Appendix

Notation: For simplicity we henceforth write the voter's net policy benefit for the incumbent policy $U(x_V; x_I, x_C)$ as $U_V \geq 0$ which is assumed to be positive, and write the agent's net benefit for the incumbent policy $U(x_S; x_I, x_C)$ as $-U_S$, where $U_S \geq 0$ denotes the agent's net utility for the challenger policy. We also write π_0^0 as just π_0, π_1^1 as π_H, and π_1^0 as π_L. Finally, we suppress the explicit dependence of $\bar{\Delta}_{\lambda_I}(\cdot)$ and $\bar{\theta}_C(\cdot)$ on other quantities.

It is first helpful to show the property that $\bar{\Delta}_H > \frac{q_H}{q_L}\bar{\Delta}_L$, which furthermore has the implication that $\bar{\Delta}_H \leq q_H \rightarrow \bar{\Delta}_L < q_L$. This eliminates much of the parameter space and several potential types of equilibria. To see this, observe that the desired property is equivalent to

$$q_H\gamma + \delta\bar{\Delta}_L\frac{q_H}{q_L}\cdot\left(\gamma\left(q_H - q_L\right)\left(1-\theta_C\right) - U_S\right) \geq 0$$

or

$$\frac{U_S - \gamma\left(q_H - q_L\right)\left(1-\theta_C\right)}{U_S + \gamma\theta_C\left(q_H - q_L\right)} \leq 1$$

which clearly always holds.

A Preliminary Analysis

Equilibrium values of e_L and e_H in conjunction with the incumbent's initial popularity imply different possible restrictions on the retention probabilities π_0, π_L, and π_H. These in turn imply different feasible pairs of (Δ_L, Δ_H). Anticipating these restrictions, we first examine several relevant feasible sets of (π_0, π_L, π_H) and their implications for (Δ_L, Δ_H). Specifically, for each type of triple (π_0, π_L, π_H) we characterize feasible Δ_L and then the feasible values of Δ_H given Δ_L. We then subsequently use this characterization in the equilibrium characterization.

In the subsequent case-by-base breakdown, (S) refers to "single mixing" (the voter mixes after one path of play) while (D) refers to "double-mixing" (the voter mixes after two paths of play).

Case S.1 ($\pi_0 \in (0,1)$, $\pi_L = \pi_H = 1$**)**, We have

$$\Delta_{\lambda_I} = 1 - \pi_0$$

Therefore feasible values of Δ_L are all $\Delta_L \in [0,1]$ and $\Delta_H = \Delta_L$

Case S.2 ($\pi_0 = 0$, $\pi_L = 0$, $\pi_H \in (0,1)$**)**. We have

$$\Delta_{\lambda_I} = q_{\lambda_I}\pi_H$$

and it straightforward to show that $\Delta_L \in [0, q_L]$ and $\Delta_H = \frac{q_H}{q_L}\Delta_L$.

Case S.3 ($\pi_0 = 0$, $\pi_L \in (0,1)$, $\pi_H = 1$**)**. We have

$$\Delta_{\lambda_I} = q_{\lambda_I} + \left(1 - q_{\lambda_I}\right)\pi_L$$

and it is straightforward to show that $\Delta_L \in [q_L, 1]$ and $\Delta_H = q_H + \left(\frac{1-q_H}{1-q_L}\right)\left(\Delta_L - q_L\right)$ which is clearly $< \frac{q_H}{q_L}\Delta_L$.

Case S.4 ($\pi_0 \in [0,1]$, $\pi_L = 0$, $\pi_H = 1$). We have

$$\Delta_{\lambda_I} = q_{\lambda_I} - \pi_0$$

so it is straightforward that $\Delta_L \in [0, q_L]$ and $\Delta_H = \Delta_L + (q_H - q_L)$.

Case D.1 ($\pi_0 \in [0,1]$, $\pi_L = 0$, $\pi_H \in [0,1]$). We have

$$\Delta_{\lambda_I} = -\pi_0 + q_{\lambda_I}\pi_H.$$

so it is straightforward that $\Delta_L \in [0, q_L]$. The potential values of Δ_H then fall in an interval that we will characterize. The minimum possible value of Δ_H occurs when $\pi_0 = 0$ which is case **S.2** and so $\Delta_H = \frac{q_H}{q_L}\Delta_L$. The maximum possible value of Δ_H occurs when when $\pi_H = 1$, which is **case S.4** and so the maximum value is $\Delta_H = \Delta_L + (q_H - q_L)$.

Summarizing, in Case D.1 we have $\Delta_L \in [0, q_L]$ and $\Delta_H \in \left[\frac{q_H}{q_L}\Delta_L, \Delta_L + (q_H - q_L)\right]$

Case D.2 ($\pi_0 \in [0,1]$, $\pi_L \in [0,1]$, $\pi_H = 1$). We have

$$\Delta_{\lambda_I} = q_{\lambda_I} + (1 - q_{\lambda_I})\,\pi_L - \pi_0$$

so it is straightforward that we may have any $\Delta_L \in [0,1]$. The minimum possible value of Δ_H occurs when $\pi_L = 1$ which implies $\Delta_H = \Delta_L$. The maximum possible value of Δ_H corresponds to the minimum possible value of π_L, which in turn depends on Δ_L. If $\Delta_L \in [0, q_L]$ then the minimum possible value of π_L is 0 and we are in case **S.4,** so $\Delta_H = \Delta_L + (q_H - q_L)$. If $\Delta_L \in [q_L, 1]$ then the minimum possible value of π_L must be > 0; the smallest feasible value corresponds with when $\pi_0 = 0$, so we are in case **S.3** and $\Delta_H = q_H + \left(\frac{1-q_H}{1-q_L}\right)(\Delta_L - q_L)$.

Summarizing, in case D.2 we have we have $\Delta_L \in [0,1]$ and

- if $\Delta_L \in [0, q_L]$ then $\Delta_H = [\Delta_L, \Delta_L + (q_H - q_L)]$
- if $\Delta_L \in [q_L, 1]$ then $\Delta_H = \left[\Delta_L, q_H + \left(\frac{1-q_H}{1-q_L}\right)(\Delta_L - q_L)\right]$

B Equilibrium Characterization

This section proceeds by enumerating all the types of equilibria and deriving existence conditions for each. After this analysis the equilibria are summarized as a function of the primitive parameters.

B.1 Pooling on Effort Equilibria

We consider when pooling on effort is an equilibrium that satisfies D1 (Cho and Kreps 1987). Observe that when the voter observes sabotage, the only information he receives is from sabotage itself (since failure is assured). Consequently, when the saboteur is believed to be pooling on effort, any off-equilibrium path belief about the incumbent's type following sabotage $\tilde{\theta}_I^{0,0}(\cdot) \in [0,1]$ satisfies sequential consistency (Kreps and Wilson 1982). Since the voter's reelection threshold $\bar{\theta}_C(\cdot)$ is $\in (0,1)$, the voter's set of mixed best responses to consistent beliefs off the equilibrium path is *any* reelection probability $\pi_0 \in [0,1]$. D1 thus requires the voter to assign probability weight 1 when one type of incumbent invites deviation for a strictly larger set of $\pi_0 \in [0,1]$.

We now analyze the four popularity conditions.

A very unpopular policy ($\bar{\theta}_C \in \left[\frac{\theta_I q_H}{\theta_I q_H + (1-\theta_I) q_L}, 1\right]$) We have $\pi_H^* = \pi_L^* = 0$ and $\Delta_L, \Delta_H \leq 0$, so it is indeed an equilibrium to pool on effort regardless of the voters off-path best response ($\pi_0^* \in [0,1]$).

A somewhat (un)popular policy ($\bar{\theta}_C \in \left[\frac{\theta_I(1-q_H)}{\theta_I(1-q_H)+(1-\theta_I)(1-q_L)}, \frac{\theta_I q_H}{\theta_I q_H + (1-\theta_I) q_L}\right]$) Then $\pi_H = 1 > \pi_L = 0$ and potential off path behavior is $\pi_0 \in [0,1]$. Now we ask what different values of π_0 imply for Δ_L and Δ_H—using case **S.3** the potential values of (Δ_L, Δ_H) are $\Delta_L \in [0, q_L]$ and $\Delta_H \in \Delta_L + (q_H - q_L)$.

If $\bar{\Delta}_L \geq q_L$ **then this is an equilibrium**; we know that this implies $\bar{\Delta}_H \geq q_H$ and so no off path beliefs can invite deviation; π_0^* may be anything.

If $\bar{\Delta}_L < q_L$, then **this is an equilibrium i.f.f.** $\bar{\Delta}_H \leq \bar{\Delta}_L + (q_H - q_L)$. In this case, the set of π_0 that invite deviation from a high type strictly contains the set that invite deviation from a low type, sabotage will be perceived as perfect good news (applying D1) and cause retention for sure so $\pi_0^* = 1$, and will therefore be undesirable.

Finally, if $\bar{\Delta}_L < q_L$ but $\bar{\Delta}_H > \bar{\Delta}_L + (q_H - q_L)$, then again applying D1 sabotage will be perceived as bad news or $\pi_0^* = 0$, implying $(\Delta_L = q_L, \Delta_H = q_H)$, the bureaucrat will want to deviate to sabotaging both types, and this is not an equilibrium.

Summarizing, for a somewhat unpopular or somewhat popular policy, pooling on effort is an equilibrium i.f.f.

- $\bar{\Delta}_L \geq q_L$ or $\bar{\Delta}_L < q_L$ and $\bar{\Delta}_H \leq \bar{\Delta}_L + (q_H - q_L)$

Equilibrium retention probabilities are $\pi_H^* = 1$, $\pi_L^* = 0$, and $\pi_0^* = 1$.

A very popular policy ($\bar{\theta}_C \in \left[0, \frac{\theta_I(1-q_H)}{\theta_I(1-q_H)+(1-\theta_I)(1-q_L)}\right]$) No news and failure leads to retention ($\pi_L = \pi_H = 1$). Then $\Delta_{\lambda_I} = 1 - \pi_0$ so $\Delta_L \in [0,1]$ and $\Delta_H = \Delta_L$, the set of π_0 that invite deviation from a bad type is always strictly larger than the set inviting deviation from a good type, sabotage should be perceived as *bad news* and cause the policy to be tossed for sure, so $\pi_0^* = 0$, $\Delta_L = \Delta_H = 1$, sabotage will be desirable for both types and this is **not an equilibrium.**

Summary Pooling on effort is an equilibrium that satisfies D1 i.f.f.

- The policy is very unpopular, so $\pi_H^* = \pi_L^* = 0$ and any π_0^*
- The policy is somewhat unpopular or somewhat popular (so $\pi_H^* = 1 > \pi_L^* = 0$), and either $\bar{\Delta}_L \geq q_L$ (with any π_0^*) or $\bar{\Delta}_L < q_L$ and $\bar{\Delta}_H \leq \bar{\Delta}_L + (q_H - q_L)$ (with $\pi_0^* = 1$)

B.2 Pooling on Sabotage Equilibria

We consider when pooling on sabotage is an equilibrium that satisfies a modification of D1 (Cho and Kreps 1987). Specifically, when considering which incumbent type is more likely to invite deviation by the saboteur, we restrict attention to the set of off-equilibrium path mixed strategies by the voter that are best responses to sequentially consistent off-path beliefs (Kreps and Wilson 1982). Unlike the standard signalling game, sequential consistency imposes some constraints on the voter's off-equilibrium-path beliefs because nature sends an additional signal (success or failure) to the voter following the saboteur's move.

It is easily verified that that when the saboteur is believed to be pooling on sabotage, any off equilibrium path belief about the incumbent's type $\tilde{\theta}_I^1(\cdot) \in [0,1]$ *prior to* the observation of success and failure satisfy sequential consistency. However, these beliefs will then be updated following success and failure using Bayes rule and the knowledge that high type incumbents succeed with probability q_H while low types succeed with probability q_L. The set of sequentially consistent beliefs following success and failure are thus

$$\tilde{\theta}_I^{1,1} = \frac{\tilde{\theta}_I^1 q_H}{\tilde{\theta}_I^1 q_H + \left(1-\tilde{\theta}_I^1\right) q_L} \text{ and } \tilde{\theta}_I^{1,0} = \frac{\tilde{\theta}_I^1 (1-q_H)}{\tilde{\theta}_I^1 (1-q_H) + \left(1-\tilde{\theta}_I^1\right)(1-q_L)}$$

for any value of $\tilde{\theta}_I^1 \in [0,1]$. It is straightforward to verify that both $\tilde{\theta}_I^{1,1}$ and $\tilde{\theta}_I^{1,0}$ may each take any value $\in [0,1]$, but $\tilde{\theta}_I^{1,1} = \tilde{\theta}_I^{1,0}$ if and only if $\tilde{\theta}_I^{1,1} = \tilde{\theta}_I^{1,0} = 1$ or $\tilde{\theta}_I^{1,1} = \tilde{\theta}_I^{1,0} = 0$; otherwise $\tilde{\theta}_I^{1,1} > \tilde{\theta}_I^{1,0}$. Consequently, when the saboteur is believed to be pooling on sabotage, the voter's off-equilibrium-path set of mixed best responses to consistent beliefs following effort and success or failure are (i) $\pi_L = 0$ and $\pi_H \in [0,1)$, or (ii) $\pi_L \in (0,1]$ and $\pi_H = 1$.

We now analyze the four popularity conditions.

An unpopular policy ($\bar{\theta}_C \leq \theta_P$) We argue pooling on sabotage is always an equilibrium. If the policy is unpopular then $\pi_0^* = 0$. Using that off-path actions are (i) $\pi_L = 0$ and $\pi_H \in [0,1)$, or (ii) $\pi_L \in (0,1]$ and $\pi_H = 1$ straightforwardly yields the contour of impact probabilities $\Delta_L \in [0, q_L]$ and $\Delta_H = \frac{q_H}{q_L}\Delta_L$, and $\Delta_L \in [q_L, 1]$ and $\Delta_H = q_H + \left(\frac{1-q_H}{1-q_L}\right)(\Delta_L - q_L)$ which is $< \frac{q_H}{q_L}\Delta_L$. Since we know $\bar{\Delta}_H > \frac{q_H}{q_L}\bar{\Delta}_L$, this implies the set of best responses inviting deviation from a high type is strictly larger than the set inviting deviation for a low type, implying effort should be interpreted as good news ($\pi_H^* = \pi_L^* = 1$) and cause the policy to be retained for sure, and is therefore an undesirable deviation, so this is an equilibrium.

A popular policy ($\bar{\theta}_C \geq \theta_P$) Then $\pi_0 = 1$ and pooling on sabotage is not an equilibrium, since sabotage gets the policy retained for sure and also destroys valence.

C (Partially) Separating Equilibria

We begin by ruling out certain types of strategy profiles.

First, we argue that $(e_L > 0, e_H = 0)$ cannot be an equilibrium (including both $e_L \in (0,1)$ and $e_L = 1$, ruling out one type of separating equilibrium). Observe that effort is perfect bad news and causes policy to be tossed for sure ($\pi_L = \pi_H = 0$), so it will be strictly desirable to exert effort for both types, contradicting $e_H = 0$.

We next argue that $(e_L = 1, e_H < 1)$ cannot be an equilibrium. Observe that sabotage is perfect good news and causes the policy to be retained for sure ($\pi_0 = 1$), so effort will weakly decrease the chance policy is retained, so again it will be strictly desirable on both types, contradicting $e_H < 1$.

Last we argue that $(e_L = 0, e_H = 1)$ cannot be an equilibrium; combined with the above this rules out all separating equilibria. If so then effort perfectly reveals the incumbent is good while sabotage perfectly reveals the incumbent is bad; then $\pi_H = \pi_L = 1$ and $\pi_0 = 0$, but then the bureaucrat will strictly prefer to sabotage a good incumbent under our assumptions, contradicting $e_H = 1$.

The remaining possible equilibrium efforts are four types of partially separating equilibria:

(P1) $e_L = 0$, $e_H \in (0,1)$: effort is "perfect good news," sabotage is "imperfect bad news"

(P2) $e_L \in (0,1)$, $e_H = 1$: effort is "perfect bad news," sabotage is "noisy good news"

(P3) $0 < e_L < e_H < 1$: effort is "noisy good news," sabotage is "noisy bad news"

(P4) $0 < e_H < e_L < 1$: effort is "noisy bad news," sabotage is "noisy good news"

We consider each and derive conditions under which it is an equilibrium satisfying D1.

(P1) $e_L = 0, e_H > 0$ Clearly $\pi_H = \pi_L = 1$. We first argue that for this to be an equilibirum requires the incumbent be popular or $\theta_I \geq \bar{\theta}_C$. If they are unpopular then $\pi_0 = 0$ and $\Delta_L = \Delta_H = 1$ and the agent will strictly prefer to sabotage a good policy, contradicting $e_H > 0$.

So suppose the incumbent is popular; we argue that it is always possible to derive an equilibrium of this form, and derive it. First, it is always possible to select e_H to generate principal indifference after sabotage generating $\pi_0 \in [0,1]$, yielding case **S.1** from the preliminary analysis. This requires that

$$\bar{\theta}_C = \frac{\theta_I (1 - e_H)}{\theta_I (1 - e_H) + (1 - \theta_I)} \rightarrow e_H^* = \frac{\theta_I - \bar{\theta}_C}{\left(1 - \bar{\theta}_C\right) \theta_I}$$

Next, in S.1 we have $\Delta_H = 1 - \pi_0$, so to generate saboteur indifference with a high-ability incumbent requires

$$\Delta_H = \bar{\Delta}_H \iff \pi_0 = 1 - \bar{\Delta}_H$$

Finally, we have $\Delta_L = \Delta_H = \bar{\Delta}_H > \bar{\Delta}_L$, so the saboteur strictly prefers to sabotage a low-ability incumbent, supporting $e_L = 0$.

(P2) $e_L \in (0,1)\,,\, e_H = 1$ We have $\pi_0 = 0$. We first argue this cannot be an equilibrium if the incumbent is very popular. If so, then $\pi_H = \pi_L = 1$ (since effort is noisy good news), and the saboteur will strictly prefer to sabotage a high-ability incumbent, contradicting $e_H = 1$.

Next suppose that the incumbent is somewhat (un)popular, implying that $\pi_H = 1$. We argue an equilibrium of this form exists in which $\pi_L \in (0,1)$ i.f.f. $\bar{\Delta}_L \in [q_L, 1]$, and derive the equilibrium. First, it is always possible to select e_L to generate principal indifference after effort and failure so that $\pi_L \in (0,1)$, yielding case **S.3** from the preliminary analysis. This requires that

$$\bar{\theta}_C = \frac{\theta_I\,(1-q_H)}{\theta_I\,(1-q_H) + (1-\theta_I)\,e_L\,(1-q_L)} \rightarrow e_L^* = \frac{\theta_I\,(1-q_H)}{(1-\theta_I)\,(1-q_L)} \Big/ \frac{\bar{\theta}_C}{1-\bar{\theta}_C}$$

Next, in S.3 we must have $\Delta_L \in [q_L, 1]$ and $\Delta_H = q_H + \left(\frac{1-q_H}{1-q_L}\right)(\Delta_L - q_L)$, which is clearly $< \frac{q_H}{q_L}\Delta_L$. So $\Delta_L = \bar{\Delta}_L \iff \bar{\Delta}_L \in [q_L, 1]$, the desired necessary condition. To derive π_L observe that

$$\bar{\Delta}_L = q_L + (1-q_L)\,\pi_L \iff \pi_L = \frac{\bar{\Delta}_L - q_L}{1-q_L}$$

Finally, $\Delta_H = q_H + \left(\frac{1-q_H}{1-q_L}\right)\left(\bar{\Delta}_L - q_L\right) < \frac{q_H}{q_L}\bar{\Delta}_L < \bar{\Delta}_H$, so the saboteur strictly prefers to exert effort for a high ability incumbent, supporting $e_H = 1$.

Finally, suppose that the incumbent is very unpopular, so e_L may be chosen to generate principal indifference after both failure ($\pi_L \in (0,1)$ and $\pi_H = 1$) or success ($\pi_L = 0$ and $\pi_H \in (0,1)$). Using the analysis in the somewhat (un)popular case, an equilibrium of the former type exists i.f.f. $\bar{\Delta}_L \in [q_L, 1]$, and the equilibrium quantities are as previously derived. We now argue that an equilibrium of the latter type exists i.f.f. $\bar{\Delta}_L \in [0, q_L]$. We must select e_L to generate principal indifference after effort and success so that $\pi_H \in (0,1)$, yielding case **S.2** from the preliminary analysis. This requires that

$$\bar{\theta}_C = \frac{\theta_I q_H}{\theta_I q_H + (1-\theta_I)\,e_L q_L} \rightarrow e_L^* = \frac{\theta_I q_H}{(1-\theta_I)\,q_L} \Big/ \frac{\bar{\theta}_C}{1-\bar{\theta}_C}$$

Next, in S.2 we must have that $\Delta_L \in [0, q_L]$ and $\Delta_H = \frac{q_H}{q_L}\Delta_L$. So $\Delta_L = \bar{\Delta}_L \iff \bar{\Delta}_L \in [0, q_L]$, the desired necessary condition. To derive π_H observe that

$$\bar{\Delta}_L = q_L \pi_H \iff \pi_H = \frac{\bar{\Delta}_L}{q_L}$$

Finally, $\Delta_H = \frac{q_H}{q_L}\bar{\Delta}_L < \bar{\Delta}_H$, so the saboteur strictly prefers to exert effort for a high ability incumbent, supporting $e_H = 1$.

Summary There exists an equilibrium with $e_H = 1$ and $e_L \in (0,1)$ i.f.f.

- The incumbent is very unpopular, somewhat unpopular, or somewhat popular and $\bar{\Delta}_L \in [q_L, 1]$. In the equilibrium

$$e_L^* = \frac{\theta_I\,(1-q_H)}{(1-\theta_I)\,(1-q_L)} \Big/ \frac{\bar{\theta}_C}{1-\bar{\theta}_C}\,,\, \pi_0 = 0 < \pi_L = \frac{\bar{\Delta}_L - q_L}{1-q_L} < \pi_H = 1$$

- The incumbent is very unpopular and $\bar{\Delta}_L \in [0, q_L]$. In the equilibrium
$$e_L^* = \frac{\theta_I q_H}{(1-\theta_I)\, q_L} \Big/ \frac{\bar{\theta}_C}{1-\bar{\theta}_C},\ \pi_0 = \pi_L = 0 < \pi_H = \frac{\bar{\Delta}_L}{q_L} < 1$$

(P3) $0 < e_L < e_H < 1$ First observe that when both $e_{\lambda_I} \in (0,1)\ \forall \lambda_I$ we cannot have $\pi_H \in (0,1)$ and $\pi_L \in (0,1)$ since voter posterior beliefs after success are always strictly higher than posteriors after failure (unless effort is perfectly informative). Thus to generate saboteur mixing for both incumbent types requires $\pi_0 \in (0,1)$ and either $0 = \pi_L < \pi_H < 1$ (case D.1) or $0 < \pi_L < 1 = \pi_H$ (case D.2).

We first argue that for an equilibrium with $0 < e_L < e_H < 1$ the following conditions are necessary and sufficient: (a) the incumbent is somewhat popular ($\bar{\theta}_C \in \left[\theta_I, \frac{\theta_I q_H}{\theta_I q_H + (1-\theta_I) q_L}\right]$), (b) reelection probabilities are as in case D.2 ($0 < \pi_L < 1 = \pi_H$), (c) $\bar{\Delta}_L \in [0, q_L]$, and (d) $\bar{\Delta}_H \in \left[\bar{\Delta}_H, \bar{\Delta}_L + (q_H - q_L)\right]$.

If instead the incumbent were very popular then $\pi_H = \pi_L = 1$, a contradiction; if the incumbent were unpopular then $\pi_0 = 0$, also a contradiction. Finally, if the incumbent is somewhat popular then $\pi_H = 1$, so reelection probabilities must be as in case D.2.

Now if the incumbent if somewhat popular then it is always possible to select (e_L, e_H) to generate principal indifference after both sabotage and effort and failure. Equilibrium effort levels solve the following system of equations:

$$\begin{aligned}\frac{\theta_I e_H (1-q_H)}{\theta_I e_H (1-q_H) + (1-\theta_I) e_L (1-q_L)} &= \frac{1}{1 + \frac{(1-\theta_I) e_L (1-q_L)}{\theta_I e_H (1-q_H)}} = \bar{\theta}_C \\ &= \frac{\theta_I (1-e_H)}{\theta_I (1-e_H) + (1-\theta_I)(1-e_L)} = \frac{1}{1 + \frac{(1-\theta_I)(1-e_L)}{\theta_I (1-e_H)}}\end{aligned}$$

which yields

$$\frac{e_L}{e_H} = \left(\frac{1-q_H}{1-q_L}\right)\left(\frac{\theta_I}{1-\theta_I} \Big/ \frac{\bar{\theta}_C}{1-\bar{\theta}_C}\right) \quad \text{and} \quad \frac{1-e_L}{1-e_H} = \frac{\theta_I}{1-\theta_I} \Big/ \frac{\bar{\theta}_C}{1-\bar{\theta}_C}$$

Solving then yields

$$e_L^* = \left(\frac{1-q_H}{q_H - q_L}\right)\frac{(\theta_I - \bar{\theta}_C)}{\bar{\theta}_C (1-\theta_I)} \quad \text{and} \quad e_H^* = \left(\frac{1-q_L}{q_H - q_L}\right)\frac{(\theta_I - \bar{\theta}_C)}{\theta_I (1-\bar{\theta}_C)}.$$

Finally, for the saboteur to mix on both types of incumbents requires that $\Delta_L = \bar{\Delta}_L$ and $\Delta_H = \bar{\Delta}_H$. We argue this implies $\bar{\Delta}_L \in [0, q_L]$, which in turn implies $\bar{\Delta}_H \in \left[\bar{\Delta}_L, \bar{\Delta}_L + (q_H - q_L)\right]$ from the preliminary analysis of case D.2. If instead $\bar{\Delta}_L \in [q_L, 1]$ then we must have $\bar{\Delta}_H \in \left[\bar{\Delta}_L, q_H + \left(\frac{1-q_H}{1-q_L}\right)(\bar{\Delta}_L - q_L)\right]$ (again from the preliminary analysis), but $\bar{\Delta}_H > \frac{q_H}{q_L}\bar{\Delta}_L > q_H + \left(\frac{1-q_H}{1-q_L}\right)(\bar{\Delta}_L - q_L)$, a contradiction. Finally, in case D.2 the retention probabilities are defined by the system $(\pi_L - \pi_0) + q_{\lambda_I}(1-\pi_L) = \bar{\Delta}_{\lambda_I}\ \forall \lambda_I$ and we have

$$\pi_L = \frac{(q_H - \bar{\Delta}_H) - (q_L - \bar{\Delta}_L)}{q_H - q_L} \quad \text{and} \quad \pi_0 = \frac{(1-q_L)(q_H - \bar{\Delta}_H) - (1-q_H)(q_L - \bar{\Delta}_H)}{q_H - q_L}$$

(P4) $0 < e_H < e_L < 1$ We first argue that: (a) the incumbent must be somewhat unpopular ($\bar{\theta}_C \in \left[\frac{\theta_P(1-q_H)}{\theta_P(1-q_H)+(1-\theta_P)(1-q_L)}, \theta_I\right]$), (b) reelection probabilities are as in case D.1 ($\pi_0 \in (0,1)$ and $0 = \pi_L < \pi_H < 1$), (c) $\bar{\Delta}_L \in [0, q_L]$, and (d) $\bar{\Delta}_H \leq \bar{\Delta}_L + (q_H - q_L)$.

As in the analysis in (P3) we must have $\pi_0 \in (0,1)$ and either $0 = \pi_L < \pi_H < 1$ (case D.1) or $0 < \pi_L < 1 = \pi_H$ (case D.2). If the incumbent were very unpopular then we would have $\pi_L = \pi_H = 0$, a contradiction; if she were popular we would have $\pi_0 = 1$, also a contradiction; she must therefore be somewhat unpopular, further implying $0 = \pi_L < \pi_H < 1$ (case D.1).

Now if the incumbent if somewhat unpopular then it is always possible to select (e_L, e_H) to generate principal indifference after both sabotage and effort and failure. Equilibrium effort levels solve the following system of equations:

$$\begin{aligned}\frac{\theta_I e_H q_H}{\theta_I e_H q_H + (1-\theta_I) e_L q_L} &= \frac{1}{1 + \frac{(1-\theta_I) e_L q_L}{\theta_I e_H q_H}} = \bar{\theta}_C \\ &= \frac{\theta_I (1-e_H)}{\theta_I (1-e_H) + (1-\theta_I)(1-e_L)} = \frac{1}{1 + \frac{(1-\theta_I)(1-e_L)}{\theta_I (1-e_H)}}\end{aligned}$$

which yields

$$\frac{e_L}{e_H} = \frac{q_H}{q_L} \cdot \left(\frac{\theta_I}{1-\theta_I} \Big/ \frac{\bar{\theta}_C}{1-\bar{\theta}_C} \right) \quad \text{and} \quad \frac{1-e_L}{1-e_H} = \frac{\theta_I}{1-\theta_I} \Big/ \frac{\bar{\theta}_C}{1-\bar{\theta}_C}$$

Solving yields the interior solution

$$e_{L:}^* = \left(\frac{q_H}{q_H - q_L} \right) \frac{\left(\bar{\theta}_C - \theta_I\right)}{\bar{\theta}_C (1-\theta_I)} \quad \text{and} \quad e_H^* = \left(\frac{q_L}{q_H - q_L} \right) \frac{\left(\bar{\theta}_C - \theta_I\right)}{\theta_I \left(1 - \bar{\theta}_C\right)}$$

Finally, for the saboteur to mix on both types of incumbents requires that $\Delta_L = \bar{\Delta}_L$ and $\Delta_H = \bar{\Delta}_H$. From the preliminary analysis of case D.1 this immediately implies $\bar{\Delta}_L \in [0, q_L]$ and $\bar{\Delta}_H \in \left[\bar{\Delta}_L, \bar{\Delta}_L + (q_H - q_L)\right]$. The retention probabilities are defined by the system $-\pi_0 + q_{\lambda_I} \pi_H = \bar{\Delta}_{\lambda_I}\ \forall \lambda_I$ which yields

$$\pi_H^* = \frac{\bar{\Delta}_H - \bar{\Delta}_L}{q_H - q_L} \quad \text{and} \quad \pi_0^* = \frac{q_L \bar{\Delta}_H - q_H \bar{\Delta}_L}{q_H - q_L}$$

D Additional Proofs

We now provide additional proofs that support stated results in the main text.

Sufficient condition for $\bar{\Delta}_H \leq \bar{\Delta}_L + (q_H - q_L)$

We prove that the inequality in the equilibrium statements for a somewhat (un)popular is a sufficient condition for both $\bar{\Delta}_L \leq q_L$ and $\bar{\Delta}_H \leq \bar{\Delta}_L + (q_H - q_L)$. These latter properties substantially simplify the equilibrium characterization by eliminating many possibilities.

From the definitions we have that

$$q_{\lambda_I} = \delta \bar{\Delta}_{\lambda_p} \left(\frac{U_S}{\gamma_S} - \left(\mathbf{1}_{\lambda_I = H} - \theta_C\right)(q_H - q_L) \right)$$

which is equivalent to

$$q_{\lambda_I} + \delta \bar{\Delta}_{\lambda_p} (q_H - q_L) = \delta \bar{\Delta}_{\lambda_p} \left(\frac{U_S}{\gamma_S} + \theta_C (q_H - q_L) \right)$$

Also observe that that $\bar{\Delta}_{\lambda_p} \leq \Delta_{\lambda_p} \iff$

$$\frac{U_S}{\gamma_S} \geq (\mathbf{1}_{\lambda_I=H} - \theta_C)(q_H - q_L) + \frac{1}{\delta}\frac{q_{\lambda_I}}{\Delta_{\lambda_p}}$$

Now define $\hat{\Delta}_H$ to be the quantity satisfying

$$q_H + \delta q_H (q_H - q_L) = \delta \hat{\Delta}_H \left(\frac{U_S}{\gamma_S} + \theta_C (q_H - q_L) \right)$$

or

$$q_H (1 + \delta (q_H - q_L)) = \delta \hat{\Delta}_H \left(\frac{U_S}{\gamma_S} + \theta_C (q_H - q_L) \right)$$

From the definitions, any value of $\hat{\Delta}_H$ corresponding to a value of $\bar{\Delta}_H < q_H$ must satisfy $\bar{\Delta}_H < \hat{\Delta}_H$. It is also straightforward to see that

$$\frac{\hat{\Delta}_H}{\bar{\Delta}_L} = \frac{q_H (1 + \delta (q_H - q_L))}{q_L} \iff \hat{\Delta}_H = \frac{q_H (1 + \delta (q_H - q_L))}{q_L} \bar{\Delta}_L$$

We now consider when we have $\hat{\Delta}_H \leq \bar{\Delta}_L + (q_H - q_L)$; this requires

$$\frac{q_H (1 + \delta (q_H - q_L))}{q_L} \bar{\Delta}_L \leq \bar{\Delta}_L + (q_H - q_L) \iff \bar{\Delta}_L \leq \frac{q_L}{1 + q_H \delta}$$

(which is stronger than $\bar{\Delta}_L \leq q_L$). From the definition this condition is equivalent to:

$$\frac{U_S}{\gamma_S} \geq \frac{1}{\delta} + (1 - \theta_C)(q_H - q_L) + q_L$$

Further, it is also easily verified that $\bar{\Delta}_H \leq q_H \iff$

$$\frac{U_S}{\gamma_S} \geq \frac{1}{\delta} + (\mathbf{1} - \theta_C)(q_H - q_L)$$

which is a weaker condition, so when the stated condition holds we have $\bar{\Delta}_H < \hat{\Delta}_H < \bar{\Delta}_L + (q_H - q_L)$ and this is sufficient for the desired properties. Finally, if we would like the condition to hold for *all* values of θ_C then we require $\frac{U_S}{\gamma_S} \geq \frac{1}{\delta} + q_H$.

Proof of Proposition 1

Sequential equilibrium (Kreps and Wilson 1982) straightforwardly implies that both on and off the equilibrium path, the voter's beliefs will be computed using Bayes' rule using nature's probabilities of success and failure and ignoring the behavior of the saboteur. Optimal behavior is thus straightforwardly described by the popularity conditions.

To see the incumbent strategy, it is straightforward that the saboteur will never sabotage when the incumbent is very (un)popular (since doing so would have no effect on the probability of retention) or when the saboteur is somewhat popular (since sabotage would be counterproductive and ensure retention).

If the incumbent is somewhat unpopular, the net benefit of exerting effort simply the expected value of the net benefit for each incumbent type:

$$(1 - \theta_I)(q_L \gamma_S + \delta q_L (V(\mathbf{0}, \theta_C; \gamma_S, q) + U(x_S; x_I, x_C)))$$
$$+\theta_I (q_H \gamma_S + \delta q_H (V(\mathbf{1}, \theta_C; \gamma_S, q) + U(x_S; x_I, x_C)))$$

and the saboteur will sabotage i.f.f. this is ≤ 0.

This expression may be rewritten as

$$((1 - \theta_I) q_L + \theta_I q_H) \left(\frac{1}{\delta} - \theta_C (q_H - q_L) + \frac{U(x_S; x_I, x_C)}{\gamma_S} \right) + \theta_I q_H (q_H - q_L) \leq 0$$

which in turn is easily rearranged to the expression in the proposition.

Proof of Propositions 2-5

By the equilibrium characterization and the assumption that $-\frac{U(x_S;x_I,x_C)}{\gamma_S} \geq \frac{1}{\delta}+q_H$ there are three equilibria satisfying D1: (a) pooling on sabotage, (b) pooling on effort, and (c) the partially separating equilibrium (P4) with $0 < e_H < e_L < 1$.

Saboteur

We first show that the saboteur prefers pooling on sabotage to (P4) to pooling on effort.

To see that the saboteur strictly prefers pooling on sabotage to (P4), observe that deviating from her P4 strategy profile to pooling on sabotage yields her equilibrium utility due to the equilibrium indifference conditions; however, this involves the incumbent retained with strictly positive probability, and is therefore strictly worse than the equilibrium with pooling on sabotage in which the incumbent is replaced for sure.

To see that the saboteur strictly prefers (P4) to pooling on effort, observe that deviating from her (P4) strategy to pooling on effort yields her equilibrium utility, but the incumbent is retained after success with probability $\pi_H^{P4} < 1$; this is thus strictly better than the equilibrium with pooling on effort in which an incumbent who succeeds is retained for sure.

Voter

We now show that the voter prefers pooling on effort to P4 to pooling on sabotage.

To see that the voter strictly prefers pooling on effort to (P4), we make a sequence of changes altering the strategy profile in (P4) to that in the pooling on effort equilibrium that each weakly increase her utility. First, changing from $\left(\pi_0^{P4} \in (0,1), \pi_L^{P4} = 0, \pi_H^{P4} \in (0,1); e_L^{P4}, e_H^{P4}\right)$ to $\left(\pi_0 = \pi_L = \pi_H = 0; e_L^{P4}, e_H^{P4}\right)$ does not change the voter's utility due to the (P4) indifference conditions. Next changing to $(\pi_0 = \pi_L = \pi_H = 0; e_L = e_H = 1)$ strictly increases the voters's utility since first period quality increases with no change in selection. Finally, changing to $(0 = \pi_L = \pi_0 < \pi_H = 1; e_L = e_H = 1)$ strictly increases the voter's utility since retention is strictly optimal after success when effort is uninformative.

To see that the voter strictly prefers (P4) to pooling on sabotage, observe that a deviation in (P4) to $\pi_0 = \pi_L = \pi_H = 0$ (always replace) does not change her utility, which involves strictly positive effort levels; this is thus strictly better for the voter than the pooling on sabotage equilibrium which also involves always replacing, but with no effort.

Proof of Proposition 6

From the equilibrium characterization, we have that:

$$\frac{e_L}{e_H} = \frac{q_H}{q_L} \cdot \left(\frac{\theta_I}{1-\theta_I} \Big/ \frac{\bar{\theta}_C}{1-\bar{\theta}_C} \right) \quad \text{and} \quad \frac{1-e_H}{1-e_L} = \frac{\bar{\theta}_C}{1-\bar{\theta}_C} \Big/ \frac{\theta_I}{1-\theta_I}$$

The ratio $\frac{e_L}{e_H} \geq 1$ reflects the extent to which effort is "bad news" while the ratio $\frac{1-e_H}{1-e_L} \geq 1$ reflects the extent to which sabotage is "good news." Now let $R\left(\bar{\theta}_C, \theta_I\right) = \frac{\bar{\theta}_C}{1-\bar{\theta}_C} \Big/ \frac{\theta_I}{1-\theta_I}$; it is easily verified that this increases from 1 to $\frac{q_H}{q_L}$ as $\bar{\theta}_C$ increases from θ_I to $\tilde{\theta}_I^1$. Rewriting we have that:

$$\frac{e_L}{e_H} = \frac{q_H/q_L}{R\left(\bar{\theta}_C, \theta_I\right)} \quad \text{and} \quad \frac{1-e_H}{1-e_L} = R\left(\bar{\theta}_C, \theta_I\right)$$

First observe by multiplying the two equations that:

$$\frac{e_L}{1-e_L} \Big/ \frac{e_H}{1-e_H} = \frac{q_H}{q_L}$$

This immediately yields that e_L and e_H must move strictly in the same direction as a function of $R\left(\bar{\theta}_C, \theta_I\right)$; otherwise the LHS could not be constant.

Next observe that $e_H = R\left(\bar{\theta}_C, \theta_I\right) \frac{q_L}{q_H} e_L$ and $1 - e_H = R\left(\bar{\theta}_C, \theta_I\right)(1 - e_L)$ so summing the equations yields:

$$1 = R\left(\bar{\theta}_C, \theta_I\right)\left(1 - \left(1 - \frac{q_L}{q_H}\right) e_L\right)$$

Thus e_L (and from the preceding e_H) are strictly increasing in $R\left(\bar{\theta}_C, \theta_I\right)$, which is in turn strictly decreasing in θ_I and strictly increasing in $\bar{\theta}_C$, which in turn is strictly increasing in θ_C and γ_V and strictly decreasing in $U\left(x_V; \cdot\right)$.

Proof of Proposition 7

Recall that

$$\bar{\Delta}_{\lambda_i} = \frac{q_{\lambda_I}}{\delta\left(B - 1_{\lambda_I = H}\left(q_H - q_L\right)\right)}$$

where $B = \delta\left(\frac{-U(x_S; x_I, x_C)}{\gamma_S} + \theta_C\left(q_H - q_L\right)\right) > q_H - q_L$ by assumption (so $\bar{\Delta}_H < 1$). Now from the equilibrium characterization we have that

$$\pi_H = \frac{\bar{\Delta}_H - \bar{\Delta}_L}{q_H - q_L} \quad \text{and} \quad \pi_0 = \frac{q_L \bar{\Delta}_H - q_H \bar{\Delta}_L}{q_H - q_L}$$

Substituting in the definitions and algebra yields that

$$\pi_H = \frac{1 + \frac{q_L}{B}}{\delta\left(B - \left(q_H - q_L\right)\right)} \text{ and } \pi_0 = \frac{q_L q_H}{\delta\left(B - \left(q_H - q_L\right)\right) B}.$$

Both quantities are straightforwardly decreasing in B and δ. It is also easily verified that

$$\pi_H - \pi_0 = \frac{1 + \frac{q_L(1-q_H)}{B}}{\delta\left(B - \left(q_H - q_L\right)\right)}$$

Thus all three quantities are straightforwardly decreasing in B and δ.

Proof of Proposition 8

Follows immediately from the equilibrium characterization.

Proof of Proposition 9

By the equilibrium characterization there are three equilibria satisfying D1: (1) pooling on effort, (2) pooling on sabotage, and (3) the partially separating equilibrium (P2) with $e_H = 1$ and $e_L \in (0, 1)$; the assumption also yields $0 = \pi_0 = \pi_L < \pi_H < 1$.

We now argue that pooling on effort is Pareto dominant. Pareto dominance of pooling on effort to pooling on sabotage is straightforward; both involve the incumbent being replaced with probability 1, and holding retention decisions fixed both players prefer higher effort to lower effort.

We next compare pooling on effort to (P2). With pooling on effort, we have $\pi_L = \pi_H = 0$ and the incumbent is always replaced. In (P2), we have equilibrium $(\pi_0^*, \pi_L^*, \pi_H^*)$ and (e_L^*, e_H^*). To see that the saboteur strictly prefers the equilibrium with pooling on effort, observe that the retention probabilities yield indifference over effort on a low quality incumbent, so the saboteur gets the same utility by deviating to pooling on effort $(e_L = 1, e_H = 1)$ with $(\pi_0^*, \pi_L^*, \pi_H^*)$, which involves retention with strictly positive probability and is therefore strictly worse.

To see that the voter strictly prefers the equilibrium with pooling on effort, observe that the voter still gets her equilibrium utility by deviating to always replace given the bureaucrat's equilibrium effort levels, which in turn is worse than always replacing with maximum effort by the bureaucrat.

Proof of Proposition 10

By the equilibrium characterization and the assumption that $-\frac{U(x_S;x_I,x_C)}{\gamma_S} \geq \frac{1}{\delta} + q_H$ there are three equilibria satisfying D1: (1) the partially separating equilibrium (P1) with $e_L = 0$ and $e_H \in (0,1)$, $0 < \pi_0 < 1 = \pi_L = \pi_H$, (2) pooling on effort $(0 = \pi_L < \pi_H = 1)$, and (3) the partially separating equilibrium (P3) with $0 < e_L < e_H < 1$ and $\pi_0 \in (0,1)$, $\pi_L \in (0,1)$, $\pi_H = 1$.

We first compare pooling on effort to (P3). For the saboteur, in (P3) a deviation to pooling on effort would still yield her equilibrium utility but with $\pi_L^* > 0$, so her equilibrium utility is strictly worse in (P3).

For the voter, we make a sequence of changes altering the strategy profile in (P3) to that in the pooling on effort that each weakly increase her equilibrium utility. First, changing from $\left(\pi_0^{P3} \in (0,1), \pi_L^{P3} \in (0,1), \pi_H = 1; e_L^{P3}, e_H^{P3}\right)$ to $\left(\pi_0 = \pi_L = \pi_H = 1; e_L^{P3}, e_H^{P3}\right)$ does not change the voter's utility due to the (P3) indifference conditions. Next changing to $(\pi_0 = \pi_L = \pi_H = 1; e_L = e_H = 1)$ strictly increases the voters's utility since first period quality increases with no change in selection. Finally, changing to $(0 = \pi_L < \pi_0 = \pi_H = 1; e_L = e_H = 1)$ strictly increases the voter's utility since replacement is strictly optimal after failure when effort is uninformative.

We next compare (P3) to (P1). For the saboteur, a deviation to the both equilibrium effort levels in (P1) would yield her (P3) equilibrium utility holding retention probabilities fixed. We next argue that the equilibrium retention probabilities in P1 are uniformly higher, implying that the saboteur is worse off in the (P1) equilibrium than in the (P3) equilibrium. Clearly retention probabilities are higher in (P1) after success and failure; we need only argue that the retention probability is also higher after sabotage. From the equilibrium characterizations we have that

$$\pi_0^{P1} = 1 - \bar{\Delta}_H \text{ and } \pi_0^{P3} = \left(q_H + (1 - q_H)\,\pi_L^{P3}\right) - \bar{\Delta}_H$$

which shows the desired property since $q_H + (1 - q_H)\,\pi_L^{P3} < 1$.

For the voter, a deviation to always retain in (P1) still yields her (P1) equilibrium utility, and a deviation to always retain in (P3) still yields her (P3) equilibrium utility. Thus, it sufficies to show $e_{\lambda_P}^{P3} > e_{\lambda_P}^{P1}\ \forall \lambda_P$. We immediately have $e_L^{P3} > 0 = e_L^{P3}$. In addition, in both equilibria $\pi_0 \in (0,1)$ requires

$$\frac{\theta_I\,(1 - e_H)}{(1 - \theta_I)\,(1 - e_L)} = \frac{\bar{\theta}_C}{1 - \bar{\theta}_C},$$

but this immediately yields $e_L^{P3} > e_L^{P1} \rightarrow e_H^{P3} > e_H^{P1}$.