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

Ian R. Turner*

Abstract

How does the scope of review affect bureaucratic policymaking incentives? To explore this question, I consider a simple policymaking environment in which an expert agency develops policy that is upheld or overturned by an overseer who may have different policy goals. The agency can affect the quality of implementation through effort investments in addition to choosing the substantive content of policy. Under procedural review the overseer only reviews the agencys effort, which allows the agency to fully utilize its expertise, but may harm effort incentives. Substantive review also tasks the overseer with judging agencies substantive policy choices, which introduces a fundamental trade-off between agency utilization of expertise and effort investment due to pathological policy choices made by the agency. The theory characterizes when less transparent oversight, procedural review, is optimal relative to more transparent, substantive review. The results speak to when agencies should be insulated from substantive review.

Keywords: Bureaucracy; Oversight; Policymaking; Formal Theory
Word count: 7853 (in-text); 3220 (supplemental appendix); 11073 (total=in-text+appendix)

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

Delegation of policymaking authority to bureaucratic agencies is often predicated on the fact that agencies possess superior policy-relevant expertise. Yet delegation raises an enduring normative concern in politics.[1] On one hand, citizens can benefit from superior bureaucratic expertise as it informs governmental policy. On the other hand, delegation also raises the specter that these 'agents' may exploit their expertise or informational advantages to pursue policies that run counter to the wishes of some political principal, be it the general public, the president, or Congress.[2] This 'political agency problem,' as its commonly referred to, is present any time the agent's preferences diverge from those of a political principal (Bendor and Meirowitz 2004; Bendor, Glazer and Hammond 2001; Gailmard and Patty 2013*a*,*b*; Miller 2005).

One of the most ubiquitous political-institutional solutions for these agency problems is subjection of the agency's actions to ex post review. That is, the agency's decisions are subject to review, and possible invalidation, from another political actor such as a court or other oversight institution (e.g., the Office of Information and Regulatory Affairs). It is thought that review institutions of this sort will at least deter the agency from making policy choices that run directly counter to one's own policy preferences. The problem is that in environments in which delegation to the agency is desirable due to the agency's superior expertise, oversight cannot overcome the potential for agency subversion unless the agency itself chooses to reveal its information to the overseer and reduce its own relative expertise advantage.

Moreover, bureaucratic agencies do more than simply develop the substance of policy, they also develop programmatic capacity – through procedural development – that helps guide the agency's on-the-ground workforce to implement policy effectively.[3] This introduces another wrinkle to ex post oversight: the overseer must not only worry about divergent substantive policy choices predicated on the agency's ability to exploit its informational advantage, she must also consider providing proper incentives for the agency to invest in high quality implementation of policy, whatever the sub-

[1] See (Gailmard and Patty 2013*b*) for a recent discussion of this dilemma.

[2] See (Gailmard 2002) for a comprehensive treatment of bureaucratic subversion in a principal-agent framework reminiscent of the models developed in this paper.

[3] This point of view is reminiscent of 'street-level bureaucracy' (Lipsky 1980). More generally, Carpenter (2001) distinguishes an agency's analytic capacity, which allows it to adequately craft the substance of policy, and an agency's programmatic capacity, which allows the agency to apply or enforce policy effectively.

stantive content (Turner 2017*b*).

The scope of review. While ex post oversight is carried out within all three branches of the United States federal government, nearly all bureaucratic actions are subject to judicial review in various forms.[4] Most, if not all, pieces of authorizing legislation contain judicial review provisions that specify who can challenge agency actions (or not), what actions are subject to review (or not), as well as the scope of judicial review.[5] These oversight provisions are also the focus, and product, of political processes in Congress while the legislation is being drafted (Shipan 1997). One major component of the role of judicial oversight is the *scope of review*.[6] The scope of review dictates what actions, and which *type* of review overseers are directed to engage in. Two major types of oversight are *procedural review* and *substantive review*. This raises the main question in this paper: how does the type of ex post review shape the incentives for both effort investments that improve the quality of policy outcomes and the willingness of the agency to utilize its superior policy-relevant information?

Procedural review entails an overseer examining whether an agency has followed all relevant guidelines, invested in the capacity to administer policy effectively, and the like without direct regard for the content of the policy itself. This could represent an agency's investment in research that allows it to better understand the contingencies of the policy environment in terms of the translation of policy choices into on-the-ground outcomes, developing procedures to ensure that policies are applied equitably across constituent populations, or, more generally, investment in the capacity to enforce its policy choices without making costly errors. Previous research suggests that courts have recently moved more toward procedural review of administrative actions (Kagan 2001; Stephenson

[4] Additionally, the executive branch reviews agency policy proposals through the OIRA, an oversight agency within the Office of Management and Budget, and, *INS v. Chadha* notwithstanding, Congress carries out ex post review through oversight hearings, annual appropriations, and invoking the Congressional Review Act of 1996, which until the present Congressional term was only exercised to invalidate ergonomics standards during the Clinton Administration.

[5] See (McCann, Shipan and Wang 2016) for a comprehensive empirical description of a significant subset of judicial review provisions in authorizing legislation.

[6] Congress utilizes these provisions to also specify rules governing citizens' abilities to challenge agency actions in court (Smith 2005, 2006), as well as which courts have jurisdictional authority over which agency actions (Chutkow 2008). For several case studies, across policy areas, suggesting that Congress anticipates the role of judicial review in the policymaking process see (Cass 1989; Light 1991; Melnick 1983, 1994; Rose-Ackerman 1995; Shipan 2000).

2006).

Substantive review entails an overseer judging the actual content of policy choices made by agencies. This, generally, relates to the idea that overseers such as courts can help to enforce bureaucratic policy choices that do not run counter to the wishes of the overseer herself or those of some political principal (Epstein and O'Halloran 1999). This dimension of review is directly connected to the agency problem highlighted above: if the overseer sits at an informational disadvantage relative to the agency, then judging the substance of agency choices is difficult unless the agency itself chooses to reveal some, or all, of its private information.

Whatever the type of review, part of the role of oversight institutions is to enforce accountability. Whether this is understood as incentivizing the agency to invest effort toward high quality policy implementation or to set policy more closely in accordance with the goals of the overseer or some other principal, oversight is thought to be effective in disciplining bureaucratic behavior by forcing agencies to operate in the shadow of review.[7] In this paper, I develop an argument that ex post review institutions, such as judicial or executive review, can harm accountability in differential ways conditional on the type of review utilized.[8] Through the analysis of two variants of a formal model of policymaking between an agency and overseer I characterize the different ways that procedural and substantive review can enhance accountability, or harm it, on both effort and substantive dimensions. In the first variant, the *procedural review model*, the overseer only observes an ex ante effort investment made by the agency that improves the implementation precision of policy outcomes. In the second variant, the *substantive review model*, the overseer observes both the agency's effort investment *and* the substantive policy choice made by the agency, potentially learning about the policy environment through the agency's policy choice.

Procedural review allows the agency to fully utilize its policy-relevant information and develop policy in line with the realities of the policy environment because the agency does not have to worry about the overseer judging the substance of its choices. The cost of this, from the overseer's

[7]Previous literature has suggested that ex post veto institutions are superior to other accountability instititions such as gatekeeping (Crombez, Groseclose and Krehbiel 2006).

[8]For related, but distinct, arguments about potential weaknesses of judicial review see Melnick (1983), Shapiro and Levy (1995), and Wagner (2012). Sunstein (1989) provides a defense of the institution in administrative law.

perspective, is not learning anything about the agency's private information, which can be undesirable as the overseer's preferences diverge from those of the agency. Additionally, procedural review can provide positive incentives that leads agencies to invest higher effort toward implementation than it would have absent review. However, it can also harm these incentives and induce the agency to invest lower effort toward implementation than it would have were it not subject to review.

Substantive review, on the other hand, allows the overseer to at times perfectly learn the agency's private information and therefore provide strong 'ideological oversight.' However, this learning is based on the agency's own substantive policy choices. The agency only chooses to reveal its private information when reversal is not too punitive from the agency's perspective. Otherwise, the agency will obfuscate with some of its substantive policy choices so as to only partially reveal its private information in order to avoid reversal. To do so, the agency foregoes following its own superior information and exaggerates the extremity of policy change that is called for given the 'facts on the ground.' This result potentially subverts the very rationale supporting delegation to expert agencies in the first place. Moreover, when the overseer judges the substance of policy there is a fundamental trade-off between the agency investing high effort and fully utilizing its technical expertise. If the agency invests high effort toward quality policy implementation then the agency is also more likely to obfuscate to avoid reversal. High effort investments make the agency more protective of its policies and more likely to avoid reversal through obfuscation because it is relatively less costly to do so, from a policy perspective, when outcomes will be implemented more precisely.

Accountability and oversight. Ex post oversight of political activity comes in many forms. In terms of enforcing accountability prevalent review mechanisms include elections,[9] presidential vetoes,[10], stakeholder 'fire alarms' or Congressional oversight,[11] and judicial review.[12] Much of the previous research demonstrates how oversight can lead to the provision of perverse incentives that induce policymaking pathologies like pandering, posturing, and persisting when policymakers have

[9]For example, Ashworth (2012), Barro (1973), Fearon (1999), Ferejohn (1986).

[10]For example, Cameron (2000), Groseclose and McCarty (2001).

[11]For example, Epstein and O'Halloran (1995), Gailmard (2009), McCubbins and Schwartz (1984).

[12]For example, Beim, Hirsch and Kastellec (2014), Bueno de Mesquita and Stephenson (2007), Clark (2016), Fox and Stephenson (2015), Fox and Vanberg (2014), Patty and Turner (2016), Turner (2017*a*,*b*), Vanberg (2001).

career or reputational concerns.[13] Pandering involves choosing policies known to be favored by the public regardless of the politician's private information. Posturing occurs when politicians pursue bold policies that are ill-advised based on their private information.[14] Persisting deals with situations in which politicians' private information suggests abandoning a policy that does not appear to be working as intended but they nonetheless continue on so as to avoid looking incompetent (Majumdar and Mukand 2004). In all of these cases the desire by politicians to remain in office, avoid being fired or demoted, or avoid having their policies vetoed leads them to disregard their superior private information due to these reputational considerations.

In line with these studies, scholars have also studied how institutions promoting transparency affect accountability. Many of these studies have highlighted how increasing the transparency of policymaking may harm accountability.[15] I extend this line of inquiry by exploring how increasing the transparency of agency actions in the review process can impact accountability negatively through a novel channel: *policy exaggeration*. In particular, I build on two closely related studies.

Turner (2017*b*) shows that procedural oversight can both strengthen and weaken agency effort incentives even when there is no preference disagreement between the reviewer and agency.[16] In contrast, I characterize how procedural review impacts agency effort incentives in the presence of preference disagreement and illustrate how both effort incentives and incentives to follow policy-relevant information are impacted when the information available to the overseer during review varies. Thus, in this paper the overseer has the opportunity, if engaged in substantive review, to potentially block policies with which she ideologically disagrees, but as I will show, this is less likely when the agency has invested high effort. When the agency has invested high effort it will often choose instead to obfuscate with its substantive policy choices by exaggerating how much policy change is called for given the underlying policy environment, thereby ignoring (and obscuring from the overseer) its private information to avoid policy reversal.

[13] For a comprehensive overview of these pathologies see Gersen and Stephenson (2014).

[14] Both pandering and posturing are throughly explored in Canes-Wrone, Herron and Shotts (2001).

[15] For example, Fox (2007), Fox and Stephenson (2011), Fox and Van Weelden (2012, 2015), Patty and Turner (2016), Prat (2005).

[16] See also Bueno de Mesquita and Stephenson (2007) for related results.

This latter result is similar to the second paper upon which this paper builds, Patty and Turner (2016). In that paper, the authors characterize when an agent will disregard policy-relevant information and "cry wolf," or propose policy changes that are more extreme than is called for by the policy environment. The authors primary focus is when the overseer would prefer to have her review powers set aside, thereby allowing the agency to enact policy unencumbered by review, to avoid the introduction of this perverse incentive. In this paper I introduce an effort dimension that improves the quality of realized policy outcomes and compare the different pathologies that arise across review institutions. In a sense, I bridge the gap across these two existing studies by looking at both effort and informational dynamics in the face of two different types of ex post oversight. Thus, while the agency will also "cry wolf," or exaggerate the level of policy change called for, I show that the perverse incentives to do so are exacerbated by high effort investments to improve implementation. This provides insights into the trade-offs between effort and expertise as well as between the two different styles of ex post oversight, which are not explored in the aforementioned studies. These trade-offs provide implications for how oversight may, or may not, provide for bureaucratic accountability.

1 The model

I analyze a two-player, non-cooperative game between a bureaucratic agency, A, that makes policy and an overseer or reviewer, R, that has the power to review and invalidate agency policy actions. The agency is an expert in the sense that it learns private policy-relevant information, and is directed by statute to make policy. The overseer is empowered to review and overturn (or, veto) agency-made policy and return policy to an exogenous status quo.

Prior to learning about the policy environment the agency chooses to invest high effort or low effort toward the quality of policy implementation. This choice is denoted by $e \in \{0,1\}$ where $e = 0$ is low effort and $e = 1$ is high effort. Investing high effort leads to a net effort cost, $\kappa > 0$. This choice can be thought of as how hard the agency works to follow procedures in place to improve policy and acquire relevant programmatic capacity to implement policy precisely on the ground.

Formally, this effort investment directly affects an implementation shock, denoted by $\varepsilon \in \mathbb{R}$. This shock is conditioned by the agency's effort choice and is distributed according to a cumulative distribution function $F_\varepsilon(e)$ with mean zero and strictly positive variance, $V_\varepsilon(e) \in (0,1)$. Having mean zero implies that the shock is centered on the agency's substantive policy choice, described below. The variance of ε when the agency invests high effort is strictly less than when low effort is invested: $0 < V_\varepsilon(1) < V_\varepsilon(0) < 1$.[17] This implies that high effort investments produce strictly more precise policy outcomes than low effort investments.

Following the agency's effort investment it learns about the policy environment by observing a true *state of the world*, denoted by $\omega \in \Omega = \{0,1,2\}$. The ex ante probability that the true state is ω is p_ω, implying a probability distribution over states given by $p = \{p_0, p_1, p_2\}$. The three different states represent whether the relevant policy environment calls for very little policy change ($\omega = 0$), moderate policy change ($\omega = 1$), or extreme policy change ($\omega = 2$). The value of ω represents the agency's sincere (expert) opinion about how much policy ought to be adjusted to match the facts on the ground.

Upon observing ω the agency sets a substantive 'policy target,' denoted by $x_A \in X = \{0,1,2\}$. This substantive policy choice can be thought of as a target because realized, agency-made, policy outcomes are conditional on realization of the implementation shock ε, which is further conditional on the agency's effort choice as described above. Following the agency's choices the overseer reviews the agency and chooses to either uphold or overturn the agency's policy. If the overseer upholds the agency then final policy is given by $x = x_A + \varepsilon$ and if the overseer overturns then final policy is $x = 0$.

I analyze two variants of the model that differ only in the information available to the overseer at the time of review. In the *procedural review model* the overseer only observes the agency's effort investment decision before making her review decision. This choice is represented by $r(e) \in \{0,1\}$ where $r(e) = 0$ implies upholding and $r(e) = 1$ overturning. In the *substantive review model* the overseer observes both the agency's effort investment and substantive policy choice. This choice is

[17] Bounding the variances above by one is inconsequential for the results. It simply streamlines the analysis by restricting implementation errors from shifting outcomes all the way to another substantive policy choice.

then represented by $r(x_A, e) \in \{0, 1\}$, where zero and one are understood in the same way. In the former case the overseer is only asked to ensure that the agency has followed all relevant procedural requirements and developed sufficient capacity for quality implementation. In the latter case the overseer not only takes the agency's investments toward implementation into account, but is also directed to judge the substance of the agency's policy.

The agency is motivated by wanting to match policy to the state and have high quality implementation, conditional on the costs of high effort, and have his policy upheld by the overseer. If the agency is overturned then it internalizes a reversal cost, denoted by $\pi \in (0, 1)$. This can represent a reputational cost, opportunity costs of time wasted on policy that will never be realized, or a direct cost such as a fine or demotions. If one understands π as a reputational cost then it can also represent a measure of agency independence insofar as this cost is negatively correlated with independence. Agencies with low independence will have higher reputational costs and highly independent or insulated agencies may worry less about reputation and therefore have lower reversal costs. Overall, the agency can be though of as "faithful" in the sense that there are no distortions in preferences associated with ideology or the like. The agency ultimately wants policy to be decided according to the state of the policy environment ω. Substantively this represents, as an example, a 'public spirited' bureaucracy that is motivated purely by the policy area rather than ideology or bias. The overseer, however, may differ in her ideal policy relative to the agency. This could be due to an ideological or political agenda, or simply an ex ante 'bias' regarding what policy choice is optimal given the state of the environment (ω). This bias is represented by $\beta \in (0, 1)$. Overseer and agency interests are captured by the following payoff functions:

$$\begin{aligned} u_R(e, x, r) &= -(\omega - \beta - (1 - r)x)^2, \\ u_A(e, x, r) &= -(\omega - (1 - r)x)^2 - \kappa e - \pi r, \end{aligned}$$

where the parameters of the problem — β, κ, π — are exogenous and common knowledge. Notice that the overseer's payoff function implies that her bias, β, induces her to prefer policy that is less

ambitious, or closer to the status quo, than the agency. Ultimately, the overseer wants policy to be as close as possible to her ideal point ($x = \omega - \beta$) and the agency wants policy to match the state ($x = \omega$) and for its policy to be upheld to avoid paying the reversal cost π. Further, both players value high effort implementation to reduce the potential impact of the implementation shock, but there is conflict between the players on this dimension since the agency is the only player that internalizes the cost of doing so (κ).

I will denote the agency's effort and substantive policy strategies as s_A^e and $x_A(\omega)$, respectively. The overseer's review strategy, $s_R(\cdot)$, varies based on the information available to her. So, $s_R(e)$ denotes the overseer's review strategy in the procedural review model where she only observes e and $s_R(x_A, e)$ denotes the analogous strategy for the substantive review model. Finally, the overseer's beliefs are denoted by $b_R(x_A)$.[18] I utilize perfect Bayesian equilibrium in weakly undominated pure strategies as my solution concept. This requires that the overseer hold correct beliefs, updated via Bayes' rule, about the state of the world and that both players make choices to maximize their subjective expected payoffs.

2 Reviewing procedure

In the procedural review model the overseer only observes the agency's effort investment, e. This implies that in equilibrium the agency always matches substantive policy to the state: $x_A^{\text{proc}}(\omega) = \omega$.[19] Put simply, since the overseer cannot condition its review decision on x_A and the substantive policy and effort are separable in the agency's payoff function, the agency is always better off setting substantive policy to match the state to minimize spatial policy losses due to its choice of x_A.[20]

The overseer then chooses between upholding and overturning the agency based on its observation of e and correct beliefs regarding the agency's substantive policy strategy $x_A^{\text{proc}}(\omega)$. If the overseer chooses to overturn the agency then final policy is set at $x = 0$. Thus, the overseer's

[18] These beliefs will only end up being applicable in the substantive review model since the overseer never has an opportunity to update her beliefs regarding ω in the procedural review model.

[19] While technically this is working out of order in terms of backward induction, acknowledging this characteristic of the equilibrium up front simplifies analysis of the overseer's optimal review strategy.

[20] This is formally shown in Lemma 1 in the supplemental appendix.

subjective expected payoff for overturning the agency is given by,[21]

$$EU_R(r=1|x_A^{\text{proc}}(\omega),e,p) \quad = \quad -p_0(\beta^2)-p_1((1-\beta)^2)-p_2((2-\beta)^2).$$

Since $x=0$ in this case, the overseer knows that it will lose $(\omega-\beta)^2$ for each ω, which is weighted by the probability that a given ω is realized.

Alternatively, the overseer could uphold the agency. In this case her subjective expected payoff is given by,

$$EU_R(r=0|x_A^{\text{proc}}(\omega),e,p) \quad = \quad -\beta^2-V_\varepsilon(e).$$

In this case the overseer knows that the agency, given $x_A^{\text{proc}}(\omega)$, will match substantive policy to the state. That means in terms of spatial losses associated with substantive policy choices the overseer only loses utility based on her bias β since she would have preferred policy be closer to the status quo than the agency. The overseer also loses expected utility based on the implementation imprecision associated with agency-made policy, $V_\varepsilon(e)$. She loses less utility if $e=1$ relative to $e=0$. Combining and rearranging these subjective expected payoffs yields the incentive compatibility constraint that must be satisfied in order for the overseer to uphold the agency, implying the following optimal review strategy:

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

The overseer requires the agency to invest sufficient effort to limit the volatility of agency-made policy in order to uphold. The potential errors in implementation, captured by $V_\varepsilon(e)$, must be low enough relative to the overseer's net substantive policy losses, given her bias, if she upholds relative to when she overturns.

[21] All derivations for the overseer's optimal review strategy in the procedural review model can be found in the proof of Lemma 2 in the supplemental appendix.

The condition to uphold the agency is more likely to be satisfied when the agency has invested high effort than low effort. Thus, there are two thresholds for upholding the agency based on the overseer's bias.[22] Specifically, rearranging the condition for the overseer to uphold shows that the overseer's bias cannot be too large in order for the agency to receive deference: $\beta \in \left[0, \frac{p_1+4p_2-V_\varepsilon(e)}{2p_1+4p_2}\right)$. Plugging in $e=1$ and $e=0$ yields these two thresholds of β. Let $\beta_1 \equiv \frac{p_1+4p_2-V_\varepsilon(1)}{2p_1+4p_2}$ and $\beta_0 \equiv \frac{p_1+4p_2-V_\varepsilon(0)}{2p_1+4p_2}$. Since $V_\varepsilon(1) < V_\varepsilon(0)$, the threshold on overseer bias is higher when the agency has invested high effort: $\beta_1 > \beta_0$. The agency can be upheld for higher levels of preference disagreement when it has invested high effort.

Where the overseer's bias lies relative to these two thresholds dictates the *review regime* the agency faces. If $\beta < \beta_0 < \beta_1$ then the overseer will always uphold the agency regardless of effort investment. This review regime is *perfectly deferential*. If $\beta > \beta_1 > \beta_0$ then the overseer will always overturn the agency regardless of effort. In this case the review regime is *perfectly skeptical*. Finally, if $\beta_0 < \beta < \beta_1$ then the overseer upholds the agency if and only if the agency invests high effort and I refer to this review regime as *conditionally-deferential*. Agency effort decisions depend crucially on which regime it is operating within.

If the agency is facing a perfectly deferential overseer then it knows that its policy choices will always be upheld. Thus, the only consideration is how much high effort improves the precision of policy outcomes relative to low effort, and the costs that must be internalized for that improvement. Consider the agency's subjective expected payoffs for high and low effort, respectively:

$$\begin{aligned} EU_A(e=1|s_R(e), x_A^{\text{proc}}(\omega)) &= -V_\varepsilon(1) - \kappa, \\ EU_A(e=0|s_R(e), x_A^{\text{proc}}(\omega)) &= -V_\varepsilon(0). \end{aligned}$$

Since the agency always matches policy to the state and will always be upheld when it invests high effort it expects to lose utility based on the variability of realized outcomes and the costs of high effort it must pay. In contrast, if the agency invests low effort then it can avoid paying the

[22] Details can be found in the appendix in the proof of Lemma 2.

effort costs, but increases the variability of policy outcomes, thereby increasing the likelihood of costly distortions in implementation. For the agency to invest high effort in this case incentive compatibility requires that $EU(e=1|s_R(e),x_A^{\text{proc}}(\omega)) \geq EU_A(e=0|s_R(e),x_A^{\text{proc}}(\omega))$, which implies that,

$$V_\varepsilon(0) - V_\varepsilon(1) \quad \geq \quad \kappa,$$

must be satisfied. The precision improvement associated with investing high effort must outweigh the costs of doing so. In other words, the more that investing high effort improves the precision of implemented policy outcomes the more likely it is that the agency will find it profitable to pay the costs of that investment.

If instead the agency is facing a perfectly skeptical overseer, it is never incentive compatible to invest high effort. When the overseer is so biased that the agency cannot 'work hard enough' to appease her high effort investment generates a net loss proportional to the costs of that effort. Since the agency is overturned with certainty whether or not it invests high effort toward implementation, it is better off avoiding paying the effort costs and investing low effort instead.

Finally, the most interesting case is when the overseer is conditionally-deferential. In this case the agency decides between investing low effort and being overturned by the overseer, which leads to internalizing the reversal cost π but avoidance of effort costs, and investing high effort and being upheld, which allows the agency to avoid the reversal cost but comes at the expense of the costs of high effort. Recall that the agency does not yet know ω when it chooses to invest high or low effort. Thus, in this case the agency must also take into account the probability distribution p over potential states of the world. Specifically, when the agency invests low effort it knows it will be overturned, but since it does not yet know the state it does not know exactly how costly doing so will be from a substantive policy perspective. The agency's subjective expected payoff for investing

low effort, given p, is given by,

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

If the agency invests low effort then, in expectation, it loses utility based on the probability of each state and the losses associated with having $x=0$ for each as well as having to pay the reversal cost π.

If instead the agency invests high effort it will be upheld and therefore be able to match policy to the state and avoid paying the reversal cost, but it will have to bear the costs of the expected imprecision of realized outcomes $V_\varepsilon(1)$ and pay the cost of effort κ:

$$EU_A(e=1|s_R(e),x_A^{\text{procedure}}(\omega)) = -V_\varepsilon(1)-\kappa.$$

Combining and rearranging these two subjective expected payoffs yields the condition that must be met in order for the agency to optimally invest high effort when facing conditional deference:

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

The left-hand side of this inequality captures the net benefits of investing high effort and being upheld while the right-hand side captures the cost, κ, of doing so. The more punitive the reversal costs the more likely it is that this expression will be satisfied. Similarly, the more precise high effort policy is (i.e., the lower is $V_\varepsilon(1)$) the more likely it is the agency will find it beneficial to invest high effort. Taken together, the preceding analysis characterizes the equilibrium to the procedural review model, stated in the following result.

Proposition 1. *The equilibrium to the procedural review model is characterized by the following collection of strategies:*

- *The overseer makes review decisions based on $s_R(e)$ (i.e., equation 1).*
- *The agency always sets substantive policy to match the state: $x_A^{proc}(\omega) = \omega$.*
- *When facing a perfectly deferential overseer (i.e., $\beta_0 < \beta_1 < \beta$) the agency invests high effort when $V_\varepsilon(0) - V_\varepsilon(1) \geq \kappa$.*
- *When facing a perfectly skeptical overseer (i.e., $\beta_0 < \beta_1 < \beta$) the agency never invests high effort.*
- *When facing a conditional-deference overseer (i.e., $\beta_0 < \beta < \beta_1$) the agency invests high effort if $p_1 + 4p_2 + \pi - V_\varepsilon(1) \geq \kappa$.*

Proof. Proofs for all in-text results can be found in the supplemental appendix. ■

Focusing on the case in which the overseer upholds the agency if and only if the agency invests high effort also illustrates a key trade-off present in the procedural review model. The presence of procedural review itself can be either positive or negative with respect to agency effort incentives. Specifically, as compared to a world in which there is no ex post oversight of agency policymaking, if $p_1 + 4p_2 + \pi - V_\varepsilon(e) > \kappa > V_\varepsilon(0) - V_\varepsilon(1)$ then the presence of procedural review is beneficial in that it induces the agency to invest high effort when it would not have done so in the absence of such oversight. This is a positive incentive effect from a policymaking perspective. In contrast, if $V_\varepsilon(0) - V_\varepsilon(1) > \kappa > p_1 + 4p_2 + \pi - V_\varepsilon(e)$ then procedural review induces the agency to invest low effort when it would have invested high effort if it were not operating in the shadow of oversight. That is, the overseer provides a form of policy insurance that deters the agency from finding investment in high quality implementation palatable.[23] Thus, while procedural review allows the agency to utilize its technical expertise, which may be normatively desirable given the oft-cited rationale for delegation to expert agencies, it may come at the cost of both substantive disagreement (due to overseer biases) and the proper provision of effort incentives.

[23] This result is similar to the results in (Bueno de Mesquita and Stephenson 2007) and (Turner 2017*b*) that show that judicial review, or ex post oversight more generally, can dissuade an agency from regulating at all or weaken effort incentives, respectively. It is also qualitatively similar to the "bail out effect" played by judicial review in (Fox and Stephenson 2011).

3 Judging substance

In the substantive review model the overseer observes both the agency's effort investment e and substantive policy choice x_A. The agency's choice of x_A potentially reveals information about ω to the overseer. Since the agency wishes to avoid having his policy choice reversed this introduces the possibility of obfuscation. The first question I address is whether and when the agency will set substantive policy 'truthfully.' Formally, a truthful policymaking strategy for the agency corresponds to behavior in a separating equilibrium and is denoted by,

$$x_A^{\text{truth}}(\omega) = \omega.$$

If the agency is truthful then the overseer learns ω perfectly. This can be thought of as a normative benchmark in the sense that if the agency was authorized to make policy due to its information or expertise, this is a case in which the agency fully utilizes those advantages. Given $x_A^{\text{truth}}(\omega)$ the overseer's review strategy is illustrated in Table 1.[24]

ω	**Aligned Preferences:**	**Conditionally Aligned Preferences:**	**Moderate Preferences:**	**Conditionally Extreme Preferences:**	**Extreme Preferences:**
0	$r(e) = 1, \forall e$	$r(e) = 1, \forall e$	$r(e) = 1, \forall e$	$r(e) = 1, \forall e$	$r(e) = 1, \forall e$
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 decisions given truthful policymaking ($x_A = \omega$) conditional on state ω, effort e, and bias β.

Table 1 illustrates that the overseer will never uphold the agency when she learns that $\omega = 0$ and that as overseer-agency preference divergence grows (i.e., as β increases) it is more difficult

[24] Proposition 2 states the key result for truthful policymaking. Overseer-agency preference alignments are characterized according to the following: preferences are *aligned* when $\beta \in \left[0, \frac{1-V_\varepsilon(0)}{2}\right)$; preferences are *conditionally aligned* when $\beta \in \left[\frac{1-V_\varepsilon(0)}{2}, \frac{1-V_\varepsilon(1)}{2}\right)$; preference divergence is *moderate* when $\beta \in \left[\frac{1-V_\varepsilon(1)}{2}, \frac{4-V_\varepsilon(0)}{4}\right]$; preferences are *conditionally extreme* when $\beta \in \left(\frac{4-V_\varepsilon(0)}{4}, \frac{4-V_\varepsilon(1)}{4}\right]$; and preference divergence is *extreme* when $\beta > \frac{4-V_\varepsilon(1)}{4}$. In the supplemental appendix, Lemma 4 derives the full review strategy for the overseer when the agency is truthful for each value of e in each state ω (along with an analogous table with the technical conditions on overseer bias) and Lemma 5 characterizes the agency's effort choices for each possibility.

for the agency to be upheld following truthful policymaking. The overseer never upholds a truthful agency following $x_A = 0$ because if she were to uphold then she would internalize a net loss proportional to the implementation imprecision of agency-made policy:

$$\Delta EU_R(r = 0 | x_A^{\text{truth}}(\omega) = 0) = -V_\varepsilon(e).$$

Substantively, this implies that when the overseer learns that the state of the world calls for maintenance of the status quo the overseer would prefer to just 'shut the agency down' and stop it from taking any new policy actions. Even if the agency invests high effort there is still the chance for errors in implementation and therefore the overseer is better off simply precluding the agency from meddling.[25] Further, it is more difficult for the agency to receive deference when the overseer's bias increases because the overseer prefers policies closer to the status quo, given ω, as β increases.

Notice that, as compared to the procedural review model, there is no case in which the overseer is *perfectly deferential* as before. However, if the overseer is extremely biased she does become *perfectly skeptical*. Since the thresholds on β listed in the table are endogenous to the agency's effort decisions there are ranges of biases in which the overseer is *conditionally-deferential* and upholds the agency if and only if $e = 1$ following particular choices of x_A – specifically, conditionally-aligned preferences and conditionally-extreme preferences. Thus, one immediate difference across the two types of review is that the overseer can partially overcome her commitment problem inherent in the procedural review model: there are no cases in which the agency, due to its superior expertise, will be upheld with certainty. This follows from the fact that substantive review leads to information being revealed to the overseer (full information in the truthful case), thereby reducing the agency's relative expertise or informational advantage.

The agency, in response, will not always set policy truthfully since the agency is not only

[25] This is stark oversight behavior driven by the fact that there is no uncertainty associated with outcomes when the overseer overturns. It is straightforward to introduce uncertainty, say $V_{\text{reverse}} > 0$, that is associated with overturning. This would capture the idea that disallowing the agency to engage in maintenance of the status quo could still lead to distortions in outcomes that arise out of, for example, the private interactions of firms and individuals without any further agency intervention. If this were the case then there would be a positive probability, or environments in which, the overseer would sometimes uphold the agency following $x_A^{\text{truth}} = 0$. While this may be useful in supporting a wider range of particular types of equilibria its preclusion does not alter the qualitative nature of the results.

driven by matching policy to the state, but also by avoiding reversal, and internalizing the reversal cost π. Accordingly, the agency's substantive policymaking strategy is contingent on the relationship between π and the costs associated with mismatching policy and the state. The agency will only truthfully set substantive policy, given the overseer's best responses in Table 1, if the reversal cost is not too punitive, which is captured formally in the following result.

Proposition 2. *There is a truthful separating equilibrium in which the agency always matches policy to the state* $\left(x_A^{truth}(\omega)\right)$ *if and only if reversal costs are not too punitive* $(V_\varepsilon(e) > \pi)$. *Furthermore,* $p_2(V_\varepsilon(0) - V_\varepsilon(1)) \geq \kappa$ *is sufficient to ensure the agency invests high effort for all ranges of overseer bias except when the agency will always be overturned* $\left(\beta > \frac{4-V_\varepsilon(e)}{4}\right)$, *in which case the agency will never invest high effort.*

The agency would rather set policy truthfully and be overturned (paying π) than obfuscate with its choice and be upheld (avoid paying π) only if the potential implementation errors lead to worse outcomes than the cost of being reversed. That is, when the reversal cost π is not very punitive — is lower than the cost of the errors possible from agency-made policy — the agency cares more about policy than it does reputation (i.e., being upheld) and therefore would prefer being overturned when it knows its capacity will lead to relatively poor implementation.

The (sufficient) condition to ensure that the agency invests high effort in all of the cases in which it will be upheld after truthful policymaking follows from the fact that the most stringent test of that effort decision is when the agency is upheld if and only if $\omega = 2$. Given that the agency will be upheld following truthful revelation of $\omega = 2$ it follows that the agency will invest high effort if the precision improvement of doing so outweighs the costs of doing so. This is then weighted by the probability that $\omega = 2$ since the agency makes its effort decision prior to learning the state. In cases in which the agency would also be upheld following $x_A = 1$ the constraint for high effort is more lenient since high effort can be supported for higher levels of effort costs. Finally, if the agency will always be overturned due to an extremely biased overseer the agency never invests high effort since that would simply lead to a net loss proportional to the cost of that effort.[26]

[26]Full details can be found in the proofs of Proposition 2 and Lemma 5 in the supplemental appendix.

The requirement to support truthfulness — that reversal costs not be too punitive — also implies a fundamental trade-off between high effort and truthful policymaking.

Corollary 1. *The incentive for the agency to obfuscate with its substantive policy choice* $\left(x_A \neq x_A^{truth}(\omega)\right)$ *is stronger when the agency invests high effort.*

The requirement for the agency to always follow the policymaking strategy $x_A^{\text{truth}}(\omega)$ is that $\pi < V_\varepsilon(e)$, the stringency of which varies with the agency's effort choice. Since $V_\varepsilon(1) < V_\varepsilon(0)$ the condition is more difficult to satisfy when the agency invests high effort in the sense that π must be less punitive than when the agency invests low effort. This also implies that there is a wider range of $\pi \in (0,1)$ in which the agency would prefer to deviate from truthful policymaking *if it has already invested high effort*. If the agency has invested high effort then it has stronger incentives to avoid being reversed because it has already paid the effort costs. Put simply, when the agency has invested in improving the quality of policy outcomes it has stronger incentives to take actions that will lead to those outcomes being realized even when that means sacrificing matching policy to the state. Thus, while the overseer benefits from the agency investing high effort this also increases the possibility that the agency deviates from truthful policymaking, which can prove costly to the overseer.

There are two environments of interest for analyzing situations in which the agency would prefer to obfuscate through policy exaggeration to induce being upheld relative to setting policy truthfully: (1) highly punitive reversal costs ($\pi > V_\varepsilon(0) > V_\varepsilon(1)$), and (2) moderately punitive reversal costs ($V_\varepsilon(0) > \pi > V_\varepsilon(1)$). Before analyzing each case specifically, the following result characterizes obfuscation equilibria for both of these environments generally.

Proposition 3. *Suppose* $\pi > V_\varepsilon(e)$. *If the overseer is moderately biased* $\left(\beta \in \left(\frac{1-V_\varepsilon(e)}{2}, \frac{4-V_\varepsilon(e)}{4}\right)\right)$ *and the need for extreme policy change is sufficiently likely relative to moderate policy change:*

$$\frac{p_1}{p_1+p_2} \leq \frac{1}{4}\left(4-4\beta-V_\varepsilon(e)\right) \tag{2}$$

then there is a pure strategy semi-pooling obfuscation equilibrium in which the agency sets substan-

tive policy at $x_A = 0$ *when* $\omega = 0$ *and* $x_A = 2$ *for both* $\omega \in \{1,2\}$*:*

$$x_A^{\textit{semi-pool}}(\omega) = \begin{cases} 0 & \textit{if } \omega = 0, \\ 2 & \textit{if } \omega \in \{1,2\}, \end{cases}$$

and the overseer upholds $x_A = 2$ *and overturns* $x_A \in \{0,1\}$*:*

$$s_R(x_A, e) = \begin{cases} \textit{Uphold: } r = 0 & \textit{if } x_A = 2, \\ \textit{Overturn: } r = 1 & \textit{if } x_A \in \{0,1\}. \end{cases}$$

In this environment obfuscation occurs when the overseer is moderately biased. This is because the agency will never deviate from truthful policymaking when $\omega = 0$ since the agency's payoff loss for that deviation (assuming it leads to being upheld) is the substantive cost of the deviation and the implementation imprecision associated with being upheld ($1 + V_\varepsilon(e)$). Since $\pi \in (0,1)$ by assumption the condition for the agency to remain truthful always holds. This is not true, however, when $\omega = 1$. When the overseer's preferences do not diverge much from the agency the overseer will already uphold $x_A^{\text{truth}}(1) = 1$ so the agency has no reason to obfuscate. However, when the overseer has moderately divergent preferences the agency may choose to deviate. In that case a deviation from $x_A^{\text{truth}}(1) = 1$ to $x_A^{\text{semi-pool}}(1) = 2$ leads to a substantive policy loss of 1, which is the same substantive loss the agency incurs if it is truthful and gets overturned (returning policy to $x = 0$, implying a policy loss of 1 for the agency). Thus, the net losses associated with deviating in this case are simply the potential for implementation errors $V_\varepsilon(e)$. So long as the costs of reversal are greater than that potential policy loss the agency would prefer to obfuscate and induce deference from the overseer, as noted above.

Of course, for this to be an equilibrium it must also be true that the overseer will uphold following observation of $x_A = 2$. Since ω is no longer being revealed perfectly the overseer updates her beliefs about ω following $x_A = 2$.[27] The left-hand side of equation 2 denotes the overseer's

[27] When the overseer observes $x_A = 0$ she knows with certainty that $\omega = 0$ since $x_A^{\text{semi-pool}}(0) = 0$ always. This is in

(posterior) belief that $\omega = 1$ given $x_A^{\text{semi-pool}}(\omega)$. The condition requires that the probability that the true state is one is low enough relative to the probability that the true state is two for the overseer to uphold. This follows from the fact that in this preference environment the overseer wants to overturn when $\omega = 1$ and uphold when $\omega = 2$. The right-hand side of the equation captures the net policy benefits associated with upholding in this case given that the state could be either $\omega = 1$ or $\omega = 2$. So long as inequality 2 holds then the overseer optimally overturns the agency following policy choices of $x_A = 0$ and $x_A = 1$ and upholds the agency any time she observes a policy choice of $x_A = 2$.

Again, recall that all of this is predicated on the idea that the agency *wants* to obfuscate with its policy choices when $\omega = 1$, which requires that $V_\varepsilon(e) > \pi$ which further depends on how punitive reversal costs are relative to agency effort. I now turn to analyzing these environments while assuming that the condition for the overseer to uphold (equation 2) is satisfied in both cases.

Highly punitive reversal. In this case being overturned is highly costly for the agency. Substantively, this could represent the policymaking environment for agencies with low levels of political independence (i.e., agencies with high reputational concerns). Formally, this is defined as an environment in which $\pi > V_\varepsilon(0) > V_\varepsilon(1)$. In this environment the agency always wants to obfuscate with its policy choice by setting $x_A^{\text{semi-pool}}(\omega) = 2$ for $\omega \in \{1,2\}$ as described in Proposition 3. The agency's effort investments in this case are characterized by the following result.

Proposition 4. *Suppose* $\pi > V_\varepsilon(0) > V_\varepsilon(1)$ *and* $\frac{p_1}{p_1+p_2} \leq \frac{1}{4}(4-4\beta-V_\varepsilon(0))$. *Then, given* $x_A^{\textit{semi-pool}}(\omega)$ *and* $s_R(x_A,e)$*, the agency invests high effort if* $(p_1+p_2)(V_\varepsilon(0)-V_\varepsilon(1)) \geq \kappa$.

Proposition 4 says that when the agency is always obfuscating by choosing $x_A = 2$ when $\omega = 1$ it will invest high effort so long as the implementation precision improvement, in the states in which the agency will be upheld (which obtain with probability $p_1 + p_2$), outweighs the effort costs associated with inducing that improvement. Since the agency must make this effort investment decision prior to learning ω, the agency weights the potential for these policy improvements by the probabilities that its policy will be upheld by the overseer.

line with the fact, described above, that the agency never wants to deviate from truthfulness when $\omega = 0$.

This is similar to the case in the procedural review model in which the agency is always upheld, except for the fact that when $\omega = 0$ the agency will not be upheld and therefore the agency does not take that state into account when making its effort investment decision.[28] This implies that the constraint for the agency to invest high effort is more stringent than in the case of a perfectly deferential overseer in the procedural review model. Moreover, in this case the agency does not enjoy the ability to match policy to the state in all cases. Thus, the agency would rather be subject to procedural review when it will always be upheld, and it will be more likely to invest high effort under procedural review, rather than have to obfuscate sometimes as is the case under substantive review. Finally, note that the agency will invest high effort for a wider range of effort costs in this equilibrium than in the truthful equilibrium when the overseer has moderately divergent preferences since $(p_1 + p_2)(V_\varepsilon(0) - V_\varepsilon(1)) > p_2(V_\varepsilon(0) - V_\varepsilon(1))$. This implies that the agency will invest high effort for higher effort costs if it obfuscates with its policy choice following that investment.

Moderately punitive reversal. In this environment being overturned is only costly enough to induce obfuscation when the agency has invested high effort: $V_\varepsilon(0) > \pi > V_\varepsilon(1)$. Thus, the agency only wants to obfuscate with its policy choice when $\omega = 1$ following high effort investments. The agency decides between investing low effort, setting policy truthfully, and being overturned and investing high effort, obfuscating with its policy choice when $\omega = 1$, and receiving deference. The next result characterizes agency effort investment behavior in this environment.

Proposition 5. *Suppose* $V_\varepsilon(0) > \pi > V_\varepsilon(1)$ *and* $\frac{p_1}{p_1+p_2} \leq \frac{1}{4}(4-4\beta-V_\varepsilon(0))$. *Then, given* $x_A^{\textit{semi-pool}}(\omega)$ *and* $s_R(x_A, e)$*, the agency invests high effort if* $p_1(\pi - V_\varepsilon(1)) + p_2(V_\varepsilon(0) - V_\varepsilon(1)) \geq \kappa$.

Proposition 5 says that the agency will invest high effort when reversal costs are moderately punitive if the benefits of doing so *and* avoiding reversal when $\omega = 1$, given that the agency will only obfuscate following $e = 1$, outweighs the cost of high effort κ. The condition for high effort when reversal costs are moderate is more stringent than when these costs are high (i.e., $p_1(\pi - V_\varepsilon(1)) + p_2(V_\varepsilon(0) - V_\varepsilon(1)) < (p_1 + p_2)(V_\varepsilon(0) - V_\varepsilon(1))$ since $\pi < V_\varepsilon(0)$), implying that high effort

[28] To see this more starkly note that $p_1 + p_2 = 1 - p_0$, implying that the precision improvements only matter to the agency when $\omega \neq 0$.

will be invested for a wider range of the parameter space as π increases. Similar to the previous environment of high reversal costs, in this case the agency will invest high effort for a wider range of effort costs than in the truthful equilibrium with moderately divergent overseer preferences since $p_1(\pi - V_\varepsilon(1)) + p_2(V_\varepsilon(0) - V_\varepsilon(1)) > p_2(V_\varepsilon(0) - V_\varepsilon(1))$. The agency is willing to invest high effort at higher cost levels when he will obfuscate and be upheld following that investment.

The preceding analysis of substantive oversight illustrates that increasing transparency in the review process introduces the possibility for obfuscation through policy exaggeration. In equilibrium, this implies that in environments conducive to semi-pooling behavior agency policymaking becomes polarized: either the agency truthfully reveals that no policy change is called for or, if any policy change is called for, it pursues extreme policy change. There is no chance for moderate changes to bring policy in line with the policy environment due to reputational considerations and the preference environment unless reversal costs are very low. This implies, understanding low reversal costs as representative of highly insulated agencies, that independent agencies are more likely to engage in truthful policymaking than highly politically accountable agencies that face much higher reputational considerations.

Incentives for the agency to match policy to the state, or set policy truthfully, are not the only important incentives affected by oversight. Procedural review allows the agency to follow its policy-relevant information and set policy to match the contingencies of the policy environment. However, procedural review may also deter the agency from investing high effort in certain circumstances. Substantive review is more likely to lead the agency to disregard policy-relevant information and instead pursue only extreme policy adjustment when it believes policy change is called for *and* this policy exaggeration is *more likely* when the agency has already invested high effort. Thus, which form of institutional oversight is more beneficial depends crucially on characteristics of the agency (e.g., effort costs, reversal costs) as well as those of the policy environment (e.g., preference arrangements, probability that policy change is called for, etc.). The next section explores these considerations from the overseer's perspective.

4 Reviewing procedure vs. judging substance

Is it always better for the overseer to have more information when she reviews the agency? That is, does substantive review always benefit the overseer relative to procedural review? To explore this question I consider the overseer's ex ante welfare in a similar policymaking situation across the two different scopes of review. The results show that either type of review can be optimal depending on the nature of preferences and the impact that agency effort has on implementation precision.

I focus on the procedural review environment in which the overseer is conditionally-deferential. The overseer only upholds the agency in this case if the agency invests high effort. I compare this with the substantive review environment characterized in Proposition 5: a moderately biased overseer that will overturn the agency if it invests low effort and chooses $x_A(1) = 1$ but will uphold the agency if it chooses $x_A^{\text{semi-pool}}(1) = 2$ and invests high effort. I will also compare the cases in which the agency invests high effort and, in the case of substantive review, obfuscates with its substantive policy choice when $\omega = 1$ by setting $x_A^{\text{semi-pool}}(\omega) = 2$. These are the most interesting environments across the two models and serve as a good comparison since the agency invests high effort in both cases. Thus, the welfare comparison comes down to when the overseer benefits from also observing x_A relative to only observing e.

The overseer's ex ante welfare when reviewing procedure is given by,

$$\begin{aligned} W_R^P(e=1, x_A^{\text{proc}}(\omega), p) &= -p_0(\beta^2 + V_\varepsilon(1)) - p_1(\beta^2 + V_\varepsilon(1)) - p_2(\beta^2 + V_\varepsilon(1)), \\ &= -\beta^2 - V_\varepsilon(1). \end{aligned}$$

Since in this case the overseer upholds the agency with certainty she can expect to lose utility based on her bias β and the imprecision of implementation given high effort $V_\varepsilon(1)$. The overseer's analogous ex ante welfare when judging substance is given by,

$$W_R^S(e=1, x_A^{\text{semi-pool}}(\omega), p) = -p_0(\beta^2) - p_1((\beta+1)^2 + V_\varepsilon(1)) - p_2(\beta^2 + V_\varepsilon(1)).$$

The overseer is better off with substantive review when $\omega = 0$ since in this case she would prefer to overturn where she could not in the procedural review model. When $\omega = 1$ she is better off under the procedural review model since $(\beta+1)^2 + V_\varepsilon(1) > \beta^2 + V_\varepsilon(1)$. Under substantive review the agency obfuscates by choosing $x_A = 2$, which leads to a larger policy loss for the overseer. When $\omega = 2$ outcomes are equivalent for the overseer so she is indifferent between the two institutions.

Combining and rearranging the two welfare expressions yields the overseer's net welfare from procedural review, relative to substantive review:

$$\begin{aligned}
\Delta W_R(\text{Procedure vs. Substance}) &= W_R(e=1, x_A^{\text{proc}}(\omega), p) - W_R(e=1, x_A^{\text{semi-pool}}(\omega), p), \\
&= -\beta^2 - V_\varepsilon(1) + p_0(\beta^2) + p_1((\beta+1)^2 + V_\varepsilon(1)) + p_2(\beta^2 + V_\varepsilon(1)), \\
&= p_1(2\beta+1) - V_\varepsilon(1)(1-p_1-p_2). \qquad (3)
\end{aligned}$$

So long as $\Delta W_R(\text{Procedure vs. Substance}) > 0$ the overseer benefits from procedural review relative to substantive review, implying that the overseer is actually made *worse off* by judging the additional information of x_A because this induces the agency to obfuscate when $\omega = 1$. When inequality 3 goes in the other direction, the overseer would prefer to be able to judge the substance of the agency's policy choice. It is more likely that the overseer benefits from procedural review as the probability that any policy change is called for (i.e., $\omega = 1$ and $\omega = 2$) increases, as her bias β increases, and as high effort implementation precision increases (i.e., as $V_\varepsilon(1)$ decreases). This follows from the fact that under substantive review the increased transparency of the agency's policy choices in the review process induces the agency to obfuscate leading to larger substantive policy losses. The only time the agency strictly benefits from substantive review is when the state is $\omega = 0$. Thus, the more likely it is that the state is either $\omega = 1$ or $\omega = 2$ the lower the likelihood the overseer will benefit from being able to overturn $x_A = 0$.

Note that $1 - p_1 - p_2 = p_0$, which implies that as the probability that $\omega = 0$ increases so does the likelihood that the overseer will benefit from substantive review. This is also true as the overseer's bias decreases and as the impact of high effort investments on quality implementation

decreases (i.e., as $V_\varepsilon(1)$ increases). In the case of the overseer's bias, it is more likely that she benefits from the extra information provided by substantive review when she is least in need of it: when her preferences are close to those of the agency.

Ultimately, when the policymaking environment is structured such that increasing transparency of agency actions will also induce the agency to obfuscate by exaggerating the need for extreme policy change the overseer only benefits from that extra information when her preferences are relatively close to those of the agent and the likelihood that the policy environment requires no policy change is high. This suggests that it is far from clear that providing overseers with more information during the review process will generate net benefits once one takes into account how that information disclosure will impact the upstream incentives for the policymakers in possession of that information. In some environments the overseer would prefer to be directed, through statutory language or the like, to only review procedure and be explicitly precluded from judging substance.

5 Discussion and conclusion

I have presented a theory of how different types of ex post oversight can produce different bundles of policymaking incentives to bureaucratic agencies. While procedural review allows the agency to utilize its informational advantage to set the substance of policy, it can harm incentives for effort investments that improve the implementation of policy on the ground. Substantive review, in contrast, can induce the agency to disregard policy-relevant information and exaggerate the need for, and magnitude of, policy change to avoid having its policies reversed. These perverse incentives are strengthened when the agency invests high effort toward implementation and when reversal costs are highly punitive. A key insight is that when the transparency of policymaking is increased there is a trade-off between effort incentives and the incentives for agencies to utilize their policy-relevant expertise. This undercuts the powerful normative rationale for delegation to expert agencies by inducing these agencies to underutilize their expertise.

Additionally, I have provided results that suggest that the overseer can benefit from *less information* in the review process when the probability that policy change is called for is high. This

suggests that it may be beneficial to shield bureaucratic policy actions from substantive review when they are asked to regulate dynamic, volatile policy environments that require substantive policy adjustments frequently. This is even more beneficial from the perspective of a more strongly biased overseer. Preference divergence with the agency, from the point of view of a political principal, like Congress or the president, with preferences similar to those of the overseer, is more likely to be harmful when the agency is subject to substantive oversight. All of these perverse effects are predicated on the fact that agencies seek to avoid the punitive costs of being reversed. Increasing the transparency of agencies' actions only intensifies those costs to the point of driving an agency to disregard private information and obfuscate with its policy choice. The scope of review that agencies are subjected to can have profoundly differential effects on the agency's policymaking incentives. These results suggest political actors designing review provisions that define the relationships between agencies and their overseers need to be cognizant of the 'ripple effect' these choices may have throughout the policymaking process.

References

Ashworth, Scott. 2012. "Electoral Accountability: Recent Theoretical and Empirical Work." *Annual Review of Political Science* 15:183–201.

Barro, Robert J. 1973. "The Control of Politicians: An Economic Model." *Public Choice* 14(1):19–42.

Beim, Deborah, Alexander V Hirsch and Jonathan P Kastellec. 2014. "Whistleblowing and Compliance in the Judicial Hierarchy." *American Journal of Political Science* 58(4):904–918.

Bendor, Jonathan and Adam Meirowitz. 2004. "Spatial Models of Delegation." *American Political Science Review* 98(2):293–310.

Bendor, Jonathan, Amihai Glazer and Thomas Hammond. 2001. "Theories of Delegation." *Annual Review of Political Science* 4(1):235–269.

Bueno de Mesquita, Ethan and Matthew C. Stephenson. 2007. "Regulatory Quality under Imperfect Oversight." *American Political Science Review* 101(3):605–620.

Cameron, Charles M. 2000. *Veto Bargaining*. New York, NY: Cambridge University Press.

Canes-Wrone, Brandice, Michael C. Herron and Kenneth W. Shotts. 2001. "Leadership and Pandering: A Theory of Executive Policymaking." *American Journal of Political Science* 45(3):532–550.

Carpenter, Daniel P. 2001. *The Forging of Bureacuratic Autonomy: Reputations, Networks, and Policy Innovation in Executive Agencies, 1862-1928*. Princeton, NJ: Princeton University Press.

Cass, Ronald A. 1989. Review, Enforcement, and Power under the COmmunications Act of 1934: Choice and Chance in Institutional Design. In *A Legislative History of the Communications Act of 1934*, ed. Max D. Paglin. New York, NY: Oxford University Press.

Chutkow, Dawn M. 2008. "Jurisdiction Stripping: Ideology, Institutional Concerns, and Congressional Control of the Court." *Journal of Politics* 70(4):1053–1064.

Clark, Tom S. 2016. "Scope and Precedent: Judicial Rule-making Under Uncertainty." *Journal of Theoretical Politics* 28(3):353–384.

Crombez, Christophe, Tim Groseclose and Keith Krehbiel. 2006. "Gatekeeping." *Journal of Politics* 68(2):322–334.

Epstein, David and Sharyn O'Halloran. 1995. "A Theory of Strategic Oversight: Congress, Lobbyists, and the Bureaucracy." *Journal of Law, Economics, and Organization* 11(2):227–255.

Epstein, David and Sharyn O'Halloran. 1999. *Delegating Powers: A Transaction Cost Politics Approach to Policy Making Under Separate Powers*. New York, NY: Cambridge University Press.

Fearon, James D. 1999. Electoral Accountability and the Control of Politicians. In *Democracy, Accountability and Representation*, ed. Adam Przeworski, Susan C. Stokes and Bernard Manin. New York, NY: Cambridge University Press.

Ferejohn, John. 1986. "Incumbent Performance and Electoral Control." *Public Choice* 50(1):5–25.

Fox, Justin. 2007. "Government Transparency and Policymaking." *Public Choice* 131(1-2):23–44.

Fox, Justin and Georg Vanberg. 2014. "Narrow versus Broad Judicial Decisions." *Journal of Theoretical Politics* 26(3):355–383.

Fox, Justin and Matthew C. Stephenson. 2011. "Judicial Review as a Response to Political Posturing." *American Political Science Review* 105(2):397–414.

Fox, Justin and Matthew C Stephenson. 2015. "The Welfare Effects of Minority-Protective Judicial Review." *Journal of Theoretical Politics* 27(4):499–521.

Fox, Justin and Richard Van Weelden. 2012. "Costly Transparency." *Journal of Public Economics* 96(1):142–150.

Fox, Justin and Richard Van Weelden. 2015. "Hoping for the Best, Unprepared for the Worst." *Journal of Public Economics* 130(2015):59–65.

Gailmard, Sean. 2002. "Expertise, Subversion, and Bureaucratic Discretion." *Journal of Law, Economics, & Organization* 18(2):536–555.

Gailmard, Sean. 2009. "Discretion Rather than Rules: Choice of Instruments to Control Bureaucratic Policy Making." *Political Analysis* 17(1):25–44.

Gailmard, Sean and John W. Patty. 2013*a*. "Formal Models of Bureaucracy." *Annual Review of Political Science* 15:353–377.

Gailmard, Sean and John W. Patty. 2013*b*. *Learning While Governing: Expertise and Accountability in the Executive Branch*. Chicago, IL: University of Chicago Press.

Gersen, Jacob E. and Matthew C. Stephenson. 2014. "Over-accountability." *Journal of Legal Analysis* 6(2):185–243.

Groseclose, Tim and Nolan McCarty. 2001. "The Politics of Blame: Bargaining before an Audience." *American Journal of Political Science* 45(1):100–119.

Kagan, Elena. 2001. "Presidential Administration." *Harvard Law Review* 114(8):2245–2385.

Light, Paul C. 1991. *Forging Legislation*. New York, NY: W.W. Norton.

Lipsky, Michael. 1980. *Street-Level Bureaucracy*. New York, NY: Russell Sage Foundation.

Majumdar, Sumon and Sharun W. Mukand. 2004. "Policy Gambles." *American Economic Review* 94(4):1207–1222.

McCann, Pamela J., Charles R. Shipan and Yuhua Wang. 2016. "Congress and Judicial Review of Agency Actions." *Working Paper. University of Southern California* .

McCubbins, Mathew D. and Thomas Schwartz. 1984. "Congressional Oversight Overlooked: Police Patrols versus Fire Alarms." *American Journal of Political Science* 28(1):165–179.

Melnick, R. Shep. 1983. *Regulation and the Courts: The Case of The Clean Air Act*. Washington, D.C.: The Brookings Institution.

Melnick, R. Shep. 1994. *Between the Lines: Interpreting Welfare Rights*. Washington, D.C.: The Brookings Institution Press.

Miller, Gary J. 2005. "The Political Evolution of Principal-Agent Models." *Annual Review of Political Science* 8:203–225.

Patty, John W. and Ian R. Turner. 2016. "Ex Post Review and Expert Policymaking: When Does Oversight Reduce Accountability?" *Unpublished Manuscript. Yale University. Presented at the 2016 Annual Meeting of the American Political Science Association* .

Prat, Andrea. 2005. "The Wrong Kind of Transparency." *American Economic Review* 95(3):862–877.

Rose-Ackerman, Susan. 1995. *Controlling Environmental Policy*. New Haven, CT: Yale University Press.

Shapiro, Sidney A. and Richard E. Levy. 1995. "Judicial Incentives and Indeterminacy in Substantive Review of Administrative Decisions." *Duke Law Journal* 44(6):1051–1080.

Shipan, Charles. 1997. *Designing Judicial Review*. Ann Arbor, MI: University of Michigan Press.

Shipan, Charles R. 2000. "The Legislative Design of Judicial Review: A Formal Analysis." *Journal of Theoretical Politics* 12(3):269–304.

Smith, Joseph L. 2005. "Congress Opens the Court Doors: Statutory Changes to Judicial Review Under the Clean Air Act." *Political Research Quarterly* 58(1):139–149.

Smith, Joseph L. 2006. "Judicial Procedures as Instruments of Political Control: Congress's Strategic Use of Citizen Suits." *Legislative Studies Quarterly* 31(2):283–305.

Stephenson, Matthew C. 2006. "A Costly Signaling Theory of "Hard Look" Judicial Review." *Administrative Law Review* 58(4):753–814.

Sunstein, Cass R. 1989. "On the Costs and Benefits of Aggressive Judicial Review of Agency Action." *Duke Law Journal* 1989(3):522–537.

Turner, Ian R. 2017*a*. "Political Agency, Oversight, and Bias: The Instrumental Value of Politicized Policymaking." *Unpublished Manuscript. Yale University* .
URL: *https://goo.gl/dw3c8I*

Turner, Ian R. 2017*b*. "Working Smart *and* Hard? Agency Effort, Judicial Review, and Policy Precision." *Journal of Theoretical Politics* 29(1):69–96.

Vanberg, Georg. 2001. "Legislative-Judicial Relations: A Game-Theoretic Approach to Constitutional Review." *American Journal of Political Science* 45(2):346–361.

Wagner, Wendy. 2012. "Revisiting the Impact of Judicial Review on Agency Rulemakings: An Empirical Investigation." *William & Mary Law Review* 53(5):1717–1795.

A Supplemental appendix

A.1 Procedural review model

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

Proof of Lemma 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.

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

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

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

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

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

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

implying a net utility loss equal to the policy choice deviation. Thus, the agency is strictly worse off by deviating from the proposed strategy when the overseer will uphold the agency.

Case 2: Agency overturned. The agency's payoff in this case is equivalent regardless of its policy choice. So long as the overseer overturns $x = 0$ and therefore the agency is (weakly) better off sticking to the proposed equilbrium strategy of $x_A^*(\omega) = \omega$.

Taken together these two cases imply that, in weakly undominated pure strategies, the agency will always choose $x_A^{\text{procedure}}(\omega) = \omega$ in the procedural review model. ■

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

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

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

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

Now, the overseer's expected payoff for reversing the agency depends on the state ω, which is unknown to the overseer in the procedural review model. The overseer's expected payoff for over-

turning if $\omega = 0$, which the overseer believes to have obtained with probability p_0, is given by,

$$\begin{aligned} EU_R(r=1|e,\beta,\omega=0) &= -(\omega-\beta-(1-r)x)^2, \\ &= -(0-\beta-(0)x)^2, \\ &= -\beta^2. \end{aligned}$$

The overseer's expected payoff for reversing the agency given that $\omega = 1$, which has occurred with probability p_1, is given by,

$$\begin{aligned} EU_R(r=1|e,\beta,\omega=1) &= -(\omega-\beta-(1-r)x)^2, \\ &= -(1-\beta-(0)x)^2, \\ &= -(1-\beta)^2. \end{aligned}$$

Finally, the overseer's expected payoff for reversing when $\omega = 2$, which the overseer believes to have obtained with probability p_2, is given by,

$$\begin{aligned} EU_R(r=1|e,\beta,\omega=2) &= -(\omega-\beta-(1-r)x)^2, \\ &= -(2-\beta-(0)x)^2, \\ &= -(2-\beta)^2. \end{aligned}$$

Combining these possibilities given the overseer's beliefs over the probability distribution of states (i.e., $p = \{p_0, p_1, p_2\}$) yields the overseer's overall expected payoff for reversing an agency that has invested effort e:

$$\begin{aligned} EU_R(r=1|e,\beta,p) &= -(\omega-\beta-(1-r)x)^2, \\ &= -p_0\left(\beta^2\right) - p_1\left((1-\beta)^2\right) - p_2\left((2-\beta)^2\right). \end{aligned}$$

Combining and rearranging these two expected payoffs (for upholding and overturning, respectively)

yields the incentive compatibility constraint that must be met in order for the overseer to uphold the agency:

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

This yields the result as stated in the lemma. ■

Now, recall the definitions derived from the overseer's incentive compatibility constraint to uphold. That is, it must be the case that $\beta \in \left(0, \frac{p_1+4p_2-V_\varepsilon(e)}{2p_1+4p_2}\right]$ for the overseer to uphold. We can define two β-thresholds based on whether the agency invested high or low effort: $\beta_1 \equiv \frac{p_1+4p_2-V_\varepsilon(1)}{2p_1+4p_2}$ and $\beta_0 \equiv \frac{p_1+4p_2-V_\varepsilon(0)}{2p_1+4p_2}$ 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 deferential*. If $\beta_1 < \beta_0 < \beta$ then the overseer always overturns and is *perfectly skeptical*. If $\beta_1 < \beta < \beta_0$ then the overseer upholds if and only if $e = 1$ and is *conditionally deferential*. The next result characterizes how the agency best responds with its effort choices conditional on these oversight regimes.

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

1. *If $\beta < \beta_1 < \beta_0$ then the overseer is **perfectly deferential** and the agency invests high effort if $V_\varepsilon(0) - V_\varepsilon(1) \geq \kappa$.*

2. *If $\beta_1 < \beta_0 < \beta$ then the overseer is **perfectly skeptical** and the agency never invests high effort.*

3. *If $\beta_1 < \beta < \beta_0$ then the overseer is **conditionally deferential** and the agency invests high effort if $p_1 + 4p_2 + \pi - V_\varepsilon(1) \geq \kappa$.*

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

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

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

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

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

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.

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

the agency's expected payoff for investing low effort in this case,

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

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

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

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

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

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

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

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

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

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

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

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

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

Proposition 1. *The equilibrium to the procedural review model is characterized by the following collection of strategies:*

- *The overseer makes review decisions based on $s_R(e)$ (i.e., equation 1).*
- *The agency always sets substantive policy to match the state: $x_A^{procedure}(\omega)=\omega$.*
- *When facing a perfectly deferential overseer (i.e., $\beta_0<\beta_1<\beta$) the agency invests high effort when $V_\varepsilon(0)-V_\varepsilon(1)\geq\kappa$.*
- *When facing a perfectly skeptical overseer (i.e., $\beta_0<\beta_1<\beta$) the agency never invests high effort.*
- *When facing a conditional-deference overseer (i.e., $\beta_0<\beta<\beta_1$) the agency invests high effort if $p_1+4p_2+\pi-V_\varepsilon(1)\geq\kappa$.*

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

A.2 Substantive review model

A.2.1 Truthful separating equilibria

In this section I prove the results for truthful separating equilibria in the substantive review model.

Optimal substantive review.

Lemma 4. *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,*

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

Proof of Lemma 4. 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:

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

With this expected payoff for upholding, for any level of $e \in \{0.1\}$, 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,

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

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

$$-\beta^2 - V_\varepsilon(e) \geq -\beta^2,$$

which is never satisfied. Thus, the overseer *always* overturns the agency ($r = 1$) when the agency sets policy truthfully and $\omega = 0$.

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,

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

For the overseer to uphold incentive compatibility requires that,

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

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

$$\begin{aligned} EU_R(r=1|x_A(2)=2,e) &= -(\omega-\beta-(1-r)x)^2, \\ &= -(2-\beta)^2, \\ &= 4\beta-\beta^2-4. \end{aligned}$$

This yields the following incentive compatibility constraint to uphold:

$$\begin{aligned} -\beta^2 - V_\varepsilon(e) &\geq 4\beta-\beta^2-4, \\ 4-4\beta &\geq V_\varepsilon(e), \\ 4(1-\beta) &\geq V_\varepsilon(e). \end{aligned}$$

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 2, which corresponds to Table 1 in the main body.

ω	**Aligned Preferences:** $\beta \in \left[0, \frac{1-V_\varepsilon(0)}{2}\right)$	**Conditionally Aligned Preferences:** $\beta \in \left[\frac{1-V_\varepsilon(0)}{2}, \frac{1-V_\varepsilon(1)}{2}\right)$	**Moderate Preferences:** $\beta \in \left[\frac{1-V_\varepsilon(1)}{2}, \frac{4-V_\varepsilon(0)}{4}\right]$	**Conditionally Extreme Preferences:** $\beta \in \left(\frac{4-V_\varepsilon(0)}{4}, \frac{4-V_\varepsilon(1)}{4}\right]$	**Extreme Preferences:** $\beta > \frac{4-V_\varepsilon(1)}{4}$
0	$r(e)=1, \forall e$	$r(e)=1, \forall e$	$r(e)=1, \forall e$	$r(e)=1, \forall e$	$r(e)=1, \forall e$
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 2: Overseer decisions given truthful policymaking ($x_A = \omega$) conditional on state ω, effort e, and bias β.

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. The next result characterizes the conditions under which the agency will *always* set policy truthfully by separating.

Proposition 2. *There is a truthful separating equilibrium in which the agency always matches policy to the state* $\left(x_A^{truth}(\omega)\right)$ *if and only if reversal costs are not too punitive* $(V_\varepsilon(e) > \pi)$. *Furthermore,* $p_2(V_\varepsilon(0) - V_\varepsilon(1)) \geq \kappa$ *is sufficient to ensure the agency invests high effort for all ranges of overseer bias except when the agency will always be overturned* $\left(\beta > \frac{4-V_\varepsilon(e)}{4}\right)$, *in which case the agency will never invest high effort.*

Proof of Proposition 2. To prove the result I derive the incentive compatibility conditions for the agency to stick with $x_A^{\text{truth}}(\omega) = \omega$, rather than deviate, for each possible state of the world. First note that any time the agency will be upheld following truthfully matching its policy choice to the state there is no incentive to deviate. Thus, we need only consider the cases in which a deviation would lead to being upheld when remaining truthful would lead to reversal. I consider each case in turn.

Case 1: $\omega = 0$**.** When the true state is $\omega = 0$ the agency must choose between setting $x_A = 0$ truthfully, revealing ω to the overseer, and being reversed and deviating to $x_A = 1$ when it would induce being upheld (which only occurs for particular ranges of overseer biases). First, consider the agency's payoff from being truthful given that it will be overturned:

$$\begin{aligned} EU_A(x_A^{\text{truth}}(0) = 0 | s_R^*(x_A, e) = 1, e) &= -(\omega - (1-r)x)^2 - \kappa e - \pi r, \\ &= -(0 - (1-1)x)^2 - \kappa e - \pi, \\ &= -\kappa e - \pi. \end{aligned}$$

Now consider the agency's payoff from deviating to $x_A = 1$, assuming that that will induce being upheld (if it simply induces being overturned then the problem is trivial since outcomes do not vary):

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

Combining and rearranging these payoffs yields the incentive compatibility constraint for agency to remain truthful even though it will lead to reversal:

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

which cannot be satisfied since $\pi < 1$ and $V_\varepsilon(e) > 1$, $\forall e \in \{0,1\}$. Thus, the agency would always prefer to truthfully reveal ω by matching policy to the state even though it will be overturned when $\omega = 0$.

Case 2: $\omega = 1$. In this case the agency would only ever 'deviate up' to $x_A = 2$ to induce being upheld since deviating down to $x_A = 0$ would lead to reversal. Thus, the agency chooses between remaining truthful and revealing $\omega = 1$, which leads to being overturned, or deviating to $x_A = 2$, which leads to being upheld (again, if it did not then thre is no incentive to deviate at all). Consider first the agency's payoff from setting $x_A^{\text{truth}}(1) = 1$,

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

Now consider the agency's payoff from deviating to $x_A = 2$ to induce the overseer to uphold,

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

Now, combining and rearranging these expressions yields the incentive compatibility constraint for

the agency to remain truthful and match policy to the state when $\omega = 1$:

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

Thus, the agency will remain truthful even when it will lead to being overturned when $\omega = 1$ if errors in implementation are more costly than the punishment associated with being overturned by the overseer.

Case 3: $\omega = 2$. In this case there is no incentive for the agency to deviate. Either $x_A(2) = 2$ is upheld, in which case there is no incentive to deviate, or $x_A(2) = 2$ is overturned. If it is overturned then it must be the case that the overseer is extremely biased. This implies that the overseer is also too biased to uphold $x_A = 1$ (or $x_A = 0$) and therefore there is again no incentive to deviate.

Now, notice that the only time the condition for the agency to remain truthful can be violated, given the restrictions of the model, is when $\omega = 1$. In this case the agency will continue to set $x_A^{\text{truth}}(\omega) = \omega$ if $V_\varepsilon(e) \geq \kappa$. Recall that $V_\varepsilon(0) > V_\varepsilon(1)$. If $pi > V_\varepsilon(0) > V_\varepsilon(1)$ then there will always be an incentive for the agency to deviate, regardless of its prior effort investment, when $x_A(1) = 2$ will lead to being upheld. If $V_\varepsilon(0) > \pi > V_\varepsilon(1)$ then the agency has an incentive to deviate when it has invested high effort ($e = 1$) and setting $x_A(1) = 2$ will lead to being upheld. Thus, $V_\varepsilon(1) > \pi$ is both necessary and sufficient to ensure that the agency never has an incentive to deviate from truthful policymaking, as stated in the result.

The final statement in the result regarding the sufficient condition for the agency to invest high effort given that it always sets substantive policy truthfully is illustrated by Lemma 5. ■

Lemma 5. *Suppose the agency always sets substantive policy truthfully. Then, conditional on the level of overseer bias, the agency makes effort investment decisions as follows:*

- *If* $\beta \in \left[0, \frac{1-V_\varepsilon(0)}{2}\right)$ *then the agency invests high effort if* $(1 - p_0)(V_\varepsilon(0) - V_\varepsilon(1)) \geq \kappa$.
- *If* $\beta \in \left[\frac{1-V_\varepsilon(0)}{2}, \frac{1-V_\varepsilon(1)}{2}\right)$ *then the agency invests high effort if* $p_1(1 + \pi - V_\varepsilon(1)) + p_2(V_\varepsilon(0) - V_\varepsilon(1)) \geq \kappa$.

- *If* $\beta \in \left[\frac{1-V_\varepsilon(1)}{2}, \frac{4-V_\varepsilon(0)}{4}\right]$ *then the agency invests high effort if* $p_2(V_\varepsilon(0)-V_\varepsilon(1)) \geq \kappa$.
- *If* $\beta \in \left(\frac{4-V_\varepsilon(0)}{4}, \frac{4-V_\varepsilon(1)}{4}\right]$ *then the agency invests high effort if* $p_2(4+\pi-V_\varepsilon(1)) \geq \kappa$.
- *If* $\beta > \frac{4-V_\varepsilon(1)}{4}$ *then the agency never invests high effort.*

Proof of Lemma 5. I derive the stated condition in each environment to illustrate the result. First, consider the first case in which $\beta \in \left[0, \frac{1-V_\varepsilon(0)}{2}\right)$. In this case the overseer reverses following observation of $x_A = 0$ and upholds $x_A \in \{1,2\}$. Since we are in an environment in which the agency always sets policy truthfully ($\pi < V_\varepsilon(1) < V_\varepsilon(0)$) the agency chooses high or low effort based on its expected utility given the probability distribution over states, $p = \{p_0, p_1, p_2\}$. Consider the agency's expected utilities for $e = 1$ and $e = 0$ in this case:

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

Incentive compatibility requires that the following holds for the agency to invest high effort:

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

Noting that $p_1 + p_2 = (1 - p_0)$ completes the first result.

Now consider the case in which $\beta \in \left[\frac{1-V_\varepsilon(0)}{2}, \frac{1-V_\varepsilon(1)}{2}\right)$. In this case the overseer reverses $x_A = 0$ and $x_A = 1$ if $e = 0$, and upholds $x_A = 1$ if $e = 1$ and $x_A = 2$. The agency's expected payoffs for $e = 1$ and $e = 0$ in this case are given by,

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

Incentive compatibility requires the following expression to hold for the agency to invest high effort,

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

as stated in the second part of the result.

Consider the case in which $\beta \in \left[\frac{1-V_\varepsilon(1)}{2}, \frac{4-V_\varepsilon(0)}{4}\right]$. In this case the overseer reverses $x_A \in \{0,1\}$ regardless of e and upholds $x_A = 2$. The agency's expected payoffs for investing high and low effort, respectively, are given by,

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

Incentive compatibility requires that the following inequality hold for the agency to invest high effort:

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

as is stated in the third piece of the result.

Consider the penultimate case in which $\beta \in \left(\frac{4-V_\varepsilon(0)}{4}, \frac{4-V_\varepsilon(1)}{4}\right]$. In this case the overseer reverses following $x_A = 0$ and $x_A = 1$ regardless of e, $x_A = 2$ if $e = 0$, and upholds if $x_A = 2$ and

$e = 1$. The agency's expected payoffs for $e = 1$ and $e = 0$ in this case are given by,

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

Incentive compatibility requires that the following inequality hold for the agency to invest high effort:

$$\begin{aligned} -p_0(\kappa+\pi) - p_1(1+\kappa+\pi) - p_2(V_\varepsilon(1)+\kappa) &\geq -p_0(\pi) - p_1(1+\pi) - p_2(4+\pi), \\ -\kappa - p_0\pi - p_1 - p_1\pi - p_2V_\varepsilon(1) &\geq -p_0\pi - p_1 - p_1\pi - 4p_2 - p_2\pi, \\ -\kappa - p_2V_\varepsilon(1) &\geq -4p_2 - p_2\pi, \\ 4p_2 + p_2\pi - p_2V_\varepsilon(1) &\geq \kappa, \\ p_2(4+\pi - V_\varepsilon(1)) &\geq \kappa, \end{aligned}$$

as stated in the fourth scenario in the result.

Finally, consider the case in which $\beta > \frac{4-V_\varepsilon(1)}{4}$. In this case the agency is always reversed by the overseer regardless of e and x_A. In this case the agency's expected payoffs given it will always be reversed are given by,

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

Thus, the net expected payoff for investing high effort is,

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

or a net loss proportional to the cost of high effort. This implies that the agency will never invest

high effort in this environment. ■

Corollary 1. *The incentive for the agency to obfuscate with its substantive policy choice* $\left(x_A \neq x_A^{truth}(\omega)\right)$ *is stronger when the agency invests high effort.*

Proof of Corollary 1. This follows from the fact that the general condition that is sufficient to ensure that the agency sets substantive policy truthfully is $V_\varepsilon(e) \geq \pi$, as derived in the proof of Proposition 2. The range of reversal penalties that would lead the agency to abandon truthful policymaking following low effort investment is $\pi \in (V_\varepsilon(0), 1)$ and the analogous range following high effort investment is $\pi \in (V_\varepsilon(1), 1)$. Since $V_\varepsilon(1) < V_\varepsilon(0)$ there is a strictly wider range of π that would cause the agency to deviate from truthful policymaking following high effort investment, which implies that the incentives for the agency to deviate from truthful policymaking are stronger following high effort investment. This further implies that the agency is more likely to deviate when it has invested high effort into policymaking to avoid being reversed, as stated in the result. ■

A.2.2 Obfuscation equilibria

Proposition 3. *Suppose* $\pi > V_\varepsilon(e)$. *If the overseer is moderately biased* $\left(\beta \in \left(\frac{1-V_\varepsilon(e)}{2}, \frac{4-V_\varepsilon(e)}{4}\right)\right)$ *and*

$$\frac{p_1}{p_1 + p_2} \leq \frac{1}{4}\left(4 - 4\beta - V_\varepsilon(e)\right) \tag{4}$$

then there is a pure strategy semi-pooling obfuscation equilibrium in which the agency sets substantive policy at $x_A = 0$ *when* $\omega = 0$ *and* $x_A = 2$ *for both* $\omega \in \{1,2\}$*:*

$$x_A^{semi\text{-}pool}(\omega) = \begin{cases} 0 & \text{if } \omega = 0, \\ 2 & \text{if } \omega \in \{1,2\}, \end{cases}$$

and the overseer upholds $x_A = 2$ *and overturns* $x_A \in \{0,1\}$*:*

$$s_R(x_A, e) = \begin{cases} \textit{Uphold: } r = 0 & \text{if } x_A = 2, \\ \textit{Overturn: } r = 1 & \text{if } x_A \in \{0,1\}. \end{cases}$$

Proof of Proposition 3. Consider the following 'pure' semi-pooling strategy for the agency:

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

If the agency employs $x_A^{\text{semi-pool}}$, then the overseer will never uphold following observation of $x_A = 0$. This is because given the agency's strategy the overseer learns with certainty that $\omega = 0$, and upholding in this case leads to a net loss. To see this, consider the overseer's utility for reversing following $x_A = 0$,

$$\begin{aligned} EU_R(r = 1 | x_A^{\text{semi-pool}} = 0) &= -(\omega - \beta - (1-r)x)^2, \\ &= -(0 - \beta - (0)x)^2, \\ &= -\beta^2. \end{aligned}$$

The overseer's analogous payoff for upholding is given by,

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

Thus, the overseer would only uphold the agency following $x_A = 0$ given $x_A^{\text{semi-pool}}$ if,

$$-\beta^2 - V_\varepsilon(e) \geq -\beta^2,$$

which can never be satisfied.

Fixing off-path beliefs for the overseer such that an observation of $x_A = 1$ induces the belief that $\omega = 1$ the overseer's payoff for upholding following $x_A = 1$ is given by,

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

The overseer's payoff for reversing in this case is given by,

$$\begin{aligned} EU_R(r = 1|x_A = 1) &= -(1-\beta-(0)x)^2, \\ &= -(1-\beta)^2. \end{aligned}$$

Incentive compatibility requires the following expression to hold for the overseer to uphold in this case,

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

This yields the lower bound on the overseer's preferences as stipulated in the result. So long as $\beta > \frac{1-V_\varepsilon(e)}{2}$ the overseer's best response in this case is to reverse.

Finally, consider the overseer's decision-making following $x_A = 2$. In this case there are two possibilities, either $\omega = 1$ or $\omega = 2$. The overseer's beliefs in this case are updated according to Bayes' rule and the prior probabilities p_1 and p_2 as follows,

$$\begin{aligned} Pr[\omega = 1|x_A^{\text{semi-pool}} = 2] &= \frac{p_1}{p_1+p_2}, \text{ and} \\ Pr[\omega = 2|x_A^{\text{semi-pool}} = 2] &= \frac{p_2}{p_1+p_2}. \end{aligned}$$

The overseer's expected payoffs for upholding and overturning following observation of $x_A = 2$ are

given by,

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

Combining these expected payoffs for upholding and overturning letting $q \equiv \frac{p_1}{p_1+p_2}$ and $(1-q) \equiv \frac{p_2}{p_1+p_2}$ (the overseer's beliefs regarding ω) yields the overseer's inventive compatibility constraint to uphold the agency given $x_A^{\text{semi-pool}}$ following observation of $x_A = 2$:

$$\begin{aligned}
-\left(q((\beta+1)^2+V_\varepsilon(e))+(1-q)(\beta^2+V_\varepsilon(e))\right) &\geq -\left(q((1-\beta)^2)+(1-q)((2-\beta)^2)\right),\\
-\beta^2-q-2\beta q-V_\varepsilon(e) &\geq -4-\beta^2-2\beta q+4\beta+3q,\\
4-4\beta-V_\varepsilon(e) &\geq 4q,\\
\frac{1}{4}(4-4\beta-V_\varepsilon(e)) &\geq \frac{p_1}{p_1+p_2}
\end{aligned}$$

Thus, the overseer will uphold the agency, given $x_A^{\text{semi-pool}}(\omega)$, following observation of $x_A = 2$ if $\frac{p_1}{p_1+p_2} \leq \frac{1}{4}(4-4\beta-V_\varepsilon(e))$. Note that $\frac{1}{4}(4-4\beta-V_\varepsilon(e)) > 0$ so long as $\beta < \frac{4-V_\varepsilon(e)}{4}$, which yields the upper bound on the overseer's preferences as stipulated in the result. That is, given $x_A^{\text{semi-pool}}(\omega)$, $s_R(x_A,e)$ is a best response as stated in the result.

To verify that $x_A^{\text{semi-pool}}(\omega)$ is a best response to $s_R(x_A^{\text{semi-pool}}, e)$ first consider the case when $\omega = 0$. In this case the agency has no incentive to deviate unless it will lead to being upheld. The agency's payoff for sticking with the posited strategy is given by,

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

If instead the agency were to deviate to $x_A = 2$, which would induce deference, its payoff is given by,

$$\begin{aligned} EU_A(x_A = 2|\omega = 0, s_R) &= -(0 - (1)(2+\varepsilon)^2 - \kappa e - \pi(0), \\ &= -(0-2)^2 - V_\varepsilon(e) - \kappa e, \\ &= -4 - V_\varepsilon(e) - \kappa e. \end{aligned}$$

Thus, the agency will stick with $x_A^{\text{semi-pool}}(0) = 0$ if,

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

which is always satisfied since $\pi \in (0,1)$.

We need only consider the case in which $\omega = 1$ to verify that choosing $x_A(\omega) = 2$ for $\omega \in \{1,2\}$ is a best response since when $\omega = 2$ the agency is getting to match policy to the state and avoid reversal. Consider the agency's payoff from choosing $x_A = 2$ when $\omega = 1$, given that this

induces being upheld,

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

The agency's payoff for instead choosing $x_A = 1$, which will lead to reversal is given by,

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

Thus, incentive compatibility requires that the following hold for the agency to stick with $x_A^{\text{semi-pool}}(\omega)$ (assuming that the agency breaks indifference with being truthful, hence the 'strictness' of the inequality),

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

as stipulated in the statement of the result. ■

Highly punitive reversal cost. In this case $\pi > V_\varepsilon(0) > V_\varepsilon(1)$ and the equilibrium outlined in Proposition 3 holds for both $e=0$ and $e=1$ so long as $\frac{p_1}{p_1+p_2} \leq \frac{1}{4}(4-4\beta-V_\varepsilon(e))$ for both $e=0$ and $e=1$. Note that $\frac{1}{4}(4-4\beta-V_\varepsilon(1)) > \frac{1}{4}(4-4\beta-V_\varepsilon(0)$ since $V_\varepsilon(0) > V_\varepsilon(1)$. This implies that a higher probability that $\omega=1$, p_1, relative to the probability that $\omega=2$, p_2, will support this obfuscation when $e=1$, relative to $e=0$. This again suggests that obfuscation of the sort described in Proposition 3 is easier to support when the agency has invested high effort. The next result characterizes when, in this environment, the agency will invest high effort rather than low effort.

Proposition 4. *Suppose* $\pi > V_\varepsilon(0) > V_\varepsilon(1)$ *and* $\frac{p_1}{p_1+p_2} \leq \frac{1}{4}(4-4\beta-V_\varepsilon(0))$. *Then, given* $x_A^{\textit{semi-pool}}(\omega)$

and $s_R(x_A, e)$, the agency invests high effort if $(p_1 + p_2)(V_\varepsilon(0) - V_\varepsilon(1)) \geq \kappa$.

Proof of Proposition 4. Given $\pi > V_\varepsilon(0) > V_\varepsilon(1)$ and $\frac{p_1}{p_1+p_2} \leq \frac{1}{4}(4 - 4\beta - V_\varepsilon(0)$ $x_A^{\text{semi-pool}}(\omega)$ and $s_R(x_A, e)$ are as described in Proposition 3 for all e. This implies that the agency will be upheld following $x_A = 2$ and overturned following $x_A \in \{0, 1\}$. Thus, we need only compare the agency's utility from investing high versus low effort given the probability distribution over potential states of the world, $p = \{p_0, p_1, p_2\}$.

Consider the agency's expected utility for $e = 1$ and $e = 0$, respectively, in this environment,

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

The agency will invest high effort, rather than low effort, if and only if,

$$\begin{aligned} -p_0(\kappa + \pi) - p_1(1 + V_\varepsilon(1) + \kappa) - p_2(V_\varepsilon(1) + \kappa) &\geq -p_0(\pi) - p_1(1 + V_\varepsilon(0)) - p_2(V_\varepsilon(0)), \\ -p_0\kappa - p_0\pi - p_1 - p1V_\varepsilon(1) - p_1\kappa - p_2V_\varepsilon(1) - p_2\kappa &\geq -p_0\pi - p_1 - p_1V_\varepsilon(0) - p_2V_\varepsilon(0), \\ -\kappa - p_0\pi - p_1 - p_1V_\varepsilon(1) - p_2V_\varepsilon(1) &\geq -p_0\pi - p_1 - p_1V_\varepsilon(0) - p_2V_\varepsilon(0), \\ -\kappa - p_1V_\varepsilon(1) - p_2V_\varepsilon(1) &\geq -p_1V_\varepsilon(0) - p_2V_\varepsilon(0), \\ (p_1 + p_2)(V_\varepsilon(0) - V_\varepsilon(1)) &\geq \kappa, \end{aligned}$$

as stated in the result. ■

Moderately punitive reversal cost. In this case $V_\varepsilon(0) > \pi > V_\varepsilon(1)$ and the equilibrium in Proposition 3 again holds so long as $\frac{p_1}{p_1+p_2} \leq \frac{1}{4}(4 - 4\beta - V_\varepsilon(e))$ for both $e = 0$ and $e = 1$. However, in this case when $\omega = 1$ and the agency has invested low effort it would rather truthfully reveal $\omega = 1$ and be overturned. Thus, the agency chooses between investing high effort and obfuscating by playing $x_A^{\text{semi-pool}}(\omega)$, which leads to being upheld, and investing low effort and being truthful by playing $x_A^{\text{truth}}(\omega)$, which leads to being reversed following observation of $x_A = 1$ (as stipulated in $s_R(x_A, e)$). This leads to the following result with respect to agency effort investment.

Proposition 5. *Suppose* $V_\varepsilon(0) > \pi > V_\varepsilon(1)$ *and* $\frac{p_1}{p_1+p_2} \leq \frac{1}{4}(4-4\beta-V_\varepsilon(0))$. *Then, given* $x_A^{semi\text{-}pool}(\omega)$ *and* $s_R(x_A,e)$, *the agency invests high effort if* $p_1(\pi - V_\varepsilon(1)) + p_2(V_\varepsilon(0) - V_\varepsilon(1)) \geq \kappa$.

Proof of Proposition 5. In this environment the agency chooses between investing low effort and setting policy truthfully and investing high effort and playing the strategy $x_A^{\text{semi-pool}}(\omega)$. The agency's corresponding expected payoffs for $e = 1$ and $e = 0$, respectively, are given by,

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

Incentive compatibility requires that the following inequality hold for the agency to invest high effort in this case:

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

as was to be shown. ■