. ■

One thing worth noting about proposition C.3 is that in the environment specified the agency could also obfuscate by appeasing the overseer ($x_A(1)=0$) rather than obfuscating through exaggeration. In that case, the overseer strictly prefers substantive review because it induces the agency to set $x_A=0$ when $\omega=1$, which benefits the overseer. To see this, note that the payoffs when $\omega=0$ and $\omega=2$ are exactly the same as in proposition C.3, but the payoff for $\omega=1$ under substantive review is now $-(1-\beta)^2-V_\varepsilon(1)$ instead of $-(\beta+1)^2-V_\varepsilon(1)$. The relevant comparison then becomes $-p_1\beta^2-V_\varepsilon(1)$ (under procedural review) and $-p_1(1-\beta)^2-V_\varepsilon(1)$ (under substantive review). Thus, if $-p_1(1-\beta)^2>-p_1\beta^2$ then substantive review is strictly preferred to procedural review, which requires that $\beta>\frac{1}{2}$. This inequality is always satisfied in this environment since $\beta>\frac{1+V_{SQ}-V_\varepsilon(1)}{2}$, the RHS of which is greater than one-half since $V_{SQ}>V_\varepsilon(1)$ (and $V_{SQ}>0$). Thus, if the agency were to obfuscate by appeasement rather than obfuscate by exaggeration then substantive review would be preferred to procedural review. Nonetheless, obfuscating through exaggeration exists in equilibrium in this environment, providing proof that procedural review can be preferred to substantive review in terms of ex ante overseer expected utility.

Proposition C.4. *Assume under procedural review the overseer is conditionally-deferential so she upholds if and only if the agency invests high effort, and under substantive review the environment is defined as in example 2 so that when the agency invests low effort obfuscation is never incentive compatible. Further, assume that* $p_1+4p_2+V_{SQ}-V_\varepsilon(1)+\pi\geq\kappa>(p_0+p_1)(V_{SQ}-V_\varepsilon(1)+\pi)+p_2(V_\varepsilon(0)-V_\varepsilon(1))$ *so that the agency invests high effort when facing procedural review but low effort under substantive review. Then procedural review is preferred to substantive review when,*

$$p_0(V_{SQ}-V_\varepsilon(1))+p_1(1-2\beta+V_{SQ}-V_\varepsilon(1))+p_2(V_\varepsilon(0)-V_\varepsilon(1))>0. \tag{2}$$

Otherwise, substantive review is preferred to procedural review. Furthermore, equation 2 is more likely to be satisfied as p_1 and β decrease.

Proof of Proposition C.4. Under procedural review the agency matches policy to the state, invests high effort, and the overseer upholds, which yields $-\beta^2 - V_\varepsilon(1)$ as the expected payoff for each potential state. Under substantive review the agency truthfully sets policy when $\omega \in \{0,1\}$ since obfuscation is not incentive compatible, and the overseer overturns those choices. The agency also sets policy truthfully when $\omega = 2$, but the overseer upholds in this case. This yields the following ex ante expected utilities for the overseer conditional on scope of review:

$$\begin{aligned} EU_R^P &= -p_0(\beta^2 + V_\varepsilon(1)) - p_1(\beta^2 + V_\varepsilon(1)) - p_2(\beta^2 + V_\varepsilon(1)), \\ EU_R^S &= -p_0(\beta^2 + V_{SQ}) - p_1((1-\beta)^2 + V_{SQ}) - p_2(\beta^2 + V_\varepsilon(0)). \end{aligned}$$

Now, for procedural review to be beneficial relative to substantive review the following inequality must be satisfied:

$$\begin{aligned} EU_R^P &> EU_R^S, \\ -p_0(\beta^2) - p_1(\beta^2) - p_2(\beta^2) - V_\varepsilon(1) &> -p_0(\beta^2 + V_{SQ}) - p_1((1-\beta)^2 + V_{SQ}) - p_2(\beta^2 + V_\varepsilon(0)), \\ p_0(V_{SQ} - V_\varepsilon(1)) + p_1(1 - 2\beta + V_{SQ} - V_\varepsilon(1)) + p_2(V_\varepsilon(0) - V_\varepsilon(1)) &> 0, \end{aligned}$$

which is sometimes satisfied and sometimes fails to be satisfied. Notice that $p_0(V_{SQ} - V_\varepsilon(1))$ and $p_2(V_\varepsilon(0) - V_\varepsilon(1))$ are both always positive since $V_\varepsilon(0) > V_{SQ} > V_\varepsilon(1)$ by assumption (and p_0 and p_2 are non-negative). In contrast, $p_1(1 - 2\beta + V_{SQ} - V_\varepsilon(1)) < 0$, since $\beta > \frac{1 + V_{SQ} - V_\varepsilon(1)}{2}$.

The direction of the inequality thus depends on whether $|p_1(1 - 2\beta + V_{SQ} - V_\varepsilon(1))| > p_0(V_{SQ} - V_\varepsilon(1)) + p_2(V_\varepsilon(0) - V_\varepsilon(1))$. If that holds then $EU_R^P < EU_R^S$ and substantive review is preferred to procedural review. If instead $|p_1(1 - 2\beta + V_{SQ} - V_\varepsilon(1))| < p_0(V_{SQ} - V_\varepsilon(1)) + p_2(V_\varepsilon(0) - V_\varepsilon(1))$ then $EU_R^P > EU_R^S$ and procedural review is preferred to substantive review.

Thus, the likelihood that procedural review is preferred to substantive review in this case is decreasing in the probability that $\omega = 1$ (p_1) – which also implies increases in p_0, p_2, or both – and overseer bias β. Conversely, as $\beta \to \frac{4 + V_{SQ} - V_\varepsilon(1)}{4}$ and as p_1 increases it is more likely that substantive review is preferred to procedural review. ■

Proposition C.5. *Assume under procedural review the overseer is conditionally deferential so she upholds if and only if the agency invests high effort, and under substantive review the environment is as in example 2 so that when the agency invests high effort it obfuscates by setting $x_A(1) = 2$ and the overseer upholds that choice. Further, assume that $\kappa > p_1 + 4p_2 + V_{SQ} - V_\varepsilon(1) + \pi > (p_0 + p_1)(V_{SQ} - V_\varepsilon(1) + \pi) + p_2(V_\varepsilon(0) - V_\varepsilon(1))$ so that the agency invests low effort in equilibrium under both procedural and substantive review. Then substantive review is always preferred to procedural review.*

Proof of Proposition C.5. Under procedural review the agency matches policy to the state, invests low effort, and the overseer overturns. Under substantive review the agency truthfully sets policy when $\omega \in \{0,1\}$ since obfuscation is not incentive compatible, and the overseer overturns those choices. The agency also sets policy truthfully when $\omega = 2$, but the overseer upholds in this case. This yields the following ex ante expected utilities for the overseer conditional on scope of review:

$$\begin{aligned} EU_R^P &= -p_0(\beta^2 + V_{SQ}) - p_1((1-\beta)^2 + V_{SQ}) - p_2((2-\beta)^2 + V_{SQ}), \\ EU_R^S &= -p_0(\beta^2 + V_{SQ}) - p_1((1-\beta)^2 + V_{SQ}) - p_2(\beta^2 + V_\varepsilon(0)). \end{aligned}$$

Now, for procedural review to be beneficial relative to substantive review the following inequality must be satisfied:

$$\begin{aligned} EU_R^P &> EU_R^S, \\ -p_0(\beta^2 + V_{SQ}) - p_1((1-\beta)^2 + V_{SQ}) - p_2((2-\beta)^2 + V_{SQ}) &> -p_0(\beta^2 + V_{SQ}) - p_1((1-\beta)^2 + V_{SQ}) - p_2(\beta^2 + V_\varepsilon(0)), \\ p_2(V_\varepsilon(0) - V_{SQ}) &> p_2(2-\beta)^2 - \beta^2, \end{aligned}$$

which is never satisfied because $\beta < \frac{4 + V_{SQ} - V_\varepsilon(0)}{4}$. Thus, in this environment the overseer always benefits from substantive review, as stated in the result. ■