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

Ian R. Turner[†]

Abstract

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

Keywords: Bureaucracy; Oversight; Policymaking; Formal Theory

[*] I would like to especially thank John Patty, Maggie Penn, Brian Rogers, Justin Fox, Randy Calvert, and Keith Schnakenberg for helpful discussions throughout the evolution on this project. I would also like to thank Deborah Beim, Andrea Aldrich, Dan Carpenter, Larry Rothenberg, Jack Paine, Scott Abramson, Scott Gehlbach, Mike Ting, and seminar audiences at the APSA, MPSA, PEPL, the University of Wisconsin, Harvard University, and the University of California, San Diego. Early work on this project was supported by NSF Grant DGE-1143954.

[†] 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*b*).

[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 (2017), Shipan (2000), Turner (2017*a*,*b*).

[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*b*) shows that procedural oversight can both strengthen and weaken agency effort incentives even when there is no preference disagreement between the reviewer and agency.[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 (2017). 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 (2017), 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.

The model

I analyze a two-player, non-cooperative policymaking game between a bureaucratic agency, A, and an overseer or reviewer, R. 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 mean zero and strictly positive variance, $V_\varepsilon(e) \in (0,1)$.[22] Mean zero implies that the shock is cen-

[21] This can be thought of as an investment in agency capacity that allows for higher quality policy implementation (Ting 2011; Turner 2017*b*). 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.

[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.

tered 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. That is, the agency learns about the policy environment by learning ω. 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

[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).

($\omega = 1$), 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.

Third, following the agency's choices the overseer reviews the agency and chooses to either uphold or overturn the agency's policy, denoted by $r(\cdot) \in \{0,1\}$ where $r = 0$ ($r = 1$) represents upholding (overturning) the agency. 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 $x = 0$ 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. Put simply, the overseer can either allow the agency to engage in new policy actions, even when the content of policy does not change very much (i.e., $x_A = 0$), or effectively tell the agency it is not allowed to intervene in the policy environment. Reversing 'shuts down' the agency's new proposed intervention and the state of the policy environment is returned to one with no substantive policy change ($x = 0$) and status quo levels of implementation quality.

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 before making her review decision. This choice is represented by $r(e) \in \{0,1\}$.[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 $r(x_A, e) \in \{0,1\}$. In the

[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.

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:

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

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. Finally, if the overseer overturns ($r = 1$) then both players must internalize the status quo level of implementation imprecision captured by V_{SQ}.

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.

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: $x_A(\omega) = \omega$. Since the overseer cannot condition her 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. 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 implementation. The answer depends crucially on the nature of procedural oversight.

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

[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.

V_{SQ}). Thus, the overseer's subjective expected payoff for overturning is given by,

$$-p_0\left(\beta^2\right) - p_1\left((1-\beta)^2\right) - p_2\left((2-\beta)^2\right) - V_{SQ}.$$

Since there is no policy change the overseer knows that she will lose $(\omega - \beta)^2 + V_{SQ}$ for each ω, which is weighted by the probability that a given ω is realized. Alternatively, the overseer could uphold the agency, in which case her expected payoff is,

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

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

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

To uphold the overseer requires that the agency invest sufficient effort to limit the volatility of agency-made policy. The implementation precision improvement induced by agency-made policy $V_\varepsilon(e)$, relative to the status quo level of implementation uncertainty V_{SQ}, must outweigh the overseer's net spatial losses, given her bias, for upholding 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. Rearranging the condition in expression (1) shows that the overseer will uphold the agency if she is not too biased: $\beta \leq \frac{p_1 + 4p_2 + V_{SQ} - V_\varepsilon(e)}{2p_1 + 4p_2}$. Let β_0 be the upper bound when the agency invested low effort and let β_1 be the upper bound on overseer bias when the agency has invested high effort. Since $V_\varepsilon(1) < V_\varepsilon(0)$ it follows that $\beta_1 > \beta_0$, 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.

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

The overseer's bias relative to these 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 $\beta_0 < \beta_1 < \beta$ the overseer will always overturn the agency; she is *perfectly skeptical.* Finally, when $\beta_0 < \beta < \beta_1$, the overseer upholds the overseer if and only if the agency invested high effort and she is therefore *conditionally deferential.* Agency effort decisions depend 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 to avoid effort costs at the expense of reversal and bearing the costs of high effort to avoid reversal. Since the agency does not know ω when investing effort it must also take into account the probability distribution over potential states. Specifically, when the agency invests low effort it will be overturned, but since it does not yet know the state it does not know exactly how costly that will be from a substantive policy perspective. The agency's subjected expected payoff for investing low effort is,

$$-p_1 - 4p_2 - V_{SQ} - \pi.$$

If the agency invests low effort then, in expectation, it loses utility based on the probability of each state obtaining and the policy losses associated with $x = 0$, as well as the status quo level of implementation imprecision V_{SQ} and 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 effort costs κ:

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

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

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

The left-hand side captures the net benefits of investing high effort and being upheld while the right-hand side captures the cost 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 relative to status quo precision (i.e., larger $V_{SQ} - V_{\varepsilon}(1)$) and the more likely policy change will be appropriate (i.e., higher p_1 and/or p_2) the more likely it is investing high effort will benefit the agency. 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 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 $p_1 + 4p_2 + V_{SQ} - V_{\varepsilon}(1) + \pi \geq \kappa$.*

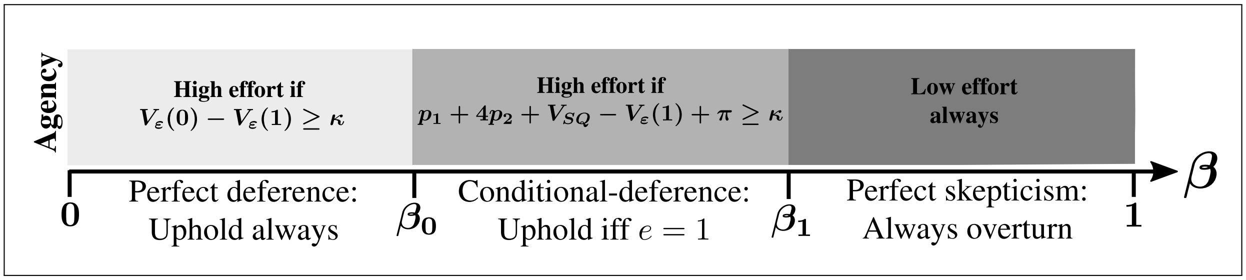

Figure 1: Equilibrium review regimes and agency effort investments

Focusing on the conditional deference case highlights a trade-off in the procedural review model. The presence of procedural review can positively or negatively affect agency effort incentives. Compared to an environment in which there is no oversight of agency policymaking, if $p_1 + 4p_2 + V_{SQ} - V_\varepsilon(1) + \pi > \kappa > V_\varepsilon(0) - V_\varepsilon(1)$ then 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 $V_\varepsilon(0) - V_\varepsilon(1) > \kappa > p_1 + 4p_2 + V_{SQ} - V_\varepsilon(1) + \pi$ then review induces the agency to invest low effort when it would have invested high effort were it not operating in the shadow of oversight. 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 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 and perverse effort incentives when the presence of oversight induces agencies to invest low effort.

Judging substance

In the substantive review model the overseer observes both the agency's effort investment and substantive policy choice. The agency's choice of x_A potentially reveals information about ω to the overseer, which introduces the possibility of obfuscation to avoid reversal. I first characterize agency substantive policy choices and overseer review decisions for a given effort investment and then turn to the connection between effort and policy choice.

The first question I address is whether and when the agency will set substantive policy 'truth-

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

fully.' 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 this is a case in which the agency fully utilizes its superior expertise. Given $x_A^{\text{truth}}(\omega)$ the overseer's optimal review strategy is illustrated in figure 2.

Figure 2 shows that the overseer will uphold $x_A = 0$ when the status quo level of implementation precision is worse than agency-made implementation precision: $V_{SQ} \geq V_\varepsilon(e)$. When the agency is setting policy truthfully and $x_A = 0$ the overseer knows $\omega = 0$ and therefore upholding and overturning both yield $x = 0$.[28] Thus, the overseer need only consider whether new agency intervention will lead to better implementation than the status quo. In addition, figure 2 shows that deference becomes less likely as preference disagreement between the agency and overseer increases. Similar to the procedural review model, when preference disagreement is sufficiently extreme any policy change, regardless of e, is overturned when the agency reveals ω.

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 associated 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 set substantive policy truthfully, and thereby reveal ω to the overseer, if the reversal cost is not too punitive.

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

The agency would rather set policy truthfully and be overturned than obfuscate and be upheld only if the potential implementation errors, relative to the status quo level of implementation imprecision, lead to worse outcomes than the cost of being reversed. That is, when the reversal

[28] In fact, regardless of the agency's policymaking strategy (e.g., pooling on $x_A = 0$, semi-pooling on $x_A = 0$) the overseer upholds $x_A = 0$ if and only if $V_{SQ} \geq V_\varepsilon(e)$, as shown by lemma B.4 in the online appendix. Moreover, lemma B.2 shows that the agency never deviates from $x_A(0) = 0$ even to avoid reversal.

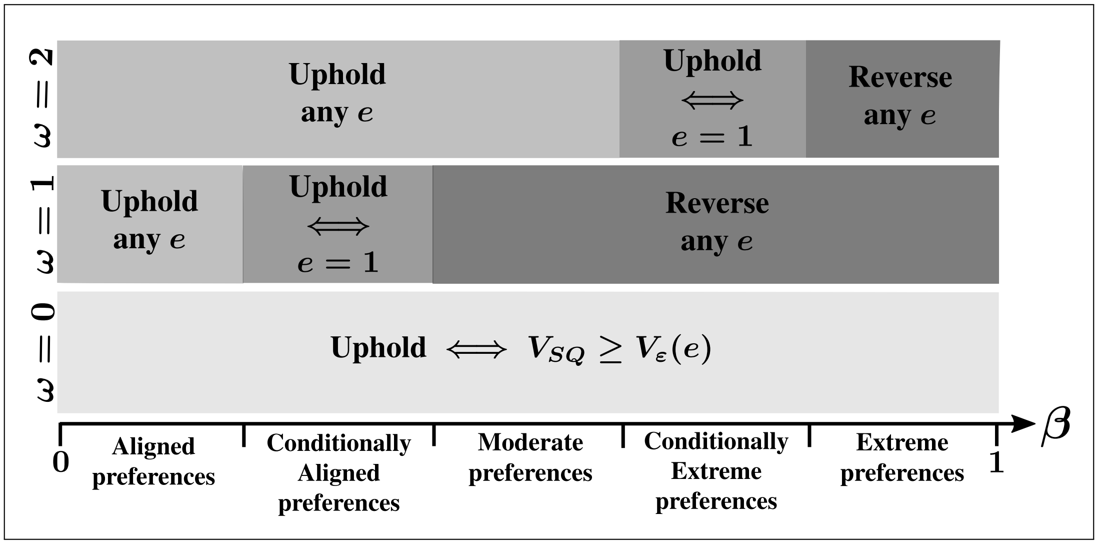

Figure 2: Overseer decisions given truthful policymaking conditional on state, effort, and bias.
Note: Preferences are *aligned* when $\beta \in [0, {}^{(1+V_{SQ}-V_{\varepsilon}(0))}/_{2})$; preferences are *conditionally aligned* when $\beta \in [{}^{(1+V_{SQ}-V_{\varepsilon}(0))}/_{2}, {}^{(1+V_{SQ}-V_{\varepsilon}(1))}/_{2})$; preference divergence is *moderate* when $\beta \in [{}^{(1+V_{SQ}-V_{\varepsilon}(1))}/_{2}, {}^{(4+V_{SQ}-V_{\varepsilon}(0))}/_{4}]$; preferences are *conditionally extreme* when $\beta \in ({}^{(4+V_{SQ}-V_{\varepsilon}(0))}/_{4}, {}^{(4+V_{SQ}-V_{\varepsilon}(1))}/_{4}]$; and preference divergence is *extreme* when $\beta > {}^{(4+V_{SQ}-V_{\varepsilon}(1))}/_{4}$.

cost π is not very punitive the agency cares more about policy than it does reputation and therefore prefers being overturned when it knows its capacity will lead to relatively poor implementation. If instead reversal costs are sufficiently punitive then the agency would instead choose to obfuscate by choosing a policy that does not match state if doing so would allow it to avoid reversal. Of course, if $\pi > V_{\varepsilon}(e) - V_{SQ}$ but there is no alternative policy choice that would lead to being upheld then the agency may continue to set policy truthfully and be overturned.

Proposition 2 has several immediate implications. First, it highlights a key 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 condition supporting truthful policymaking shows that when the agency invests high effort it is more likely that reputational considerations will drive the agency to obfuscate with its policy choice to avoid reversal. Since $V_{\varepsilon}(1) < V_{\varepsilon}(0)$ there is a wider range of π in which the agency would prefer to deviate from truthful policymaking *if it has already invested high effort*.[29] When the agency has invested in improving the quality of policy outcomes it has stronger incentives to take actions that lead to those outcomes being realized even at the cost of matching policy to the state. High effort investment benefits the overseer but it also strengthens the incentives for the agency to deviate from truthful policymaking to avoid reversal, which may prove costly to the overseer.

Another implication of proposition 2 is that when the overseer will uphold $x_A = 0$ for a given e, which requires $V_{SQ} \geq V_\varepsilon(e)$, there are ranges of preference disagreement in which there is no truthful separating equilibrium. This follows from the fact that $V_{SQ} \geq V_\varepsilon(e)$ implies that $\pi > V_\varepsilon(e) - V_{SQ}$ for any $\pi \in (0,1)$. Thus, when $V_{SQ} \geq V_\varepsilon(e)$ there are preference environments where the agency would prefer to deviate from truthful policymaking when it is being overturned to a policy choice that will avoid reversal. This profitable deviation is always available since the overseer will uphold $x_A = 0$. For example, if the overseer is moderately biased and will overturn a truthful policy choice of $x_A = 1$ then the agency benefits by instead setting $x_A = 0$ to avoid being reversed.

This logic further illustrates the observation in corollary 1. If the agency is operating in a *high volatility* policy environment in which low effort implementation is still better than the status quo, $V_{SQ} > V_\varepsilon(0) > V_\varepsilon(1)$, then the agency will always obfuscate to be upheld when possible since $V_\varepsilon(e) - V_{SQ} > \pi$ for any e. If instead the agency is tasked with policymaking in either a *low volatility*, $V_\varepsilon(0) > V_\varepsilon(1) > V_{SQ}$, or a *moderate volatility*, $V_\varepsilon(0) > V_{SQ} > V_\varepsilon(1)$, policy environment then it is possible that reversal aversion may not be high enough in some cases to induce obfuscation.

In these latter policy environments reversal costs can be either *highly punitive* so that $\pi > V_\varepsilon(e) - V_{SQ}$ for any e or *moderately punitive* so that $V_\varepsilon(0) - V_{SQ} > \pi > V_\varepsilon(1) - V_{SQ}$. The key difference in equilibrium is that when reversal costs are moderately punitive the agency will obfuscate to avoid reversal only when it has invested high effort whereas highly punitive reversal implies that the agency will obfuscate to avoid being overturned any time that is a feasible option. The agency may still set policy truthfully when $\pi > V_\varepsilon(e) - V_{SQ}$, but this is only true when there is no reason to deviate, either because it is being upheld or there is no alternative policy that would avoid reversal.

[29] This follows from the fact that $V_\varepsilon(0) > V_\varepsilon(1)$, which implies that $V_\varepsilon(0) - V_{SQ} > V_\varepsilon(1) - V_{SQ}$.

Specifically, if agency-overseer ideal points are very close together then the agency can continue to set policy truthfully. Any substantive choice that moves policy from the status quo and reveals ω to the overseer will be upheld regardless of effort investment and the agency never benefits by deviating from $x_A = 0$ when $\omega = 0$ even if it means accepting reversal. Oversight has ‘no bite’ in this setting. Additionally if, given effort, there is no obfuscatory policy choices that would avoid reversal then the agency will continue to set policy truthfully and accept being overturned because it has no other option.

In contrast, when an obfuscatory policy choice is available to avoid reversal and $\pi > V_\varepsilon(e) - V_{SQ}$ the agency will pursue that option. When preferences are extremely divergent but $V_{SQ} \geq V_\varepsilon(e)$ there is a pooling equilibrium in which the agency sets $x_A(\omega) = 0$ for all ω and the overseer upholds that choice. Thus, when $\pi > V_\varepsilon(e) - V_{SQ}$ and preferences are sufficiently aligned or there is no obfuscatory policy available then the agency can continue to set policy truthfully in equilibrium, and when preferences are extreme but the overseer would uphold $x_A = 0$ the agency can always retain the status quo and avoid reversal.

This leaves the intermediate regions of preference disagreement where the agency may set policy using a semi-pooling equilibrium strategy to avoid being overturned. This is optimal when truthfully setting $x_A = 1$ will lead to reversal. The first possibility is when the overseer will uphold $x_A = 0$ because $V_{SQ} \geq V_\varepsilon(e)$, characterized in the following result.

Proposition 3. *Assume $\beta \in \left[\frac{1+V_{SQ}-V_\varepsilon(e)}{2}, \frac{4+V_{SQ}-V_\varepsilon(e)}{4}\right]$ 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 $x_A = 2$.*

When the overseer will overturn $x_A^{\text{truth}} = 1$ but uphold $x_A = 0$ the agency can obfuscate by setting $x_A = 0$ when $\omega = 1$ to be upheld. This is a strategy where the agency is able to avoid reversal by *obfuscating to appease* the overseer. When $\omega = 1$ the overseer would prefer the agency to instead set $x_A = 0$, all else equal, so maintaining the status quo rather than moderately shifting policy is a way for the agency to avoid being overturned. This is consistent with many previous theories of bureaucratic oversight in which review of agency policy choices leads the agency to moderate its choices to satisfy the overseer (e.g., Epstein and O'Halloran 1999; Shipan 1997; Wiseman 2009). Substantively, this equilibrium implies that policy will be overly conservative in the sense that in environments where the agency's private information suggests it should moderately shift policy it instead appeases the overseer by foregoing policy change due to reputational concerns.

The agency may also be able to employ another strategy to avoid reversal: *Obfuscation through exaggeration*ol, which is characterized in the following result.

Proposition 4. *Assume $\beta \in \left[\frac{1+V_{SQ}-V_{\varepsilon}(e)}{2}, \frac{4+V_{SQ}-V_{\varepsilon}(e)}{4}\right]$ 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 $x_A(0) = 0$ and $x_A(\omega) = 2$ for $\omega \in \{1,2\}$, and the overseer upholds $x_A = 0$ if $V_{SQ} \geq V_{\varepsilon}(e)$, overturns $x_A = 1$, and upholds $x_A = 2$.*

Proposition 4 shows that when the overseer will overturn $x_A^{\text{truth}} = 1$ the agency may be able to obfuscate and avoid reversal by exaggerating the need for extreme policy change. When the agency sets $x_A(\omega) = 2$ for $\omega \in \{1,2\}$ the overseer will uphold $x_A = 2$ so long as, ex ante, $\omega = 2$ is sufficiently likely relative to $\omega = 1$: $\frac{p_1}{p_1+p_2} \leq \frac{1}{4}(4 - 4\beta + V_{SQ} - V_{\varepsilon}(e))$. When this condition is met the agency is able to exaggerate the need for policy change by setting $x_A = 2$ when $\omega = 1$ to avoid being overturned. Similar to the logic in corollary 1 the agency is more likely to be able to obfuscate by setting $x_A = 2$ when $\omega = 1$ if it invested high effort.[30]

This highlights a potential 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 the aforementioned theories of bureaucratic oversight. 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

[30] This follows from the fact that $V_{\varepsilon}(1) < V_{\varepsilon}(0)$ implies $\frac{1}{4}(4 - 4\beta + V_{SQ} - V_{\varepsilon}(0)) < \frac{1}{4}(4 - 4\beta + V_{SQ} - V_{\varepsilon}(1))$.

large policy loss through policy-state mismatch and increases the utility of allowing the agency to intervene in the environment, which jointly increases the overall expected payoff for upholding.

Notice that the equilibrium in proposition 4 is possible whether or not the overseer would uphold $x_A = 0$. This implies that when $V_{SQ} \geq V_{\varepsilon}(e)$ and $\frac{p_1}{p_1+p_2} \leq \frac{1}{4}(4-4\beta+V_{SQ}-V_{\varepsilon}(e))$ both semi-pooling equilibria are possible. The agency is indifferent between choosing to either appease or exaggerate when $\omega = 1$ as long as both lead to being upheld. However, when $V_{SQ} < V_{\varepsilon}(e)$ and the conditions in proposition 4 are satisfied, the only obfuscation strategy for the agency is to exaggerate by setting $x_A = 2$ when $\omega = 1$. When neither of those possibilities are available the agency continues to set policy truthfully in all states, which leads to reversal when $\omega \in \{1,2\}$ and deference when $\omega = 2$. In either semi-pooling equilibrium there is 'polarized' policymaking in the sense that only extreme policy change is pursued, otherwise the status quo is maintained. There are no moderate, incremental policy changes in this environment.

Obfuscation is predicated on the agency's need and opportunity to do so. These considerations are, in turn, functions of the environment and the agency's effort investment. If the agency is operating in a highly volatile policy environment, $V_{SQ} > V_{\varepsilon}(0) > V_{\varepsilon}(1)$, then it can always obfuscate by appeasing (proposition 3) when being overturned and would always choose to do so since $\pi > V_{\varepsilon}(e) - V_{SQ}$ for all e. If the agency is operating in a moderate volatility environment, $V_{\varepsilon}(0) > V_{SQ} > V_{\varepsilon}(1)$ then the agency will never obfuscate following $e = 0$ since $\pi < V_{\varepsilon}(0) - V_{SQ}$. However, following high effort the agency can and will obfuscate to be upheld when possible since $\pi > V_{\varepsilon}(1) - V_{SQ}$. Finally, if the agency is tasked with regulating a low volatility environment, $V_{\varepsilon}(0) > V_{\varepsilon}(1) > V_{SQ}$ then it is still possible that the agency will always obfuscate when possible if reversal costs are highly punitive, $\pi > V_{\varepsilon}(e) - V_{SQ}$ for all e, obfuscate when possible only when $e = 1$ if reversal costs are moderately punitive, $V_{\varepsilon}(0) - V_{SQ} > \pi > V_{\varepsilon}(1) - V_{SQ}$, or never obfuscate if reversal costs are very low, $\pi < V_{\varepsilon}(e) - V_{SQ}$ for all e.

In the interest of clarity I focus on general intuition underlying equilibrium effort investment and present two illustrative examples in-text.[31] Since the parameters that dictate the type of pref-

[31] Lemmas B.5 – B.9 in the online appendix present the complete set of equilibrium effort results.

erence environment (β), policy environment ($V_{SQ}, V_\varepsilon(0), V_\varepsilon(1)$), and reversal aversion level (π) are fixed and known the agency's effort investment essentially dictates what type of substantive policymaking strategy it will subsequently play in each potential state, given the overseer's equilibrium review strategy. Of course, when the agency makes this investment it does not yet know what state will be realized. Thus, the agency makes its effort investment decision based on its expectations about which state will be realized, and the subsequent policymaking strategy and payoff for each potential state. In some states the agency may be able to set policy truthfully and be upheld, while in others it may not be able to obfuscate and instead has to accept being overturned, while in still others it choose to obfuscate to be upheld following its effort choice.

In states where the agency will be upheld regardless of its effort investment it invests effort solely based on whether the implementation improvement from high effort relative to low effort is worth it. In states where any substantive policy choice will be overturned, even if the agency invests high effort, the agency never wants to invest high effort since it would not change the outcome. High effort can also be the difference between obfuscating and setting policy truthfully, and being overturned or being able to obfuscate to be upheld.

When high effort allows for truthful policymaking and low effort involves obfuscating the agency considers the net benefit of being able to match policy to the state and the implementation improvement from high versus low effort. When high effort allows the agency to obfuscate and be upheld rather than be overturned it considers the costs of mismatching policy and the state, the implementation improvement from high effort relative to status quo implementation uncertainty, and the benefit of avoid reversal and paying π. In different environments, different states are associated with these different strategies and payoffs. Thus, each possibility is weighted by the probability that that state will be realized in a given preference environment. So long as the overall expected payoffs outweigh the costs of high effort then the agency is willing to pay for that investment.

To further illustrate these effort investment dynamics, I analyze two particular examples. The first example illustrates an environment in which high effort investment allows the agency to set policy truthfully rather than obfuscate when it invests low effort. The second example analyzes a setting where low effort leads to reversal, while high effort allows the agency to obfuscate and avoid being overturned. Full calculations can be found in the online appendix.

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

This setting is one in which the agency can set policy truthfully in all states of the world and be upheld if it invests high effort. If instead the agency invests low effort it will truthfully set $x_A(0) = 0$ and be overturned, it will obfuscate by setting $x_A = 2$ when $\omega = 1$ to be upheld since $\pi > V_\varepsilon(0) - V_{SQ}$ (semi-pooling strategy in proposition 4), and will truthfully set policy $x_A(2) = 2$ and be upheld since preferences are conditionally aligned. Whether the agency invests high effort in equilibrium depends on whether the net benefits of being able to set policy truthfully and be upheld, which requires high effort investment, are large enough to offset the cost κ to obtain those benefits.

Plugging the parameter values into the relevant incentive compatibility condition for high effort reveals that the agency will invest high effort in this environment if and only if $\kappa \leq \frac{11}{16}$. When this condition is satisfied the agency will invest high effort, which will allow it to subsequently set policy truthfully in all states and the overseer will always uphold. If instead $\kappa > \frac{11}{16}$ then the agency will truthfully set $x_A(0) = 0$, which is overturned by the overseer, obfuscate through exaggeration by setting $x_A = 2$ when $\omega = 1$ to avoid reversal, and truthfully set $x_A(2) = 2$, which is upheld. □

Example 1 shows how high effort investment produces the joint benefit of being able to match policy to the state and avoid reversal. The next example illustrates a setting in which high effort leads the agency to obfuscate to avoid reversal, which is the outcome following low effort.

Example 2. *(High effort to obfuscate)* Let $\beta \in \left(\frac{1+V_{SQ}-V_\varepsilon(1)}{2}, \frac{4+V_{SQ}-V_\varepsilon(0)}{4}\right]$ so that preferences are *moderate* and 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*.

This environment is one in which when the agency invests high effort it is upheld for truthfully setting policy when either $\omega = 0$ or $\omega = 2$, and it obfuscates by setting either $x_A = 0$ or $x_A = 2$ (when the conditions for proposition 4 are satisfied) when $\omega = 1$, which the overseer upholds. 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 obfuscate 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 more precise implementation – outweigh the costs of those benefits.

Plugging in the parameter values to the agency's incentive compatibility condition to invest high effort in this setting reveals that the agency will invest high effort if and only if $\kappa \leq \frac{7}{16}$. When the inequality holds the agency will invest high effort, which allows it to subsequently set policy truthfully and be upheld when $\omega = 0$, obfuscate to be upheld when $\omega = 1$ (by either appeasement, which is always available, or exaggeration, if $\frac{p_1}{p_1+p_2} \leq \frac{1}{4}(4 - 4\beta + V_{SQ} - V_\varepsilon(1))$, and truthfully set $x_A = 2$ when $\omega = 2$, which is upheld by the overseer. If instead $\kappa > 7/16$ then the agency sets policy truthfully when $\omega \in \{0, 1\}$ and accepts being overturned, since obfuscation is not incentive compatible, and is upheld following truthful policymaking when $\omega = 2$. □

In example 2 high effort allows the agency to obfuscate to avoid reversal, as in propositions 3 and 4, when low effort leads to being overturned. This illustrates how high effort can indirectly lead to the overseer receiving obscured information through the direct effect on agency obfuscation incentives. In short, when the agency has invested high effort its motivations to avoid being reversed are also intensified because it does not want to let that investment go to waste.

Overall, the preceding analysis shows that increasing transparency in the review process introduces the possibility for obfuscation through either policy appeasement or policy exaggeration. In equilibrium, this implies that in environments conducive to semi-pooling behavior agency policymaking becomes polarized: In either case the agency only retains the status quo or 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. If one takes low reversal costs as representative of highly insulated agencies, this implies 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’ 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 either introduce status quo bias when moderate changes are called for or pursue only extreme policy adjustment when it believes *any* policy change is called for. Further, some form of obfuscation is more likely if the agency has invested high effort. Thus, which form of institutional oversight is more beneficial depends crucially on the nature of the political environment.

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 these questions I consider the overseer’s ex ante expected utility across the two different scopes of review. A reasonable conjecture is that the increased control over specific policy choices that is provided by substantive review can only benefit the overseer. In some cases, that conjecture is true and substantive review does benefit the overseer. However, in other environments *less* information can prove beneficial and procedural review is ex ante preferred by the overseer.

Low and high policy disagreement

When agency-overseer preferences are either closely aligned or extremely misaligned substantive review is weakly preferred to procedural review. In the former case, the agency is always upheld under procedural review and any policy change is upheld under substantive review while $x_A = 0$ is upheld if $V_{SQ} \geq V_{\varepsilon}(e)$. If the overseer would also uphold retention of the status quo when judging substance then there is no difference at all between the review institutions. However, if she would overturn that same choice under substantive review then substantive review is strictly beneficial due to the increased control over that single state, which is absent under procedural review. Thus, given sufficiently aligned preferences the scope of review is either inconsequential or substantive review is strictly preferred. In the latter case, the agency is always overturned under procedural review and the overseer overturns any $x_A \in \{1,2\}$ and upholds $x_A = 0$ if $V_{SQ} \geq V_{\varepsilon}(e)$ under substantive review. If the overseer overturns the agency no matter what across the review institutions then scope of review is again inconsequential. However, if the overseer would benefit from upholding $x_A = 0$ then substantive review dominates procedural review since that increased control when $\omega = 0$ will improve her payoffs (by upholding $x_A(0) = 0$ or inducing pooling on $x_A = 0$ for all ω).[32]

Intermediate policy disagreement

When preference disagreement is moderate so that procedural oversight leads to conditional-deference and substantive oversight may induce the agency to obfuscate, *how* the agency obfuscates dictates which form of review is optimal. If the agency obfuscates by appeasing the overseer and setting $x_A = 0$ when $\omega = 1$, as in proposition 3, then the overseer benefits from the increased control over moderate policy changes that substantive review provides. Given moderate preference divergence the overseer would prefer $x_A = 0$ when $\omega = 1$, which can only be induced through substantive review. Thus, the place to look for an environment where procedural review may be preferred to substantive review is one in which the agency obfuscates through exaggeration.

Specifically, I focus on the environment outlined in example 2 above, which corresponds to the equilibrium policymaking behavior in proposition 4. I compare this substantive review setting, in which the overseer is moderately biased, to the conditional-deference case under procedural review. This implies that under substantive review the agency sets policy truthfully when $\omega \in \{0,2\}$ and obfuscates by exaggerating when $\omega = 1$ so that $x_A(1) = 2$ when it has invested high effort. In this case the overseer upholds $x_A = 1$ and $x_A = 2$ and overturns $x_A = 1$. If the agency instead invests low

[32] Details can be found in section C.1 in the online appendix.

effort then it sets policy truthfully in every state, but the overseer overturns $x_A = 0$ and $x_A = 1$ and upholds $x_A = 2$. Under procedural review the overseer upholds if and only if the agency invests high effort and the agency always matches substantive policy to the state.

The condition for the agency to invest high effort is more lenient under procedural review. That is, there are values of κ for which the agency would invest high effort if the overseer is reviewing procedure but not when she is judging substance. This implies that, in this environment, there are three possibilities: The agency invests high effort under both systems, it invests high effort only if it is under procedural review, and it invests low effort under both types of review. I discuss each possibility and how they affect the overseer's preferences over review type.[33]

Low effort under both review systems. When effort costs are so high that the agency would invest low effort regardless of the review system the overseer always benefits from judging substance. If $\omega = 0$ or $\omega = 1$ then the overseer's expected payoffs are the same under both systems since in both situations the agency is overturned. If instead $\omega = 2$ is realized the overseer benefits from the extra information provided by substantive review because in this preference environment she would prefer to uphold $x_A = 2$, regardless of effort investment, but cannot do so under a procedural review system. Thus, her ability to uphold $x_A = 2$ under substantive review benefits her from an ex ante perspective, which implies that she would be better off in this case with substantive review.

High effort only if procedural review. When effort costs are intermediate so that the agency invests high effort under procedural review and low effort under substantive review the overseer's ex ante preference over review systems depends further on her relative bias and how likely it is that each state is realized. Under procedural review the agency always matches policy to the state and invests high effort so the overseer upholds. Under substantive review the agency invests low effort, sets policy truthfully, and the overseer overturns $x_A \in \{0,1\}$ and upholds $x_A = 2$. Given this equilibrium behavior, the overseer prefers more information when $\omega = 1$ because she prefers to overturn in that case, which she cannot if she is reviewing procedure. On the other hand, in the event that either $\omega = 0$ or $\omega = 2$ she would prefer procedural review because the substantive

[33] Details can be found in the section C.2 in the online appendix.

outcome in those cases does not vary across institutions but the implementation uncertainty does. Because the agency invests high effort in those cases under procedural review the overseer is better off than under substantive review where the agency is overturned (when $\omega = 0$) or upheld following low effort (when $\omega = 2$). Thus, the likelihood that the overseer prefers procedural review in this setting is increasing as the probability that moderate policy change is called for decreases, implying increases in p_1 and/or p_2, and as her bias decreases, because the negative impact on her expected utility when $\omega = 1$ is smaller when β is smaller. Conversely, the more likely that $\omega = 1$ will be realized and the more biased the overseer is relative to the agency the more likely it is she will prefer substantive review in order to have better control over policymaking when $\omega = 1$.

High effort under both review systems. Finally, when effort costs are so low that the agency will invest high effort under both review systems the overseer always benefits from procedural review. This environment is, in some respects, an optimal comparison: Under both systems the agency invests high effort and the overseer upholds equilibrium policy choices. Thus, the only point of departure between the two institutions is that under procedural review the agency can always match policy to the state and under substantive review the agency obfuscates by exaggeration. Otherwise, the agency invests high effort and matches policy to the state both when retention of the status quo and extreme policy change is called for under both systems of oversight. This implies that the overseer's payoffs are equivalent under either review institution when $\omega = 0$ or $\omega = 2$: The substantive outcome is policy matching the state with high effort implementation precision. However, when moderate policy change is called for (i.e., $\omega = 1$ is realized) the overseer is worse off when judging substance because the agency obfuscates by setting $x_A(1) = 2$ and exaggerating the need for policy change to avoid reversal. The overseer would prefer, if she had complete information, to overturn the agency when $\omega = 1$ but in the absence of being able to do so she is better off when the agency is able to match policy to the state (i.e., $x_A(1) = 1$) than when the agency obfuscates, which is only possible under procedural review. The upshot is that the overseer is always better off with less information in the review process because procedural review serves as an institutional commitment that allows the agency to moderately shift policy when called for, which in turn obviates the perverse policymaking incentives that cause obfuscation by exaggeration.

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 may benefit from less information for the purposes of review. This suggests that it is far from clear that providing overseers with more information during the review process will generate net benefits once one takes into account how that information disclosure 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.

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 high. A key insight is that when the transparency of policymaking is increased there is a trade-off between effort incentives and the incentives for agencies to utilize their policy-relevant expertise. This undercuts the powerful normative rationale for delegation to expert agencies by inducing these agencies to underutilize their expertise.

Additionally, I have provided results that suggest that the overseer can benefit from *less information* in the review process. 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. All of the perverse effects are a product of 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 potentially exaggerate with its policy choice. This suggests that in certain political environments insulating the agency from reputational considerations can improve the incentives for the agency to work hard and utilize its expertise. Overall, the scope of review that agencies are subjected to can have profoundly differential effects on the agency's policymaking incentives. 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 Greg 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* .

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. 2017. "Ex Post Review and Expert Policymaking: When Does Oversight Reduce Accountability?" *Unpublished Manuscript. Yale University. Presented at the 2016 Annual Meeting of the American Political Science Association* .

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

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*a*. "Political Agency, Oversight, and Bias: The Instrumental Value of Politicized Policymaking." *Unpublished Manuscript. Yale University* .

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

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

Contents

A Procedural review **2**

A.1 Agency substantive policy choice 2

A.2 Optimal procedural oversight 3

A.3 Agency effort investment 4

B Substantive review **6**

B.1 Truthful equilibrium 6

B.2 Obfuscation equilibria 14

B.2.1 Substantive policy choices 15

B.2.2 Effort choices 18

B.2.3 In-text examples 30

C Comparing Review Institutions **33**

C.1 Aligned and extreme preferences 33

C.2 Moderate preference examples 35

A Procedural review

A.1 Agency substantive policy choice

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

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

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

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

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

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

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

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

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

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

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

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

A.2 Optimal procedural oversight

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

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

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

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

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

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

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

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

as stated in the lemma. ■

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

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

A.3 Agency effort investment

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

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

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

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

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

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

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

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

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

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

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

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

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

Conditional-deference. In this case the overseer upholds the agency if and only if the agency invests high effort. The agency's expected payoff for investing high effort, which induces being upheld, is,

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

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

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

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

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

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

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

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

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

B Substantive review

B.1 Truthful equilibrium

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

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

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

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

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

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

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

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

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

Thus, the overseer upholds the agency when $x_A^{\text{truth}}(0) = 0$ so long as $V_\varepsilon(e) \leq V_{SQ}$.
Case 2: $\omega = 1$. The overseer's expected payoff for reversing the agency when $\omega = 1$ and $x_A(1) = 1$, for a given e, is given by,

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

For the overseer to uphold incentive compatibility requires that,

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

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

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

This yields the following incentive compatibility constraint to uphold:

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

Combining the cases analyzed above yields the result. ■

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

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

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

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

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

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

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

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

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

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

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

Proof of Proposition 2. Lemma B.2 shows that the agency is always truthful when $\omega = 0$ so we derive the incentive compatibility conditions for the agency to truthfully reveal $\omega \in \{1,2\}$. First, consider the case in which $\omega = 1$. If the agency is upheld when it truthfully sets $x_A(1) = 1$ then there is no incentive to deviate. Similarly, if the agency is overturned when $x_A(1) = 1$ and also overturned whenever $x_A = 0$ and $x_A = 2$ then there is no reason to deviate. Thus, we need only check whether the agency would deviate when $x_A(1) = 1$ is overturned but either $x_A = 0$ or $x_A = 2$ would be upheld. In either case the agency deviates spatially by one so expected utility is equivalent for deviating to $x_A = 0$ and $x_A = 2$ when $\omega = 1$ so we only show the derivations for deviating to $x_A = 0$ noting that the calculations are equivalent when $x_A = 2$ is the deviation. The agency's expected utilities for setting $x_A = 1$ truthfully and deviating by one to be upheld are given by,

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

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

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

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

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

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

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

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

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

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

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

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

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

I will plug these general expected utilities into the relevant incentive compatibility conditions to analyze each case given $r(x_A,e)$ from Lemma B.1.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

B.2 Obfuscation equilibria

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

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

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

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

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

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

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

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

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

B.2.1 Substantive policy choices

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

B.2.2 Effort choices

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

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

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

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

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

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

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

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

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

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

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

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

That is, in this environment the agency will be reversed in states zero and one if it invests low effort, but it invests high effort it will be upheld for truthful policymaking in state zero and will obfuscate to be upheld in state one. In state two the agency will be upheld for truthful policymaking for both low and high effort. Thus, conditional on states zero or one obtaining (probability $p_0 + p_1$) the agency invests high effort if the implementation improvement from doing so ($V_{SQ} - V_{\varepsilon}(1)$) and the benefit of avoiding reversal (π) are high enough, and conditional on state two obtaining (probability p_2) the agency wants to invest high effort if the implementation improvement between high and low effort investment are high enough ($V_\varepsilon(0) - V_\varepsilon(1)$). As long as those collective potential benefits are higher than the cost of investing high effort then the agency will do so. Equivalent derivations produce analogous incentive compatibility conditions for each preference environment conditional on the volatility of the policy environment (ordering of $V_{SQ}, V_\varepsilon(0)$, and $V_\varepsilon(1)$) and how punitive reversal is (ordering of π and $V_\varepsilon(e) - V_{SQ}$ for both e), captured in the next set of results.

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

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

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

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

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

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

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

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

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

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

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

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

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

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:

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

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

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

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

Now let $V_\varepsilon(0) - V_{SQ} > \pi > V_\varepsilon(1) - V_{SQ}$. Note that $V_{SQ} > V_\varepsilon(0) > V_\varepsilon(1)$ can not be true in this setting. Because preference are aligned agency-overseer equilibrium behavior is the same: There is no reason to obfuscate since the agency is upheld for setting policy truthfully when $\omega \in \{1,2\}$ and lemma B.2 implies the agency also never deviates when $\omega = 0$. Thus, the incentive compatibility conditions for high effort are equivalent to those derived for the $\pi > V_\varepsilon(e) - V_{SQ}$ for all e cases above, as stated in the result. ■

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Thus, the agency will invest high effort when,

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

which yields the following incentive compatibility condition for high effort,

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

This derives all the results in the lemma. ■

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

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

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

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

Proof of Lemma B.8. First, let $\pi > V_\varepsilon(e) - V_{SQ}$ for all e. When $V_{SQ} > V_\varepsilon(0) > V_\varepsilon(1)$ the agency will match policy to the state and be upheld when $\omega \in \{0,2\}$ and obfuscate to $x_A = 0$ or $x_A = 2$ (if possible) and be upheld when $\omega = 1$. After $e = 0$ the agency will pool by setting policy at $x_A = 0$ for all ω and be upheld. This yields the following expected utilities for high and low effort,

$$\begin{aligned} EU_A(e=1) &= -p_0V_\varepsilon(1) - p_1(1+V_\varepsilon(1) - p_2V_\varepsilon(1) - \kappa, \\ EU_A(e=0) &= -p_0V_\varepsilon(0) - p_1(1+V_\varepsilon(0)) - p_2(4+V_\varepsilon(0)), \end{aligned}$$

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

$$\begin{aligned} -p_0V_\varepsilon(1) - p_1(1+V_\varepsilon(1) - p_2V_\varepsilon(1) - \kappa &\geq -p_0V_\varepsilon(0) - p_1(1+V_\varepsilon(0)) - p_2(4+V_\varepsilon(0)), \\ (p_0+p_1)(V_\varepsilon(0) - V_\varepsilon(1)) + p_2(4+V_\varepsilon(0) - V_\varepsilon(1)) &\geq \kappa. \end{aligned}$$

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,

$$\begin{aligned} EU_A(e=1) &= -p_0V_\varepsilon(1) - p_1(1+V_\varepsilon(1)) - p_2V_\varepsilon(1) - \kappa, \\ EU_A(e=0) &= -p_0V_{SQ} - p_1(1+V_{SQ}) - p_2(4+V_{SQ}) - \pi, \end{aligned}$$

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

$$\begin{aligned} -p_0V_\varepsilon(1) - p_1(1+V_\varepsilon(1)) - p_2V_\varepsilon(1) - \kappa &\geq -p_0V_{SQ} - p_1(1+V_{SQ}) - p_2(4+V_{SQ}) - \pi, \\ 4p_2 + V_{SQ} - V_\varepsilon(1) + \pi &\geq \kappa. \end{aligned}$$

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

$$\begin{aligned} EU_A(e=1|\text{obfuscate}) &= -p_0(V_{SQ}+\pi) - p_1(1+V_\varepsilon(1)) - p_2(V_\varepsilon(1)) - \kappa, \\ EU_A(e=1|\text{not obfuscate}) &= -p_0V_{SQ} - p_1(1+V_{SQ}) - p_2(4+V_{SQ}) - \pi, \\ EU_A(e=0|\text{obfuscate}) &= -p_0V_{SQ} - p_1(1+V_{SQ}) - p_2(4+V_{SQ}) - \pi, \\ EU_A(e=0|\text{not obfuscate}) &= -p_0V_{SQ} - p_1(1+V_{SQ}) - p_2(4+V_{SQ}) - \pi. \end{aligned}$$

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

B.2.3 In-text examples

Example 1. *(High effort to tell the truth)* Let $\beta\in\left[\frac{1+V_{SQ}-V_\varepsilon(0)}{2},\frac{1+V_{SQ}-V_\varepsilon(1)}{2}\right)$ so that preferences are *conditionally aligned.* Further, fix the following parameter values: $\beta={}^7/_{16}$, $p_0={}^1/_4$, $p_1={}^1/_4$, $p_2={}^1/_2$, $V_{SQ}={}^1/_4$, $V_\varepsilon(0)={}^1/_2$, $V_\varepsilon(1)={}^1/_8$, $\pi={}^1/_2$. These parameter values further imply that $\frac{p_1}{p_1+p_2}\leq\frac{1}{4}(4-4\beta+V_{SQ}-V_\varepsilon(0))$ holds so that the conditions for semi-pooling characterized in Proposition 4 are satisfied, $V_\varepsilon(0)>V_{SQ}>V_\varepsilon(1)$ so that the policy environment is *moderately volatile*, and that $\pi>V_\varepsilon(0)-V_{SQ}>V_\varepsilon(1)-V_{SQ}$ so that reversal costs are *highly punitive*.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

C Comparing Review Institutions

C.1 Aligned and extreme preferences

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

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

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

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

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

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

$$\begin{aligned} EU_R^S(r(0,e)=1) &= -p_0((0-\beta-(0)x)^2+V_{SQ})-p_1((1-\beta-(1)(1+\varepsilon))^2)-p_2((2-\beta-(1)(2+\varepsilon))^2), \\ &= -p_0(\beta^2+V_{SQ})-p_1(\beta^2+V_\varepsilon(e))-p_2(\beta^2+V_\varepsilon(e)), \\ &= -\beta^2-p_0V_{SQ}-(p_1+p_2)V_\varepsilon(e). \end{aligned}$$

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

$$\begin{aligned} EU_R^P(r(e)=0) &> EU_R^S(r(0,e)=1), \\ -\beta^2-V_\varepsilon(e) &\geq -\beta^2-p_0V_{SQ}-(p_1+p_2)V_\varepsilon(e), \\ p_0(V_{SQ}-V_\varepsilon(e)) &\geq 0, \end{aligned}$$

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

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

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

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

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

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

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

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

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

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

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

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

C.2 Moderate preference examples

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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