ECONOMIC AND ECOLOGICAL RULES FOR WATER QUALITY TRADING

Richard D. Horan and James S. Shortle

ABSTRACT: Emissions trading in textbook form uses markets to achieve pollution targets cost-efficiently. This result is accomplished in markets that regulators can implement without knowing pollution abatement costs. The theoretical promise of emissions trading, along with real-world success stories from air emissions trading, has led to initiatives to use trading for water pollution control. Yet, trading, particularly when it involves non-point sources of pollution, requires significant departures from the textbook concept. This paper explores how features of water quality problems affect the design of markets for water pollution control relative to textbook emissions markets. Three fundamental design tasks that regulators must address for pollution trading to achieve an environmental goal at low cost are examined: (1) defining the point and non-point commodities to be traded, (2) defining rules governing commodity exchange, and (3) setting caps on the commodity supplies so as to achieve an environmental target. We show that the way in which these tasks are optimally addressed for water quality markets differs significantly from the textbook model and its real-world analogs. We also show that the fundamental appeal of emissions trading is lost in the case of realistic water quality markets, as market designs that reduce the costs of achieving water quality goals may no longer be implementable without the regulatory authority having information on abatement costs.

(KEY TERMS: water quality economics; point source pollution; nonpoint source pollution; environmental regulations; water quality trading; environmental markets.)


INTRODUCTION

... the atmospheric pollution control authority’s responsibilities...will not have to include the guesswork involved in attempting to estimate individual emitter and receptor preference functions [Crocker (1966, p. 81), in describing the benefits of emissions trading]

Emissions trading programs, in textbook form, utilize the power of private market transactions to achieve pollution targets cost-effectively. This result is accomplished in markets that environmental regulators can construct easily and without substantial information requirements. In particular, environmental regulators can construct cost-minimizing emissions markets without knowing the abatement costs of individual agents. This theoretical capacity of
trading to minimize abatement costs without regulatory authorities knowing individual polluters’ abatement costs makes trading economically appealing in comparison with traditional regulatory tools that are significantly more complex and information-intensive (Crocker, 1966; Hanley et al., 1997).

Water quality trading was the focus of John Dales’ (1968) seminal book recommending markets for environmental management, but has not until recently been a focus of applications of the tool. The theoretical promise of emissions trading, along with real-world success stories from air emissions trading programs and encouragement from the U.S. Environmental Protection Agency (USEPA), has led to substantial interest by state and regional authorities to use water quality trading to overcome the challenges faced in implementing total maximum daily loads (TMDLs). In particular, water quality trading is seen as a way to encourage nonpoint sources of pollution to voluntarily undertake steps to reduce their emissions, while at the same time cost-effectively allocating emissions controls across point and nonpoint sources within a watershed (Shortle and Horan, 2006). Dozens of United States (U.S.) water quality trading initiatives have been attempted or are underway for nutrients, sediments, selenium, temperature, and metals, and involve various combinations of point and nonpoint sources (USEPA, 2010). Reining in nonpoint emissions is vital to satisfying TMDLs in many watersheds. This is because water pollution has largely been regulated through nontradable, technology-based effluent permits applied to point sources of water pollution, whereas nonpoint sources have been largely unregulated (Ribaudo, 2001). The environmental consequence is that water quality goals are not achieved where uncontrolled nonpoint sources are significant causes of water quality problems. The economic consequence is that the pollution control that is achieved, because it does not consider the relative costs of alternative point and nonpoint sources, is overly costly (Davies and Mazurek, 1998; Ribaudo, 2001; USEPA, 2001). U.S. water quality trading initiatives are intended to help address these problems, as are programs in other parts of the world. Australia, Canada, and the Netherlands have active trading programs. Regions in which trading programs are under study or in development include Chile, China, New Zealand, and the nations surrounding the Baltic Sea.

But can water quality trading programs fulfill the theoretical promise of the textbook models? The national scale cap-and-trade markets that have been so successful for controlling air pollutants have been successful because they look like and “...work roughly as the textbooks describe” (Joskow et al., 1998). Yet, water pollution trading, particularly when it involves nonpoint sources of pollution, requires significant departures from textbook markets (Shortle, 1987, 1990; Malik et al., 1993; Horan, 2001). Fundamental features of textbook markets are that emissions (1) can be accurately metered for each regulated emitter and (2) are substantially under control of the emitter, and that (3) the spatial location of emissions is not relevant to the attainment of the environmental target (e.g., Sterner, 2003; Ellerman, 2005). These requirements are not the characteristic of water quality management problems. To the contrary, there is routine uncertainty about sources and levels of emissions, about the response of emissions to abatement effort, and about water quality impacts of emissions from different sources. Moreover, the spatial location of emissions is important to water quality impacts.

The purpose of this paper is to explore how the characteristic features of water quality problems affect the design of water quality markets relative to textbook emissions markets. We restrict our attention to the types of markets that have been implemented in practice, as they are realistic and impose simplifications that reduce transactions costs relative to more complex markets that would promise greater efficiency in theory but that would be impractical in application (Shortle and Horan, 2008). We then examine three fundamental design tasks that regulators must address for pollution trading to achieve an environmental goal at low cost: (1) defining the point and non-point commodities to be traded, (2) defining rules governing exchanges of the commodities, and (3) setting caps on the commodity supplies so as to achieve an environmental target. Simply addressing these tasks does not guarantee the emergence of a market that can fully exploit potential gains from trade. We show that the way in which these tasks are optimally addressed for water quality markets differs significantly from the textbook model and its real-world analogs. Moreover, the fundamental appeal of emissions trading is lost in the case of realistic water quality markets, as market designs that reduce the costs of achieving water quality goals may no longer be implementable without the regulatory authority having information about individual firms’ abatement costs.

**TEXTBOOK EMISSIONS TRADING**

We begin by outlining the basic features of a textbook emissions trading market. All three market design tasks are considered, but we focus particular attention on defining trading rules and the setting of the cap. To keep things simple, we do not formally analyze how transactions costs may affect these choices, other than to note these costs are an impor-
tant reason for our focus on implementing simple market and regulatory structures. Broadly defined, transactions costs associated with emissions trading include exchange costs (Stavins, 1995) and monitoring costs (Malik et al., 1993), and they generally increase with market and regulatory complexity. The water quality markets we examine later on are both simple and realistic, and in many ways they correspond to the idealized markets described in textbooks. We show that, in the case of water quality, emissions trading that follows the textbook case may be ineffective at satisfying water quality goals. Rather, water quality markets may only be effective under market and regulatory structures that imply significant transactions costs. This means transactions costs hamper the appeal of water quality trading.

Task (a): Defining the Tradable Commodity

Market-based approaches to pollution control entail exchanges of property rights, where exchanges take place via trades in a commodity that represents these rights. How should the tradable commodities be defined? Both ecological and economic criteria must be considered. To perform well from an ecological perspective, the commodities must be reasonably accurate predictors of environmental impacts on the time and spatial scales selected for management. Economic criteria require that the tradable commodity be observable at low cost, so that trading is enforceable and not unduly deterred by transactions costs, and that the commodity is under the control of the polluter, so that the polluter can be held responsible for noncompliance (Shortle and Horan, 2008).

Emissions, when they can be metered and controlled deterministically, are the appropriate compliance measure or basis for environmental policy instruments (Oates, 1995). For example, in $\text{SO}_2$ allowance trading, an allowance is an authorization (right) held by a source to emit one ton of $\text{SO}_2$ during a given year or any year thereafter. At the end of each year, the source must hold an amount of allowances at least equal to its annual emissions. Allowances are fully marketable commodities in that they may be bought, sold, traded, or banked for use in future years. The choice of commodity is more complex in the nonpoint case, as we describe below.

Task (b): Defining Trading Rules

Trading rules are used to prevent trades from resulting in a net increase in ambient pollution levels. These rules are typically implemented in the form of trading ratios that specify how much emissions must be reduced at one source to offset a 1-unit increase in emissions at another source. A trading ratio equal to the marginal rate of substitution of emissions between two sources ensures no net change in ambient pollution. For instance, suppose the environmental goal is to limit ambient pollution below some predetermined value. If ambient pollution is proportionate to total emissions, then the environmental goal or constraint can be written as

$$
\sum_{i=1}^{n} e_i \leq a_{\text{max}},
$$

where $e_i$ represents emissions by firm $i = 1, \ldots, n$, and where $a_{\text{max}}$ represents the maximum allowable level of pollution due to human sources (i.e., net of background pollution). In this case, the marginal rate of substitution between any two sources is constant and equal to one, meaning that each source’s emissions are perfect substitutes for one another. The trading ratio that ensures trades does not violate the environmental goal from unity, so that emissions are traded on a one-for-one basis. This is the basic textbook case, and it corresponds to how trades occur within $\text{SO}_2$ and other air quality markets.

The textbook case is easily expanded to include cases where different sources’ emissions are not perfect substitutes in their impacts on environmental conditions. A more general specification of condition (1), consistent with zonal air quality trading programs [e.g., California’s Regional Clean Air Incentives (RECLAIM) program], is

$$
\sum_{i=1}^{n} \beta_i e_i \leq a_{\text{max}},
$$

where $\beta_i$ is a constant delivery coefficient that indicates the proportion of source $i$’s emissions that are delivered to an area of concern (e.g., a nonattainment area). Differences in the $\beta$s may arise due to local differences among the sources. In this case, the marginal rate of substitution of firm $i$’s emissions for firm $j$’s emissions is $|d e_j / d e_i| = \beta_j / \beta_i$, which would also be the appropriate trading ratio between these two sources. Hence, there would be different ratios applied to trades involving different combinations of firms (or firms in different “zones” if a number of firms within a particular “zone” have the same $\beta$). This is the only added complexity of this type of “spatial” trading program, but it is not a significant one. As we describe below, the market essentially works the same when the trading ratios differ from unity. Moreover, estimating the $\beta$s does not require any firm-specific cost information that might be difficult.
to attain. Rather, the $\beta$s could be estimated using the same ecosystem process models that would be employed to predict environmental impacts regardless of whether there are differences in the $\beta$s.

**Task (c): Limiting (Capping) the Aggregate Supply of the Tradable Commodity**

The third task in market design is to limit the aggregate supply of the commodities such that market allocations of polluting emissions do not violate the environmental goal. Reflecting the integration of the three tasks, the cap must be determined jointly along with the commodity definitions and trading rules. This is because trading rules are aggregation rules for commodities, and the cap is a limit on the resulting aggregation. Again, the textbook case is simple: as the tradable commodity is a firm’s emissions and as these are traded on a one-for-one basis, the aggregate supply of permits simply equals aggregate emissions – the left-hand side of Equation (1).

Accordingly, the aggregate supply condition is the same as the environmental constraint, and so the cap on emissions should be set equal to $a_{\text{max}}$. The design of this textbook market is therefore quite simple: allow firms to trade emissions-based permits on a one-for-one basis, after having capped the total number of permits at the level $a_{\text{max}}$.

**A Graphical Representation of the Textbook Case**

We analyze the choices of the trade ratio and the permit cap graphically to provide additional insight. The basic graphical analysis, here and in later sections on water quality trading (i.e., Figures 1 and 2), is based on Shortle (1987, 1990) and Malik et al. (1993) and assumes no transactions costs and that no firm is able to exert market power within the permit market. However, although those studies focused primarily on setting the trade ratio, we focus on setting both the trade ratio and the cap, and also on the information required to make these choices so that the market functions properly.

The textbook tradable emissions permit market is depicted in Figure 1 for the case of two firms. Line segment $T$ represents the environmental goal. Only combinations of emissions lying on or below the segment $T$ are feasible; all others violate the environmental goal (Equation 1). Segment $T$ also represents the set of feasible trades, as determined by the permit supply relation. The slope of segment $T$ is the trade ratio (unity), and the horizontal and vertical intercepts of segment $T$ (both set at $a_{\text{max}}$) ensure that any combination of emissions along this segment always equals $a_{\text{max}}$. An initial allocation of permits must lie somewhere on segment $T$, and the trade ratio ensures that the final allocation of permits will also lie on $T$ and will therefore satisfy the environmental goal.

Given these relations, what will be the market equilibrium outcome for this trading market? The answer requires the consideration of firms’ abatement costs. Abatement costs are illustrated by isocost curves such as $C'$ and $C_0$, each of which represents combinations of emissions levels that result in a

---

**FIGURE 1. The Textbook Case of Emissions Trading.**

**FIGURE 2. Point-Nonpoint Trading With No Cost Uncertainty.**
particular level of aggregate costs (i.e., abatement costs for Firms 1 and 2). For instance, suppose firm $i$’s abatement costs are given by $c_i(e_{i0} - e_i)^2$, where $c_i$ is a cost parameter and $e_{i0}$ is the firm’s (fixed) unregulated level of emissions (so that $e_{i0} - e_i$ is abatement). Then total costs for two firms is given by $C = c_1(e_{10} - e_1)^2 + c_2(e_{20} - e_2)^2$. Various combinations of $e_1$ and $e_2$ can produce a given value of $C$. Holding $C$ fixed, the formula for an isocost curve is $e_2 = e_{20} - [(C - c_1(e_{10} - e_1)^2)/c_2]^{0.5} = F(C, e_1)$. The relation $F$ is decreasing and convex in $e_1$, as in Figure 1. It is also decreasing in $C$, which means a decrease in $C$ shifts the isocost curve to the right in Figure 1: lower abatement costs correspond to more emissions. This means that isocost curve $C^*$ represents lower aggregate abatement costs than does $C_0$.

The equilibrium outcome is determined by individual firms seeking to minimize their pollution control costs, given the market parameters. For instance, suppose the initial permit allocation occurs at $e_{10}$ and $e_{20}$ on isocost curve $C_0$. From here, overall costs can be reduced to $C_1$ if a trade allows Firm 2 to increase its emissions to $e_{21}$, whereas Firm 1 must decrease its emissions to $e_{12}$ to satisfy the environmental goal $T$. Firms will have additional incentives to trade as long as trades along the curve $T$ continue to reduce the overall costs. Total abatement costs will be at their lowest level on $T$ at the point to $e_{12}$ and $e_{21}$ on isocost curve $C^*$. Specifically, firms will choose to trade until the isocost curve is tangent to the permit supply curve, at which point the ratio of their marginal abatement costs equals the trade ratio (Shortle, 1990; Malik et al., 1993).

It is important to note that the results in Figure 1 are easily extended to the case where different sources’ emissions are imperfect substitutes, as in Equation (2). The only difference in Figure 1 would be that the slope of segment $T$ would change to reflect the new trading ratio, as would the intercept terms. However, all trades would still occur along the curve, and any point on the curve would still satisfy the environmental constraint. Moreover, the two firms would take the trading ratio into account when making their trading decisions, so that the equilibrium outcome would still occur at the point of tangency between the permit supply curve and the isocost curve.

Note that the regulatory authority does not require firm-specific cost information to design the market to attain the least cost outcome: the commodity choice (emissions), the trade ratio choice (the marginal rate of substitution, which is constant), and the permit cap choice ($\alpha_{max}$) are all chosen based on the ecological constraint. This is the power of textbook markets. The regulatory authority only has to understand the ecology of the problem, whereas firms only have to understand the economics of their own problem (own abatement costs and the incentives provided by the regulatory authority’s ecologically based regulatory choices). We now turn to the case of water quality trading, where the basic features of the problem significantly alter these results.

WATER QUALITY TRADING

Water quality trading that involves both point sources and nonpoint sources of pollution becomes significantly more complex than the textbook case described above.

Task (a): Defining the Tradable Commodity for Nonpoint Sources

Emissions remain the preferred tradable commodity for point sources of pollution, but trades involving metered emissions are not plausible for nonpoint sources. Nonpoint emissions by individual agents are highly stochastic (e.g., due to weather events that drive nutrients and other emissions off farm fields) and imperfectly observable (i.e., observable with considerable error, or largely unobservable, for instance due to emissions leaving farm fields diffusely). The stochastic nature means that nonpoint sources cannot perfectly control actual emissions; they can only control the probability distribution of their emissions. Largely unobservable emissions, combined with imperfect control over emissions, make imposing limits on actual emissions impossible. These features imply that the tradable nonpoint commodity cannot be defined in terms of measurable emissions. In consequence, some other observable construct must serve as the basis for defining the tradable commodity.

The choice of basis for nonpoint instruments, whether for trading or other incentive or regulatory mechanisms, has been much discussed in the economic literature on nonpoint pollution (Shortle and Horan, 2001). The choices that have drawn the most interest as a tradable commodity are permits based on observable inputs or practices that affect nonpoint pollution (e.g., fertilizer or manure applications to farm land, farming practices that affect sediment or nutrient runoff), and emissions estimates constructed from observations on input and practices (e.g., estimates of pollution delivery from farm fields using observations of inputs and practices as data).

Input trading makes the most sense when the flow of nonpoint externalities can be closely tied to a single input or practice. A good example is the Dutch
manure quota, which was introduced at the farm level in the late 1980s and made tradable in the 1990s (Wossink, 2004). In that case, negative air and water quality externalities were directly tied to massive volumes of animal manure from intensive livestock production. In general, however, numerous choices combine to control the distribution of non-point pollution loads from a given location. For example, nitrogen pollution from a farm will depend on the amounts, timing, and manner that fertilizer or animal manures are applied, the crops grown, tillage practices, and the use of conservation practices that intercept runoff. The complexity of designing a trading program that specifically targets individual inputs and practices would be great, as would be the complexity and transactions costs of participation. This suggests the use of commodity definitions that aggregate over individual choices into a smaller set of environmental performance indicators.

Emerging trading programs in the U.S. take this approach. Specifically, estimated emissions trading is the method of choice in U.S. water quality trading programs. For example, Pennsylvania’s nutrient credit trading program is designed to reduce nitrogen and phosphorous loads from point and agricultural nonpoint sources to the Chesapeake Bay. The state’s Department of Environmental Protection has developed a spreadsheet that farmers can use to calculate nitrogen or phosphorous reduction credits from the implementation of agricultural best management practices (BMPs) from an approved list (PADEP, 2006). The credits are estimates of the steady-state annual average reduction in the levels delivered to the Bay from a farm. The spreadsheet uses estimates of the nitrogen reduction efficiencies of BMPs to calculate the reduction of nitrogen loads at the farm, and applies two factors from the USEPA’s Chesapeake Bay model to estimate the proportion of nutrients that move from farms to the Bay. More generally, EPA trading guidelines (USEPA, 2007) focus on estimated reductions in emissions. In what follows, we also focus our attention on point-nonpoint trading programs in which estimated emissions are defined as the tradable nonpoint source commodity.

An important aspect of this approach should be noted in comparison with trading actual metered emissions. There is enormous uncertainty about the actual water quality outcomes of individual trades based on modeled emissions. This exists because the prediction errors for water quality models are known to be quite large (e.g., Reckow, 1994). Water quality policy makers are concerned with the likelihood that water quality objectives are achieved (Lichtenberg and Zilberman, 1988). The U.S. TMDL program requires, for example, sufficiently stringent load limits to achieve water quality goals with a margin of safety. Addressing the risk requires explicit consideration of the uncertainty about water quality outcomes of trades related to model error, the inherent variability of nonpoint loads, and management factors that influence the actual results of the application of BMPs. Strongly assuming that estimated emissions are unbiased, trading estimated emissions directly controls the average emissions level but not the variability of emissions. When trading estimated emissions, farmers, or other nonpoint sources, will choose the set of practices that minimize their cost of producing reductions in the estimate. Nonpoint sources have no inherent incentive to expend resources to provide a reduction in the estimated load with any degree of reliability.

Emerging trading programs address this risk using trading ratios between point and nonpoint sources that are intended to produce a margin of safety in trades with nonpoint sources. We will take the design of trade ratios up in the next section. But we note here that trade ratios that are applied uniformly across nonpoint source trades cannot communicate information about the risk associated with individual trades. Thus, two nonpoint sources who offer up an equivalent number of credits at an equivalent price will be viewed as perfect substitutes in a market that trades based on estimated emissions, even if the variability of the associated pollution reductions are quite different. In consequence, risk will be suboptimally addressed. This feature will not be apparent in our two-source graphical analysis below, as that analysis includes a single nonpoint source, but it will be an issue for markets involving multiple nonpoint sources.

Task (b): Defining Trading Rules

In the textbook case, trading rules are defined as trading ratios that prevent trades from violating ambient pollution targets. Given the highly stochastic nature of nonpoint emissions, the effects of trade on the ambient environmental conditions must be viewed probabilistically rather than deterministically. Prior work on water quality trading has focused on probabilistically defined goals based on damages or ambient pollution (e.g., Shortle, 1987, 1990; Malik et al., 1993; Horan, 2001; Shortle and Horan, 2008). For instance, an approach that addresses the risk associated with nonpoint uncertainty, as well as other factors that may lead to stochastic ambient water quality, is a probabilistic “safety-first” goal of the form \( \Pr (a \leq a_{\text{max}}) \geq \alpha \), where \( \alpha \in (0,1) \) is the minimum acceptable probability of achieving the goal \( a_{\text{max}} \) (Beavis and Walker, 1983; Lichtenberg and Zilberman, 1988; Lichtenberg et al.,...
1989; Shortle et al., 1999). The water quality goal can be written as:

\[
\Pr \left( \sum_i \beta_i e_i + \sum_j \lambda_j r_j(x_j, \omega_j) \leq a_{\max} \right) \geq \alpha,
\]

where \( r_j \) represents actual emissions by nonpoint source \( j \), which depend on source \( j \)'s vector of input choices \( x_j \) (e.g., farm management practices and applications of chemicals like fertilizer and pesticides) and stochastic variables that influence nonpoint pollution \( \omega_j \) (e.g., rainfall), and \( \lambda_j \) represents the delivery coefficient for emissions by nonpoint source \( j \) (which is stochastic and possibly covaries with \( r_j \)). This “safety-first” approach is consistent with the U.S. TMDL program, as a TMDL specifies that ambient pollution (total loads) should remain below an upper bound, \( a_{\max} \), with a high probability, \( \alpha \).

Assuming that nonpoint permits are defined in terms of estimated emissions, we express the nonpoint commodity as \( r_j = E[r_j(x_j, \omega_j)] \). Nonpoint sources will make their input decisions to minimize their costs of attaining the level of \( r_j \) that their permit holdings allow (Shortle and Horan, 2008). Accordingly, a nonpoint source’s input choices, \( x_j \), depend on its permitted estimated emissions level, \( \bar{r}_j \). We write this relation between inputs and permitted emissions as \( x_j(\bar{r}_j) \). In turn, the realized nonpoint emissions level can be expressed as a stochastic function of the permitted emissions, \( r_j(\bar{r}_j, \omega_j) = r_j(x_j(\bar{r}_j), \omega_j) \). Hence, Equation (3) becomes

\[
\Pr \left( \sum_i \beta_i e_i + \sum_j \lambda_j r_j(\bar{r}_j, \omega_j) \leq a_{\max} \right) \geq \alpha.
\]

The environmental objective (Equation 4) is therefore conditional on the levels of the tradable commodities for each source type: it depends on point source emissions and on nonpoint source estimated emissions. Moreover, this environmental objective is nonlinear in \( e_i \) and \( r_j \), for all \( i \) and \( j \), which means the marginal rate of substitution between the estimated emissions from different sources will be nonconstant: the value will depend on the current levels of estimated emissions produced by each source, and there will be (except in trivial limiting cases) a different marginal rate of substitution for each combination of sources involved in a trade.

The marginal rate of substitution is trade-contingent because the tradable commodity (estimated emissions) is imperfectly related to the environmental objective (Equation 4). With the tradable nonpoint commodity being estimated (or mean) emissions, nonpoint sources have incentives to adjust their input choices to control their mean emissions. But these input choices also influence the variability of emissions, which influence the ability to attain Equation (4). The problem is that a market based on estimated emissions does not reward (penalize) firms who do (not) consider impacts of their choices on the variability of emissions, and so firms have no incentive to consider these impacts. In turn, these unintended variability impacts result in a nonlinear relationship between estimated emissions and the environmental objective (Equation 4), resulting in a nonconstant marginal rate of substitution (Shortle and Horan, 2008).

It may help to restate the significance of these results. The tradable commodity that we have described is practical and realistic, as the regulatory agency judges a firm’s compliance only based on that firm’s mean emissions – not its emissions variability. However, choosing not to monitor variability to determine compliance makes it harder to ensure the water quality goal (Equation 4), which is also realistic and which does depend on variability, is satisfied. The result is that choosing a simple, practical compliance measure means some other aspect of the trading program must become more complex if we are to ensure the water quality goal is satisfied. In particular, it is no longer possible to adopt a fixed trading ratio in the manner outlined above. This is because the marginal rate of substitution of (estimated) emissions between two sources is not fixed. Moreover, we describe below that trades based on estimated emissions and a constant trade ratio could violate the goal. As the agency cares about whether the water quality goal is achieved, the trading ratio must be determined in an alternative way. We emphasize that these arguments have nothing to do with economic efficiency – at this point, we are only describing what is necessary to achieve the environmental goal, given the choice of a simple measure of compliance and firm responses to this choice.

Fixed trading ratios can still be used in this setting to specify how much emissions must be reduced at one source to offset a 1-unit increase in emissions at another source. But determining the appropriate value for the trading ratio is now more complex. We can no longer simply set a trading ratio equal to a unique ecologically determined marginal rate of substitution of (estimated) emissions between two sources, because this rate is no longer unique. Additional economic information is required to determine the appropriate rate. Specifically, the required information is that which allows us to predict how firms might respond in water quality markets having different values of the trading ratio. Indeed, the prediction of firm responses is needed to ensure the chosen trade ratio will lead to
satisfaction of the water quality goal. Only firm-specific cost information will enable us to make such predictions, as is illustrated graphically below.

A key feature of textbook emissions trading programs is that the regulatory authority could construct an emissions market using only ecological data. Information on firm costs was not required, which is beneficial because firms are often unwilling to share their private cost information. But ecological data alone are insufficient to design a water quality market—data on economic costs are also required to make policy choices. This means that, in the case of water quality trading, emissions trading programs lose a key advantage relative to other types of policy tools for which cost information is also required to make good policy choices.

Task (c): Limiting (Capping) the Aggregate Supply of the Tradable Commodities

The third task in market design is to limit the aggregate supply of the commodities to ensure the environmental goal is not violated. In the textbook case described above, we said that the cap must be determined jointly with the trading rules, and the same is true here. But there is a fundamental difference in the case of water quality markets: the aggregate market supply condition now takes the form:

$$e_1 + \sum_{i \neq 1}^m \frac{e_i}{t_i} + \sum_j^m \frac{f_j}{t_j} \leq Z,$$

where $t_i^j$ is the trading ratio between Point Source 1 and point source $i$ (i.e., $|d e_i / d e_1| = t_i^j$), $t_j^j$ is the trading ratio between Point Source 1 and nonpoint source $j$, and $Z$ is the cap on aggregate estimated emissions. The trading ratios adjust for the differential environmental impacts of each source’s (real or estimated) emissions, so that the aggregate supply condition is denominated in terms of Point Source 1’s emissions (any other source could have been used to define the units after an appropriate rescaling of the condition).

Note that the trade ratio for trades between nonpoint source $j$ and point source $i \neq 1$ is

$$\frac{|d f_j^j|}{|d e_i|} = t_j^j / t_i^j.$$  \hspace{1cm} (6)

In contrast to the textbook case, the aggregate supply condition (Equation 5) now differs markedly from the environmental constraint (Equation 4). Comparison of these conditions indicates that cap on estimated emissions, $Z$, should not be set equal to environmental target $a_{max}$. Rather, the cap will be determined jointly along with the trading ratios to ensure that a market equilibrium satisfies the environmental constraint (Equation 4). Further, as we stated earlier, the trading ratio will depend on cost information, so too will the choice of the cap. This was not true in the textbook case.

Prior extensions of the textbook permit markets have found that, when market structures are not perfectly competitive, cost data may be necessary to design optimal markets capable of producing emissions reductions at least cost (e.g., Hahn, 1984). However, we have said nothing yet about the optimality of the market parameter choices (trade ratio and cap). Our only result so far is that ecological information alone is insufficient for designing a market that ensures the water quality goal is met. This is not the case with the idealistic textbook framework (e.g., Montgomery, 1972), nor with extensions of that work. Of course, cost information will be of additional value if we aim to construct an optimal permit market. We now turn to a graphical representation to illustrate water market design choices.

A Graphical Representation of Water Quality Trading

The water quality trading market is depicted in Figure 2 for the special case of two firms, a nonpoint source and a point source, assuming the regulatory authority has perfect knowledge of firm costs (we relax this latter assumption below). It is important to note that our results are generalizable to more than two firms, but we cannot depict that case in two dimensions (i.e., we would need $m + n$ dimensions to depict the case of $n$ nonpoint and $m$ point sources). In Figure 2, curve $T$ represents the environmental goal. Only combinations of emissions lying on or below the curve $T$ are feasible; all others violate the environmental goal (Equation 4). This curve exhibits a nonlinear relationship, in contrast to Figure 1. Although we have drawn the curve to be strictly concave, this may not be the case (e.g., Beavis and Walker, 1983). However, only in special cases, which would be unlikely to arise when there are more than two firms, would the curve be linear. The nonlinearity of $T$ reflects the nonconstant marginal rate of substitution between the estimated emissions of the two sources. Specifically, the marginal rate of substitution is given by the slope of $T$, which varies along $T$. We noted above that the trading ratio should be set equal to the marginal rate of substitution between the estimated emissions from these two sources in satisfying the environmental goal (Equation 4). Is that still the case? And, if so, which rate, among the many options, should be chosen? Clearly the curve $T$, which reflects the ecological aspects of the problem, does not provide enough information to make this determination.
Another noteworthy aspect of $T$ is that the vertical intercept, implicitly defined by the relation $Pr(x_t(r_j) | x_j(r_j)) \leq a_{\text{max}} \geq z$, is unknown without information on nonpoint abatement costs. The reason is that the nonpoint source’s response function, $x_j(r_j)$, depends on the source’s abatement costs. Uncertainty in this response function can be captured by our probabilistic water quality goal, Equation (4), though greater uncertainty will generally imply a more stringent relation.

The environmental goal and the permit supply relation coincided in Figure 1. However, as indicated above, these relations do not coincide in the present case if we adopt fixed trading ratios. The supply relation is given by Equation (5), which is linear. With two sources, the relation can be written as:

$$r = t'(Z - e). \tag{7}$$

Two potential supply relations are denoted by the segments $S_1$ and $S^*$ in Figure 2. Each is downward sloping with a slope equal to the point-nonpoint trading ratio, $t'$. The horizontal intercept is the permit cap, $Z$.

Pollution sources have no incentives to consider $Z$ directly; they will only exchange pollution rights in response to the supply relation, along which initial and post-trade permit allocations must lie. Accordingly, the supply relation must lie everywhere below $T$, as is the case of $S_1$, to ensure that the water quality goal is satisfied for all feasible permit allocations. The trade ratio and emission cap can be chosen to locate $S_1$ below $T$. In our discussion of Figure 1, we indicated a market equilibrium will occur where the firms’ isocost curve is tangent to the permit supply relation. The same is true here. The market equilibrium associated with supply relation $S_1$ is point $A$, which implies a cost of $C_1$. As drawn, this equilibrium over-controls pollution relative to the water quality goal $T$, and therefore implies excessive abatement costs relative to policies that could just satisfy $T$. Note that greater uncertainty about the nonpoint source’s abatement costs (and hence its response function, $x_j(r_j)$) will generally result in a smaller vertical intercept of $T$, which implies $S_1$ must rotate downward (counterclockwise from the fixed horizontal intercept) if we are to ensure $Z$ is everywhere satisfied. This can be accomplished via a smaller trade ratio, though it will imply even greater control costs.

Consider a market design that theoretically just satisfies the goal $T$ (still assuming abatement costs are known). Specifically, the supply curve must be chosen, via the choices of the trading ratio and the cap, to ensure firms will trade their way to an equilibrium point on $T$. As described above, firms will trade until the isocost curve is tangent to the supply relation. This market equilibrium will only arise on $T$ if the supply relation is also a tangent to $T$ at the equilibrium point. This market equilibrium, denoted by point $B$ in Figure 2, is attained by setting the trading ratio, $t'$, equal to the slope of the tangent of curves $C'$ and $T$ at point $B$, and by setting the cap then satisfies the condition (in Equation 5) as $Z' = e' + r' / t'$. Finally, as the isocost curve is tangent to $T$ at point $B$, this outcome satisfies the water quality goal at least cost overall. The market design that produces this outcome is the optimal or least cost design.

Although the literature on point-nonpoint trading (e.g., Shortle, 1990; Malik et al., 1993; Shortle and Horan, 2008) generally promotes the least cost market design, there are several things to note about equilibrium $B$ in Figure 2. First, point $B$ is the only outcome out of all allowable trades, as indicated by the supply relation $S'$, that satisfies the water quality goal. Any initial permit allocation other than $B$, which must lie on $S'$, will fail to satisfy the water quality goal! Prior work has not addressed this issue. That all but one permit allocation violates the water quality goal is not a concern if, as we have assumed, there are no transactions costs and the regulatory authority has perfect information on firm costs. These assumptions imply the regulatory authority can perfectly predict the equilibrium outcome in the trading market, and that outcome $B$ will emerge so as to satisfy the water quality goal. If trades generate transactions costs, however, then trading is unlikely to result in an outcome of $B$, and the water quality goal will be violated. Herein lies the tradeoff between the environmental goal and abatement costs: the goal will be achieved at high cost using supply relation $S_1$, whereas the goal will be violated, though at low costs, when using the supply relation $S'$. In either case, the appeal of emissions trading is reduced.

If the regulatory authority has imperfect information on firm costs, then this cost uncertainty will impact upon the optimal market design problem via impacts to both the isocost curves and the water quality goal relation. Consider increased uncertainty about the isocost curve, which means it will be more difficult to predict the point of tangency with the supply relation. A regulatory solution that accounts for this additional uncertainty will result in the permit supply curve intersecting curve $T$, as in Figure 3. Specifically, the supply of permits, $Z'$, is set at a lower level than $Z$ and the trading ratio is reduced to $t'$ as precautions to prevent firms from trading to an allocation that violates $T$. This increases the likelihood that the resulting trading equilibrium will satisfy the environmental goal, though the expected costs of doing so will increase.
Greater uncertainty about the nonpoint source's abatement costs also affects the water quality goal via the nonpoint source's response function, \( x_j(\bar{r}_j) \). This uncertainty is akin to an increase in the uncertainty associated with nonpoint source emissions. Starting with our baseline model in Figure 2, what is the effect of an increase in the variability of nonpoint emissions? The answer depends on whether the mean and variance of emissions move together. Assuming they do, then an increase in nonpoint risk would tend to rotate curve \( T \) inward, to \( T' \) in Figure 3, as greater reductions in nonpoint emissions are required to ensure the environmental goal is satisfied. Accordingly, we may expect some combination of reductions in \( Z \) and \( t \) to move the permit supply curve inward as well, reinforcing the regulatory responses described above for uncertainty about the isocost curve. The opposite result arises if the mean and variance of nonpoint emissions covary negatively. A positive covariance seems more likely. For instance, as mean nonpoint emissions are driven to zero then the variance would also have to approach zero (as negative emissions are not possible).

Comparison of Figures 2 and 3 indicates that increases in abatement cost uncertainty are likely to: (1) reduce the permit cap, from \( Z^* \) to \( Z' \), thereby encouraging more stringent controls by all sources so as to mitigate the increased risk, and (2) reduce the trading ratio, from \( t^* \) to \( t' \), thereby reducing the cost of purchasing allowances from nonpoint sources, encouraging more nonpoint source controls that reduce both the mean and the variability of these emissions. Both of these recommendations run counter to what has been implemented in practice, however. In practice, trading ratios are generally increased (often significantly) in response to nonpoint risk, and caps are generally not initially applied to nonpoint sources. Both of these practices further limit both the efficiency and environmental gains that can be made via water quality markets.

**CONCLUSIONS**

We began this paper with a question: can water quality trading programs fulfill the theoretical promise of the textbook models? Even under the best of circumstances, it may be unreasonable to expect pollution markets to achieve theoretical optima given bounded rationality of traders, transactions costs, and other factors. But even disregarding these issues, we have identified a fundamental barrier to designing water quality markets that trade estimated emissions to improve the efficiency of pollution control allocations across sources. That barrier is the potentially high regulatory costs associated with the regulator's need to know polluters' abatement costs in order to design markets that can improve the efficiency of achieving environmental outcomes. This information need stems from the indirect management of the variability of stochastic emissions that results when the tradable commodity is defined in terms of estimated emissions. In this case, the information needed is no different from the information needed for efficient technology-based emissions (or estimated emission) limits.

This result diminishes, but does not eliminate the appeal of trading within the context of existing Clean Water Act regulations. For example, agreements that allow point sources to meet permit requirements through offsite reductions in nonpoint pollution rather than onsite effluent reductions may lead to efficiency gains. But it is clearly incorrect to assume that the gains promised by idealized textbook models will simply materialize in similarly constructed water quality trading markets: if you build it, they may not come. Gains will only materialize in water quality settings if the regulatory choices are made appropriately. Ecological information alone is insufficient to make these choices. Both economic and ecological information is required, and the choices should be made based on a sound conceptual framework that properly integrates this information.

We also hope that our analysis prompts interest in new models of water quality trading. The limitations we identify stem from a loss of degrees of freedom in the management of nonpoint pollution that result...
from defining the nonpoint commodity in terms of estimated emissions. We speculate that these problems may be alleviated by defining the nonpoint commodity in terms of multiple attributes that better indicate the effect on the probability of impacts on water quality conditions. Such a market would consist of multiple nonpoint commodities that improve upon the ability to meet the environmental goal. For instance, nonpoint sources might make trades in both the mean and the variance of their emissions – in some sense trading both the level of control and the quality of that control. A multiple attribute market might also alleviate current concerns over post-trade liability. Currently, point sources may be held liable for a permit violation in cases where a trade with a nonpoint source failed to produce the controlled environmental improvement. It could be that a nonpoint source undertakes appropriate actions to reduce mean emissions, but a violation still arises because variation is insufficiently reduced. Being able to trade in the variance of emissions, in addition to the mean, might reduce the probability of permit violations and increase the gains from trading for both point sources and for the broader society.

LITERATURE CITED


