Point-Nonpoint Heresy?!

An Endogenous Risk Explanation for Point-Nonpoint Trading Ratios Less than One

Richard D. Horan, Professor, Department of Agricultural, Food, and Resource Economics, Michigan State University, horan@msu.edu

James S. Shortle, Distinguished Professor, Department of Agricultural Economics, Sociology, and Education, Pennsylvania State University, jshortle@psu.edu

Selected Paper prepared for presentation at the Agricultural & Applied Economics Association's 2014 AAEA Annual Meeting, Minneapolis, MN, July 27-29, 2014.

Introduction

Water quality trading initiatives to foster cost-effective water quality improvements are found in the United States, Canada, Australia, and New Zealand, and interest is emerging elsewhere (Selman et al. 2008; Fisher-Vanden and Olmstead 2013). Many initiatives encourage trading between point and nonpoint sources, which facilitates the integration of pollution control across source types to improve efficiency and effectiveness (Shortle and Horan 2013).

A key question for point-nonpoint trading programs is how to deal with nonpoint source emissions uncertainties, which are substantial in comparison to point source emissions uncertainties (e.g., US EPA 2003, 2014). Point source emissions, and abatement of these emissions, are considered to be relatively deterministic since point source abatement technologies are well-developed and allow substantial control over emissions. These emissions are also largely observable since they exit through a pipe or other conveyance. In contrast, nonpoint emissions are highly stochastic due to weather processes driving pollution events, and are unobservable due to the complex and diffuse pathways of these pollutants (Griffin and Bromley 1983; Shortle and Dunn 1986). Stochasticity creates *objective* uncertainty about the effectiveness of nonpoint controls. There is also *subjective* uncertainty about control effectiveness due to limited information about how specific control practices will perform in different locations. Unlike point source controls, the effectiveness of nonpoint controls is highly site specific due to the highly heterogeneous site-specific factors (e.g., soils, topography, geology, stream networks, weather, land use practices) influencing nonpoint emissions. This heterogeneity along with the unobservability of nonpoint emissions combine to create substantial subjective uncertainty about nonpoint control effectiveness. The unobservability and uncertainties associated with nonpoint emissions means point-nonpoint trades cannot be based on changes in actual nonpoint emissions. Programs have instead based trades on estimated reductions in nonpoint emissions (U.S. EPA 2003).

Trading relatively certain point source emissions reductions for highly uncertain estimates of nonpoint reductions is a major concern that is addressed with a point-nonpoint trading ratio, sometimes referred to as the uncertainty ratio (U.S. EPA 2003, 2014). The key risk management variable in extant programs, the trading ratio is the rate at which estimated nonpoint emissions must be reduced for point sources to increase their emissions by one unit. Trade ratios correct for imperfect substitution between the traded point and nonpoint commodities in terms of achieving environmental goals, and thereby guide the allocation of controls among the sources.[1] Ratios less than one encourage substitution of nonpoint reductions for point source reductions, while ratios in excess of one do the opposite.

Economic research on optimal trade ratio design provides theoretical and some empirical support for ratios greater than or less than one depending on how nonpoint emissions uncertainties respond to trading (Shortle 1987, 1990; Malik et al. 1993; Horan 2001; Horan and Shortle 2005, 2011; Hennessy and Feng 2008).[2] For instance, if reducing mean nonpoint emissions also reduces objective uncertainty (i.e., the natural variability of nonpoint emissions and of ambient pollution), then risk is reduced and so nonpoint controls should be encouraged via a smaller trade ratio. Alternatively, subjective uncertainty may be increasing in the use of nonpoint abatement measures, creating risk that is optimally offset via a larger trade ratio. Which effects dominate and whether any particular trade ratio should be greater or less than one is an empirical question (Malik et al. 1993) that depends on site-specific factors and also on the

[1] Trading ratios can also be adjusted to address spatial heterogeneities among different sources, but we keep our focus here on the role of uncertainty.

[2] This is a bit of a simplification, as correlations between key environmental and cost relationships also matter, as do transactions costs (Horan and Shortle 2011; Malik et al. 1993).

design of the trading program. Program design is important because environmental risks endogenously depend on nonpoint polluter behaviors made in response to the program.

In contrast to the theoretical results from economic research, the received wisdom among program developers and the governmental and nongovernmental consulting communities that advise them is that trading ratios should be in excess of one to provide an adequate margin of safety for nonpoint risk. Indeed, trade ratios in practice universally exceed one, often substantially so, with the justification being that this addresses subjective uncertainties (e.g., MDEQ, 2000; NWF 1999; Wiedelman and Jones 2000; US EPA 2003, 2014). These choices are not derived from formal economic analyses and appropriate framing of risk (Selman et al. 2008; Horan 2001). Rather, the *a priori* assumption that subjective risk dominates is akin to treating risk as a fixed, exogenous measure rather than as an endogenous one that responds to policy-induced behavioral changes.

The implied perspective that risk is exogenous suggests the endogenous nature of risk has not been adequately conveyed in the point-nonpoint setting, which should not be surprising. Prior analyses only present economically optimal equilibrium outcomes, whereby the optimal trade ratio is simply presented as the slope at the tangency of an iso-abatement cost curve and an iso-curve measuring either expected damage costs or a physical water quality target. Such a presentation does not clearly illustrate out-of-equilibrium tradeoffs involving abatement costs and environmental risks. As a result, the endogenous nature of risk – particularly the relative magnitudes of objective and subjective uncertainty – is obscured.

We develop a new approach that illustrates the optimal design of trade ratios in response to tradeoffs between abatement costs and endogenous environmental risks. The approach utilizes two new concepts – the *relative* marginal benefit curve and the *relative* marginal environmental cost curve. The relative marginal benefit curve measures the relative marginal cost savings, or benefits, from reallocating abatement from point sources to nonpoint sources. This curve exhibits the properties of a downward-sloping demand curve for nonpoint abatement. Moreover, assuming the permit market clears regardless of whether the trade ratio is optimally set, the height of this curve relates to the magnitude of the trade ratio. This allows us to examine economic tradeoffs and abatement outcomes associated with different trade ratios (not just the unique, optimal equilibrium ratio), unlike prior studies.

The relative marginal environmental cost curve measures the relative increase in expected marginal damage costs from reallocating abatement from point sources to nonpoint sources. This curve exhibits the properties of an upward-sloping marginal environmental cost curve for nonpoint abatement that relates to the marginal variance of nonpoint abatement – a measure of nonpoint risk. Moreover, the curve exhibits an important relation with objective and subjective risk that helps us to identify which type of risk effect dominates in different outcomes.

Together, the relative marginal benefit and cost curves allow us to examine economic and environmental tradeoffs involving trading ratios, nonpoint abatement, and the various types of risk. Specifically, we illustrate how abatement costs and environmental variability endogenously determine the optimal level and dominate type (objective vs. subjective) of risk and hence the optimal trading ratio. Our results show that an optimally designed trading program reallocates abatement to nonpoint sources to reduce abatement costs or to reduce environmental risks from nonpoint sources, but not both. This outcome is in direct contrast to the stated goals of the EPA's national trading rules (US EPA 2003).

Our methodology is also shown to be useful for examining second-best program design. Unlike the first-best programs typically considered by economic theory, nonpoint source participation in most extant trading programs is not incentivized through regulations on nonpoint emissions. This feature can impact the design of second-best trading ratios (Horan and Shortle 2005), but the role of uncertainty has not been fully explored. While prior work has shown that second-best considerations tend to increase the trading ratio, other things equal, we find that risk-effects are not equal relative to the first-best case. Once we account for these endogenously-changing risk effects, we find the prior recommendations are reversed. This means that second-best ratios are more likely to be less than one, in contrast to all extant programs, due to the endogenous nature of risk.

Model

Consider a lake or other water body polluted by deterministic point source emissions, e, and stochastic nonpoint source emissions, r. We simplify the analysis by assuming one point source and one nonpoint source (not an uncommon modeling approach; see Malik et al. 1993; Hennessy and Feng 2008; Horan and Shortle 2005), and that emissions are uniformly mixed so that actual emissions are perfect substitutes in terms of generating economic damages, $D(e + r)$, with D' , $D'' > 0$. These assumptions enable us to focus on the role of uncertainty rather than the role of spatial heterogeneity.

Runoff is stochastic due to random weather events that drive pollutants off fields. Because of this stochasticity, which is represented by the random variable α, the nonpoint source can only control the distribution of runoff. We assume this control occurs through the scalar choice of abatement, z. In addition to being stochastic, actual runoff is unobservable and so it is impossible to estimate the true distribution of r. This means there is subjective uncertainty over r, which we model as uncertainty about how r responds to z. Subjective uncertainty is assumed to be increasing in the level of abatement, as greater deviations from the status quo (i.e., no abatement, or $z = 0$) are deemed more uncertain. These assumptions are consistent with Malik et al. (1993), who adopt a general per acre runoff function. In this analysis we impose some structure on the runoff model, trading off generality for new insights.

Consider the following specific subjective form of r:

(1) $r(z,\alpha,\beta) = (r_0 - z[1+b\beta])(1+a\alpha)$,

where r_0 is mean runoff prior to abatement, a and b are parameters, and β is a random variable indicating the subjective uncertainty associated with abatement effort, z. Assume α and β are uncorrelated, with α having a mean of zero and a variance of σ_α, and with β having a mean of zero and a variance of σ_β. We restrict abatement effort so that expected runoff, $E\{r\} = r_0 - z$, is strictly positive, i.e., $z \in [0, r_0)$. Note that $E\{\partial r/\partial z\} = -1$, implying that abatement effort is scaled such that a unit increase in effort cause a unit decrease in expected runoff

Shortle (1987, 1990) has shown the effect of pollution effort on the variance of runoff has important implications for the design of a point-nonpoint trading program. The variance of runoff in this case is derived from (1) as

(2)
$$\begin{aligned} \text{var}\{r(z,\alpha,\beta)\} &= (r_0 - z)^2 a^2 \sigma_\alpha + z^2 b^2 \sigma_\beta + z^2 a^2 b^2 \sigma_\alpha \sigma_\beta \\ &= (r_0 - z)^2 a^2 \sigma_\alpha + z^2 b^2 \sigma_\beta (1 + a^2 \sigma_\alpha) \end{aligned}$$

From (2) we can derive

(3)
$$\frac{\partial \,\text{var}\{r(z,\alpha,\beta)\}}{\partial z} = -2(r_0 - z) a^2 \sigma_\alpha + 2 z b^2 \sigma_\beta (1 + a^2 \sigma_\alpha)$$

The first right-hand-side (RHS) term in (3) is negative, as abatement reduces the natural variability of nonpoint emissions. The second RHS term in (3) is positive, as the effectiveness of abatement is uncertain. The relative magnitudes of these two terms depend on the level of abatement. For instance, the first RHS term goes to zero as $z \to r_0$. The second RHS term goes to zero as $z \to 0$, but it increases as $z \to r_0$. Define $\hat{z}$ as the critical value of z such that expression (3) vanishes for $z = \hat{z}$. Hence, expression (3) will be positive for $z > \hat{z}$, and it will be negative for $z < \hat{z}$. While not a general result, this does suggest the risk effects of abatement may change depending on how much regulation occurs. In contrast, prior theoretical work generally discusses $\partial \text{var}(r)/\partial z$ as if it has a constant (although ambiguous) sign, whereas numerical analyses often make modeling assumptions that effectively require $\partial \text{var}(r)/\partial z$ to take on a particular sign. A key contribution of the current analysis is to explore the more general case of a non-constant marginal variance and how this feature may affect the design of trading programs.

The permit market

We model trading program that caps point source emissions plus expected nonpoint emissions. Point source permits are denoted $\hat{e}$, and nonpoint source permits $\hat{r}$. The former are denominated in terms of emissions while the latter are denominated in terms of expected runoff. With cross-type trading allowed, firms must have a combination of both types at least equal to their emissions, in the case of point sources, or expected runoff in the case of nonpoint sources. Emissions and expected runoff are imperfect substitutes, and so they are not necessarily traded on a one-for-one basis. Point source permits are traded at a price of q, and nonpoint source permits are traded at a price of p. The cross-category trading ratio is denoted t. Specifically, $t = | d\hat{r} / d\hat{e} |$ represents the required reduction in expected runoff in order for the point source to increase emissions by one unit. Prior work (e.g., Horan and Shortle 2005) has shown $t = q/p$ in a market equilibrium. As noted previously, the received wisdom and practice in extant markets is to set $t > 1$, but we do not require this. Rather, we explore the optimal choice of t, given the two types of uncertainty.

Denote point source abatement costs by $c(e)$, with $c'(e) < 0$ and $c''(z) > 0$. Nonpoint source abatement costs are denoted $g(z)$, with $g(0) = g'(0) = 0$, $g'(z) > 0$, and $g''(z) > 0$. Let $\hat{e}^0$ and $\hat{r}^0$ be the initial point and nonpoint source permits available, respectively, and assume $\hat{e}^0$ is held by point sources and $\hat{r}^0$ is held by nonpoint sources.

Given its initial permit holding, the point source will choose emissions levels, e, point source permit holdings, $\hat{e}_{ps}$, and nonpoint source permit holdings, $\hat{r}_{ps}$, to minimize net costs, $C = c(e) + q[\hat{e}_{ps} - \hat{e}^0] + p\hat{r}_{ps}$, given that its total emissions cannot be greater than its permit holdings, $e \leq \hat{e}_{ps} + (1/t)\hat{r}_{ps}$. The term $(1/t)\hat{r}_{ps}$ represents the emissions the firm can generate based on its expected runoff permits. Assuming the emissions constraint is satisfied as an equality, and using the market equilibrium result that $t = q/p$, we can eliminate $\hat{e}_{ps}$ as a choice variable so that $C = c(e) + q[e - \hat{e}^0_{ps}]$. Similarly, we can derive net costs facing nonpoint sources as $G = g(z) + p[E\{r(z,\alpha,\beta)\} - \hat{r}^0]$. Necessary conditions for emissions and abatement are

(4) $\partial C / \partial e = c'(e) + q = 0 \Rightarrow e(q)$

(5) $\partial G / \partial z = g'(z) + pE\{\partial r / \partial z\} = g'(z) - p = 0 \Rightarrow z(p)$

where $e(q)$ is the point source's derived demand for emissions and $z(p)$ is the nonpoint source's supply of abatement. Given these relations, the market solution is determined by the relation $t = q/p$ along with the market clearing condition

(6) $Q = \hat{e}^0 + (1/t)\hat{r}^0 \geq e(q) + (1/t)E\{r(z(p),\alpha,\beta)\}$,

where Q is the aggregate number of permits (denominated in terms of point source emissions).

Market design choices

We assume the regulatory authority seeks to design a point-nonpoint permit market to minimize the expected social costs of pollution and its control, $SC = c(e) + g(z) + E\{D(e + r)\}$, subject to polluters' market behavior. The authority's ability to minimize SC will depend on the degree to which it has control over the relevant policy variables: the trading ratio, t, and initial permit levels, $\hat{e}^0$ and $\hat{r}^0$. A first-best optimum would involve no binding constraints on these variables, whereas a second-best outcome results from binding constraints on one or more variables. For instance, extant programs generally operate with constraints on the initial permit levels. The initial number of point source permits, $\hat{e}^0$, are often already regulated through the National Pollutant Discharge Elimination System (NPDES). Nonpoint sources are generally not regulated initially, and so their initial permit allocation corresponds to their expected runoff levels when they do not abate, i.e., $\hat{r}^0 = E\{r(0,\alpha,\beta)\}$.[3] Trades therefore involve the nonpoint source selling credits that it generates through the program.

We begin by presenting the second-best trading ratio, as this form of the trading ratio is quite general and subsumes the first-best form as a special case. The second-best ratio minimizes SC, subject to market behavior and also subject to the market clearing condition (6) being constrained by the initial permit allocation. A shadow value of $\lambda > 0$ is associated with the constraint (6) in the most likely case where too few permits are allocated and the constraint is binding. The first-best outcome arises with $\lambda = 0$ in the special (but unlikely) case where the

[3] The only exception to this is when the nonpoint source has previously enrolled in a program requiring abatement. Denoting such an abatement level by z^0, initial permits in this case are defined by $\hat{r}^0 = E\{r(z^0,\alpha,\beta)\} < E\{r(0,\alpha,\beta)\}$. This is the principle of additionality, which requires credits be generated in addition to those already promised through other programs. Regardless of whether the nonpoint source has sold credits via other programs, the fact that initial permit levels are not chosen optimally prevents the trading outcome from being first-best.

allocation equals the first-best allocation. Horan and Shortle (2005) derive the second-best trading ratio, which using our notation is

$$(7)\qquad t^{**}=\frac{E\{D'^{**}\}-\lambda^{**}s_e^{**}\varepsilon_{qe}^{**}-\lambda^{**}s_r^{**}\varepsilon_{pr}^{**}}{E\{D'^{**}\}+\dfrac{\operatorname{cov}\{D'^{**},\partial r^{**}/z\}}{E\{\partial r^{**}/\partial z\}}}=\frac{E\{D'^{**}\}-\lambda^{**}s_e^{**}\varepsilon_{qe}^{**}-\lambda^{**}s_r^{**}\varepsilon_{pr}^{**}}{E\{D'^{**}\}-\operatorname{cov}\{D'^{**},\partial r^{**}/z\}}$$

where ε_{qe}^{**}<0 is the point source's inverse elasticity of demand for emissions, ε_{pr}^{**} <0 is the nonpoint source's inverse elasticity of demand for expected pollution loads, $s_e=(e^{**}-\hat{e}^0)/e^{**}$ is the point source's proportional excess demand for point source permits, $s_r = (\hat{r}^0 - E\{r^{**}\})/$ $E\{r^{**}\}$ is the nonpoint source's proportional excess supply of nonpoint permits, and the superscript ** indicates that all variables are set at their second-best optimal values.

Departing now from Horan and Shortle (2005), we rewrite relation (7) and also use (4) and (5) to yield

$$(8)\qquad 1-\frac{1}{t^{**}}=\frac{-c'(e^{**})-g'(z^{**})}{-c'(e^{**})}=\Lambda^{**}+\frac{\operatorname{cov}\{D'^{**},\partial r^{**}/z\}}{E\{D'^{**}\}}$$

where $\Lambda^{**}=-[\lambda^{**}s_e^{**}\varepsilon_{qe}^{**}+\lambda^{**}s_r^{**}\varepsilon_{pr}^{**}]/[t^{**}E\{D'^{**}\}]>0$. The first equality always holds in market equilibrium (even for suboptimal values of *t*, *e*, and *z*). The second equality holds only when *t* is chosen optimally. As noted above, in the special case where $\hat{e}^0$ and $\hat{r}^0$ are also chosen optimally, then (8) holds with $\lambda = \Lambda = 0$ (i.e., all variables are evaluated at their first-best levels, indicated below with a superscript *). For each case, expression (8) shows that the trading ratio (in relation to unity) depends on the relative marginal costs and benefits of control between the two sources. We explore the possibilities in subsequent sections, starting with the first-best case as a benchmark to facilitate our understanding of the second-best case.

First-best design

Disentangling the tradeoffs implicit in (8) is complicated by the large number of variables in the model. We simplify by switching from our use of the ordinary permit demand relations $z(p)$ and $e(q)$ (from (4) and (5), respectively), to inverse relations that depend on z. This allows us to explore tradeoffs in the nonpoint source abatement dimension.

To proceed with z as our independent variable, first set e according to the market-clearing relation $e(z,\hat{e}^0,\hat{r}^0,t) = \hat{e}^0 + (1/t)(\hat{r}^0 - r_0 + z)$. Next, use the ratio of the polluters' first order conditions (4) and (5) to obtain

$$(9) \qquad t = \frac{-c'(e(z,\hat{e}^0,\hat{r}^0,t))}{g'(z)}$$

Equation (9) implicitly defines $t(z,\hat{e}^0,\hat{r}^0)$, which is the real price of emissions permits (since $t = p/q$), conditional on the choice of z. This yields the conditional residual demand for emissions as $e(z,\hat{e}^0,\hat{r}^0) = e(z,\hat{e}^0,\hat{r}^0,t(z,\hat{e}^0,\hat{r}^0))$. Using the relations $t(z,\hat{e}^0,\hat{r}^0)$ and $e(z,\hat{e}^0,\hat{r}^0)$ in what follows ensures the emissions market clears when nonpoint sources choose z, given the initial permit levels $\hat{e}^0$ and $\hat{r}^0$.

Next, define the relation $MC_{sav}(z,\hat{e}^0,\hat{r}^0) = [-c'(e(z,\hat{e}^0,\hat{r}^0)) - g'(z)]/[-c'(e(z,\hat{e}^0,\hat{r}^0))]$. This function, when positive (negative), represents the relative marginal cost savings (increase) of reallocating abatement from point sources to nonpoint sources. In other words, MC_{sav} represents the relative marginal benefits of nonpoint abatement. This function is illustrated in Fig. 1 for the case where $\hat{e}^0 = \hat{e}^{0*}$ and $\hat{r}^0 = \hat{r}^{0*}$ (where $\hat{e}^{0*}$ and $\hat{r}^{0*}$ could be any combination of $\hat{e}^0$ and $\hat{r}^0$ that satisfy the relation $\hat{e}^0 + (1/t^*)\hat{r}^0 = e^* + (1/t^*)z^*$). $MC_{sav}(z,\hat{e}^{0*},\hat{r}^{0*})$ is declining in z since $\partial MC_{sav}/\partial z = [c'(e(z))g''(z) - g'(z)c''(e(z))(1/t)]/[c'(e(z))]^2 < 0$. In particular, for any given t

and Q, $MC_{sav} = 1$ when $z = 0$ and $MC_{sav} < 0$ when z is sufficiently large. This means nonpoint sources have a cost advantage prior to trading, which is usually assumed to be the case, but this advantage disappears once nonpoint sources take on a sufficient share of abatement. Note that market clearing ensures $MC_{sav}(z, \hat{e}^0, \hat{r}^0)$ always equals $1 - 1/t(z, \hat{e}^0, \hat{r}^0)$, and so the height of $MC_{sav}(z, \hat{e}^{0*}, \hat{r}^{0*})$ is positively correlated with $t(z, \hat{e}^{0*}, \hat{r}^{0*})$. Also recall that MC_{sav} is an inverse relation. In reality, the regulatory authority chooses t, which will produce a market response for z such that $MC_{sav}(z, \hat{e}^{0*}, \hat{r}^{0*}) = 1 - 1/t$.

Finally, define $MEC(z, \hat{e}^0, \hat{r}^0) = \text{cov}\{D'(e(z, \hat{e}^0, \hat{r}^0) + z), \partial r(z)/\partial z\}/E\{D'(e(z, \hat{e}^0, \hat{r}^0) + z)\}$. The function $MEC(z, \hat{e}^0, \hat{r}^0)$ is of the same sign as $\text{cov}\{D'(e(z, \hat{e}^0, \hat{r}^0) + z), \partial r(z)/\partial z\}$, which reflects the environmental risk impacts of nonpoint abatement. In turn, this covariance term is of the same sign as $\partial\,\text{var}\{r(z, \alpha, \beta)\}/\partial z$ (Shortle 1990): risk is greater (lesser) when z increases (decreases) the variability of nonpoint runoff. Accordingly, when $MEC(z, \hat{e}^0, \hat{r}^0)$ is positive (negative), this relation represents the relative marginal environmental costs (benefits) of reallocating abatement to nonpoint sources, in terms of increasing (reducing) environmental risk.[4] The curve MEC is illustrated in Fig. 1 for the case where $\hat{e}^0 = \hat{e}^{0*}$ and $\hat{r}^0 = \hat{r}^{0*}$. $MEC(z, \hat{e}^{0*}, \hat{r}^{0*})$ is an upward sloping marginal environmental cost curve with a horizontal intercept of $\hat{z}$, since we know from equation (3) that $\partial\text{var}\{r(z,\alpha,\beta)\}/\partial z$ and hence $\text{cov}\{D', \partial r/\partial z\}$ are negative for $z < \hat{z}$, zero for $z = \hat{z}$, and positive otherwise. The negative portion of MEC means nonpoint source abatement is initially risk-reducing, i.e., an environmental benefit.

It is easily verified that the intersection of $MC_{sav}(z, \hat{e}^{0*}, \hat{r}^{0*})$ and $MEC(z, \hat{e}^{0*}, \hat{r}^{0*})$ yields the

[4] Recall that emissions by point and nonpoint sources have the same marginal impacts on expected damages, i.e., the denominator of this term. Hence, the only difference these sources have on expected damages, at the margin, is in the risk effects of nonpoint runoff.

first-best outcome, as z^* solves the following first-best equivalent of (8):

$$(10)\qquad 1-\frac{1}{t(z,\hat{e}^{0*},\hat{r}^{0*})}=MC_{sav}(z,\hat{e}^{0*},\hat{r}^{0*})=MEC(z,\hat{e}^{0*},\hat{r}^{0*})\,.$$

By construction, the first-best value of *z* also yields the first-best values e^* = $e(z^*,\hat{e}^{0*},\hat{r}^{0*})$ and t^* = $t(z^*,\hat{e}^{0*},\hat{r}^{0*})$. Equation (10) indicates the first-best trading ratio: (i) is less than one when society benefits from less environmental risk at the margin, at the expense of relatively more expensive nonpoint controls (i.e., $t^* < 1$ when $MC^*_{sav} = MEC^* < 0$ with $z^* < \hat{z}$; Fig. 1a); (ii) exceeds one when society benefits from relatively cheaper nonpoint controls at the margin, but at the expense of greater environmental risk (i.e., $t^* > 1$ when $MC^*_{sav} = MEC^* > 0$ with $z^* > \hat{z}$; Fig. 1b). Either case is possible. These relations between the trading ratio and environmental risk are consistent with prior work (e.g., Shortle 1987; Malik et al. 1993). However, considering both the MC_{sav} and *MEC* terms together illustrates more clearly than prior work that the social gains from point sources trading for more nonpoint abatement come from either a reduction in environmental risks or reduced overall abatement costs, but not both. Additionally, Fig. 1 serves as a graphical benchmark for our analysis of the second-best case below.

Second-best design

Now consider the second-best case in which initial permit levels for one or more sources is too large, so that $\hat{e}^0+(1/t^*)\hat{r}^0 > e^*+(1/t^*)z^*$ and condition (8) holds with $\lambda, \Lambda > 0$. Without loss, we focus on the case where $\hat{e}^0 = e^*$ and $\hat{r}^0 > \hat{r}^{0*}$, as nonpoint sources typically go unregulated (or, under-regulated) prior to their participation in the permit market. With $\hat{r}^0$ being exogenous, $\hat{e}^0 = e^*$ is unlikely to be optimal and could therefore be considered exogenous as well. Horan and Shortle (2005) describe that this situation results in t having two roles: determining the real emissions price and the number of permits Q. Unfortunately, t cannot perform both tasks very efficiently, implying there are social adjustment costs associated with changes in t. These additional marginal costs are represented by the term Λ in equation (8), so that $MEC + \Lambda$ is now the marginal social cost.

Consider how the marginal benefit and marginal social cost curves are likely to change in response to a larger $\hat{r}^0$, starting with $MC_{sav}(z,\hat{e}^0,\hat{r}^0)$. As $t(z,\hat{e}^0,\hat{r}^0)$ ensures that $MC_{sav}(z,\hat{e}^0,\hat{r}^0)$ = $1 - 1/t(z,\hat{e}^0,\hat{r}^0)$ always holds (even in the second-best case), we can derive

(11) $$\frac{\partial MC_{sav}}{\partial \hat{r}^0} = \frac{1}{t^2}\frac{\partial t}{\partial \hat{r}^0}.$$

Condition (11) means the change in $MC_{sav}(z,\hat{e}^0,\hat{r}^0)$ is of the same sign as $\partial t / \partial \hat{r}^0$. This derivative is derived from (9). To simplify the exposition, we rewrite (9) as

(12) $$t = \frac{-c'(Q-(1/t)[r_0 - z])}{g'(z)},$$

where Q is the total number of emissions permits, derived in (6) as $Q = \hat{e}^0 + (1/t)\hat{r}^0$. Note that Q depends on $\hat{r}^0$ directly and also via its impacts on t. Using (12), we can derive

(13) $$\frac{dt}{d\hat{r}^0} = \frac{-c''(e)}{g'(z)}\left[1 + \frac{c''(e)}{g'(z)}\frac{1}{t^2}[r_0 - z]\right]^{-1}\frac{dQ}{d\hat{r}^0}$$

The coefficient on $dQ/d\hat{r}^0$ is negative, and so the sign of $dt/d\hat{r}^0$ is opposite the sign of $dQ/d\hat{r}^0$. The sign of $dQ/d\hat{r}^0$ is technically ambiguous. However, the sign of $dQ/d\hat{r}^0$ will realistically be positive (i.e., under-regulation of the overall market, relative to the first-best) when initial nonpoint permit levels exceed their first-best values. The alternative, that decreasing in nonpoint permit stringency would lead to over-regulation, does not make sense. Assuming

$dQ/d\hat{r}^0 > 0$, a larger $\hat{r}^0$ rotates the curve MC_{sav} clockwise, with the vertical intercept remaining at unity.

Now consider how a larger $\hat{r}^0$ affects the marginal social cost curve, which was *MEC* in the first-best case. There are two effects. First, the change in $\hat{r}^0$ causes *MEC* to rotate around the fixed point $\hat{z}$ (but remain upward sloping), since *MEC* always crosses through this point which is unaffected by changes in $\hat{r}^0$. As our primary interest is in whether the change in $\hat{r}^0$ will move the system across the horizontal axis (so that *t* changes from greater than one to less than one, or vice versa), this rotation in *MEC* has no qualitative impact on the results. We therefore do not rotate *MEC* in our graphical analysis.

The second effect is that $MEC + \Lambda$ is now the marginal social cost, as described above. In principle, we can write Λ as a function of *z* and the model parameters such as initial permit levels, i.e., $\Lambda = \Lambda(z, \hat{e}^0, \hat{r}^0)$. However, the relation between Λ and *z* cannot be easily characterized. We describe how Λ is graphed in the next section.

Changes relative to the equilibrium in Fig. 1a

We first examine how Fig. 1a is likely to change in response to a larger $\hat{r}^0$, denoted $\hat{r}_1^0 > \hat{r}^{0*}$. Fig. 2a illustrates the resulting rotation of MC_{sav} from the first-best marginal benefit curve, $MC_{sav}(z, \hat{e}^{0*}, \hat{r}^{0*})$ (with associated outcomes z^* and t^*), to $MC_{sav}(z, \hat{e}^{0*}, \hat{r}_1^0)$. For the social marginal cost curve, $MEC + \Lambda$, we simplify matters by graphing the relation $MEC(z, \hat{e}^{0*}, \hat{r}_1^0)$ as if it were the same as the first-best curve $MEC(z, \hat{e}^{0*}, \hat{r}^{0*})$. As described above, this is of no consequence since the two curves share essentially the same properties.

Graphing the relation $\Lambda(z, \hat{e}^{0*}, \hat{r}_1^0)$ is more problematic because we do not know how this expression varies in relation to *z*. We simplify matters and graph Λ as a fixed value, in accordance with the following conceptual exercise. As $\Lambda(z, \hat{e}^{0*}, \hat{r}_1^0)$ can only be fixed if this expression is evaluated at a fixed value of *z*, we first choose a candidate value for the second-best optimum, say $z = z^{\#}$. If the resulting marginal social cost curve, $MEC(z, \hat{e}^{0*}, \hat{r}_1^0) + \Lambda(z^{\#}, \hat{e}^{0*}, \hat{r}_1^0)$, happens to intersect $MC_{sav}(z, \hat{e}^{0*}, \hat{r}_1^0)$ at $z^{\#}$, then we know our candidate solution was correct, i.e., $z^{\#} = z^{**}$. Otherwise, we must choose an alternative value of *z*. This exercise only helps us to visualize the result; it does not actually help us to compute z^{**}. However, this visualization is valuable because it does help us to rule out certain types of solutions, thereby shedding light on the set of viable candidates for z^{**}.

Suppose the candidate solution $z^{\#}$ produces the dashed-dotted marginal social cost curve in Fig. 2a. This curve suggests a second-best solution at the point z_1, but this will only be the true second-best outcome if $z_1 = z^{\#}$, with $MC_{sav} > 0$ and $MEC < 0$. We demonstrate in the Appendix that such an outcome cannot be optimal. The intuition is straightforward: prior to accounting for social adjustment costs (i.e., Λ), there are only benefits from reducing *t*, and thereby increasing *z*, in this case since increased nonpoint abatement yields both cost savings (since $MC_{sav} > 0$) and environmental benefits in the form of reduced environmental risks (since $MEC < 0$). An outcome at z_1 actually implies a negative shadow value, i.e., λ, $\Lambda < 0$, as there are additional gains to be made before the constraint on the permit allocation becomes binding. This means $MEC(z, \hat{e}^{0*}, \hat{r}_1^0) + \Lambda(z^{\#}, \hat{e}^{0*}, \hat{r}_1^0)$ cannot intersect $MC_{sav}(z, \hat{e}^{0*}, \hat{r}_1^0)$ at $z^{\#}$, and so we must choose another candidate for *z*.

As noted in the Appendix, the true solution for values of $z < \hat{z}$ (for which $MEC < 0$)

must occur where $MC_{sav} < 0$. A possible solution is indicated at point z^{**}, with $t^{**} < t^{*}$. But it is also possible that the solution could involve a slightly larger Λ^{**} and a slightly lower z^{**} with $t^{**} \in (t^{*},1)$; greater model specification is required to know which outcome is more likely. The result that $t^{**} < t^{*}$ is possible adds insight to Horan and Shortle's (2005) results. They found the addition of the term Λ in the second-best case causes t to increase, holding *MEC* fixed. Here, we demonstrate that t may actually fall relative to the first-best case, owing to the fact that all else is not equal: marginal cost savings shift (as do *MEC*, although this is not depicted) as $\hat{r}^0$ increases, reducing the equilibrium value of *MEC*. A sufficient reduction in this value, relative to the first-best case, could support a smaller t^{**}.

Although the relation between t^{**} and t^{*} is ambiguous, we can say for certain that $t^{**} < 1$ whenever $t^{*} < 1$. This makes intuitive sense. If the first-best case results in a societal willingness to incur relatively greater nonpoint abatement costs in return for reduced environmental risks, then a second-best case involving fewer overall controls must yield even greater marginal benefits to controlling environmental risks. The trading ratio optimally penalizes environmental risks by making it inexpensive to remove the riskier source of pollution, which is that produced by nonpoint sources.

Changes relative to the equilibrium in Fig. 1b

Now consider how the outcome in Fig. 1b is likely to change in response to a larger $\hat{r}^0$. The increase in $\hat{r}^0$ from $\hat{r}^{0*}$ to $\hat{r}_a^0$ is depicted as a clockwise rotation in the curve MC_{sav}, from $MC_{sav}(z, \hat{e}^{0*}, \hat{r}^{0*})$ to $MC_{sav}(z, \hat{e}^{0*}, \hat{r}_a^0)$. The new equilibrium is depicted to occur at z^{**} and to yield $t^{**} < t^{*}$. This is only one possibility. More generally, the relative values of z and t depend on the extent of rotation in MC_{sav} and on the magnitude of Λ^{**}. For instance, suppose the horizontal intercept of $MC_{sav}(z, \hat{e}^{0*}, \hat{r}_a^0)$ lies to the right of $\hat{z}$. Then a second-best solution (if it exists) must involve $z^{**} > \hat{z}$ and $t^{**} \in (1, t^{\#})$, where $t^{\#}$ solves $1 - 1/t = MC_{sav}(\hat{z}, \hat{e}^{0*}, \hat{r}_a^0)$. The upper bound on t arises because *MEC* must be positive when $MC_{sav} > 0$ (see the Appendix), and so $MEC + \Lambda^{**}$ cannot intersect $MC_{sav}(z, \hat{e}^{0*}, \hat{r}_a^0)$ to the left of $\hat{z}$. These results imply $t^{**} < t^{*}$ is possible, provided $t^{\#} < t^{*}$. However, such an outcome becomes more unlikely the larger is the rotation in MC_{sav} (either due to a larger difference $\hat{r}^{0*} - \hat{r}_a^0$ and/or a greater responsiveness of MC_{sav} to changes in $\hat{r}^0$). Finally, $z^{**} < \hat{z}$ and $t^{**} < 1$ is the only possible outcome when the horizontal intercept of $MC_{sav}(z, \hat{e}^{0*}, \hat{r}_a^0)$ lies to the left of $\hat{z}$, as $MEC < 0$ in this case and so MC_{sav} must be negative as well (see the Appendix).

Overall, our results suggest that inefficient regulation, via a sub-optimally large $\hat{r}^0$, makes it more likely that second-best trading ratios will be less than one. Such an outcome at least seems plausible for current point-nonpoint trading programs, given that nonpoint sources have historically gone unregulated and account for a significant amount of nutrient loadings (see Shortle et al. 2012). The intuition for the small ratio is that the relative lack on enforceable nonpoint controls, prior to trading, has the effect of limiting incentives for nonpoint abatement to the extent that concerns over environmental risk dominate subjective concerns over abatement risks. The resulting marginal benefits of controlling environmental risks leads to a comparatively greater allocation of nonpoint controls, facilitated by a trading ratio less than unity.

Conclusion

Point-nonpoint trading can be a useful approach for cost-effectively improving water quality, but only if the programs are designed to realize the gains. Extant programs have universally adopted large uncertainty trade ratios as a way of addressing risks associated with nonpoint sources, but these choices are not based on formal economic analysis. Rather, they appear rooted in the assumption that subjective uncertainties associated with nonpoint controls dominates objective uncertainties, regardless of program design. This exogenous take on risk is potentially problematic. While some programs may be (accidentally) correct in their choice of ratio, we suspect the ratios are too large in many extant programs, for the reasons outlined in this analysis.

Inefficiently large ratios generate two problems, which can be gleaned from considering allocations to the left of the equilibrium in Fig. 1. First, an excessive ratio increases the differences in marginal abatement costs between point and nonpoint sources, reducing the cost savings that come from trade. Indeed, large ratios tend to discourage trading in extant programs where point sources are the only ones to face enforceable abatement requirements prior to trading. Second, an inefficiently large trade ratio does not adequately manage environmental risks, resulting in excessive risks. These risks arise due to a lack of trade, as the natural variability of nonpoint emissions creates risk even when trades do not occur. Some amount of trade can actually reduce these risks under realistic conditions. It is, perhaps, ironic that there continues to be so much interest in using point-nonpoint trading to improve water quality, when the economic science that underpins this concept is not formally applied to individual programs and when the chosen program designs features limit the size of the market and cost-reducing trading activity.

References

Fisher-Vanden, K., and S. Olmstead. 2013. “Moving Pollution Trading from Air to Water:

Potential Problems, and Prognosis." *Journal of Economic Perspectives* 27(1): 147-171.

Griffin, R.C., and D.W. Bromley. 1982. "Agricultural Runoff as a Nonpoint Externality: A Theoretical Development." *American Journal of Agricultural Economics* 64:547-552.

Hennessy, D.A., and H. Feng. 2008. "When Should Uncertain Nonpoint Emissions be Penalized in a Trading Program." *American Journal of Agricultural Economics* 90(1): 249-255.

Horan, R.D. 2001. "Differences in Social and Public Risk Perceptions and Conflicting Impacts on Point/Nonpoint Trading Ratios." *American Journal of Agricultural Economics* 83(4):934-941.

Horan, R.D., and J.S. Shortle. 2005. "When Two Wrongs Make a Right: Second-Best Point-Nonpoint Trading Ratios." *American Journal of Agricultural Economics* 87(2): 340-352.

Horan, R.D., and J.S. Shortle. 2011. "Economic and Ecological Rules for Water Quality Trading." *Journal of the American Water Resources Association* 47: 59-69.

Malik, A.S., D. Letson, and S.R. Crutchfield. 1993. "Point/Nonpoint Source trading of Pollution Abatement: Choosing the Right Trading Ratio." *American Journal of Agricultural Economics* 75:959-967.

Michigan Department of Environmental Quality (MDEQ). 2000. "Part 30. Water Quality Trading." *Proposed Rules as Published in the Michigan Register and put on Public Notice.*

National Wildlife Federation. 1999. *A New Tool for Water Quality: Making Watershed-Based Trading Work for You.* Montpelier, VT: National Wildlife Federation.

Selman, M., S. Greenhalgh, E. Branosky, C. Jones, and J. Guiling. 2009. *Water Quality Trading Programs: An International Overview*. World Resources Institute Issue Brief. Washington DC.

Shortle, J.S. 1987. "Allocative Implications of Comparisons between the Marginal Costs of Point and Nonpoint Source Pollution Abatement." *Northeastern Journal of Agricultural and Resource Economics* 16:17-23.

Shortle, J.S. 1990. "The Allocative Efficiency Implications of Water Pollution Abatement Cost Comparisons." *Water Resources Research* 26(5):793-797.

Shortle, J.S., and J.W. Dunn. 1986. "The Relative Efficiency of Agricultural Source Water Pollution Control Policies." *American Journal of Environmental Economics* 68(3) 668-677.

Shortle, J.S., and R.D. Horan. 2013. "Policy Instruments for Water Quality Protection." *Annual Review of Resource Economics* 5: 6.1-6.28.

Shortle, J.S., M.O. Ribaudo, R.D. Horan, and D. Blandford. 2012. "Reforming Agricultural Nonpoint Pollution Policy in an Increasingly Budget-Constrained Environment." *Environmental Science & Technology* 46(3): 1316-1325.

U.S. Environmental Protection Agency. 2003. "Water Quality Trading Policy; Issuance of Final Policy", *Federal Register*: January 13, 2003 (Volume 68, Number 8), Page 1608-1613.

U.S. Environmental Protection Agency. 2014. *Accounting for Uncertainty in Offset and Trading Programs*. EPA Technical Memorandum, Feb. 12.

Wiedelman, A. and C. Jones. 2000. *Nutrient Trading for the Chesapeake Bay Program*.

Appendix: Deriving the second-best outcome using residual demand relations

We can use the residual demand relations $t(z,\hat{e}^0,\hat{r}^0)$ and $e(z,\hat{e}^0,\hat{r}^0)$ to construct a measure of social costs as a function of nonpoint abatement:

(A1) $SC = c(e(z,\hat{e}^0,\hat{r}^0)) + g(z) + E\{D(e(z,\hat{e}^0,\hat{r}^0) + r(z))\}$

As we are focusing on the second-best case, in which $\hat{e}^0$ and $\hat{r}^0$ are treated as fixed parameters, we subsequently drop these terms from our notation. *SC* is minimized by choosing *z* to ensure that $dSC/dz = 0$, or

(A2)
$$\begin{aligned}\frac{dSC}{dz} &= c'(e(z))e'(z) + g'(z) + E\{D'(e(z)+r(z))[e'(z)+r'(z)]\}\\ &= [g'(z) + c'(e(z))] + \mathrm{cov}\{D',r'\} + [c'(e(z)) + E\{D'\}][e'(z) - 1]\\ &= \left(-c'(e(z)) + E\{D'\}\right)\left(-MC_{sav} + MEC + \Phi[e'(z) - 1 + MC_{sav} + MEC]\right)\end{aligned}$$

where $\Phi = [c'(e(z)) + E\{D'\}]/[-c'(e(z)) + E\{D'\}]$. Considering $-c'(e(z)) = E\{D'\}$ in the first-best outcome, then with too few permits in the second-best case we might expect $-c'(e(z)) < E\{D'\}$. Accordingly, in the second best case, $\Phi \in (0,1)$ and also dSC/dz will be of the same sign as

(A3) $-MC_{sav} + MEC + \Phi[e'(z) - 1 + MC_{sav} + MEC]$.

The second-best solution is determined by setting (A3) to zero. Hence, comparison of (A3) with expression (8) indicates the following relation holds

(A4) $\Lambda = \Phi[e'(z) - 1 + MC_{sav} + MEC]$

where $\Lambda > 0$ when evaluated at the second-best optimum.

We can simplify matters by recalling that $e(z) = \hat{e}^0 + (1/t(z))[\hat{r}^0 - r_0 + z]$. With $\hat{r}^0 = r_0$ in extant programs where nonpoint sources have not yet abated, then $e'(z) = (1/t(z))[1 - \varepsilon_{tz}]$, where

$\varepsilon_{tz} = t'(z)z/t(z)$ is the elasticity of *t* with respect to abatement. Further, recall that MC_{sav} always equals 1 – 1/*t*(*z*) in a market equilibrium, so that $t'(z) = t^2 \partial MC_{sav} / \partial z < 0$. We can use this relation to derive

(A5) $\varepsilon_{MC,z} = \varepsilon_{tz} / (t(z) - 1)$,

where $\varepsilon_{MC,z}$ is the elasticity of MC_{sav} with respect to *z*. Note that $\varepsilon_{MC,z} > 0$ when t > 1 (and hence $MC_{sav} > 0$), and $\varepsilon_{MC,z} < 0$ when t < 1 (and hence $MC_{sav} < 0$).

Using relation (A5), we can rewrite expression (A4) as

(A6) $\Lambda = \Phi[-MC_{sav}(1+\varepsilon_{MC,z}) + MC_{sav} + MEC] = \Phi[-MC_{sav}\varepsilon_{MC,z} + MEC]$

An optimal outcome requires $\Lambda > 0$ when (A3) vanishes, i.e.,

(A7) $\Lambda = -MC_{sav}\varepsilon_{MC,z} + MEC > 0$, when $-MC_{sav} + MEC + \Lambda = 0$.

The first part of condition (A7) is clearly violated (so that $\Lambda < 0$) when $MEC < 0$ and $MC_{sav} > 0$ (and hence $\varepsilon_{MC,z} > 0$), and hence such an outcome cannot be an optimal solution. The intuition is that there are only benefits from reducing *t*, and thereby increasing *z*, in this case since increased nonpoint abatement yields both cost savings and environmental benefits (via a reduction in risk). Using the notation of equation (8) in the main text, such an outcome implies a negative shadow value, i.e., $\lambda < 0$: there are additional gains to be made before the constraint on the permit allocation becomes binding. Finally, we cannot rule out condition (A7) being satisfied when *MEC* and MC_{sav} (and hence $\varepsilon_{MC,z}$) are of the same sign and $MEC < MC_{sav}$.

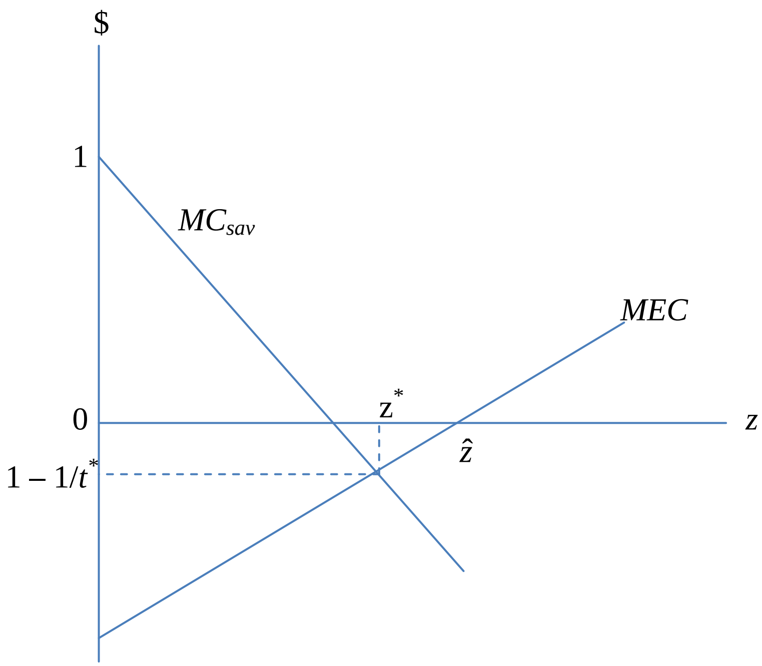

Fig. 1a. First-best outcome when $z^* < \hat{z}$

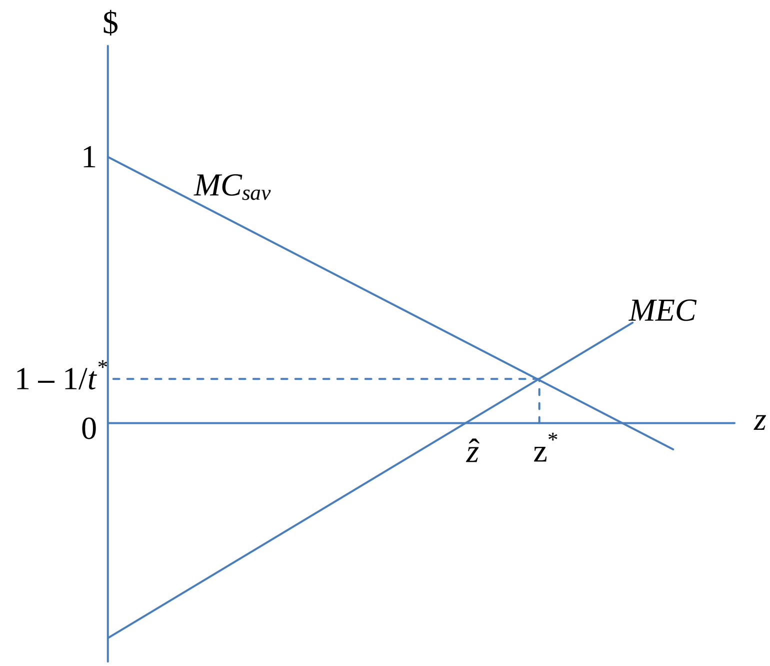

Fig. 1b. First-best outcome when $z^* > \hat{z}$

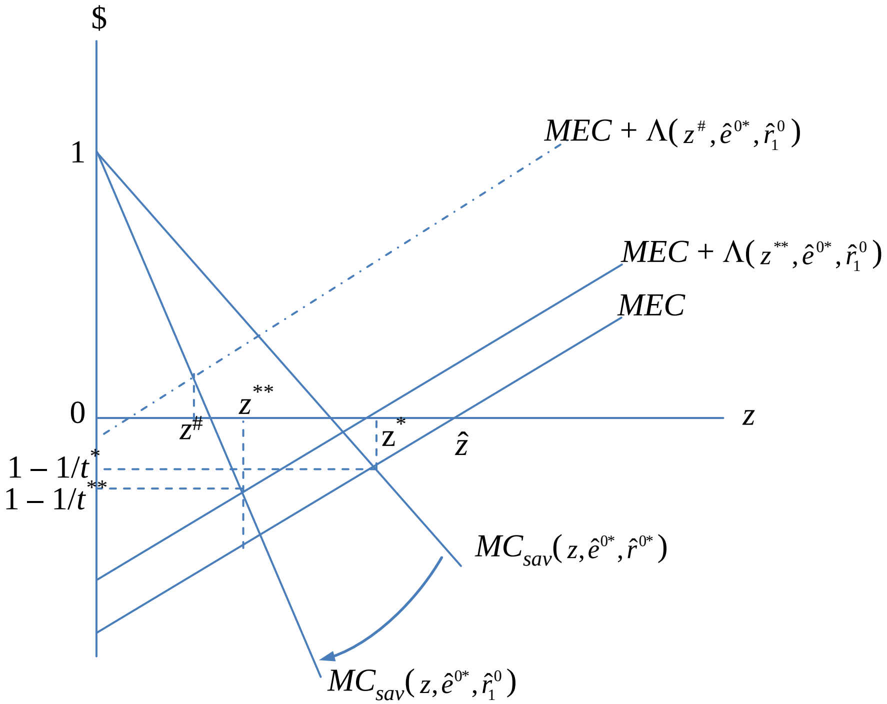

Fig. 2a. Rotation of MC_{sav} when $z^* < \hat{z}$

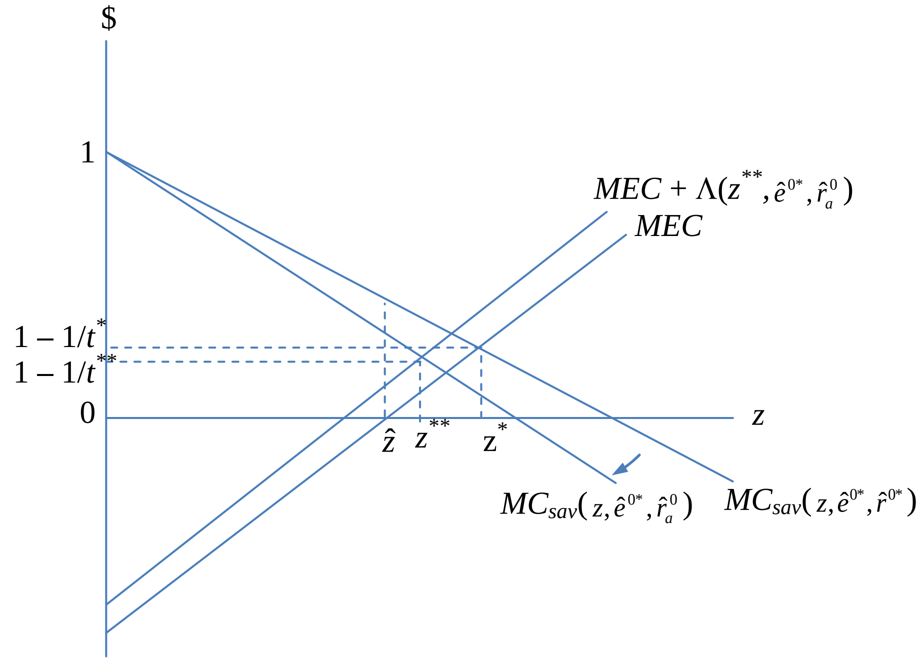

Fig. 2b. Rotation of MC_{sav} when $z^* > \hat{z}$