Unpublished Mathematical Appendix for: Public Pressure and the Heterogeneous Effects of Voluntary Pollution Abatement

Ruohao Zhang[a,*], Neha Khanna[b]

[a] *Department of Agricultural Economics, Sociology, and Education, The Pennsylvania State University, University Park, PA*

[b] *Department of Economics, Binghamton University, Binghamton, NY*

June 14, 2023

1 Baseline Model

1.1 Model Overview

We characterize each firm by θ, which represents firm characteristics such as managerial skill, size, energy type, equipment or ownership, and is correlated with the firm's emissions and participation choice. Let a firm's abatement cost be a function of firm characteristics and emissions, i.e, $F(\theta, e)$, where e is the firm's emissions due to its choice of input bundle. Let the firm's marginal abatement cost function be $f(\theta, e) = \frac{\partial F(\theta,e)}{\partial e}$. Abatement costs decrease with emissions, so $f(\theta, e) < 0$. Marginal abatement cost decreases as the amount of abatement decreases, which is equivalent to assuming the marginal abatement cost function is an increasing function of emissions, so $\frac{\partial f(\theta,e)}{\partial e} > 0$. For simplicity of illustration, θ is a scalar and denotes the "abatement efficiency" of the firm (Zhou et al., 2020). Specifically, a higher value of θ means lower total abatement cost and marginal abatement cost at any emission level, so $\frac{\partial F(\theta,e)}{\partial \theta} < 0$ and $\frac{\partial^2 F(\theta,e)}{\partial e \partial \theta} > 0$.[1] Because lower e means more abatement, these assumptions imply that both total

* Corresponding author. E-mail: rjz5261@psu.edu.

[1] See Appendix section 1.4.1 for details. Because $\frac{\partial F(\theta,e)}{\partial e} < 0$, a lower marginal abatement cost means $\frac{\partial F(\theta,e)}{\partial e}$ becomes less negative, so $\frac{\partial^2 F(\theta,e)}{\partial e \partial \theta} > 0$.

abatement cost and marginal abatement cost increase with every one additional unit reduction of emission.

The firm's total pollution cost function is

$$C(x, z, e, p) = R(x, e) + M(z, e, p) \tag{1}$$

where p is the firm's participation status with $p = 1$ representing participation in the voluntary program and $p = 0$ representing non-participation. $R(x, e)$ is the emission cost due to regulation pressure, and $M(z, e, p)$ is the emission cost from public pressure. e is the firm's emission level, x is a vector of firm and regulator characteristics determining regulation pressure, and z is a vector of industry and stakeholders' characteristics related to public pressure, which includes industry reputation, media and press releases, market share of green consumers and investors, and local communities' characteristics.[2] Similar to θ, for simplicity of model illustration, x and z are scalars representing the stringency of regulation pressure and public pressure.[3] A greater value of x and z means the firm faces more stringent regulators and stakeholders with regards to its environmental performance. For the regulation component of emission cost, we assume that $\frac{\partial R}{\partial x} > 0$ and $\frac{\partial R}{\partial e} > 0$. For the public component of emission cost, we assume that $\frac{\partial M}{\partial z} > 0$ and $\frac{\partial M}{\partial e} > 0$. The relationship between $M(z, e, p = 0)$ and $M(z, e, p = 1)$ is ambiguous depending on the value of e.

We disaggregate $M(z, e, p)$ so that $C(x, z, e, p)$ becomes:

$$C(x, z, e, p) = R_0(x, e) + M_0(z, e) + p(M_P(e) + G(e)), \tag{2}$$

where $R_0(x, e)$ and $M_0(z, e)$ are the pre-program (also non-participation) emission costs due to regulation and public pressure, respectively. $M_P(e)$ is the change in the public pressure component of emission cost associated with the positive environmental signal provided by the

[2] This is a simplified model. In reality, the elements in θ, x and z may contain overlapping elements. In the absence of pre-existing regulation pressure, x takes the value such that $R(x, e) = 0$ for all e.

[3] We assume θ, x and z are scalars because it ensures independent variations across the three functions. Without independent variation, suppose there is a single variable k that affects both the marginal emission cost and the marginal abatement cost. Then a marginal change in k causes both the marginal emission cost and the marginal abatement cost curves to move together, and the marginal effect of k on emissions is ambiguous. In the theoretical model we assume that such variables are fixed so that we can derive the comparative statistics. In the empirical analysis we condition our outcome on such variation (i.e., HAP/TRI Ratio).

firm's participation in the program, so $M_P(e) \leq 0$. The information disclosure engendered by the voluntary pollution abatement program can be leveraged by public scrutiny and raises the public pressure on participating firms if their environmental performance is worse than the public expectation. This additional change in the emission cost is captured by $G(e)$, which is non-negative and increases in e.

We make the following assumptions regarding each component of the total pollution cost function. First, we assume that the pre-program total and marginal emission cost from regulation pressure increase both in the effectiveness of regulation pressure x and in the firm's emissions e, so that $\frac{\partial R_0(x,e)}{\partial x} > 0$, $\frac{\partial r_0(x,e)}{\partial x} > 0$, $\frac{\partial R_0(x,e)}{\partial e} > 0$, and $\frac{\partial r_0(x,e)}{\partial e} > 0$, where $r_0(x,e) = \frac{\partial R_0(x,e)}{\partial e}$. Second, we assume that the pre-existing total and marginal emission cost from public pressure increases in the effectiveness of public pressure z and the firm's emission level e, so $\frac{\partial M_0(z,e)}{\partial z} > 0$, $\frac{\partial m_0(z,e)}{\partial z} > 0$, $\frac{\partial M_0(z,e)}{\partial e} > 0$, and $\frac{\partial m_0(z,e)}{\partial e} > 0$, where $m_0(z,e) = \frac{\partial M_0(z,e)}{\partial e}$. Third, we assume that, holding constant the risk of being label a greenwasher, there is a decline in public pressure if the firm chooses to participate in the program (a downward shift to MC'_P in the main paper Figure 1). However, the decline is smaller when emissions are low than when emissions are high and it gradually converges to zero with emissions, so that $\frac{\partial M_P(e)}{\partial e} < 0$ and $\frac{\partial m_P(e)}{\partial e} > 0$, where $m_P(e) = \frac{\partial M_P(e)}{\partial e}$. Fourth, letting $g(e) = \frac{\partial G(e)}{\partial e}$, we have $\frac{\partial G(e)}{\partial e} > 0$ and $\frac{\partial g(e)}{\partial e} > 0$, because the change in the cost of emissions due to the additional public scrutiny of participants is an increasing function of firm emissions. All the assumptions are subject to $\frac{\partial C(x,z,e,p)}{\partial e} \geq 0$, since the marginal emission cost is always non-negative. Therefore, the marginal emission cost can be written as

$$\begin{aligned} c(x,z,e,p) =& r(x,e) + m(z,e,p) \\ =& r_0(x,e) + m_0(z,e) + p(m_P(e) + g(e)). \end{aligned} \tag{3}$$

Based on the above properties we have, $\frac{\partial c(x,z,e,p)}{\partial e} > 0$, regardless of the participation status p. That is the marginal emission cost is increasing in emissions regardless of whether the firm participates in the program or not. Should the firm choose to participate in the voluntary pollution abatement program, the portion of marginal emission cost generated by the public scrutiny of its environmental performance increases rapidly with emissions, such that it eventually exceeds the difference between MC_N and MC'_P in the main paper Figure 1 at some threshold emission

level (labeled by e'' in the main paper Figure 2). Accordingly, we have the following property:

$$\frac{\partial g(e)}{\partial e} > -\frac{\partial m_P(e)}{\partial e}, \tag{4}$$

and $c(x, z, e, p = 1) > c(x, z, e, p = 0)$ if and only if $e > e''$.

The firm makes participation and emission decisions to minimize its total environmental cost described above. As mentioned in the introduction of the paper, we define a participating firm to be a free-rider of the voluntary pollution abatement program if its emissions are higher as a participant than as a non-participant, shown as the type 2 firms in the main paper Figure 2.

1.2 Optimal Emissions

A firm seeks to minimize its total environmental cost $D(\theta, x, z, e, p)$, which is the sum of total abatement cost, total emission cost, and a fixed cost if the firm decides to participate in the voluntary pollution abatement program:

$$\min_{e,p} D(\theta, x, z, e, p) = F(\theta, e) + C(x, z, e, p) + pc_F, \tag{5}$$

where c_F is the fixed cost of participation. In the first stage the firm identifies its optimal emissions, $e^*(\theta, x, z, p)$, under both participation and non-participation, according to the first order condition

$$f(\theta, e) + c(x, z, e, p) = 0, \ \Rightarrow \ e^* = e^*(\theta, x, z, p). \tag{6}$$

The optimal emission level $e^*(\theta, x, z, p)$ is determined by continuous variables x, z, θ and binary participation status variable p. Based on our assumptions and model properties, we have the following comparative statistics describing the marginal effect of each variable on the optimal emission level. First, for the continuous variables x, z and θ, it can be shown that $\frac{\partial e^*}{\partial x} < 0$, $\frac{\partial e^*}{\partial z} < 0$, and $\frac{\partial e^*}{\partial \theta} < 0$ (see Appendix 1.4.2). In other words, our model indicates that regardless of participation status, firm's emissions decline when there is an increase in regulation pressure, public pressure, or abatement efficiency. The impact of the binary participation variable p on the optimal emission level is labeled the participation effect, and depends on the threshold

emission level e'' at which MC_P crosses MC_N, so that

$$c(x, z, e'', p = 1) = c(x, z, e'', p = 0), \tag{7}$$

which is equivalent to $m_P(e'') + g(e'') = 0$. Let $e^N = e^*(\theta, x, z, p = 0)$ and $e^P = e^*(\theta, x, z, p = 1)$ be the two potential optimal emission levels for the same firm under different participation status, and $\Delta e^*(\theta, x, z) = e^*(\theta, x, z, p = 1) - e^*(\theta, x, z, p = 0)$ be the difference between these two potential emission levels or the participation effect. It can be shown that $e^N > e''$ if and only if $e^P > e''$, and $e^N \leq e''$ if and only if $e^P \leq e''$ (see Appendix 1.4.3). Let $\tilde{e}(\theta, x, z)$ be the observed emission level regardless of the firm's participation decision. In particular, $\tilde{e}(\theta, x, z) = e^P$ if the firm participates, and $\tilde{e}(\theta, x, z) = e^N$ if the firm does not participate. Because $\tilde{e}(\theta, x, z)$ is a special case of $e^*(\theta, x, z, p)$, it has the same properties as $e^*(\theta, x, z, p)$: $\frac{\partial \tilde{e}}{\partial x} < 0$, $\frac{\partial \tilde{e}}{\partial z} < 0$, and $\frac{\partial \tilde{e}}{\partial \theta} < 0$. It can be shown that $\Delta e^*(\theta, x, z) < 0$ if and only if $\tilde{e}(\theta, x, z) > e''$, and $\Delta e^*(\theta, x, z) \geq 0$ if and only if $\tilde{e}(\theta, x, z) \leq e''$ (see Appendix 1.4.3 for proofs). These results lead to our first proposition:

Proposition 1. There exists a threshold emission level e'' such that firms with emissions lower than e'' emit more pollution if participating than not participating in the voluntary pollution abatement program, and vice versa (see Appendix 1.4.3 for proofs).

Proposition 1 identifies the free-riders of the voluntary pollution abatement program: these are the firms with pre-program emissions less than e'' that have higher emissions if they choose to participate than if they do not.

1.3 Participation Incentives

A firm makes its participation decision by comparing the optimal total costs under participation versus non-participation. If the firm participates, the optimal total environmental cost is

$$\begin{aligned} D^P &= F(\theta, e^P) + C(x, z, e^P, p = 1) + c_F \\ &= F(\theta, e^*(\theta, x, z, p = 1)) + C(x, z, e^*(\theta, x, z, p = 1), p = 1) + c_F. \end{aligned} \tag{8}$$

If the firm does not participate, the optimal total environmental cost is

$$\begin{aligned} D^N &= F(\theta, e^N) + C(x, z, e^N, p = 0) \\ &= F(\theta, e^*(\theta, x, z, p = 0)) + C(x, z, e^*(\theta, x, z, p = 0), p = 0). \end{aligned} \tag{9}$$

Denoting ΔD^* as the difference in the optimal cost between non-participation and participation, we have

$$\Delta D^* = \Big(F(\theta, e^N) + C(x, z, e^N, p = 0)\Big) - \Big(F(\theta, e^P) + C(x, z, e^P, p = 1)\Big) - c_F. \tag{10}$$

A firm will participate if $\Delta D > 0$, and will not participate if $\Delta D \leq 0$. A larger ΔD implies a greater participation incentive. Recall that $\tilde{e}(\theta, x, z)$ is the firm's emissions regardless of participation status. Applying the envelope theorem, we obtain the following results. If $\tilde{e} > e''$, we have $\frac{\partial \Delta D}{\partial x} > 0$, $\frac{\partial \Delta D}{\partial \theta} > 0$, and $\frac{\partial \Delta D}{\partial z} > 0$. If $\tilde{e} \leq e''$, we have $\frac{\partial \Delta D}{\partial x} \leq 0$, $\frac{\partial \Delta D}{\partial \theta} \leq 0$, and $\frac{\partial \Delta D}{\partial z} \leq 0$ (see Appendix 1.4.4 for details). In conclusion, our model gives the following proposition:

Proposition 2. If the outcome emission level $\tilde{e} \leq e''$, then a marginal decrease in "abatement efficiency", regulation pressure or public pressure, *ceteris paribus*, increases the probability of participation, and vice versa.

1.4 Proof of Propositions

1.4.1 Properties of abatement cost function

Let a be the quantity of abatement, and $\bar{e}$ be the maximum emissions. Then $e = \bar{e} - a$ and the abatement cost function can be written as

$$F(\theta, e) = F(\theta, \bar{e} - a).$$

The abatement cost increases in abatement but decreases in "abatement efficiency", so that $\frac{\partial F(\theta, \bar{e}-a)}{\partial a} > 0$ and $\frac{\partial F(\theta, \bar{e}-a)}{\partial \theta} < 0$. Furthermore, we assume the marginal abatement cost increases in abatement, but decreases in "abatement efficiency", so that $\frac{\partial^2 F(\theta, \bar{e}-a)}{\partial a^2} > 0$, $\frac{\partial^2 F(\theta, \bar{e}-a)}{\partial a \partial \theta} < 0$.

Because $e = \bar{e} - a$, $\frac{\partial a}{\partial e} < 0$, we have the following equations:

$$\begin{aligned} f(\theta, e) = & \frac{\partial F(\theta, e)}{\partial e} = \frac{\partial F(\theta, \bar{e} - a)}{\partial a} \frac{\partial a}{\partial e} < 0, \\ & \frac{\partial f(\theta, e)}{\partial e} = \frac{\partial^2 F(\theta, \bar{e} - a)}{\partial a^2} \Big(\frac{\partial a}{\partial e}\Big)^2 > 0, \\ & \frac{\partial F(\theta, e)}{\partial \theta} = \frac{\partial F(\theta, \bar{e} - a)}{\partial \theta} < 0, \\ & \frac{\partial^2 F(\theta, e)}{\partial e \partial \theta} = \frac{\partial^2 F(\theta, \bar{e} - a)}{\partial a \partial \theta} \frac{\partial a}{\partial e} > 0. \end{aligned}$$

1.4.2 Comparative statistics for continuous variables

Taking the first derivative of equation 6 with respect to continuous variables (x, z, θ,), we get:

$$\begin{aligned} \theta: & \ \frac{\partial c(x, z, e^*, p)}{\partial e^*} \frac{\partial e^*}{\partial \theta} + \frac{\partial f(\theta, e^*)}{\partial e^*} \frac{\partial e^*}{\partial \theta} + \frac{\partial f(\theta, e^*)}{\partial \theta} = 0 \\ x: & \ \frac{\partial c(x, z, e^*, p)}{\partial e^*} \frac{\partial e^*}{\partial x} + \frac{\partial f(\theta, e^*)}{\partial e^*} \frac{\partial e^*}{\partial x} + \frac{\partial r_0(x, e^*)}{\partial x} = 0 \\ z: & \ \frac{\partial c(x, z, e^*, p)}{\partial e^*} \frac{\partial e^*}{\partial z} + \frac{\partial f(\theta, e^*)}{\partial e^*} \frac{\partial e^*}{\partial z} + \frac{\partial m_0(z, e^*)}{\partial z} = 0. \end{aligned} \tag{11}$$

This leads to the following results:

- $\frac{\partial c(x,z,e,p)}{\partial e} > 0$, $\frac{\partial f(\theta,e)}{\partial e} > 0$, and $\frac{\partial f(\theta,e)}{\partial \theta} > 0$. $\rightarrow \frac{\partial e^*}{\partial \theta} < 0$.
- $\frac{\partial c(x,z,e,p)}{\partial e} > 0$, $\frac{\partial f(\theta,e)}{\partial e} > 0$, and $\frac{\partial r_0(x,e)}{\partial x} > 0$. $\rightarrow \frac{\partial e^*}{\partial x} < 0$.
- $\frac{\partial c(x,z,e,p)}{\partial e} > 0$, $\frac{\partial f(\theta,e)}{\partial e} > 0$, and $\frac{\partial m_0(z,e)}{\partial z} > 0$. $\rightarrow \frac{\partial e^*}{\partial z} < 0$.

1.4.3 Comparative statistics for participation status and proof of Proposition 1

Because p is a binary variable, we use the following integral to show how each term in equation 6 changes due to the change of p between 0 and 1:

1. $p = 0 \rightarrow p = 1$:

$$\int_{e^N}^{e^P} \frac{\partial c(x, z, e, p = 0)}{\partial e} de + \int_{e^N}^{e^P} \frac{\partial f(\theta, e)}{\partial e} de + \big[c(x, z, e^P, p = 1) - c(x, z, e^P, p = 0)\big] = 0,$$

2. $p = 1 \rightarrow p = 0$:

$$\int_{e^P}^{e^N} \frac{\partial c(x, z, e, p = 1)}{\partial e} de + \int_{e^P}^{e^N} \frac{\partial f(\theta, e)}{\partial e} de + \big[c(x, z, e^N, p = 0) - c(x, z, e^N, p = 1)\big] = 0,$$

(12)

Because $\frac{\partial c(x,z,e,p)}{\partial e} > 0$ and $\frac{\partial f(\theta,e)}{\partial e} > 0$, $\Delta e^*(\theta,x,z) = e^P - e^N < 0$ if and only if

$$\begin{aligned} &\int_{e^N}^{e^P} \frac{\partial c(x,z,e,p=0)}{\partial e} de + \int_{e^N}^{e^P} \frac{\partial f(\theta,e)}{\partial e} de < 0, \\ &\textbf{or} \\ &\int_{e^P}^{e^N} \frac{\partial c(x,z,e,p=0)}{\partial e} de + \int_{e^P}^{e^N} \frac{\partial f(\theta,e)}{\partial e} de > 0, \end{aligned} \tag{13}$$

which is equivalent to

$$\begin{aligned} &c(x,z,e^P,p=1) - c(x,z,e^P,p=0) \\ &= m_P(e^P) + g(e^P) > 0, \\ &\textbf{or} \\ &c(x,z,e^N,p=0) - c(x,z,e^N,p=1) \\ &= -m_P(e^N) - g(e^N) < 0. \end{aligned} \tag{14}$$

When emissions are at the threshold e'', we have $g(e'') = -m_P(e'')$. Besides, $\frac{\partial g(e)}{\partial e} > -\frac{\partial m_P(e)}{\partial e}$. Therefore, for any $\hat{e} > e''$ it is always true that $g(\hat{e}) > -m_P(\hat{e})$, and vice versa. It can be proved that $e^N > e''$ if and only if $e^P > e''$, and $e^N \leq e''$ if and only if $e^P \leq e''$, because the two conditions in equation 13 and 14 are equivalent to each other.

Since $e^N > e''$ and $e^P > e''$ are equivalent, $e^N \leq e''$ is also equivalent to $e^P \leq e''$. Regardless of the observed emission level $\tilde{e}$, we always have $\Delta e^*(\theta,x,z) < 0$ if and only if $\tilde{e} > e''$, and $\Delta e^*(\theta,x,z) \geq 0$ if and only if $\tilde{e} \leq e''$. In other words, $e^N > e^P$ if and only if $\tilde{e} > e''$, and $e^N \leq e^P$ if and only if $\tilde{e} \leq e''$.

1.4.4 Proof of Proposition 2

Using the envelope theorem, we have the following properties:

- $\frac{\partial D^*}{\partial \theta} = \frac{\partial F(\theta,e^*)}{\partial \theta} + [f(\theta,e^*) + c(x,z,e^*,p)]\frac{\partial e^*}{\partial \theta} = \frac{\partial F(\theta,e^*)}{\partial \theta} < 0,$
- $\frac{\partial D^*}{\partial x} = \frac{\partial C(x,z,e^*,p)}{\partial x} + [f(\theta,e^*) + c(x,z,e^*,p)]\frac{\partial e^*}{\partial x} = \frac{\partial R_0(x,e^*)}{\partial x} > 0,$
- $\frac{\partial D^*}{\partial z} = \frac{\partial C(x,z,e^*,p)}{\partial z} + [f(\theta,e^*) + c(x,z,e^*,p)]\frac{\partial e^*}{\partial z} = \frac{\partial m_0(z,e^*)}{\partial z} > 0,$

Let $\Delta D^* = D^*(\theta,x,z,P=0) - D^*(\theta,x,z,P=1)$. Then we have

$$\frac{\partial \Delta D^*}{\partial x} = \frac{\partial R_0(x,e^N)}{\partial x} - \frac{\partial R_0(x,e^P)}{\partial x}, \tag{15}$$

$$\frac{\partial \Delta D^*}{\partial \theta} = \frac{\partial F(\theta, e^N)}{\partial x} - \frac{\partial F(\theta, e^P)}{\partial x}, \tag{16}$$

$$\frac{\partial \Delta D^*}{\partial z} = \frac{\partial M_0(z, e^N)}{\partial z} - \frac{\partial M_0(z, e^P)}{\partial z}, \tag{17}$$

We show in Appendix A.3 that $\tilde{e} > e''$ is equivalent to $e^N > e^P$, and $\tilde{e} \leq e''$ is equivalent to $e^N \leq e^P$. Thus we get the following results:

- Because $\frac{\partial^2 R_0(x,e)}{\partial x \partial e} > 0$, we have $\frac{\partial \Delta D^*}{\partial x} > 0$ if $\tilde{e} > e''$, and $\frac{\partial \Delta D^*}{\partial x} \leq 0$ if $\tilde{e} \leq e''$.
- Because $\frac{\partial^2 F(\theta,e)}{\partial \theta \partial e} > 0$, we have $\frac{\partial \Delta D^*}{\partial \theta} > 0$ if $\tilde{e} > e''$, and $\frac{\partial \Delta D^*}{\partial \theta} \leq 0$ if $\tilde{e} \leq e''$.
- Because $\frac{\partial^2 M_0(z,e)}{\partial z \partial e} > 0$, we have $\frac{\partial \Delta D^*}{\partial z} > 0$ if $\tilde{e} > e''$, and $\frac{\partial \Delta D^*}{\partial z} \leq 0$ if $\tilde{e} \leq e''$.

2 Model Extension 1: Reallocation of Regulation Resources

Under a reallocation of regulation resources, the regulation pressure experienced by a firm changes, regardless of whether the firm participates in the voluntary pollution abatement program or not (though the direction and the magnitude of the change may be different). The firm's total pollution cost function then becomes

$$\begin{aligned} C(x, z, e, p) =& R_0(x, e) + M_0(z, e) \\ &+ p(R_P(e) + M_P(e) + G(e)) + (1 - p)R_N(e). \end{aligned} \tag{18}$$

$R_P(e)$ and $R_N(e)$ represent the change in emission cost due to the change in regulation pressure. Under participation, the firm experiences a decrease in regulation pressure (a downward shift from $\widetilde{MC}_P$ to MC_P in Online Appendix Figure A.1), and the shift is greater when emissions are lower and gradually converges to zero as emissions go to infinity. So we have $R_P(e) \leq 0$, $\frac{\partial R_P(e)}{\partial e} \leq 0$, and $\frac{\partial r_P(e)}{\partial e} \geq 0$, where $r_P(e) = \frac{\partial R_P(e)}{\partial e}$. Recall that under non-participation, the firm will experience an increase in regulation pressure (a upward shift from $\widetilde{MC}_N$ to MC_N Online Appendix Figure A.1), and the shift is larger when emissions are higher. So we have $R_N(e) \geq 0$, $\frac{\partial R_N(e)}{\partial e} \geq 0$, and $\frac{\partial r_N(e)}{\partial e} \geq 0$, where $r_N(e) = \frac{\partial R_N(e)}{\partial e}$. Again, these additional assumptions are subject to $\frac{\partial C(x,z,e,p)}{\partial e} \geq 0$, since the marginal emission cost is always non-negative. Therefore,

the marginal emission cost can be written as

$$\begin{aligned} c(x,z,e,p) =& r(x,e,p) + m(z,e,p) \\ =& r_0(x,e) + m_0(z,e) + p(r_P(e) + m_P(e) + g(e)) \\ &+ (1-p) r_N(e). \end{aligned} \tag{19}$$

Let $\Delta MC''(e) = r_N(e) - r_P(e) - m_P(e)$, then we have the following property:

$$\frac{\partial g(e)}{\partial e} > \frac{\partial \Delta MC''(e)}{\partial e}, \tag{20}$$

and $c(x,z,e,p=1) > c(x,z,e,p=0)$ if and only if $e > e''$.

A firm seeks to minimize its total environmental cost

$$\min_{e,p} D(\theta,x,z,e,p) = F(\theta,e) + C(x,z,e,p) + pc_F, \tag{21}$$

and the first order condition is

$$f(\theta,e) + c(x,z,e,p) = 0, \ \Rightarrow \ e^* = e^*(\theta,x,z,p). \tag{22}$$

Similar to Appendix 1.4.3, $\Delta e^*(\theta,x,z) = e^P - e^N < 0$ if and only if

$$\begin{aligned} & c(x,z,e^P,p=1) - c(x,z,e^P,p=0) \\ & = r_P(e^P) + m_P(e^P) + g(e^P) - r_N(e^P) > 0, \\ & \textbf{or} \\ & c(x,z,e^N,p=0) - c(x,z,e^N,p=1) \\ & = r_N(e^N) - r_P(e^N) - m_P(e^N) - g(e^N) < 0. \end{aligned} \tag{23}$$

Recall that we use $\Delta MC''(e)$ to represent the gap between marginal emission cost under participation and non-participation without considering the cost of being labeled a greenwasher.

Equation 23 is equivalent to

$$\begin{aligned} & g(e^P) > \Delta MC''(e^P), \\ & \textbf{or} \\ & g(e^N) > \Delta MC''(e^N). \end{aligned} \tag{24}$$

When emissions are at the threshold e'', we have $g(e'') = \Delta MC''(e'')$. Besides, $\frac{\partial g(e)}{\partial e} - \frac{\partial MC''(e)}{\partial e} > 0$. Therefore, for any emission level $\hat{e} > e''$ it is always true that $g(\hat{e}) > \Delta MC''(\hat{e})$, and vice versa. It can be proved that $e^N > e''$ if and only if $e^P > e''$, and $e^N \leq e''$ if and only if $e^P \leq e''$, because of two conditions in equations 13, 14 and 24 are equivalent to each other. Therefore, Proposition 1 holds in this extension.

A firm's participation incentive is

$$\Delta D^* = \Big(F(\theta, e^N) + C(x, z, e^N, p = 0)\Big) - \Big(F(\theta, e^P) + C(x, z, e^P, p = 1)\Big) - c_F. \tag{25}$$

We have the same result as the baseline model about the participation incentives, and proposition 2 holds (proof is same as Appendix 1.4.4).

3 Model Extension 2: Abatement Technology Innovations and Spillovers

In the baseline model, θ represents firm characteristics related to the firm's abatement technology. Instead of assuming θ is given exogenously, here we allow it to take different values depending on the firm's participation status: $\theta \in \{\theta_0, \theta_N, \theta_P\}$, which represent the firm's ex-ante, non-participation and participation abatement technology, respectively. We assume $\theta_P > \theta_N > \theta_0$.

Technology improvement affects the marginal abatement cost, but does not affect the marginal emission cost. Since $\Delta e^*(\theta, x, z) = e^*(\theta_P, x, z, p = 1) - e^*(\theta_N, x, z, p = N)$, the two equivalent necessary and sufficient conditions that $\Delta e^*(\theta, x, z) = e^P - e^N < 0$ become

$$\begin{aligned} & g(e^P) - \Delta MC''(e^P) > f(\theta_N, e^P) - f(\theta_P, e^P), \\ & \textbf{or} \\ & \Delta MC''(e^N) - g(e^N) < f(\theta_P, e^N) - f(\theta_N, e^N). \end{aligned} \tag{26}$$

This implies the emission e'', i.e., the emissions at X_{NP} in Figure 2 in the main paper, is no longer the threshold level that differentiates participating plants' emission changes (as it was in the baseline model without technology spillover). Instead, there is a different emission level e''_S, which makes $\Delta MC''(e''_S) - g(e''_S) = f(\theta_N, e''_S) - f(\theta_P, e''_S)$ and is the relevant threshold that differentiate participating plants' emission changes. Because $f(\theta_N, e) > f(\theta_P, e)$ at all emission levels, $MC''(e''_S) > g(e''_S)$ and $e''_S < e''$.[4] 1.4.3). Empirically, if the model does not differentiate between θ_N and θ_P (as in the baseline model), then the grid search used in estimating the main paper equation 1 will identify e''_S rather than e''. This will result in a mis-classification of plants between types, and we will observe that type 2 plants with emissions between e''_S and e'' also decrease emissions when participating, therefore being mistakenly categorized as type 1 plants.

References

Zhou, R., X. Bi, and K. Segerson (2020). Evaluating voluntary environmental programs with spillover effects. *Journal of the Association of Environmental and Resource Economists* 7(1), 145–180.

[4]This is because $MC''(e'') = g(e'')$ and $\frac{\partial\left(MC''(e)-g(e)\right)}{\partial e} < 0$. The value of e''_S depends on the difference between participating and non-participating marginal abatement cost curves (proof is similar to Appendix section