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

Ian R. Turner†

August 2019

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.

Word count: 8487

*I would like to especially thank John Patty, Maggie Penn, Brian Rogers, Justin Fox, Randy Calvert, Keith Schnakenberg, Andrea Aldrich, Deborah Beim, Dan Carpenter, Jack Paine, Tiberiu Dragu, Sean Gailmard, Mike Ting, and Ryan Hübert for helpful feedback throughout the evolution on this project. I would also like to thank seminar audiences at the Political Economy and Public Law Conference at University of Southern California, University of Wisconsin, University of Rochester, and Harvard University for providing exceptionally insightful comments. Early work on this project was supported by NSF Grant DGE-1143954. Of course, all errors are solely my own.

†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 it is commonly referred to, is present any time the agent's preferences diverge from those of a political principal.[3]

One ubiquitous political-institutional solution 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 deter the agency from making policy choices that run directly counter to one's own policy preferences. However, 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.[4] This introduces another wrinkle to ex post oversight: The overseer must not only worry about divergent substantive policy choices based 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 (Turner 2017).

[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.

[3] Comprehensive overviews of political agency, from different angles, are provided by Bendor and Meirowitz (2004), Bendor, Glazer and Hammond (2001), Gailmard and Patty (2013*a*,*b*), and Miller (2005).

[4] 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.

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.[5] 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.[6] 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*.[7] The scope of review dictates what actions, and which *type* of review overseers are directed to engage. Two major types of oversight are *procedural review* and *substantive review*. The main question in this paper is: How does the type of review shape the incentives for 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 focus on the content of the policy. 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.[8]

For example, FEMA's provision of emergency housing for evacuees following Hurricanes Katrina and Rita was challenged in federal court on the basis that the system developed to evaluate assistance applications was not satisfactory and led to too many (avoidable) erroneous decisions.[9]

[5] 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 recently was only exercised to invalidate ergonomics standards during the Clinton Administration.

[6] See McCann, Shipan and Wang (2016) for a comprehensive description of legislative judicial review provisions.

[7] For several case studies, across policy areas, suggesting that Congress anticipates the role of judicial review in the policymaking process see Light (1991); Melnick (1983, 1994).

[8] Previous work argues courts have recently moved more toward procedural review of administrative actions (Stephenson 2006). Moreover, there is some evidence to suggest that when the scope of review has been explicitly outlined by statute that the Supreme Court attempts to honor that statutory mandate (Verkuil 2002).

[9] See *Association of Community Organizations For Reform Now (ACORN), et. al. v. Federal Emergency Management Agency (FEMA),* 463 F. Supp. 2d 26 (D.D.C. 2006).

No one involved questioned the actual standards to receive assistance, nor was FEMA's authority to make these decisions in question. However, the court ultimately ruled against FEMA because the agency had not developed sufficient capacity to effectively allocate assistance without making costly errors, in this case leaving many citizens homeless. For the purposes of my argument, the key feature is that oversight was not focused on the content of the policy itself, but rather on how well the agency would be able to implement the policy on the ground.[10]

Substantive review entails an overseer also 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.

Questions of permissible search and seizure fit reasonably well within substantive review as it is conceptualized here. Both implementation capacity and the substantive content of these sorts of policies are salient to oversight. The content of policy might be thought of as what the 'correct' or permissible standard is to establish probable cause. A move to either more lax or more stringent standards may signal that policy needs to be recalibrated to adapt to environmental conditions.[11] Yet, even holding fixed the substantive standards, the development of clear, effective procedures is of equal importance in terms of realized policy outcomes. For any permissible standard if there was not sufficient investment in the procedural framework adopted to train and guide street-level police officers who make on-the-ground decisions about when the standards are satisfied, we might reasonably expect that the policy will be applied in highly variable ways, leading to overall worse, and often impermissible, policy outcomes. Thus, even when the substance of policy seems permissible it may be that ineffective implementation leads an overseer to step in.

[10]Huber (2007) also discusses issues involving OSHA and workplace safety that hinge on implementation capacity.

[11]Of course, policy shifts must be constitutional as well. But, if there is any level of discretion in setting these standards then the general connections I draw here follow.

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. 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.[12] Through the analysis of two variants of a formal model of policymaking 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.[13] 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 because it does not have to worry about the overseer judging the substance of its choices. The cost of this, from the overseer's 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. Procedural review can provide positive incentives that lead 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 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

[12]For related, but distinct, arguments about potential weaknesses of judicial review see Melnick (1983), Shapiro and Levy (1995), and Wagner (2012).

[13]Following the effort investment aspect of the model, the theory of policymaking developed here complements the literature on policy development and valence, which spans political and institutional contexts (e.g., Callander 2011; Callander and Martin 2017; Hirsch and Shotts 2015, 2018; McCarty 2017). In a sense, this article provides an applied microfoundation for policy valence in a bureaucratic setting, similar to how Hitt, Volden and Wiseman (2017) endogenize policy valence in the context of legislative policymaking.

agency's own substantive policy choices. The agency only chooses to reveal its private information when reversal is not too punitive from its perspective. Otherwise, the agency will obfuscate with some of its substantive policy choices to avoid reversal by only partially revealing its private information. To do so, the agency foregoes following its own superior information and either appeases the overseer by choosing less ambitious policy than it believes is required or exaggerates the extremity of policy change that is called for given the 'facts on the ground.' This dynamic 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 high 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 by obfuscating because it is relatively less costly to do so, from a policy perspective, when outcomes will be implemented more precisely.

Accountability and oversight. Oversight comes in many forms. In terms of enforcing political accountability prevalent review mechanisms include elections,[14] presidential vetoes,[15] stakeholder 'fire alarms' or Congressional oversight,[16] and judicial review.[17] Much of the previous research demonstrates how oversight can lead to the provision of perverse incentives that induce policymaking pathologies like pandering when policymakers have career or reputational concerns.[18] 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 reputational considerations. Relatedly, scholars have also studied how institutions promoting transparency affect accountability. Many of these studies have highlighted how increasing the transparency of policy-

[14] Ashworth (2012) provides a comprehensive overview of research on electoral accountability.

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

[16] For example, Gailmard (2009), McCubbins and Schwartz (1984).

[17] For example, Bueno de Mesquita and Stephenson (2007), Clark (2016), Fox and Stephenson (2015), Fox and Vanberg (2014), Patty and Turner (2019), Shipan (2000), Turner (2019, 2017).

[18] For a comprehensive overview of these pathologies see Gersen and Stephenson (2014). Also see, for example, Canes-Wrone, Herron and Shotts (2001) and Majumdar and Mukand (2004).

making may harm accountability.[19] 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*. To that end, I build on related existing studies.

Turner (2017) shows that procedural oversight can both strengthen and weaken agency effort incentives even when there is no preference disagreement between the reviewer and agency.[20] 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 affected when review institutions vary. 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 to obfuscate, thereby ignoring (and obscuring from the overseer) its private information to avoid reversal.

This latter result is similar to another related study, Patty and Turner (2019). 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 sometimes "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 opens the door for the possibility that the overseer can benefit from less information in the review process (i.e., benefit from procedural rather than substantive review). Specifically, in

[19] For example, Fox (2007), Fox and Stephenson (2011), Fox and Van Weelden (2012, 2015), Patty and Turner (2019), and Prat (2005).

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

political environments in which the overseer would like to overturn moderate policy changes but would uphold the agency sticking with the status quo or radically shifting policy in response to extreme changes in the underlying policy environment the agency may obfuscate by exaggerating the need for extreme policy changes when moderate change would suffice. The agency does so when its concern for its own reputation is more powerful than its intrinsic policy concerns, which is more likely to be the case when the agency has already invested in implementation capacity. When this is the case the overseer would be better off if she could commit to not stepping in to shut down moderate policy change, a commitment that is facilitated by the institution of procedural review. Ultimately, these results provide insight into the trade-offs between effort and expertise as well as between the two different styles of oversight. In turn, these trade-offs provide implications for how oversight may, or may not, provide for bureaucratic accountability and how one might optimally design the scope of review to promote high quality policymaking.

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.

Sequence of play. The agency first invests high or low effort toward the quality of policy implementation,[21] denoted by $e \in \{0,1\}$ where $e = 0$ ($e = 1$) is low effort (high effort). High effort leads to a net effort cost, $\kappa > 0$. This captures 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, effort investment directly affects an implementation shock, denoted by $\varepsilon \in \mathbb{R}$. The shock is conditioned by the agency's effort choice and is distributed according to $F_\varepsilon(e)$ with

[21] This can be thought of as an investment in agency capacity that allows for higher quality policy implementation (Ting 2011; Turner 2017). More generally, this is conceptually similar to models of policy valence (e.g., Hirsch and Shotts 2015), and what Carpenter (2001) refers to as programmatic capacity.

mean zero and strictly positive variance, $V_\varepsilon(e) \in (0,1)$.[22] 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 less than when low effort is invested, $V_\varepsilon(1) < V_\varepsilon(0)$. This ensures that high effort investment produces more precise policy outcomes than low effort investments.

It is worth taking a moment to connect effort investments to agency policymaking procedures conceptually. Procedural review largely focuses on ensuring agencies are making decisions that respect fairness criteria, due process, and overall equal application of law. Doing so involves the agency expending effort in designing procedures that help to guide, for example, street-level bureaucrats to uniformly apply substantive policy standards or, more generally, investing in capacity through an expanded workforce, improved technology, or updated processes to aid in high quality application of policy. Examples include developing clear guidelines for interacting with the public, an expanded workforce to conduct inspections to ensure workplace safety (Huber 2007), or improving logistical capacity to accurately assess applications for assistance. In all of these cases the effectiveness of realized policy outcomes depends crucially on the agency's ability to implement policies in line with the values noted above. This ability, in turn, is often either improved or harmed based on the level of effort (or, more generally, productive investment) the agency allocates toward these goals. Targeting these issues in the oversight process most often involves assessing the procedures and processes of enforcement developed by the agency and whether they are sufficient to ensure that errors will be minimized in the application of policy, which depends on the agency's investment in these processes.[23] In this way, procedural choices affect the *realized* substantive impact of policy, the quality of which is impacted by the effort exerted, even while holding the substantive content of policy fixed. The variance described above captures this dynamic formally.

Second, following the agency's effort investment, Nature reveals a true *state of the world*, $\omega \in \Omega = \{0,1,2\}$, to the agency. The agency learns about the policy environment by learning ω.

[22] Bounding the variances above by one does not affect the results. It simply streamlines the analysis by restricting implementation errors from shifting outcomes all the way to another substantive policy choice.

[23] A salient recent example involves recent state-level voter identification laws. Many of the court-mandated injunctions induced by adoption of these laws centered primarily on the determination that states had not adequately demonstrated that they would be able to enforce the laws fairly (or efficiently) given the procedures they had designed to do so (e.g., *Applewhite, et. al. v. Commonwealth of Pennsylvania, et. al.*, 330 M.D. 2012).

The ex ante probability that the true state is ω is p_ω. The three different states represent whether the relevant policy environment calls for little to no 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.

Finally, 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 $x = 0$. Allowing the agency to pursue new policy may bring things more in line with the facts of the policy environment through adjusting x_A, but new agency actions also come with some implementation uncertainty from the realization of ε. Reversing 'shuts down' the agency's new proposed intervention and the state of the policy environment goes back to that of the known status quo situation. Once the overseer's review decision is made ε is realized according to $F_\varepsilon(e)$ and final policy is generated.

Information and oversight. 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.[24] In the *substantive review model* the overseer observes both the agency's effort investment and substantive policy choice. This choice is then represented by $r(x_A, e) \in \{0,1\}$, where zero and

[24]Of course, in reality when review occurs the actual policy, not just procedural language, is publicly available. One should not take the model to imply that when overseers are directed to ignore substance that they literally cannot observe either that the agency took action or the action itself. Rather, this information structure captures realistic environments in which the content of policy is very clearly within the purview of the agency and therefore not under question, the overseer is a generalist (e.g., courts) assessing highly technical actions take by bureaucratic agencies and therefore unable to adequately judge content, or simply environments in which overseers take seriously the scope of review they are asked to adhere to and therefore do not render judgments based on content (Verkuil 2002). The comparison of institutions at the heart of this article does not depend on one interpretation of the information structure since this set-up captures any of the aforementioned variants of oversight.

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.

Preferences and equilibrium. The agency is motivated to match policy to the state and have high quality implementation, conditional on the costs of high effort, and have its 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 where π 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. 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 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.

The agency's effort and substantive policy strategies are 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 $b_R(x_A)$.[25] Perfect Bayesian equilibrium in weakly undominated strategies is the solution concept, which requires that the overseer hold correct beliefs updated via Bayes' rule on the path of play 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. This implies that in equilibrium the agency always matches substantive policy to the state: $x_A^P(\omega) = \omega$. 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 minimizing spatial policy losses by setting substantive policy to match the state. So, given that the agency is always able to target policy at matching the state, the question in the procedural review model is under what conditions the agency will invest high effort to improve the quality of policy implementation. The answer will depend crucially on the nature of procedural oversight, to which I now turn.

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 $x_A^P(\omega)$. If the overseer chooses to overturn the agency then final policy is set at $x = 0$. Thus, the overseer's subjective expected payoff for overturning the agency is given by,

$$-p_0\left(\beta^2\right) - p_1\left((1-\beta)^2\right) - p_2\left((2-\beta)^2\right).$$

[25] These beliefs are only applicable in the substantive review model since the overseer never has an opportunity to update her beliefs regarding ω in the procedural review model.

Since there is no policy change 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,

$$-\beta^2 - V_\varepsilon(e).$$

The overseer knows that the agency will match substantive policy to the state. That means in terms of 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. The overseer also loses utility based on the implementation imprecision associated with agency-made policy, $V_\varepsilon(e)$. She loses less utility when the agency invested high effort due to lower expected implementation errors (i.e., $V_\varepsilon(1) < V_\varepsilon(0)$). Combining and rearranging these subjective expected payoffs yields the following optimal review strategy:

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

To uphold the overseer requires the agency to invest sufficient effort to limit the volatility of agency-made policy. 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 overturning.

The condition to uphold the agency is more likely to be satisfied when the agency has invested high effort. Thus, there are two thresholds for upholding the agency based on the overseer's bias. 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: $\beta \in \left[0, \frac{p_1+4p_2-V_\varepsilon(e)}{2p_1+4p_2}\right)$. The upper bound of overseer bias in which she will still uphold the agency is a function of agency effort (as it feeds into $V_\varepsilon(e)$). Let β_0 be the upper bound when the agency has invested low effort and β_1 represent the upper bound when the agency has invested high effort. Since implementation variance is lower when the agency invests high effort $\beta_1 > \beta_0$, implying that oversight is more stringent when

the agency has invested low effort.[26] That is, the agency will be upheld at 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. When $\beta < \beta_0 < \beta_1$ the overseer will always uphold the agency regardless of effort investment; she is *perfectly deferential*. On the other extreme, when $\beta > \beta_1 > \beta_0$ the overseer will always overturn the agency regardless of effort; she is *perfectly skeptical*. Finally, when $\beta_0 < \beta < \beta_1$ the overseer upholds the agency if and only if the agency invests high effort and the review regime is *conditionally-deferential*. Agency effort decisions depend crucially on these review regimes.

Perfectly deferential review. When the agency faces a perfectly deferential overseer it will be upheld whether or not it invests high effort. Thus, the only consideration from the agency's perspective is how much investing high effort will improve implementation relative to low effort, and whether that improvement is worth the cost of doing so:

$$\underbrace{V_\varepsilon(0) - V_\varepsilon(1)}_{\text{precision improvement}} \geq \underbrace{\kappa.}_{\text{effort cost}}$$

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 bear the cost of that investment.

Perfectly skeptical review. If the agency is facing a perfectly skeptical overseer it never invests 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 equal to the costs of that effort. Since the agency is overturned with certainty whether or not it invests high effort, the status quo will remain in place either way. Thus, the agency is better off avoiding high effort costs and investing low effort instead.

Conditional-deference review. Finally, the most interesting case is when the overseer is conditionally-deferential. In this case the agency decides between investing low effort and avoiding the effort costs at the expense of the reversal cost, and investing high effort, which leads to being upheld and avoiding the reversal cost but comes at the expense of high effort costs. Since the agency does not yet

[26] The upper bound on overseer bias when the agency invests low effort is $\beta_0 \equiv \frac{p_1+4p_2-V_\varepsilon(0)}{2p_1+4p_2}$ and the upper bound when the agency invests high effort is $\beta_1 \equiv \frac{p_1+4p_2-V_\varepsilon(1)}{2p_1+4p_2}$. The fact that $V_\varepsilon(1) < V_\varepsilon(0)$ implies $\beta_0 < \beta_1$.

know ω when it chooses effort it 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. With this in mind, the agency's subjective expected payoff for investing low effort is given by,

$$-p_1 - 4p_2 - \pi.$$

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 $x = 0$ 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:

$$-V_\varepsilon(1) - \kappa.$$

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 (i.e., higher π) the more likely it is that this expression will be satisfied. Similarly, the more precise high effort policy is (i.e., lower $V_\varepsilon(1)$) 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 and represented graphically in Figure 1.

Proposition 1. *In the equilibrium of the procedural review model the overseer makes review deci-*

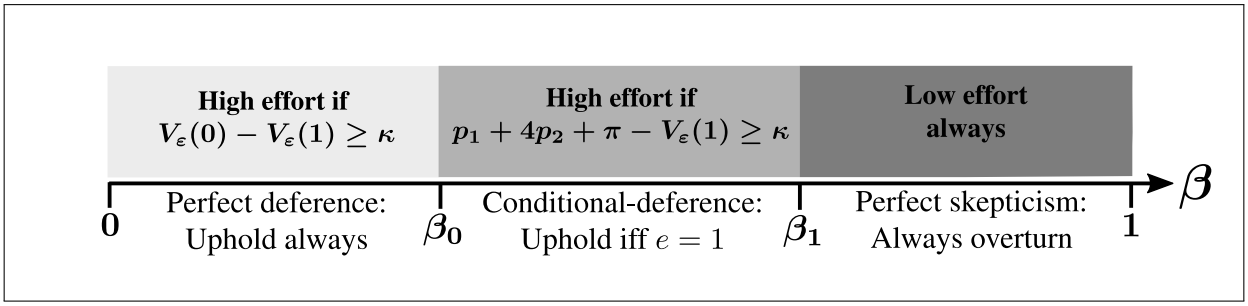

Figure 1: Equilibrium review regimes and agency effort investments

sions according to $s_R(e)$ (equation 1), the agency always sets substantive policy to match the state and invests effort, conditional on review regime, as follows:

- *When facing a perfectly deferential overseer (i.e., $\beta_0 < \beta_1 < \beta$) the agency invests high effort when $V_\varepsilon(0) - V_\varepsilon(1) \geq \kappa$.*
- *When facing a perfectly skeptical overseer (i.e., $\beta_0 < \beta_1 < \beta$) the agency never invests high effort.*
- *When facing a conditional-deference overseer (i.e., $\beta_0 < \beta < \beta_1$) the agency invests high effort if $p_1 + 4p_2 + \pi - V_\varepsilon(1) \geq \kappa$.*

Focusing on the conditional-deference case highlights a key trade-off in the procedural review model. The presence of procedural review itself can positively or negatively affect agency effort incentives. Compared to a world in which there is no oversight of agency policymaking, if $p_1 + 4p_2 + \pi - V_\varepsilon(1) > \kappa > V_\varepsilon(0) - V_\varepsilon(1)$ 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 absent oversight. In contrast, if $V_\varepsilon(0) - V_\varepsilon(1) > \kappa > p_1 + 4p_2 + \pi - V_\varepsilon(1)$ 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 investing in high quality implementation.[27] Thus, procedural review allows the agency to utilize its

[27]This is similar to results in Bueno de Mesquita and Stephenson (2007) and Turner (2017) 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" in Fox and Stephenson (2011).

technical expertise freely, which may be normatively desirable given the oft-cited rationale for delegation. However, it may come at the cost of both substantive policy disagreement, based on the level of preference divergence, and perverse effort incentives when the presence of oversight induces the agency to invest low effort.

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.' 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 Figure 2.[28]

Figure 2 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 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, i.e., $-V_\varepsilon(e)$. 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

[28]Proposition 2 below states the key result for truthful policymaking. In the Online Supporting Information, Lemma 4 (pg. 7) derives the full review strategy for the overseer when the agency is truthful for each value of e in each state ω and Lemma 5 (pg. 11) characterizes the agency's effort choices for each possibility.

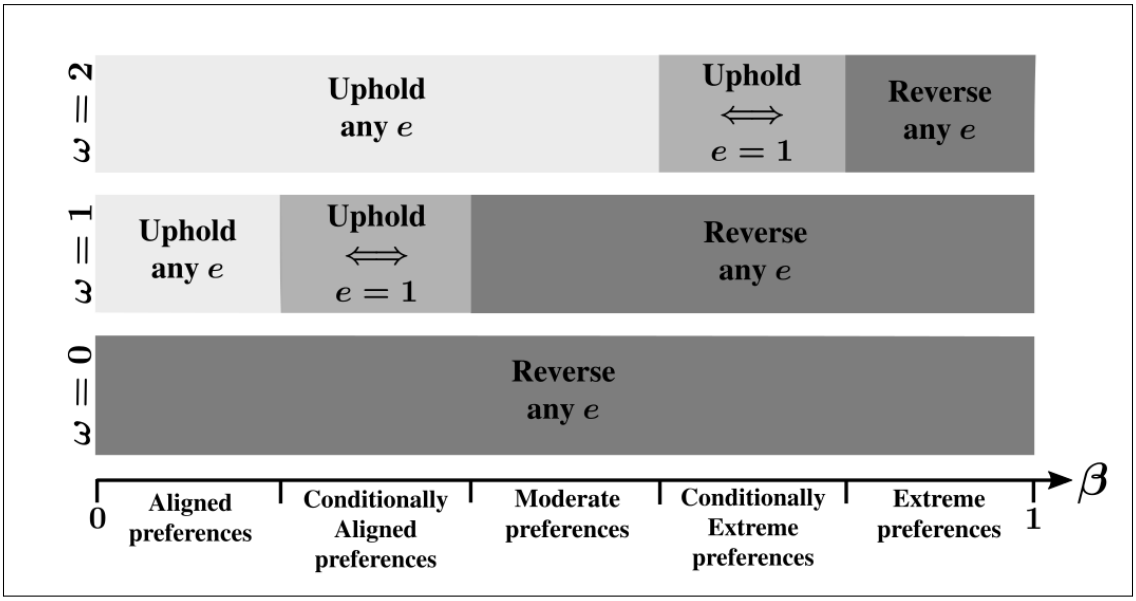

Figure 2: Overseer decisions given truthful policymaking conditional on state, effort, and bias.

Note: 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 $\beta > \frac{4-V_\varepsilon(1)}{4}$.

actions.[29] 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 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

[29]This result depends on the fact that outcomes following reversal have no additional uncertainty. This assumption is relaxed in the Online Supporting Information (Appendix B, beginning on pg. 19), which greatly complicates the analysis and expands the set of possibilities but the main substantive insights presented in the main text continue to hold.

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, thereby reducing the agency's relative expertise advantage.

The agency, in response, will not always set policy truthfully since it is not only driven by matching policy to the state, but also by avoiding reversal and 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 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 if and only if reversal costs are not too punitive (i.e., $V_\varepsilon(e) > \pi$). Further, $p_2(V_\varepsilon(0) - V_\varepsilon(1)) \geq \kappa$ is sufficient to ensure the agency invests high effort for all ranges of overseer bias except when the agency will always be overturned, in which case the agency never invests 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. If instead reversal costs are sufficiently punitive ($\pi > V_\varepsilon(e)$) then the agency may instead choose to obfuscate by choosing a policy that does not match the state to avoid reversal.

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 extreme policy change is called for ($\omega = 2$). Since 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. 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. Of course, if the agency will always be overturned the agency never invests high effort since that would simply lead to a net loss proportional to the cost of that effort.

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 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. That is, 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*. 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 is 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's equilibrium strategy, $x_A^{semi\text{-}pool}(\omega)$*, sets substantive policy such that* $x_A = 0$ *when* $\omega = 0$ *and* $x_A = 2$ *for both* $\omega \in \{1,2\}$ *and the overseer upholds* $x_A = 2$ *and overturns* $x_A \in \{0,1\}$.

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 $x_A = 2$.[30] The left-hand side of equation 2 denotes the overseer's (posterior) belief that $\omega = 1$ given $x_A^{\text{semi-pool}}(\omega)$. The condition requires that the probability that moderate change is called for (p_1) is low enough relative to the probability that extreme change is called for (p_2) 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 given that the state could be either $\omega = 1$ or $\omega = 2$. So long as inequality 2 holds then the overseer optimally overturns

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

the agency following maintenance of the status quo or moderate policy change ($x_A = 0$ or $x_A = 1$) and upholds the agency any time she observes extreme policy change ($x_A = 2$).

This highlights a key problem with allowing the overseer more information during review. Once the substance of policy is judged the agency may have incentive to exaggerate the need for policy change by pursuing extreme policy change when its private information suggests moderate change would suffice. This runs counter to many previous theories of bureaucratic oversight in which review of agency policy choices leads the agency to moderate its choices to appease the overseer (e.g., Epstein and O'Halloran 1999; Shipan 1997; Wiseman 2009). Rather than appease the overseer by shading policy toward the status quo, the agency exaggerates the need for policy change because it signals to the overseer that she runs the risk of large policy losses if she 'shuts the agency down.' Exaggerating in this way both suggests that overturning will lead to large policy loss through policy-state mismatch and increases the utility of allowing the agency to intervene in the environment, which increases the expected utility from upholding.

All of this is predicated on the idea that the agency *wants* to exaggerate with its policy choice when $\omega = 1$, which depends on how punitive reversal costs are relative to the impact of agency effort on implementation errors. 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 represents policymaking environments where agencies have 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 through exaggeration 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 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.

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.[31] 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 having to obfuscate when moderate policy change is called for. 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 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

[31] 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 $\omega \neq 0$.

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 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 exaggerate 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 reputational concerns are very weak. This implies, understanding low reversal costs as representative of highly insulated agencies, that independent agencies are more likely to engage in truthful policymaking than more politically accountable agencies that face stronger reputational considerations.

Incentives for the agency to match policy to the 'facts on the ground,' 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 *any* 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 character-

istics 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 appropriate). 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 overseer can be made better off with less information – procedural review – when the agency will exaggerate by pursuing extreme policy change when only moderate change is called for when the overseer also judges the content of agency policy – substantive review. More generally, 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,

$$W_R^P = -\underbrace{p_0\left(\beta^2 + V_\varepsilon(1)\right)}_{\text{payoff if } \omega = 0} - \underbrace{p_1\left(\beta^2 + V_\varepsilon(1)\right)}_{\text{payoff if } \omega = 1} - \underbrace{p_2\left(\beta^2 + V_\varepsilon(1)\right)}_{\text{payoff if } \omega = 2}.$$

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,

$$W_R^S = - \underbrace{p_0\left(\beta^2\right)}_{\text{payoff if } \omega=0} - \underbrace{p_1\left((\beta+1)^2+V_\varepsilon(1)\right)}_{\text{payoff if } \omega=1} - \underbrace{p_2\left(\beta^2+V_\varepsilon(1)\right)}_{\text{payoff if } \omega=2}.$$

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 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:

$$\begin{aligned} \Delta W_R(\text{Procedure vs. Substance}) &= W_R^P - W_R^S, \\ &= p_1(2\beta+1) - V_\varepsilon(1)(1-p_1-p_2). \end{aligned} \tag{3}$$

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 exaggerate 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 *not* altering policy is optimal 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). Interestingly, this suggests that the overseer is more likely to benefit 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 and/or when it is more likely that no policy change is called for.

Ultimately, when the policymaking environment is structured so 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 any policy change is low. This suggests that it is not clear that expanding the scope of review, by providing overseers with more information during the review process, yields net benefits once one takes into account how that information disclosure alters 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 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 under-utilize 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 exaggerate 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.

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.

Callander, Steven. 2011. "Searching For Good Policies." *American Political Science Review* 105(4):643–662.

Callander, Steven and Gregory J. Martin. 2017. "Dynamic Policymaking with Decay." *American Journal of Political Science* 61(1):50–67.

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.

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

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.

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.

Hirsch, Alexander V. and Kenneth W. Shotts. 2015. "Competitive Policy Development." *American Economic Review* 105(4):1646–1664.

Hirsch, Alexander V. and Kenneth W. Shotts. 2018. "Policy-Development Monopolies: Adverse Consequences and Institutional Responses." *Journal of Politics* 80(4):1339–1354.

Hitt, Matthew P., Craig Volden and Alan E. Wiseman. 2017. "Spatial Models of Legislative Effectiveness." *American Journal of Political Science* 61(3):575–590.

Huber, Gregory A. 2007. *The Craft of Bureaucratic Neutrality: Interests and Influence in Government Regulation of Occupational Safety*. New York, NY: Canbridge University Press.

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* .
URL: *http://goo.gl/EXuIGU*

McCarty, Nolan. 2017. "The Regulation and Self-Regulation of a Complex Industry." *Journal of Politics* 79(4):1220–1236.

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. 2019. "Ex Post Review and Expert Policymaking: When Does Oversight Reduce Accountability?" *Journal of Politics* .

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

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: Interest Groups, Congress, and Communications Policy*. 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.

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

Ting, Michael M. 2011. "Organizational Capacity." *Journal of Law, Economics, & Organization* 27(2):245–271.

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

Turner, Ian R. 2019. "Political Agency, Oversight, and Bias: The Instrumental Value of Politicized Policymaking." *Journal of Law, Economics, & Organization* .

Verkuil, Paul R. 2002. "An Outcomes Analysis of Scope of Review Standards." *William & Mary Law Review* 44(2):679–735.

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

Wiseman, Alan E. 2009. "Delegation and Positive-Sum Bureaucracies." *Journal of Politics* 71(3):998–1014.

Online Supporting Information

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

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

Contents

A In-text model results

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:* $x_A^P(\omega) = \omega$.

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,

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

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 $x_A^P(\omega) = \omega$ in the procedural review model. ■

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

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

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 overturning 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:

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

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

- When facing a perfectly deferential overseer (i.e., $\beta_0 < \beta_1 < \beta$) the agency invests high effort when $V_\varepsilon(0) - V_\varepsilon(1) \geq \kappa$.
- When facing a perfectly skeptical overseer (i.e., $\beta_0 < \beta_1 < \beta$) the agency never invests high effort.
- When facing a conditional-deference overseer (i.e., $\beta_0 < \beta < \beta_1$) the agency invests high effort if $p_1 + 4p_2 + \pi - V_\varepsilon(1) \geq \kappa$.

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,

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

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

With the overseer's review strategy in hand, I now turn to analysis of when the agency will truthfully set policy, and the accompanying effort investments in those cases. The next result char-

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 1: Overseer decisions given truthful policymaking ($x_A = \omega$) conditional on state ω, effort e, and bias β.

acterizes 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 if and only if reversal costs are not too punitive (i.e., $V_\varepsilon(e) > \pi$). Further, $p_2(V_\varepsilon(0) - V_\varepsilon(1)) \geq \kappa$ is sufficient to ensure the agency invests high effort for all ranges of overseer bias except when the agency will always be overturned, in which case the agency never invests 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.

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:

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

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,

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

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:

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

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:

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

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 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 the need for extreme policy change is sufficiently likely relative to moderate policy change,*

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

then there is a pure strategy semi-pooling obfuscation equilibrium in which the agency's equilibrium strategy, $x_A^{\textit{semi-pool}}(\omega)$, sets substantive policy such that $x_A = 0$ when $\omega = 0$ and $x_A = 2$ for both $\omega \in \{1,2\}$ and the overseer upholds $x_A = 2$ and overturns $x_A \in \{0,1\}$.

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 $b_R(1) = Pr[\omega = 1|x_A = 1] = 1$ 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,

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

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,

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

Combining these expected payoffs for upholding and overturning, and defining $q \equiv \frac{p_1}{p_1+p_2}$ 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 $x_A = 2$:

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

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 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,

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

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 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:

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

as was to be shown. ■

B Robustness of main insights

B.1 Alternative model with status quo variability

In this section I provide comprehensive analysis of an alternative model in which there is inherent variability associated with the status quo. Everything else is identical to the model in the main text. That is, I relax the assumption that when the overseer reverses the agency there is no implementation uncertainty analogous to $V_\varepsilon(e)$ when the agency is upheld. Specifically, if the overseer upholds the agency then final policy is given by $x = x_A + \varepsilon$ (as in the main text) and if the overseer overturns then final policy is $x = 0$ but there is still residual uncertainty captured by the strictly positive variance $V_{SQ} \in (0,1)$. This variance captures the fact that even maintenance of the status quo requires some level of action, which carries with it the potential for inefficiencies or errors. This leads to the following alternative utility functions:

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

illustrating that both players must internalize reversal variance V_{SQ} when the agency is overturned, $r = 1$.[1] With the alternative model in hand, I now turn to analyzing it below. I present the analysis in the same order as that for the main text model above.

[1] An alternative way to model this would be to include a second implementation shock, $\varepsilon_{SQ} \sim G(0, V_{SQ})$, that obtains only when the agency is reversed. This would ultimately lead to V_{SQ} falling out of derivations in the same way that $V_\varepsilon(e)$ shows up following the agency's being upheld. To incorporate V_{SQ} as simply as possible, I chose to add it directly to utility but the two approaches are substantively similar.

B.2 Procedural review

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

Proof of Lemma B.1. At the point in the game at which the agency makes its substantive policy choice, x_A, its effort investment e is a sunk cost. Thus, e and $V_\varepsilon(e)$ are fixed. Additionally, since x_A is not observed by the overseer the overseer's review decision is invariant to the agency's choice. Thus, there are two cases to check: (1) the agency will be upheld and (2) the agency will be overturned.

Agency upheld. The agency's expected payoff for the proposed strategy is given by,[2]

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

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

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

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

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

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

Agency overturned. The agency's payoff is equivalent in this case since the agency's choice of x_A will not change whether it is overturned and by this point in the game the only oversight-relevant choice, e, has been chosen. Thus, there is no incentive for the agency to deviate from setting policy so that $x_A(\omega) = \omega$. Taken together these two cases imply that, in weakly undominated strategies,

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

the agency will always choose $x_A(\omega) = \omega$ in the procedural review model. ■

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

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

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

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

Now, the overseer's expected payoff for reversing the agency depends on the state ω, which is unknown to the overseer in the procedural review model. The overseer's expected payoff for overturning given p_ω for each ω, is given by,

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

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

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

as stated in the lemma. ■

We can rearrange the condition to uphold in terms of overseer bias to yield an upper bound for upholding on β: $\beta \in \left(0, \frac{p_1 + 4p_2 + V_{SQ} - V_\varepsilon(e)}{2p_1 + 4p_2}\right]$. We can further define two β-thresholds based on whether the agency invested high or low effort: Let $\beta_1 := \frac{p_1 + 4p_2 + V_{SQ} - V_\varepsilon(1)}{2p_1 + 4p_2}$ and $\beta_0 := \frac{p_1 + 4p_2 + V_{SQ} - V_\varepsilon(0)}{2p_1 + 4p_2}$ where $\beta_0 < \beta_1$ since $V_\varepsilon(1) < V_\varepsilon(0)$. If $\beta < \beta_1 < \beta_0$ then the overseer always upholds and is *perfectly 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 section characterizes how the agency best responds with its effort choices conditional on these oversight regimes.

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

1. *If $\beta < \beta_1 < \beta_0$ then the overseer is perfectly deferential and the agency invests high effort if $V_\varepsilon(0) - V_\varepsilon(1) \geq \kappa$.*
2. *If $\beta_1 < \beta_0 < \beta$ then the overseer is perfectly skeptical and the agency never invests high effort.*
3. *If $\beta_1 < \beta < \beta_0$ then the overseer is conditionally deferential and the agency invests high effort if $V_{SQ} - V_\varepsilon(1) + p_1 + 4p_2 + \pi \geq \kappa$.*

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

Perfect deference. In this case the agency knows that it will be upheld regardless of its choice of e. The agency's expected payoff, given it will be upheld for sure, for investing low effort is given by,

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

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

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

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

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

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

Perfect skepticism. In this case the agency will be reversed by the overseer with certainty, regardless of its choice of e. The agency will never invest high effort in this case. Policy outcomes are the same regardless of the agency's effort choice ($x = 0$) so any high effort investment simply produces a net cost κ. Thus, it is never incentive compatible for the agency to invest high effort given that it will overturned by the overseer with certainty. This is case 2 in the result.

Conditional-deference. In this case the overseer upholds the agency if and only if the agency invests

high effort. The agency's expected payoff for investing high effort, which induces being upheld, is,

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

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

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

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

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

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

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

- *When facing a perfectly deferential overseer (i.e., $\beta_0 < \beta_1 < \beta$) the agency invests high effort when $V_\varepsilon(0)-V_\varepsilon(1) \geq \kappa$.*
- *When facing a perfectly skeptical overseer (i.e., $\beta_0 < \beta_1 < \beta$) the agency never invests high effort.*
- *When facing a conditional-deference overseer (i.e., $\beta_0 < \beta < \beta_1$) the agency invests high effort if $V_{SQ}-V_\varepsilon(1)+p_1+4p_2+\pi \geq \kappa$.*

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

B.3 Substantive review

B.3.1 Truthful equilibrium

Lemma B.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,*

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

Proof of Lemma B.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:

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

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

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

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

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

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

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

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

for a given e, is given by,

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

For the overseer to uphold incentive compatibility requires that,

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

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

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

This yields the following incentive compatibility constraint to uphold:

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

Combining the cases analyzed above yields the result. ■

The oversight rule derived above leads to five cases based on the level of effort the agency invests earlier in the game. The cases, along with the technical conditions on β, are displayed in Table 2, which corresponds to Figure 1 in the main body.

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

Table 2: Overseer best responses given truthful policymaking, conditional on ω, e, and 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. First, the following lemma is useful for the rest of the analysis.

Lemma B.5. *When $\omega = 0$ the agency always sets $x_A = 0$.*

Proof of Lemma B.5. First, note that there is no reason for the agency to deviate from $x_A = 0$ when $\omega = 0$ if it will be upheld by the overseer. Thus, we need only check whether the agency would benefit by deviating from $x_A = 0$ when it will be overturned. First, suppose that $r(0,e) = 1$ and $r(1,e) = 0$ so that deviating to $x_A = 1$ would lead to being upheld. The agency's expected utilities from $x_A = 0$ and $x_A = 1$ in this case are given by,

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

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

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

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

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

Proof of Proposition B.2. Lemma B.5 shows that the agency is always truthful when $\omega = 0$ so we derive the incentive compatibility conditions for the agency to truthfully reveal $\omega \in \{1,2\}$. First, consider the case in which $\omega = 1$. If the agency is upheld when it truthfully sets $x_A(1) = 1$ then there is no incentive to deviate. Similarly, if the agency is overturned when $x_A(1) = 1$ and also overturned whenever $x_A = 0$ and $x_A = 2$ then there is no reason to deviate. Thus, we need only check whether

the agency would deviate when $x_A(1) = 1$ is overturned but either $x_A = 0$ or $x_A = 2$ would be upheld. In either case the agency deviates spatially by one so expected utility is equivalent for deviating to $x_A = 0$ and $x_A = 2$ when $\omega = 1$ so we only show the derivations for deviating to $x_A = 0$ noting that the calculations are equivalent when $x_A = 2$ is the deviation. The agency's expected utilities for setting $x_A = 1$ truthfully and deviating by one to be upheld are given by,

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

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

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

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

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

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

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

Taken together, the agency never deviates from $x_A = 0$ when $\omega = 0$ even if it leads to being overturned and when $\omega \in \{1,2\}$ the agency will remain truthful (and separate) if and only if $V_\varepsilon(e) - V_{SQ} \geq \pi$, as stated in the result. The statement regarding the sufficient condition for high effort in the result follows from Lemma B.6. ■

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

Lemma B.6. *Assume $V_\varepsilon(e) - V_{SQ} \geq \pi$ for all e, which requires that $V_\varepsilon(0) > V_\varepsilon(1) > V_{SQ}$, so that the agency always sets policy truthfully. Conditional on $s_R(x_A, e)$ the agency makes effort investments as follows. The agency invests high effort when the overseer is aligned if $(p_1 + p_2)(V_\varepsilon(0) - V_\varepsilon(1)) \geq \kappa$, when the overseer is conditionally aligned if $p_1(1 + V_{SQ} - V_\varepsilon(1) + \pi) + p_2(V_\varepsilon(0) - V_\varepsilon(1)) \geq \kappa$, when the overseer is moderately biased if $p_2(V_\varepsilon(0) - V_\varepsilon(1)) \geq \kappa$, when the overseer is conditionally extreme if $p_2(4 + V_{SQ} - V_\varepsilon(1) + \pi) \geq \kappa$, and never invests high effort when the overseer is extremely biased.*

Proof of Lemma B.6. First, note that if $V_{SQ} > V_\varepsilon(0) > V_\varepsilon(1)$ then there cannot be a truthful separating equilibrium since for any e, $V_\varepsilon(e) - V_{SQ} < \pi$. Similarly, we set aside the case in which $V_\varepsilon(0) > V_{SQ} > V_\varepsilon(1)$ since at the moment we are interested in effort incentives assuming that the agency will subsequently set policy truthfully, which cannot hold for all ranges of overseer biases under this ordering following $e = 1$. Thus, the only case we need to characterize is when $V_{SQ} < V_\varepsilon(1) < V_\varepsilon(0)$. I will derive the condition for high effort in each of the agency-overseer preference environments assuming this ordering.

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

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

I will plug these general expected utilities into the relevant incentive compatibility conditions to analyze each case given $r(x_A, e)$ from Lemma B.4.

Consider an aligned overseer: $\beta \in \left[0, \frac{1 + V_{SQ} - V_\varepsilon(0)}{2}\right)$. In this case $r(0, e) = 1$ and $r(x, e) = 0$

for $x \in \{1,2\}$. The agency's expected utilities for investing high and low effort, respectively, are:

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

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

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

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

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

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

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

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

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

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

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

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

given by,

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

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

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

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

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

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

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

Proposition B.3. *Suppose preferences are aligned: $\beta < \frac{1+V_{SQ}-V_\varepsilon(0)}{2}$. Then the agency sets policy truthfully and the overseer upholds the agency following $x_A=0$ if $V_{SQ} \geq V_\varepsilon(e)$ and $x_A \in \{1,2\}$ for any e. Furthermore, when $V_{SQ} > V_\varepsilon(0) > V_\varepsilon(1)$ the agency will invest high effort if $V_\varepsilon(0) - V_\varepsilon(1) \geq \kappa$; when $V_\varepsilon(0) > V_{SQ} > V_\varepsilon(1)$ the agency invests high effort if $p_0(V_{SQ} - V_\varepsilon(1) + \pi) + (p_1 + p_2)(V_\varepsilon(0) - V_\varepsilon(1)) \geq \kappa$; and when $V_\varepsilon(0) > V_\varepsilon(1) > V_{SQ}$ the agency invests high effort if $(p_1 + p_2)(V_\varepsilon(0) - V_\varepsilon(1)) \geq \kappa$.*

Proof of Proposition B.3. From Lemma B.5 it follows that $x_A(0)=0$ regardless of whether the agency will be upheld or not. Moreover, we know from Lemma B.4 that the overseer will uphold a truthful choice of $x_A=0$ if and only if $V_{SQ} \geq V_\varepsilon(e)$ and any truthful policy change when $\beta < \frac{1+V_{SQ}-V_\varepsilon(0)}{2}$ even if $e=0$. Note that this holds for any ordering of agency and reversion variances: $V_{SQ} > V_\varepsilon(0) > V_\varepsilon(1)$, $V_\varepsilon(0) > V_{SQ} > V_\varepsilon(1)$, and $V_\varepsilon(0) > V_\varepsilon(1) > V_{SQ}$.

In terms of the sufficient condition for effort, consider first the case in which $V_{SQ} > V_\varepsilon(0) > V_\varepsilon(1)$ so that $r(0,e)=0$, $r(1,e)=0$, and $r(2,e)=0$ given truthful policymaking (Lemma B.4). The

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

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

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

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

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

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

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

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

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

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

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

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

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

B.3.2 Obfuscation equilibria

First, I establish two results that are useful in constructing obfuscation equilibria: (1) Following lemma B.5 the only possible pooling equilibrium involves the agency setting $x_A(\omega)=0$ for all ω, and (2) Regardless of the agency's policy strategy (e.g., pooling, semi-pooling, etc.) the overseer upholds any observation of $x_A=0$ if and only if $V_{SQ} \geq V_\varepsilon(e)$.

Corollary B.2. *If there is a pooling equilibrium then it involves the agency choosing $x_A(\omega)=0$ for all ω.*

Proof of Corollary B.2. This follows straightforwardly from Lemma B.5. If the agency will never deviate from $x_A(0)=0$, even when doing so would avoid reversal, then every agency policymaking strategy must involve $x_A(0)=0$ implying that if there is a pooling equilibrium then it involves $x(\omega)=0, \forall\omega$. ■

Lemma B.7. *For any agency policymaking strategy, the overseer upholds $x_A=0$, given e, if and only if $V_{SQ} \geq V_\varepsilon(e)$.*

Proof of Lemma B.7. Consider any (possibly mixed) agency policymaking strategy that involves setting $x_A=0$ with positive probability. (Note that Lemma B.5 ensures that so long as $p_0>0$ any agency strategy involves setting $x_A=0$ with positive probability.) Let $q_0 \coloneqq Pr(\omega=0|x_A=0)$, $q_1 \coloneqq Pr(\omega=1|x_A=0)$, and $q_2 \coloneqq Pr(\omega=2|x_A=0)$ represent the overseer's posterior beliefs over ω given observation of $x_A=0$. For instance, if the agency's equilibrium strategy is $x_A(\omega)=\omega$ then $q_0=1$ and $q_1=q_2=0$ or if $x_A(\omega)=\omega$ for $\omega \in \{0,1\}$ and $x_A(2)=2$ (in equilibrium) then $q_0=\frac{p_0}{p_0+p_1}$, $q_1=\frac{p_1}{p_0+p_1}$, and $q_2=0$. The overseer's expected utilities for upholding and overturning

following $x_A = 0$ are given by,

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

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

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

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

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

Proposition B.4. *Suppose preference disagreement is such that any agency choice to change policy, $x_A \in \{1,2\}$, is overturned given e. Then there is a pooling equilibrium in which the agency sets $x_A(\omega) = 0$ for all ω and the overseer upholds $x_A = 0$ if $V_{SQ} \geq V_\varepsilon(e)$.*

Proof of Proposition B.4. First, Lemma B.7 ensures that the overseer is best responding to the agency's pooling policymaking strategy by upholding only when $V_{SQ} \geq V_\varepsilon(e)$.

Lemma B.5 ensures that the agency always sets $x_A(0) = 0$. Now assume $x_A = 1$ leads to being overturned but $x_A = 0$ leads to being upheld, implying that $V_{SQ} \geq V_\varepsilon(e)$. The agency's expected utilities in that case are given by,

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

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

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

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

Now assume $V_{SQ} < V_\varepsilon(e)$ so that the overseer overturns $x_A = 0$. Since the overseer also overturns $x_A = 1$ and $x_A = 2$ due to being extremely biased there is no benefit to the agency for deviating from always setting $x_A = 0$. Finally, note that upholding $x_A = 0$ if and only $V_{SQ} \geq V_\varepsilon(e)$ given that the agency is pooling on $x_A = 0$ follows from Lemma B.7. ■

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

Proposition B.5. *Assume the overseer is moderately biased, $\beta \in \left[\frac{1+V_{SQ}-V_\varepsilon(e)}{2}, \frac{4+V_{SQ}-V_\varepsilon(e)}{4}\right]$ and $\pi > V_\varepsilon(e) - V_{SQ}$ so that a truthful separating equilibrium does not exist. If $V_{SQ} \geq V_\varepsilon(e)$ then there is a pure strategy semi-pooling equilibrium in which the agency sets $x_A(\omega) = 0$ for $\omega \in \{0,1\}$ and $x_A(2) = 2$, and the overseer upholds $x_A = 0$, overturns $x_A = 1$, and upholds $x_A = 2$.*

Proof of Proposition B.5. Suppose the agency sets $x_A(\omega) = 0$ for $\omega \in \{0,1\}$ and $x_A(2) = 2$. The derivation of overseer best responses to truthful policymaking in Lemma B.4 and the assumption that the overseer is moderately biased ensure that the overseer is best responding by upholding $x_A(2) = 2$. The equilibrium also follows from overseer off-path beliefs such that observation of $x_A = 1$ induces belief $b_R^*(\omega = 1 | x_A = 1) = Pr(\omega = 1 | x_A = 1) = 1$, which is consistent with PBE and leads the overseer to overturn $x_A = 1$ due to being moderately biased (per the overseer best responses derived in Lemma 4). The overseer's posterior beliefs following observation of $x_A = 0$ given the agency's strategy are given by,

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

By Lemma B.7 the overseer upholds $x_A = 0$ if and only if $V_{SQ} \geq V_\varepsilon(e)$, as stated in the result.

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

Proposition B.6. *Assume the overseer is moderately biased,* $\beta \in \left[\frac{1+V_{SQ}-V_\varepsilon(e)}{2}, \frac{4+V_{SQ}-V_\varepsilon(e)}{4}\right]$ *and* $\pi > V_\varepsilon(e) - V_{SQ}$ *so that a truthful separating equilibrium does not exist. If* $\omega = 2$ *is sufficiently likely relative to* $\omega = 1$*:* $\frac{p_1}{p_1+p_2} \leq \frac{1}{4}(4-4\beta+V_{SQ}-V_\varepsilon(e))$ *then there is a pure strategy semi-pooling equilibrium in which the agency sets* $x_A(\omega) = \omega$ *for* $\omega \in \{0,2\}$ *and* $x_A = 2$ *when* $\omega = 1$*, and the overseer upholds* $x_A = 0$ *if* $V_{SQ} \geq V_\varepsilon(e)$*, overturns* $x_A = 1$*, and upholds* $x_A = 2$*.*

Proof of Proposition B.6. Suppose that the agency sets policy according to the following strategy:

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

From Lemma B.7, the overseer upholds $x_A(0) = 0$ if and only if $V_{SQ} \geq V_\varepsilon(e)$ and, from Lemma B.4, does not uphold $x_A(1) = 1$ for any e given moderately biased preferences. In this equilibrium the overseer's off-path beliefs are such that observation of $x_A = 1$ induces the belief $b_R^*(\omega = 1|x_A = 1) = Pr(\omega = 1|x_A = 1) = 1$, which is consistent with PBE. Thus, we need only check whether $r(x_A = 2, e) = 0$ is incentive compatible. First, note that the overseer's equilibrium posterior beliefs about ω following $x_A = 2$ are given by,

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

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

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

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

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

as stated in the result.

The agency is best responding when $x_A = 0$ since Lemma B.5 shows that the agency never benefits by deviating from $x_A = 0$ when $\omega = 0$. Moreover, since $\pi > V_\varepsilon(e) - V_{SQ}$ truthful separating is not a best response when $\omega = 1$ if there is a deviation that will lead to being upheld. Thus, the deviation from $x_A(1) = 1$ to $x_A(1) = 2$ is optimal in this case so long as $r(2,e) = 0$ which requires that $\frac{p_1}{p_1+p_2} \leq \frac{1}{4}(4 - 4\beta + V_{SQ} - V_\varepsilon(e))$. Finally, the agency clearly has no reason to deviate from $x(2) = 2$ since the overseer is upholding $x_A = 2$ in equilibrium. ■

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

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

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

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

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

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

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

These expressions are used in the proofs below to derive overall incentive compatibility

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

Consider an environment in which $\beta \in \left[\frac{1+V_{SQ}-V_{\varepsilon}(1)}{2}, \frac{4+V_{SQ}-V_{\varepsilon}(0)}{4}\right]$ (moderate preference disagreement), the policy environment is moderately volatile, $V_{\varepsilon}(0) > V_{SQ} > V_{\varepsilon}(1)$, and reversal costs are moderately punitive, $V_{\varepsilon}(0) - V_{SQ} > \pi > V_{\varepsilon}(1) - V_{SQ}$. Using the general expression above for each possible state given that when $e = 1$ the agency is upheld for setting policy truthfully if $\omega = 0$ or $\omega = 2$ and will obfuscate by setting $x_A = 0$ or $x_A = 2$ (if the conditions for Proposition B.6 hold) when $\omega = 1$ while when $e = 0$ the agency is overturned when $\omega = 0$ because it truthfully sets $x_A = 0$ and preferences are moderate, does not deviate when $\omega = 1$ to be upheld and instead sets policy truthfully and accepts being overturned since reversal costs are only moderately punitive, and is upheld for truthfully setting $x_A = 2$ when $\omega = 2$. These cases dictate which expression from above applies to each possible state conditional on the agency's effort investment. This leads to the following expected utility expressions for $e = 1$ and $e = 0$:

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

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

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

That is, in this environment the agency will be reversed in states zero and one if it invests low effort, but it invests high effort it will be upheld for truthful policymaking in state zero and will obfuscate to be upheld in state one. In state two the agency will be upheld for truthful policymaking for both low and high effort. Thus, conditional on states zero or one obtaining (probability $p_0 + p_1$) the agency invests high effort if the implementation improvement from doing so ($V_{SQ} - V_{\varepsilon}(1)$) and the benefit of avoiding reversal (π) are high enough, and conditional on state two obtaining (probability p_2) the agency wants to invest high effort if the implementation improvement between high and low effort investment are high enough ($V_{\varepsilon}(0) - V_{\varepsilon}(1)$). As long as those collective potential benefits are higher

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

Lemma B.8. *Assume preferences are aligned: $\beta < \frac{1+V_{SQ}-V_\varepsilon(0)}{2}$. If reversal costs are highly punitive so that the agency always obfuscates to be upheld when possible, $\pi > V_\varepsilon(0) - V_{SQ} > V_\varepsilon(1) - V_{SQ}$, it invests high effort as follows:*

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

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

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

Proof of Lemma B.8. The result follows from plugging in the relevant general expression from above that correspond to the appropriate potential state and the subsequent equilibrium oversight and policymaking behavior.

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

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

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

in equilibrium,

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

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

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

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

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

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

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

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

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

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

Now let $V_\varepsilon(0)-V_{SQ} > \pi > V_\varepsilon(1)-V_{SQ}$. Note that $V_{SQ} > V_\varepsilon(0) > V_\varepsilon(1)$ can not be true in this setting. Because preference are aligned agency-overseer equilibrium behavior is the same: There is

no reason to obfuscate since the agency is upheld for setting policy truthfully when $\omega \in \{1,2\}$ and lemma B.5 implies the agency also never deviates when $\omega = 0$. Thus, the incentive compatibility conditions for high effort are equivalent to those derived for the $\pi > V_\varepsilon(e) - V_{SQ}$ for all e cases above, as stated in the result. ■

Lemma B.9. *Assume preferences are conditionally aligned:* $\beta \in \left[\frac{1+V_{SQ}-V_\varepsilon(0)}{2}, \frac{1+V_{SQ}-V_\varepsilon(1)}{2}\right)$. *If reversal costs are highly punitive so that the agency will always obfuscate to be upheld when possible,* $\pi > V_\varepsilon(0) - V_{SQ} > V_\varepsilon(1) - V_{SQ}$, *it invests high effort as follows:*

- *When* $V_{SQ} > V_\varepsilon(0) > V_\varepsilon(1)$ *the agency invests high effort if* $p_1 + V_\varepsilon(0) - V_\varepsilon(1) \geq \kappa$.
- *When* $V_\varepsilon(0) > V_{SQ} > V_\varepsilon(1)$ *the agency invests high effort if* $p_0(V_{SQ} - V_\varepsilon(1) + \pi) + p_1 + (p_1 + p_2)(V_\varepsilon(0) - V_\varepsilon(1)) \geq \kappa$ *assuming proposition B.6 conditions hold so it can obfuscate with* $x_A(1) = 2$ *to be upheld and invests high effort if* $(p_0 + p_1)(V_{SQ} - V_\varepsilon(1) + \pi) + p_1 + p_2(V_\varepsilon(0) - V_\varepsilon(1)) \geq \kappa$ *when proposition B.6 conditions do not hold.*
- *When* $V_\varepsilon(0) > V_\varepsilon(1) > V_{SQ}$ *the agency invests high effort if* $p_1 + (p_1 + p_2)(V_\varepsilon(0) - V_\varepsilon(1)) \geq \kappa$ *assuming proposition B.6 conditions hold so it can obfuscate with* $x_A(1) = 2$ *to be upheld and invests high effort if* $p_1(1 + V_{SQ} - V_\varepsilon(1) + \pi) + p_2(V_\varepsilon(0) - V_\varepsilon(1)) \geq \kappa$ *when proposition B.6 conditions do not hold.*

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

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

Proof of Lemma B.9. To derive the results I plug in the relevant general expressions from above that correspond to the appropriate potential state and the subsequent equilibrium oversight and policy-making behavior.

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

effort in this setting:

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

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

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

When $V_\varepsilon(0) > V_{SQ} > V_\varepsilon(1)$ and the agency invests high effort it can again set policy truthfully and be upheld in all states, while if $e=0$ it will be overturned for truthfully setting policy when $\omega = 0$, obfuscate to $x_A = 2$ when $\omega = 1$ if the conditions in proposition B.6 are satisfied, truthfully set policy and be reversed if $\omega = 1$ and the conditions in proposition B.6 are not satisfied, and set policy truthfully and be upheld when $\omega = 2$. This yields the following expected utility expressions:

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

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

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

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

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

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

$\omega = 2$. If it invests low effort then it sets policy truthfully when $\omega = 0$ and is overturned, obfuscates by setting $x_A = 2$ when $\omega = 1$ and is upheld when the proposition B.6 conditions are satisfied and sets policy truthfully when $\omega = 1$ and is reversed when those conditions are not satisfied, and sets policy truthfully and is upheld when $\omega = 2$. Together this yields the following expected utility expressions:

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

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

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

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

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

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

Now let $V_\varepsilon(0) - V_{SQ} > \pi > V_\varepsilon(1) - V_{SQ}$. The incentive compatibility conditions are the same in this case as those above when the agency cannot obfuscate. The difference here in equilibrium is that the agency would never obfuscate, so it's not a matter of whether the relevant conditions are satisfied (as in proposition B.6). This yields all the conditions as stated in the result. ■

Lemma B.10. *Assume preferences are moderately divergent:* $\beta \in \left[\frac{1+V_{SQ}-V_\varepsilon(1)}{2}, \frac{4+V_{SQ}-V_\varepsilon(0)}{4}\right]$. *If reversal costs are highly punitive so that the agency will always obfuscate to be upheld when possible,* $\pi > V_\varepsilon(0) - V_{SQ} > V_\varepsilon(1) - V_{SQ}$, *it invests high effort as follows:*

- *When* $V_{SQ} > V_\varepsilon(0) > V_\varepsilon(1)$ *the agency invests high effort if* $V_\varepsilon(0) - V_\varepsilon(1) \geq \kappa$.
- *When* $V_\varepsilon(0) > V_{SQ} > V_\varepsilon(1)$ *the agency invests high effort if* $p_0(V_{SQ} - V_\varepsilon(1) + \pi) + (p_1 + p_2)(V_\varepsilon(0) - V_\varepsilon(1)) \geq \kappa$ *assuming proposition B.6 conditions hold so it can obfuscate with* $x_A(1) = 2$ *to be upheld and invests high effort if* $(p_0 + p_1)(V_{SQ} - V_\varepsilon(1) + \pi) + p_2(V_\varepsilon(0) - V_\varepsilon(1)) \geq \kappa$ *when proposition B.6 conditions do not hold.*

- *When $V_\varepsilon(0) > V_\varepsilon(1) > V_{SQ}$ the agency invests high effort if $(p_1+p_2)(V_\varepsilon(0)-V_\varepsilon(1)) \geq \kappa$ assuming proposition B.6 conditions hold so it can obfuscate with $x_A(1)=2$ to be upheld and invests high effort if $p_2(V_\varepsilon(0)-V_\varepsilon(1)) \geq \kappa$ when proposition B.6 conditions do not hold.*

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

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

- *When $V_\varepsilon(0) > V_\varepsilon(1) > V_{SQ}$ the agency invests high effort if $p_1(V_{SQ}-V_\varepsilon(1)+\pi)+p_2(V_\varepsilon(0)-V_\varepsilon(1)) \geq \kappa$ assuming proposition B.6 conditions hold so it can obfuscate with $x_A(1)=2$ to be upheld and invests high effort if $p_2(V_\varepsilon(0)-V_\varepsilon(1)) \geq \kappa$ when proposition B.6 conditions do not hold.*

Proof of Lemma B.10. I derive the conditions for each case stated in the result.

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

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

Thus, the agency will invest high effort when,

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

When $V_\varepsilon(0) > V_{SQ} > V_\varepsilon(1)$ and the agency invests high effort it can set policy truthfully and be upheld if $\omega \in \{0,2\}$ and will obfuscate to either $x_A=0$ or $x_A=2$ (if possible) when $\omega=1$ and be upheld. If $e=0$ it will be overturned for truthfully setting policy when $\omega=0$, obfuscate to $x_A=2$ when $\omega=1$ if the conditions in proposition B.6 are satisfied, truthfully set policy and be reversed if $\omega=1$ and the conditions in proposition B.6 are not satisfied, and set policy truthfully and be upheld

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

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

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

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

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

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

When $V_\varepsilon(0) > V_\varepsilon(1) > V_{SQ}$ the agency sets policy truthfully and is overturned when $\omega = 0$ since $V_{SQ} < V_\varepsilon(1)$, obfuscates to $x_A = 2$ when $\omega = 1$ to be upheld when proposition B.6 conditions are satisfied and truthfully sets policy and is overturned when they are not, and truthfully sets policy and is upheld when $\omega = 2$ for both effort levels. If it invests low effort then policymaking choices and review choices are the same as when $e = 1$. Together this yields the following expected utility expressions:

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

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

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

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

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

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

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

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

which yields the following incentive compatibility condition for high effort,

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

This derives all the results in the lemma. ■

Lemma B.11. *Assume preferences are conditionally extreme: $\beta \in \left[\frac{4+V_{SQ}-V_\varepsilon(0)}{4}, \frac{4+V_{SQ}-V_\varepsilon(1)}{4}\right]$. If reversal costs are highly punitive so that the agency will always obfuscate to be upheld when possible, $\pi > V_\varepsilon(0) - V_{SQ} > V_\varepsilon(1) - V_{SQ}$, it invests high effort as follows:*

- *When $V_{SQ} > V_\varepsilon(0) > V_\varepsilon(1)$ the agency invests high effort if $4p_2 + V_\varepsilon(0) - V_\varepsilon(1) \geq \kappa$.*
- *When $V_\varepsilon(0) > V_{SQ} > V_\varepsilon(1)$ the agency invests high effort if $4p_2 + V_{SQ} - V_\varepsilon(1) + \pi \geq \kappa$.*
- *When $V_\varepsilon(0) > V_\varepsilon(1) > V_{SQ}$ the agency invests high effort if $4p_2 + (p_1+p_2)(V_{SQ} - V_\varepsilon(1) + \pi) \geq \kappa$ assuming proposition B.6 conditions hold so it can obfuscate with $x_A(1) = 2$ to be upheld and invests high effort if $p_2(4 + V_{SQ} - V_\varepsilon(1) + \pi) \geq \kappa$ when proposition B.6 conditions do not hold.*

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

- *When $V_\varepsilon(0) > V_{SQ} > V_\varepsilon(1)$ the agency invests high effort if $4p_2 + V_{SQ} - V_\varepsilon(1) + \pi \geq \kappa$.*
- *When $V_\varepsilon(0) > V_\varepsilon(1) > V_{SQ}$ the agency invests high effort if $4p_2 + (p_1+p_2)(V_{SQ} - V_\varepsilon(1) + \pi) \geq \kappa$ assuming proposition B.6 conditions hold so it can obfuscate with $x_A(1) = 2$ to be upheld and invests high effort if $p_2(4 + V_{SQ} - V_\varepsilon(1) + \pi) \geq \kappa$ when proposition B.6 conditions do not hold.*

Proof of Lemma B.11. First, let $\pi > V_\varepsilon(e) - V_{SQ}$ for all e. When $V_{SQ} > V_\varepsilon(0) > V_\varepsilon(1)$ the agency will match policy to the state and be upheld when $\omega \in \{0,2\}$ and obfuscate to $x_A = 0$ or $x_A = 2$ (if possible) and be upheld when $\omega = 1$. After $e = 0$ the agency will pool by setting policy at $x_A = 0$ for all ω and be upheld. This yields the following expected utilities for high and low effort,

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

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

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

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

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

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

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

When $V_\varepsilon(0) > V_\varepsilon(1) > V_{SQ}$ and the agency invests high effort then it sets policy truthfully when $\omega \in \{0,2\}$ and is overturned when $x_A = 0$ and upheld when $x_A = 2$, and it obfuscates when $\omega = 1$ by setting $x_A = 2$ if it can (proposition B.6 conditions hold) and otherwise sets policy truthfully and accepts reversal. If it invests low effort then it can never obfuscate and sets policy truthfully and is reversed. This yields the following expected utilities,

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

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

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

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

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

Lemma B.12. *Assume preferences are extreme: $\beta > \frac{4+V_{SQ}-V_\varepsilon(1)}{4}$. If reversal costs are highly punitive so that the agency will always obfuscate to be upheld when possible, $\pi > V_\varepsilon(0) - V_{SQ} > V_\varepsilon(1) - V_{SQ}$, it invests high effort as follows:*

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

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

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

Proof of Lemma B.12. Let $\pi > V_\varepsilon(e) - V_{SQ}$ for all e so that the agency always obfuscates when possible to avoid reversal. When $V_{SQ} > V_\varepsilon(0) > V_\varepsilon(1)$ the agency pools on $x_A = 0$ for all ω to avoid reversal when $\omega \in \{1,2\}$ for both effort levels (i.e., it obfuscates by setting $x_A = 0$ when $\omega \in \{1,2\}$ and is subsequently upheld since $V_{SQ} > V_\varepsilon(e)$ for all e). This yields the following expected utilities:

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

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

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

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

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

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

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

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

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

Next, I provide some parameterized examples to illustrate some of the results from above.

Example B.1. *(High effort to tell the truth)* Let $\beta \in \left[\frac{1+V_{SQ}-V_\varepsilon(0)}{2}, \frac{1+V_{SQ}-V_\varepsilon(1)}{2}\right)$ so that preferences are *conditionally aligned*. Further, fix the following parameter values: $\beta = 7/16$, $p_0 = 1/4$, $p_1 = 1/4$, $p_2 = 1/2$, $V_{SQ} = 1/4$, $V_\varepsilon(0) = 1/2$, $V_\varepsilon(1) = 1/8$, $\pi = 1/2$. These parameter values further imply that $\frac{p_1}{p_1+p_2} \leq \frac{1}{4}(4-4\beta+V_{SQ}-V_\varepsilon(0))$ holds so that the conditions for semi-pooling characterized in Proposition B.6 are satisfied, $V_\varepsilon(0) > V_{SQ} > V_\varepsilon(1)$ so that the policy environment is *moderately volatile*, and that $\pi > V_\varepsilon(0) - V_{SQ} > V_\varepsilon(1) - V_{SQ}$ so that reversal costs are *highly punitive*.

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

overturned for truthfully setting $x_A = 0$ when $\omega = 0$ but will nonetheless do so,[3] it will obfuscate by setting $x_A = 2$ when $\omega = 1$ to avoid reversal since $\pi > V_\varepsilon(0) - V_{SQ}$ (the semi-pooling equilibrium in Proposition B.6), and can set policy truthfully when $\omega = 2$ and be upheld since preferences are conditionally aligned. Thus, whether the agency invests high effort in equilibrium depends on whether the net benefits of being able to always set policy truthfully with high effort implementation are large enough to offset the costs to obtain those benefits.

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

$$EU_A(e=1) = -\underbrace{p_0 V_\varepsilon(1)}_{\substack{\text{payoff if } \omega=0,\, e=1 \\ \text{since } x_A^{\text{truth}}(0) \text{ upheld}}} - \underbrace{p_1 V_\varepsilon(1)}_{\substack{\text{payoff if } \omega=1,\, e=1 \\ \text{since } x_A^{\text{truth}}(1) \text{ upheld}}} - \underbrace{p_2 V_\varepsilon(1)}_{\substack{\text{payoff if } \omega=2,\, e=1 \\ \text{since } x_A^{\text{truth}}(2) \text{ upheld}}} - \underbrace{\kappa,}_{\text{effort cost}}$$

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

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

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

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

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

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

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

[3] This follows from Lemma B.5.

This environment is one in which when the agency invests high effort it is upheld for truthfully setting policy when either $\omega = 0$ or $\omega = 2$, and it obfuscates by setting either $x_A = 0$ or $x_A = 2$ (when the conditions for Proposition B.6 are satisfied) when $\omega = 1$ to avoid reversal. When the agency invests low effort it accepts being overturned for setting policy truthfully when $\omega \in \{0,1\}$ since $\pi < V_\varepsilon(0) - V_{SQ}$ implies that it is never incentive compatible to deviate from truthful policymaking (from Proposition B.2) and is upheld for truthfully setting policy when $x_A = 2$. Thus, whether the agency invests high effort involves whether the net benefits from doing so – being upheld when $\omega = 0$, obfuscating to be upheld when $\omega = 1$, and being upheld when $\omega = 2$ with lower implementation variance – outweigh the costs of those benefits (κ).

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

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

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

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

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

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

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

B.4 Comparing Review Institutions

This section provides results from the alternative model that compare procedural and substantive review in different environments.

Proposition B.7. *When preferences are sufficiently aligned so that the agency is always upheld following procedural review and under substantive review truthful policymaking leads the overseer to uphold $x_A = 0$ only if $V_{SQ} \geq V_\varepsilon(e)$ and uphold $x_A \in \{1,2\}$ for any e the overseer weakly prefers substantive review from an ex ante perspective.*

Proof of Proposition B.7. Consider the overseer's ex ante utility under procedural review when the agency is always upheld (for any e),

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

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

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

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

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

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

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

which can never be satisfied since $V_{SQ} < V_\varepsilon(e)$ is required to ensure $r(0,e) = 1$. Thus, in this case

the overseer benefits from substantive review due to increased control over the agency when $\omega = 0$. Overall, then, the overseer is either indifferent between procedural review and substantive review or strictly benefits from substantive review when preferences are sufficiently aligned. ■

Proposition B.8. *When preferences are so extreme that the agency is always overturned following procedural review and $x_A = 0$ is upheld only if $V_{SQ} \geq V_\varepsilon(e)$ and $x_A \in \{1,2\}$ are both overturned for all e following substantive review procedural review is weakly preferred by the overseer in terms of ex ante utility.*

Proof of Proposition B.8. Consider first the overseer's ex ante utility for procedural review when the agency is never upheld regardless of e:

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

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

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

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

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

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

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

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

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

Lemma B.13. *When the overseer is conditionally-deferential under procedural review and the environment is as defined in example B.2 under substantive review the threshold on κ to support high effort investment is higher under procedural review:*

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

Proof of Lemma B.13. The relevant inequality is derived from the incentive compatibility constraints for each relevant environment (lemma B.3 for procedural review and lemma B.10 for substantive review). The fact that the inequality in the result is always satisfied is straightforward given the assumptions of the model:

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

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

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

Proposition B.9. *Assume under procedural review the overseer is conditionally deferential so she upholds if and only if the agency invests high effort, and under substantive review the environment is as in example B.2 so that when the agency invests high effort it obfuscates by setting $x_A(1) = 2$ and the overseer upholds that choice. Further, assume that $p_1 + 4p_2 + V_{SQ} - V_\varepsilon(1) + \pi > (p_0 + p_1)(V_{SQ} - V_\varepsilon(1) + \pi) + p_2(V_\varepsilon(0) - V_\varepsilon(1)) \geq \kappa$ so that the agency invests high effort in equilibrium under both procedural and substantive review. Then procedural review is always preferred to substantive review.*

Proof of Proposition B.9. Under procedural review the agency invests high effort and is upheld. The agency matches policy to the state and implementation uncertainty is given by $V_\varepsilon(1)$ for all ω. Under substantive review the agency truthfully sets $x_A(1) = 1$ and is upheld, obfuscates when $\omega = 1$ and sets $x_A(1) = 2$ and is upheld, and truthfully sets $x_A(2) = 2$ and is upheld. In all cases, implementation uncertainty is given by $V_\varepsilon(1)$. This yields the following ex ante expected utility expressions for the

overseer conditional on scope of review:

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

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

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

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

One thing worth noting about proposition B.9 is that in the environment specified the agency could also obfuscate by appeasing the overseer ($x_A(1)=0$) rather than obfuscating through exaggeration. In that case, the overseer strictly prefers substantive review because it induces the agency to set $x_A=0$ when $\omega=1$, which benefits the overseer. To see this, note that the payoffs when $\omega=0$ and $\omega=2$ are exactly the same as in proposition B.9, but the payoff for $\omega=1$ under substantive review is now $-(1-\beta)^2-V_\varepsilon(1)$ instead of $-(\beta+1)^2-V_\varepsilon(1)$. The relevant comparison then becomes $-p_1\beta^2-V_\varepsilon(1)$ (under procedural review) and $-p_1(1-\beta)^2-V_\varepsilon(1)$ (under substantive review). Thus, if $-p_1(1-\beta)^2>-p_1\beta^2$ then substantive review is strictly preferred to procedural review, which requires that $\beta>\frac{1}{2}$. This inequality is always satisfied in this environment since $\beta>\frac{1+V_{SQ}-V_\varepsilon(1)}{2}$, the RHS of which is greater than one-half since $V_{SQ}>V_\varepsilon(1)$ (and $V_{SQ}>0$). Thus, if the agency were to obfuscate by appeasement rather than obfuscate by exaggeration then substantive review would be preferred to procedural review. Nonetheless, obfuscating through exaggeration exists in equilibrium in this environment, providing proof that procedural review can be preferred to substantive review in terms of ex ante overseer expected utility.

Proposition B.10. *Assume under procedural review the overseer is conditionally-deferential so she upholds if and only if the agency invests high effort, and under substantive review the environment is defined as in example B.2 so that when the agency invests low effort obfuscation is never incentive compatible. Further, assume that* $p_1+4p_2+V_{SQ}-V_\varepsilon(1)+\pi\geq\kappa>(p_0+p_1)(V_{SQ}-V_\varepsilon(1)+\pi)+p_2(V_\varepsilon(0)-V_\varepsilon(1))$ *so that the agency invests high effort when facing procedural review but low effort under substantive review. Then procedural review is preferred to substantive review when,*

$$p_0(V_{SQ}-V_\varepsilon(1))+p_1(1-2\beta+V_{SQ}-V_\varepsilon(1))+p_2(V_\varepsilon(0)-V_\varepsilon(1))>0. \tag{3}$$

Otherwise, substantive review is preferred to procedural review. Furthermore, equation 3 is more likely to be satisfied as p_1 and β decrease.

Proof of Proposition B.10. Under procedural review the agency matches policy to the state, invests high effort, and the overseer upholds, which yields $-\beta^2 - V_\varepsilon(1)$ as the expected payoff for each potential state. Under substantive review the agency truthfully sets policy when $\omega \in \{0,1\}$ since obfuscation is not incentive compatible, and the overseer overturns those choices. The agency also sets policy truthfully when $\omega = 2$, but the overseer upholds in this case. This yields the following ex ante expected utilities for the overseer conditional on scope of review:

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

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

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

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

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

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

Proposition B.11. *Assume under procedural review the overseer is conditionally deferential so she upholds if and only if the agency invests high effort, and under substantive review the environment is as in example B.2 so that when the agency invests high effort it obfuscates by setting $x_A(1) = 2$ and the overseer upholds that choice. Further, assume that $\kappa > p_1 + 4p_2 + V_{SQ} - V_\varepsilon(1) + \pi > (p_0 + p_1)(V_{SQ} - V_\varepsilon(1) + \pi) + p_2(V_\varepsilon(0) - V_\varepsilon(1))$ so that the agency invests low effort in equilibrium under both procedural and substantive review. Then substantive review is always preferred to procedural review.*

Proof of Proposition B.11. Under procedural review the agency matches policy to the state, invests low effort, and the overseer overturns. Under substantive review the agency truthfully sets policy when $\omega \in \{0,1\}$ since obfuscation is not incentive compatible, and the overseer overturns those choices. The agency also sets policy truthfully when $\omega = 2$, but the overseer upholds in this case. This yields the following ex ante expected utilities for the overseer conditional on scope of review:

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

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

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

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