HHS Public Access

Author manuscript

J Policy Anal Manage. Author manuscript; available in PMC 2021 June 08.

Published in final edited form as:

J Policy Anal Manage. 2015 ; 34(3): 537–566. doi:10.1002/pam.21836.

Does Collaboration Make Any Difference? Linking Collaborative Governance to Environmental Outcomes

Tyler Scott [Assistant Professor]

Assistant Professor in the School of Public and International Affairs at the University of Georgia, 204 Baldwin Hall, Athens, GA 30602.

Abstract

This paper addresses two research questions: (1) Does collaborative environmental governance improve environmental outcomes? and (2) How do publicly supported collaborative groups with different levels of responsibility, formalization, and representativeness compare in this regard? Using a representative watershed quality data series, the EPA's National Rivers and Streams Assessment and Wadeable Streams Assessment, in conjunction with a watershed management regime database coded for this analysis, I test the relationship between collaborative governance and watershed quality for 357 watersheds. Since these are observational data, a multilevel propensity score matching method is used to control for selection bias. Using an augmented inverse propensity weighted estimator, I estimate the average treatment effect on the treated for six different water quality and habitat condition metrics. Collaborative watershed groups are found to improve water chemistry and in-stream habitat conditions. I then use hierarchical linear regression modeling to examine how group responsibilities, membership diversity, and formalization affect the predicted impact of a collaborative group. Groups that engage in management activities (in comparison to coordination or planning) are found to achieve greater environmental gains. Limited differentiation is found with regards to the presence of a group coordinator, increased goal specificity, or greater stakeholder diversity.

INTRODUCTION

"Collaborative governance" and "collaborative management" are normatively popular concepts that have been widely employed in environmental policy applications worldwide (Ansell & Gash, 2008; Hall & O'Toole, 2000, 2004; Innes & Booher, 2004; McGuire, 2006; Newig & Fritsch, 2009a). Collaboration has been shown to enhance cooperation and foster belief change among stakeholders (Leach et al., 2013; Lubell, 2004a), generate funds and support for alternative policy measures when problems are too diffuse or difficult to address via regulation (Margerum, 2011), and increase the implementation success of policies and programs (Agranoff & McGuire, 2003; Meier, 2005). However, we still know very little about the relationship between collaboration and environmental outcomes (Carr et al., 2012; Koontz & Thomas, 2006) or how the environmental outcomes of collaborative approaches compare to those of other policy alternatives (Margerum, 2011; Schneider et al., 2003).

(tyler.scott@uga.edu).

This issue, whether collaborative environmental governance improves environmental outcomes, is the primary research question addressed in this analysis. My second research question builds upon the first, asking: What design and implementation characteristics make groups more or less effective at improving environmental outcomes? Research shows that collaboration alone does not necessarily yield improved outcomes (Newig & Fritsch, 2009a), but there is little existing evidence informing how policymakers might best wield collaborative governance as a strategic, context-appropriate policy tool. I address these questions using a common application of collaborative environmental governance—collaborative watershed planning and management groups (Gerlak et al., 2012; Grafton & Hussey, 2011; Hoornbeek et al., 2012; Imperial, 2005; Lubell, 2004a, 2004b, 2004c; Mazmanian & Kraft, 2009; Sabatier, 2005; Thomas & Koontz, 2011). The next section describes how this analysis fits within—and builds upon—the existing literature. I then specify the empirical approach used in this analysis and explain how it is appropriate given the data and research questions. Subsequent sections then detail my data collection process and coding scheme, present model results, and discuss findings.

THEORETICAL RATIONALE

As a wide and growing body of synthesis literature attests, public policy scholars are interested in studying the role and impact of collaborative governance in a variety of policy sectors (Bingham & O'Leary, 2008; Carr et al., 2012; Donahue & Zeckhauser, 2011; Huxham, 2003; Innes & Booher, 2010; McGuire, 2006; O'Leary & Bingham, 2009; Sabatier, 2005). However, much of this work concerns the quality of the collaborative process (Ansell & Gash, 2008; Coglianese, 1997; Frame et al., 2004; Leach, 2006; Leach et al., 2002; Langbein, 2000; Lubell, 2005; Sabatier & Shaw, 2009) or addresses changes in intermediate outcomes such as: (1) stakeholder cooperation and consensus (Collins et al., 2007; Fuller, 2009; Lubell, 2004a, 2005; McGuire & Silvia, 2010; Schively, 2007; Susskind, 1996; Weible et al., 2004); (2) the production of plans and other outputs (Beierle, 2002; Biddle & Koontz, 2014; Innes, 1996; Innes & Booher, 1999; Leach & Sabatier, 2005; Lubell, 2005; Margerum, 2011; Newig & Fritsch, 2009b; Wondolleck & Yaffee, 2000); and (3) stakeholder perceptions of outcomes (Leach, 2006; Leach et al., 2002; Lubell, 2004c; Provan & Milward, 1995; Ulibarri, 2015) that result from collaborative approaches (Carr et al., 2012; Koontz & Thomas, 2006). For instance, Lubell (2005) shows how collaborative groups with strong procedures and well-codified practices can enhance stakeholder trust and collective action beliefs, thereby increasing support for collaborative policy efforts.

Procedural and intermediate outcomes can be significant in their own right, but it is important to recognize that policymakers use collaborative governance as a tool for improving policy outcomes (Hoornbeek et al., 2012; Koontz et al., 2004). In other words, policymakers purposefully choose to engage in collaborative planning and management (Huxham, 2003; Vangen & Huxham, 2003) as a means by which to "make or implement public policy or manage public programs or assets" (Ansell & Gash, 2008, p. 544). Relatively few works have focused on the role of government in initiating and supporting collaborative groups in this fashion (Huxham, 2003; Koontz et al., 2004; Mandell, 1999, 2001; Schneider et al., 2003; Vangen & Huxham, 2003).

Initiating and maintaining collaborative governance takes time and effort; accordingly, for policymakers there are "trade-offs associated with participating in… collaborative efforts that divert scarce resources from other activities" (Layzer, 2008, p. 290). These trade-offs naturally raise a question of efficacy: Does collaborative governance improve environmental outcomes? While there are many in-depth case studies that point to specific outcomes (Ansell & Gash, 2008; Margerum, 2011; Newig & Fritsch, 2009b), there is little systematic evidence in this regard (Carr et al., 2012; Koontz & Thomas, 2006). Collaborative governance is believed to help facilitate decisionmaking, better address interrelated problems, carry greater legitimacy, and improve implementation (Sabatier et al., 2005). Other benefits attributed to collaboration include access to information, implementation support, and reduced conflict (Gigone & Hastie, 1993; Hill & Lynn, 2003; Moreland et al., 1993; Sabatier et al., 2005; Susskind et al., 1999). At the same time, collaborative processes can be time consuming and difficult (Margerum, 2011), and there are legitimate concerns about whether collaborative institutions are, as asked by Lubell (2004b), "all talk and no action." My primary hypothesis (H1) is that collaborative watershed governance results in improved environmental outcomes:

H1: Collaborative watershed governance results in improved environmental outcomes.

Policymakers not only face the general choice of whether to support collaborative governance, but also regarding the specific form that their collaborative efforts will take. The current literature contains several typologies and theoretical frameworks that characterize collaborative groups in terms of (1) conceptual themes such as geographic scale, institutional scale, inclusiveness, or stakeholder incentives (Ansell & Gash, 2008; Cheng & Daniels, 2005; Emerson et al., 2012; Margerum, 2011); or (2) comparisons between agency-led and independent collaborative institutions (Bidwell & Ryan, 2006; Moore & Koontz, 2003). For instance, (Margerum, 2008, 2011) distinguishes between the institutional scales on which collaboration occurs, while Moore and Koontz (2003) characterize groups in terms of seating (e.g., agencybased or stakeholder-based). However, none of these typologies pertain specifically to the choices policymakers face when designing and implementing a collaborative group within a given institutional context.

Ansell and Gash (2008) and Emerson et al. (2012) each pose prominent theoretical frameworks that identify key variables, such as participatory inclusiveness and stakeholder incentives, which mediate outcomes. While these frameworks speak broadly to institutional design, they do not distinguish between specific group characteristics. Thus, along with testing the direct "treatment effect" of a collaborative watershed group, I operationalize this literature by comparing collaborative groups in terms of the concrete design and implementation choices public managers must make, such as designating group tasks or inviting group members.

Specifically, I test three collaborative group attributes believed to be key drivers of group impact: (H2) the level of management responsibility accorded to the collective (Group Responsibility); (H3) diversity of representation in the group (Stakeholder Representation); (H4) group formalization (Group Formalization). In the remainder of this section, I provide a brief overview of each subhypothesis and orient each within the literature.

Group Responsibility—Group Responsibility (H2) contrasts groups that serve as coordinating bodies or engage in outreach, monitoring, or planning from groups that engage in management activities, such as serving as the lead entity for salmon recovery actions or managing land use in the watershed. Groups conducting management activities presumably engage in more intensive and ongoing collaboration. Incentives to manipulate and act cooptively are checked in situations in which actors expect to engage in ongoing cooperation (Ansell & Gash 2008, p. 560). Repeated interactions influence the willingness of organizations to collaborate (Innes, 1998; Moreland et al., 1993), and more intensive collaborative processes are shown to increase information exchange and produce higher quality decisions (Beierle, 2002). However, increasing the intensity of interactions (e.g., from information sharing to planning to joint implementation) requires greater stakeholder engagement and investment (Margerum, 2011; Sabatier et al., 2005; Wondolleck & Yaffee, 2000). Along with requiring greater time and effort (Hill & Lynn, 2003; Sabatier et al., 2005), higher intensity collaborative efforts necessitate increased power sharing among participants (Margerum, 2011). Lubell et al. (2002) find that as these types of transaction costs increase, it is more likely that actual collaboration will be supplanted by nominal, in-name-only collaboration. Thus, more group responsibility might not result in a larger impact if groups are unable to adequately fulfill such a role.

H2: Increased responsibility for a collaborative group is associated with beneficial environmental outcomes.

Stakeholder Representation—Collaborative endeavors are theorized to be more effective when they incorporate a broader range of information and perspectives (Burby, 2003; Innes & Booher, 1999; Margerum, 2011; Wondolleck & Yaffee, 2000) because this increased breadth facilitates better decisionmaking (Dryzek, 1997; Gregory et al., 2001; Smith, 2004), improved compliance (Sabatier et al., 2005), and more effective policy implementation (Burby, 2003; Carlson, 1999). While Anderson et al. (2013) demonstrate that being more responsive to stakeholders does preclude technically sound management, the literature expresses concern that attempting to incorporate the interests and knowledge of all relevant stakeholders potentially results in diluted—and thus ineffectual—plans and policies (Coglianese, 1997, 1999; Koontz et al., 2004). Further, an increased number of organizations can make it more difficult to develop key linkages (Alexander, 1995; Gray, 1989), and incorporating additional jurisdictional levels (horizontally and hierarchically) can make group actions less tractable (Margerum, 2011). To examine this, Stakeholder Representation (H3) considers the extent to which a group is comprised solely of local governments (cities, counties, and special districts) or also includes higher level public organizations (e.g., state and Federal agencies), tribal governments, and external organizations such as businesses, agricultural interests, nongovernmental organizations (NGOs), and universities.

H3: Diverse representation in a collaborative group is associated with beneficial environmental outcomes.

Group Formalization—Formalization (H4) distinguishes between collaborative efforts that are more ad hoc and those that have a stronger institutional presence (Alexander, 1993; Huxham & Vangen, 2005; Imperial, 2005; Margerum, 2011). While formal group structures and processes are found to enhance collaborative group function and longevity (Ferguson,

2004; Margerum & Born, 2000)—and increased resource support in general is found to enhance group efficacy (Curtis & Byron, 2002; Parker et al., 2010; Yaffee et al., 1996)—it remains unclear how specific resource expenditures, such as hiring a dedicated coordinator or producing more specific plans and agendas, affect group impact. I compare groups on two aspects of formalization: (1) the presence of a dedicated coordinator and (2) whether a group has itemized goals or objectives.

In some cases, a coordinator can provide key administrative support and ease group tensions (Imperial, 2005; Huxham & Vangen, 2000; Margerum, 2002; Susskind & Cruikshank, 1987; Susskind et al., 1999). Likewise, better-specified goals and objectives can "help motivate groups to resolve conflicts" (Margerum, 2011, p. 121; see also Mattessich et al., 2001; Susskind & Cruikshank, 1987), enable groups to better assess their efficacy and focus their efforts (Anderson, 1995; Hoch, 2000; Innes & Booher, 1999; Levy, 2013; Margerum, 2011; Wondolleck & Yaffee, 2000), and clearly allocate responsibilities (Margerum&Holland, 2001). On the other hand, coordinators are not free, and there can be significant opportunity costs associated with efforts to further formalize group processes or better specify plans (Margerum, 2011; Wood & Gray, 1991). Nonetheless, I hypothesize that more formalized groups will be more strongly associated with improved water quality.

H4: Increased formalization of a collaborative group is associated with beneficial environmental outcomes.

METHODOLOGY

Estimating the Effect of Collaborative Watershed Groups

A direct comparison between the treatment group (watersheds with an active collaborative group) and control group (watersheds without an active collaborative group) is inappropriate, since self-selection into the treatment group is attributable to characteristics that also affect watershed conditions. I address the issue of selection bias using a matching method (Rosenbaum & Rubin, 1983) that estimates the average treatment effect (ATE)[1] (Cameron & Trivedi, 2005) using an augmented inverse propensity weighted estimator (AIPW) (Glynn & Quinn, 2010).

The AIPW estimator $\left(\widehat{ATE}_{AIPW}\right)$ ((see also Robins et al., 1994; Scharfstein et al., 1999) involves two basic elements: (Step 1) fitting a model that estimates the probability of "treatment" (in this case, the presence of an active collaborative group) as a function of relevant observables (i.e., a propensity score, or the estimated probability that a given observation falls in the treatment group [Cameron & Trivedi, 2005; Rosenbaum & Rubin, 1983]); and (Step 2) fitting two models that estimate the outcome variable[2] of interest under treatment and control conditions, respectively, and weighting each outcome estimate by the propensity scores estimated in Step 1 in order to produce a weighted average of the two regression estimators (Glynn & Quinn, 2010). Essentially, the two regression models fit in

[1]The ATE is defined theoretically as $ATE = E[Y(1) - Y(0)]$.
[2]As the presence of a collaborative group can predate both the Wadeable Streams Assessment (WSA) and the National Rivers and Streams Assessment (NRSA), I do not model the change in outcomes between the WSA and NRSA, since both the WSA and the NRSA present potentially relevant "post-treatment" outcomes.

Step 2 are used to estimate a contrast between what would happen if every observation were put in the control group and what would happen if every observation were put in the treatment group (Freedman & Berk, 2008; Robins & Rotnitzky, 1995). This adjustment is applied to the standard inverse propensity weight (IPW) estimator (which simply estimates the ATE as the average difference between the treatment and control groups after weighting each observation by its corresponding propensity score) to take advantage of the information in the conditioning set (the data used to estimate the propensity scores) and to improve the small sample properties of the IPW estimator (Glynn & Quinn, 2010). I specify the $\widehat{ATE}_{AIPW}$ estimator and describe the technical details of this approach, in particular the analytical advantages of the AIPW estimator relative to the IPW estimator, in Appendix A.[3]

The $\widehat{ATE}_{AIPW}$ estimator only removes selection bias if it suitably accounts for the factors that motivate selection into the treatment group (Cameron & Trivedi, 2005, p. 873). For this analysis, this assumption is well founded, as Lubell et al. (2002) provide a comprehensive analysis of the contextual factors that motivate the formation of collaborative watershed groups. By including variables in the propensity score model that Lubell et al. (2002) identify as key drivers, I am confident that this model removes a great deal of the omitted variable bias. The multilevel logistic regression model used to estimate propensity scores (Pr[Z = 1]) is specified:

$$\Pr(Z_i = 1 \mid X_i) = \text{logit}^{-1}\left(\gamma_{e[i]} + \theta_{o[i]} + X_i\beta\right) \quad (1)$$

where the probability of being in the treatment group is modeled as a function of covariate vector X, which includes the variables identified by Lubell et al. (2002) as important predictors of group presence. Specifically, for each observation i, X includes developed, forested, and agricultural land cover, population density, active National Pollutant Discharge Elimination System (NPDES) permits (for a five-year period prior to the WSA or NRSA), the ratio of NPDES enforcement actions to permits (within the same five-year period), watershed area, and median income.[4] To allow for the possibility that groups occur more frequently in particular geographic regions and become more prevalent over time, I estimate propensity scores using a multilevel logistic regression model that fits random intercept terms $\gamma_{e[i]}$ for each Omernik Level II Ecoregion e and $\theta_{t[i]}$] for each year t.[5]

[3]Though it is also possible to estimate the treatment effect by including relevant covariates and the estimated propensity scores directly in a standard regression model, an advantage of the AIPW estimator is that it relaxes the linearity assumption of a regression model, instead differencing the outcomes of collaborative watersheds and the weighted matched noncollaborative watersheds (Black & Smith, 2004). All appendices are available at the end of this article as it appears in JPAM online. Go to the publisher's Web site and use the search engine to locate the article at http://www3.interscience.wiley.com/cgi-bin/jhome/34787.

[4]p(Treatment)_Agriculture + % forest + % developed + watershed area + pop. density + median income + NPDES permits + NPDES enf. ratio + state + year (ecoregion and year are random effects).

[5]I use ecoregion instead of state as a geographic grouping indicator because state-level random effects result in overfitting. For a few states in the data, there are either no treatment (Massachusetts, New Hampshire, New Jersey, Oklahoma, and Kansas) or control (Georgia) observations. In reality, however, the "population" of watersheds in each state includes watersheds with and without an active collaborative group. While the state in which an observation occurs is an important predictor of selection, in this case the state variable is too good of a predictor for the propensity score model (since predictions are based solely on observed data). Ecoregion is an excellent proxy, because the nine different Level II Omernik ecoregions in these data are able to capture geographic context (political, social, and environmental variables that might influence selection and make the presence of a collaborative group more likely) without being subject to sampling zeros that greatly increase the number of estimated propensity scores at or near 0 or 1.

Two important empirical considerations for the propensity score estimation model are (1) that the "conditioning set," that is, the variables with which propensity scores are estimated, are relatively similar between the treatment and control groups; and (2) that the distributions of estimated propensity scores for the treatment and control groups generally encompass the same range so as to provide common support (Glynn & Quinn, 2010; King & Zeng, 2006). For instance, if estimated propensity scores for observations in the control group are between 0.05 and 0.88, a treatment observation with a propensity score of 0.95 is not adequately supported by the model, since the model was fit without any observations with a propensity score greater than 0.88. Since the AIPW estimator weights observations in accordance to their observed similarity, the propensity score distributions do not need to be perfectly congruent, but it is at least important that they sufficiently overlap. Appendix B[6] examines the covariate balance between the treatment and control groups in greater detail, demonstrating that the selection model has common support, that is, that the estimated propensity scores for the observations in the treatment and control groups span a similar range (and that, while not identical, the frequency distribution of scores for each group largely overlap).

For Step 2 of the AIPW estimation procedure, I use a pair of multilevel models to estimate the water quality outcomes under treatment (collaborative governance) and control conditions. Each model includes (1) observed covariates at the individual observation level to minimize omitted variable bias[7]; and (2) models group-level random effects so as to adjust for lack of independence among samples taken in multiple time periods from the same site or from different sites in the same geographic region.[8]

Each model groups observations by state, four-digit Hydrologic Unit Code (HUC4) subbasin, and year, as well as by the two points of randomization in the WSA and NRSA sampling design: Level II Omernik Ecoregion and Strahler stream order (both described in the Data section).[9] At the first level of the model I estimate water quality outcomes for individual stream-year i in sub-basin w, state s, year t, ecoregion e, and stream order o (equation (2))[10] :

[6]All appendices are available at the end of this article as it appears in JPAM online. Go to the publisher's Web site and use the search engine to locate the article at http://www3.interscience.wiley.com/cgi-bin/jhome/34787.

[7]Note that these covariates do not need to be identical to the conditioning set used to estimate the propensity scores in Step 1 (Glynn & Quinn, 2010); thus, in Step 1, I specify only those covariates identified by Lubell et al. (2002) as being key predictors of collaborative governance, and in Step 2, I include some of these same covariates but also additional variables that are related directly to water quality outcomes.

[8]The advantage, relative to a more common fixed effects approach, is that a multilevel model accounts for uncertainty associated with each group-level adjustment (Gelman & Hill, 2006; Raudenbush & Bryk, 2001) by shrinking the adjustment toward the overall sample mean as the size of the group decreases. In other words, as the within-group sample size decreases, the model places more credence upon the whole sample estimate, and vice versa. This "partial pooling" takes advantage of more available information (Greenland et al., 1991; Poole, 1991) and avoids overstating differences between groups (Gelman, 2006; Gelman & Hill, 2006). For data in which individual observations are nested within higher level groupings, a multilevel model produces more reasonable inferences (Gelman, 2006) and more reliable estimates (Gelman et al., 2012).

[9]Since these groups are "non-nested," such that two observations can be in the same HUC4 sub-basin but different states, or vice versa, the model is a "cross-classified" model (Gelman & Hill, 2006; Raudenbush & Bryk, 2001).

[10]Empirically, each model is specified as: Outcome Metric_Site Disturbance + % agriculture + % forest + % developed + pop. density + median income + road density + HUC4 + state + year + stream order + ecoregion (HUC4, state, year, stream order, and ecoregion are random effects).

$$Y_i = \alpha_{w[i]} + \lambda_{s[i]} + \tau_{t[i]} + \gamma_{e[i]} + \theta_{o[i]} + \sum_l \delta_l \, Site_{\,i} + \varepsilon_{iwsteo} \quad (2)$$

where Y_i represents the dependent variable, a given quality metric (e.g., nitrogen level) for sample i. Accordingly, $\alpha_{w[i]}$ represents the conditional intercept estimate for i given that it is in HUC4 basin w; similarly, $\lambda_{s[i]}$ represents the conditional intercept estimate for state s, $\tau_{t[i]}$ the conditional intercept estimate for year t, $\gamma_{e[i]}$ the conditional intercept estimate for ecoregion e, and $\theta_{o[i]}$ the conditional intercept estimate for stream order o. Next, δ_l represents a vector of control parameters 1 to l for a given site ($Site_{l[i]}$) (listed in Footnote 9). Finally, ε_{iwsteo} represents the random error associated with observation i. Note that each of the random intercepts are themselves modeled; for instance, HUC4 groups are modeled as:

$$\alpha_w = \alpha_0 + \mu_w \quad (3)$$

in which α_0 represents the average outcome across basins and basin-level random error is denoted as μ_w. State, year, ecoregion, and stream order random effects are modeled in the same way (i.e., the group level outcome as a function of the across group outcome and group-level random error); these equations are omitted for space considerations.[11] Even though the AIPW estimator has many advantages, it remains possible that resultant ATE estimates are biased upwards due to unobserved factors that are positively related to both the presence of a group and water quality outcomes. These data remain observational in nature, and accordingly these results should not be considered to necessarily provide an unbiased causal estimate. Nonetheless, conditioning on observed variables identified in the literature as being key to selection likely absorbs most of the influence of unobserved nonrandom drivers. Even if omitted variable bias remains, it is likely to be small, and in the absence of more rigorous experimental designs these estimates provide better evidence than currently exists for policymakers considering initiation or support of a collaborative watershed group.

Comparing Different Types of Groups

The second part of this analysis aims to compare the predicted effects of different types of collaborative watershed groups. To estimate how group characteristics affect predicted group impact, I fit a single multilevel model that expands upon equation (2) by adding three additional terms:

$$\begin{aligned} Y_i = \alpha_{w[i]} + \lambda_{s[i]} + \tau_{t[i]} + \gamma_{e[i]} + \theta_{o[i]} + p(C_i) + C_i \\ + \sum_k \beta_k Collab_{k[i]} C_i + \sum_l \delta_l Site_i + \varepsilon_{iwsteo} \end{aligned} \quad (4)$$

Equation (4) adds three elements to equation (2): the propensity score ($p[C_i]$) for each observation as estimated in equation (1), a main effect for collaborative group presence (C_i),

[11] In discussing multilevel models, it is important to note that the standard heuristics applied to fitting parameters in ordinary least squares regression and similar (e.g., logistic) models, statistical significance, is inappropriate for determining which group-level indicators to leave in and which to leave out (Gelman & Hill, 2006, p. 271). For instance, the model includes a grouping indicator for each group, not just the indicators found to be statistically significant. This is because the focus of the analysis is not on examining intergroup differences, but rather on generating the best possible estimate.

and a summation term $(\Sigma_k \beta_k)$ representing a vector of the predicted change in water quality for each group characteristic 1 to k for observation i $(Collab_{k[i]})$, conditional on the presence of a group (C_i). In other words, each observation i is associated with a binary variable (C_i) reflecting whether that observation is within a watershed with an active collaborative group, and then a series of interaction terms $(Collab_{k[i]} C_i)$, which model the difference for groups with and without a given characteristic (e.g., group coordinator). Having an active group is an obvious prerequisite for having a group coordinator or any other group characteristic. Accordingly, these interaction terms provide a more meaningful—and empirically grounded—interpretation, since the potential impact of any specific management characteristic rightfully should be expressed as altering the predicted impact of a collaborative group and not independently. For observations without an active collaborative group, each interaction term automatically cancels out (since $C_i = 0$). In the next section, I describe the data used to fit these models.

DATA

Dependent Variables

The data used to assess water quality outcomes come from two national surveys, the WSA and the NRSA. The WSA, conducted in 2004 to 2005, sampled 1,392 stream sites that were randomly selected from all streams of a given size within an ecological region. In other words, the sampling was stratified by ecological region and stream size.[12] The probability-based design used stratification to generate a random representative sample by ecoregion and EPA region. This presents a unique opportunity for empirical research since most research on collaborative governance selects observations based upon on either the independent (management characteristics) or dependent (outcome) variables. The NRSA conducted in 2008 and 2009 resampled 357 original WSA sites. These 357 sites form the basis of this analysis. Figure 1 shows the location of each site.

The WSA and NRSA assess the ecological condition of each site according to a series of measurements of chemical stressors, metrics of physical condition, and biological indicators. From these data, six variables are selected to provide a holistic representation of stream condition and water quality: total phosphorus content and total nitrogen content (chemical stressors caused by human activities such as mining or agriculture), water turbidity and in-stream natural habitat (physical indicators reflect more proximate habitat destruction), and indices of riparian vegetation and benthic community abundance (biological indicators of condition).[13],[14] In order to measure water quality and stream condition holistically, two variables are specifically chosen from each broader category (chemical, physical, and

[12]The WSA surveyed only perennial, wadeable streams. Perennial refers to streams that flow year round under conditions of normal precipitation. The WSA sampling protocol is stratified by Strahler stream order. “Wadeable streams,” that is, those that can be sampled without using a boat, are generally considered to be of orders 1 through 5. However, Strahler ordering does not directly correspond to stream size; rather, the Strahler protocol orders models streams as directed graphs, analogous to a tree. Ordering proceeds in reverse from bottom to top, thus a “leaf” stream, that is, one that has no tributaries, is of order 1. The Ohio River is an eighth-order stream, the Mississippi River is a 10th-order stream, and the Amazon River in South America is a 12th-order stream. The sample was also stratified by the nine (of 15 total) Omernik North American Level II ecoregions that occur in the continental United States, such as the Great Plains and Mediterranean California.

[13]Total phosphorus and total nitrogen content are both measured in absolute terms, using micrograms per liter (μg/L) as units. Turbidity is measured in nephelometric turbidity units, using a tool called a nephelometer, which gauges the amount of light reflected by the particles in water. In-stream habitat complexity and riparian cover are both calculated using line-transect surveys, which

biological). The particular indicators used were selected on the basis of presence in both the WSA and NRSA and completeness of the data. Along with the preceding footnotes, Table 1 provides more detail about the dependent variables.

To facilitate comparison across outcomes, each outcome metric is logtransformed (to achieve a more normal distribution), and then mean centered and divided by two standard deviations (Gelman, 2008). This is particularly important for phosphorus, nitrogen, and turbidity, which, as shown in Table 1, each have an extremely positive skew (i.e., a few observations have very high values). While this method of standardization makes direct interpretation more difficult than does using untransformed or log-transformed inputs (which can be interpreted as simple elasticities), it offers three advantages for this analysis. For the ATE estimates, standardized effects can be compared across metrics that are originally on different scales. Further, with regards to the regression models used in the second part of the analysis, this method of standardization renders continuous variables on a similar scale to untransformed binary variables (Gelman, 2008). This allows for comparison between binary or categorical variables of interest (related to the presence of a collaborative group) and other model inputs. Finally, the parameter estimates for transformed continuous inputs compare predicted change associated with said variable moving from a low or high value (or vice versa), as the coefficient reflects the change in outcome predicted by a two standard deviation change in the input.

Covariates

Publicly available external data are also incorporated into this analysis, primarily for modeling propensity scores as described above. Watershed land cover data (the percentages of HUC8 land cover that are impervious, used for agricultural purposes, or covered by wetlands and forests) are obtained from the National Land Cover Database (NLCD). Income and population data are obtained from the American Community Survey (ACS). Government spending data are procured from the U.S. Census Bureau (stemming from the Census of Governments). NPDES permitting and enforcement data are obtained from the U.S. EPA. These data, and the scripts used to produce these data, are available by request.

calculated the summed areal proportion of each cover type. For instance, to calculate habitat complexity the surveyor assesses coverage at specific points in a 10-m by 20-m littoral plot. These data are then used to estimate the areal proportion of the reach that contains natural cover for fish and other aquatic fauna. Because this metric is a summation of the proportion of the reach that is covered by several different kinds of cover, including boulders, large woody debris, and overhanging vegetation, this value can be greater than 1. In the data used for this analysis, sites range in value for the variable from 0 to 2.58. Similarly, because riparian cover is a summation of the proportion of the streamside riparian area that is covered by canopy, midlayer, and ground-level cover, this value can be greater than 1 as well. Sites range in value for this metric from 0 to 2.18 in the data.

[14]The benthic condition index is more complicated. To assess benthic condition, the WSA and NRSA generate an index for macroinvertebrate assemblage by assessing “least disturbed” sites in each ecoregion, using these sites as the basis of comparison for assessing stream conditions. There are numerous ways to assess the condition of a macroinvertebrate community, including abundance, composition, diversity, and various submetrics related to particular taxa. Further, the appropriateness and significance of these various metrics can differ by region. Thus, for each of the nine ecoregions within which sampling was stratified, a particular subset of six benthic community metrics were chosen upon which to generate a macroinvertebrate multimetric index (MMI) for each ecoregion (each individual metric is scored on a 1 to 10 scale, after which all six metrics are summed and then normalized to a 0 to 100 scale). Metrics were chosen on the basis of sensitivity to human disturbance, commonness among sites, independence of candidate metrics, and applicability across ecoregions (EPA, 2013). For instance, in the Xeric ecoregion (composed of the Great Basin, much of Southern California, and the Intermountain West), the MMI incorporates metrics for noninsect percent distinct taxa, percent individuals in top five taxa, scraper richness, clinger percent distinct taxa, EPT richness distinct taxa, and tolerant percent distinct individuals. In total, there are 21 different metrics that are part of the MMI for at least one ecoregion.

Independent Variables

In order to develop a watershed management database, data were collected from (1) legislative documents that allocate management responsibilities and funds to groups; (2) group reports, mission statements, membership lists, and constitutional documents; and (3) watershed management plans (specifically the portion of each plan that discusses the use and role of public involvement) for each of the 357 watersheds that were sampled for both the NRSA and WSA. In very few cases are the majority of these sources available for a given watershed, so a primary challenge is to apply a uniform coding scheme to diverse sources.

The coding process for each watershed begins at the EPA's "Surf Your Watershed" site for the HUC8 designation associated with the observation.[15] This page provides background information including the state(s) and county(ies) with land area in the watershed, the primary watershed name, and links to various monitoring Web sites and in some cases local watershed organizations. I then proceeded to search for the documentation described above, starting with links provided on the EPA page, proceeding to state and local government documentation and databases, and finally conducting an extensive Google search using keywords (e.g., "watershed council," "river management group," etc.) and local geographic names to find groups and data sources without a presence in official channels. All sources used to develop these data are available from the author. This multisource approach, taking advantage of the various resources available on state and Federal agency Web sites and databases, is quite similar (though expanded) to the approach used by Moore and Koontz (2003) to identify and survey watershed groups in the state of Ohio.

In determining whether a watershed is considered for the purpose of this analysis to be managed collaboratively, only groups in which at least one governmental entity participates are included. Since the focus of this research is on the use of collaborative governance for public purposes, cases of interest are those in which public agencies act as "initiators and instigators of collaborative governance" (Ansell & Gash, 2008, p. 545) by devoting time and resources to the group. The ultimate question then is whether such public expenditure improves policy outcomes, in this case water quality. This coding strategy proves inclusive, encompassing a wide variety of interorganizational collaborative institutions with the exception of local citizen groups. These advocacy-oriented groups are not of interest in this particular study given my specific focus on collaborative efforts that are initiated or supported by public managers (i.e., instances in which a public entity has chosen to devote resources toward collaborative governance).

Variables of interest are coded as follows.

Dedicated Coordinator—Groups were coded "1" if the group does have a designated coordinator or director and "0" otherwise. This variable does not reflect the coordinator's effort level.

Objective Formalization—Anderson (1995) and Margerum (2011) distinguish between three ways in which groups codify their aims and purposes: (1) "mission statements": a

[15] http://cfpub.epa.gov/surf/locate/index.cfm.

broadly conceived sentence (or paragraph) that provides a general statement about the impetus and aims of the group; (2) “goals”: itemized, but unspecific, tenets that “describe a desired future condition” (Margerum, 2011, p. 126)—for example, “I. Improve water quality in river; II. Increase awareness about environmental behavior in community”; and (3) “objectives”: itemized statements that outline specific actions intended by the group or specific metrics by which the group is able to measure its output or outcomes (Anderson, 1995). In practice, the distinctions between goals and objectives are somewhat blurry; perhaps most problematically, the list of aims published by a group often contains a mix of both goals and objectives (i.e., some items are specific and measurable and some are not). Thus, since this analysis does not delve deeply into the content of group goals and objectives, it is most appropriate for this analysis to code a binary variable comparing groups that only publish a mission statement (0) and groups that develop an itemized list of goals and objectives (1). This facilitates a comparison between groups that more clearly codify their purposes by developing an itemized list of motives and tasks and those that do not.

Diversity of Representation—As specified above, since this study concerns publicly supported collaborative governance efforts, the baseline requirement for a watershed group to be coded as such is that the group includes a public institution as a member. Thus, the “null value” for a group’s diversity is a group that is comprised solely of local governmental representatives. Groups are scored for the presence of tribal governments, businesses, local stakeholders (e.g., advocacy organizations), NGOs (e.g., Nature Conservancy), research or educational organizations such as universities or colleges, agricultural interests, Federal agencies, and state agencies. A group receives either a “1” (present) or “0” (absent) reflecting membership by each other type of organization. These values are then summed. Thus, if a group is constituted solely from representatives of local government, tribes, and the business community, then said group’s score for the number of stakeholder types included is a “2” (since membership by local government is requisite for inclusion in the analysis).

Group Responsibility—In order to develop a comprehensive coding scheme for the types of responsibility policymakers accord to a collaborative group, seven general categories of tasks that emerge inductively from the data are employed: planning, management, outreach, monitoring, coordination, projects, and education. Collaborative group activities such as joint policy implementation are more intensive than activities such as information sharing because they entail greater transaction costs (Margerum, 2007; Wondolleck & Yaffee, 2000). Practical distinctions between many of these activities are not always concrete (for instance, a group that uses “restoration projects” for “education” and for “outreach”). Even without more detailed data concerning group activities and responsibilities, a general contrast emerges between groups that engage in management activities and those that do not. Many groups serve as information sharing forums or conduct restoration, education, or outreach projects; others engage in management activities such as overseeing endangered species recovery efforts or land use planning and management. For this variable, a group is coded as “1” if it has management responsibilities (e.g., the group itself is the lead entity on an environmental restoration plan or for Endangered Species Act recovery actions, or a group manages land use in the watershed) and “0” if it does not have such responsibilities.

Table 2 summarizes the distribution of these variables across groups. In the data collection process, it became apparent that groups vary considerably in terms of their "presence" in gray literature (e.g., agency reports) and on the Internet. Some group Web sites contain an archival section from which past yearly reports and older documents are accessible, or a specific page that references staff or organizational members. For other groups, a more deductive approach is necessary. For instance, a group resolution might be cosigned at the bottom by group members. These data would then be used to record membership of different stakeholder types. This heterogeneity increases the potential for Type II error, either the conclusion that a group does not exist (or more likely) overlooking a specific group characteristic simply because a given document or textual reference is not found or is not available.

While I am unable to eliminate this potential source of bias, I am confident that any bias is likely to be quite small for several reasons. First, I employ a consistent data discovery and coding protocol to limit bias due to collection methods. Second, Leach et al. (2002) show that concerted efforts in small geographic areas are successful in identifying additional groups and group characteristics; investigator time and resource limitations, in that I am only able to devote a few hours of time to any one observation and am unable to visit any sites, are thus the main cause of Type II error. For this reason, I expect that underidentification is random, meaning that it increases standard errors but does not bias the results (Lubell et al., 2002). As with Lubell et al. (2002), this analysis sacrifices the level of detail that would be affordable with a regional approach in favor of national generalizability. Third, it is possible that a group's choice not to maintain an active public presence and provide up-to-date records is not randomly distributed. While this could also bias treatment estimates, the implications that such a choice holds for a group's environmental impact is unclear, and there is not compelling rationale that would indicate this significantly biases the results. Finally, the data collection process I employ is similar to methods that have been used—and published—in the past (e.g., Leach et al., 2002; Lubell et al., 2002; Moore & Koontz, 2003).

The coding process itself is similar to that of qualitative document analysis (QDA) (Altheide et al., 2008; Altheide & Schneider, 2012), often used in political science. Since QDA involves the qualitative coding of textual sources for meaning, precision, and impartiality are primary methodological concerns (Guba & Lincoln, 1994). This analysis is concerned with manifest structures, rather than latent concepts. Thus, coding in this case is primarily a question of identification, rather than one of subjective interpretation. For instance, if a group document lists an individual as being a "coordinator" or "executive director," then a group is coded as having a coordinator. Thus, I do not believe that partiality is a significant concern in this analysis. To address precision, I provide an "audit trail" (Platt, 1981) in Appendix C[16] that presents the coding protocol applied to each textual resource. This provides an overview of the analytical process applied to each data source. Likewise, in adherence to the recommendation of Guba and Lincoln (1994) to provide full access to data so that findings can be replicated and verified, the author intends to make available the data

[16]AH appendices are available at the end of this article as it appears in JPAM online. Go to the publisher's Web site and use the search engine to locate the article at http://www3.interscience.wiley.com/cgi-bin/jhome/34787.

sources employed (including group Web sites, plans, reports, etc.) for each assessed watershed. These are available on request.

RESULTS

Collaborative Group Presence

In evaluating these results one should be concerned not only with the statistical significance of the parameters of interest, but also with how the estimated effect of a variable behaves across all six outcome metrics. Colloquially, one might interpret increased levels of nitrogen, phosphorus, and suspended solids and decreased vegetation, in-stream habitat, and benthic abundance as “bad for the environment” and the converse as “good for the environment.” In interpreting ATE estimates and model coefficients, it is important to note that the dependent variables are not uniform in directionality. So that each estimated parameter reflects the direction of predicted change in the outcome variable, the directionality of each variable is kept “as-is.”

Generally, if collaborative watershed management improves environmental outcomes, one might expect to observe a negative ATE for the phosphorus, nitrogen, and turbidity level models, and a positive ATE for benthic community health, riparian cover, and habitat complexity models. The same holds true for subsequent regression models. However, not all policies and programs will affect all of these variables simultaneously. For instance, a program that targets sources of nonpoint pollution such as fertilizer use might significantly affect water chemical content but have no bearing on riparian habitat. Thus, while using six metrics in concert provides a holistic conception of water quality, one should not necessarily expect any effect to perform in a wholly consistent way across all six outcome metrics. I discuss this issue in greater detail in the context of the model results below.

Of the 357 sites sampled under both the WSA and NRSA, 124 are found to have a collaborative watershed management group at the time of the WSA sample, and 167 are found to have a collaborative watershed management group at the time of the NRSA sample. However, one issue regarding the assignment of watersheds into the “treatment” group is that the various outputs of a collaborative group (such as plans or joint projects) do not likely have an immediate effect on on-the-ground conditions; instead, it is likely that any such effect would take time to be realized. For this reason, it makes little pragmatic sense to model a sample taken in the same year in which a group was formed as being in the treatment group.

While there is limited evidence about how long it takes for group actions to manifest, Leach et al. (2002) and Leach and Sabatier (2005) find that perceived success (on the part of participants) increases after groups have been active for approximately four years (of course, as discussed previously, it is unclear how perceived success relates to actual outcomes). Based on these results of Leach et al. (2002) and Leach and Sabatier (2005), I model all watersheds in which a collaborative group has been active for at least four years prior to the sample date as being in the “treatment” group. This results in a treatment group size of 233 (87 WSA samples and 146 NRSA samples), with 481 observations in the control group. The treatment estimates obtained using the AIPW estimator are shown in Table 3. Standard

errors for each ATE are estimated via bootstrapping (Funk et al., 2011; Glynn & Quinn, 2010).

Table 3 presents bootstrapped confidence intervals for each ATE estimate (the average effect of a collaborative group that has been active for at least four years prior to the observation). These bounds represent 95 percent confidence intervals for each ATE as estimated by 500 bootstrap samples. I label as significant any ATE estimate for which the bootstrapped 95 percent confidence interval does not contain zero. Four of the six ATE estimates (phosphorus, nitrogen, turbidity, and in-stream habitat complexity) are thus found to be significant with 95 percent confidence. All four of these ATE estimates also have a sign suggesting that collaborative groups engender environmental improvement.

For interpretation, it is helpful to think of the ATE estimates as if they are each a regression coefficient associated with a binary treatment variable, in this case a collaborative watershed management group that has been active for at least four years. Again, each outcome metric is log-transformed and then standardized by mean centering and then dividing by two standard deviations (see Gelman, 2008). Thus, the expected phosphorus level for a watershed in the treatment group (i.e., treatment = 1 vs. treatment = 0) is 21.5 percent less than a watershed in the control group. Since the standardized unit is two standard deviations of the log-transformed phosphorus variable (the standard deviation of which equals 1.52), we can multiply the coefficient by twice the standard deviation, and then exponentiate the result to produce a multiplicative effect estimate of 0.78 ($\exp[-0.08 \times 1{:}52 \times 2] = 0.78$). This predicts that a watershed with a collaborative group will have a phosphorus level 22 percent below that of an untreated watershed. Similarly, the suggested effects on nitrogen and turbidity are a reduction of 23 percent (SD = 1.29) and 21 percent (SD = 1.65), respectively. In-stream habitat complexity is predicted to increase by 15 percent (SD = 0.62). The suggested effects on benthic community health and riparian cover are both negligible and insignificant.

These results can perhaps be explained by considering the extent to which a collaborative watershed group might have influence over each of these metrics. Of these six metrics, riparian cover is most subject to the influence of the landowner directly proximate to the sample site; it is not likely that actions elsewhere in the watershed meaningfully influence riparian cover at the site. Thus, finding a significant increase in riparian cover is perhaps a “hard case,” in that it would require the group to exert some form of influence directly on that plot of land. Conversely, land use and management actions taken elsewhere in the watershed that reduce net erosion or chemical pollution are likely to indirectly affect stream conditions at the sample site. Simply put, one might say that riparian cover more closely depends on actions taken at the sample site, whereas in-stream vegetation, turbidity, or phosphorus content to a greater extent depend on actions taken somewhere in the watershed.

The negligible predicted difference in benthic health is perhaps explained by the link between riparian cover and benthic health, as benthic heath is shown to be sensitive to proximate conditions such as riparian cover (Sweeney et al., 2004). Further, the impacts of upstream logging and other disturbances on benthic community health are shown to resonate up to 40 years after such behavior has ceased (Zhang et al., 2009); thus, it is possible that

benthic conditions change on a much longer time scale and thus most groups have simply not been active long enough for there to be a detectable effect.

Figure 2 presents the results of a sensitivity analysis that supports this interpretation. In Figure 2, the cutoff for an active group (e.g., all groups active at least four years or six years) varies along the *x*-axis for each outcome metric, while the ATE estimate associated varies on the y-axis. Generally, the parameter estimates remain fairly consistent, supporting the use of the four-year cutoff that I explore most fully in this analysis. As expected, one result that occurs as the cutoff is raised (i.e., requiring a group to have been active for more years to be considered part of the treatment group) is that the confidence interval surrounding the ATE estimate becomes slightly wider due to the decreasing sample size of the treatment group as the cutoff becomes more stringent.[17]

The most interesting finding emerging from the sensitivity analysis is that the ATE for benthic health increases steadily as the cutoff for an active group increases. When the ATE is estimated using only groups that have been active at least eight years or more, the presence of a collaborative group is predicted to have a statistically significant positive effect on benthic health (shown in the panel as the confidence interval does not span the dashed line representing an estimate of zero effect). Since benthic community health is perhaps slowest to respond to new management practices, this lends further support for the contention that collaborative governance does have a beneficial effect on water quality overall. I further address these results in the Discussion section.

Table 4 presents the multilevel regression models used to test group characteristics. Each outcome metric is shown in a separate column. Control variables that are not substantively interesting, specifically the propensity scores used to control for selection bias, are not included in Table 4. All continuous numeric inputs to each model are standardized via the method described above. Table 4 also does not present the random intercept adjustments modeled for HUC4, state, and year groups (fit to account for spatial and temporal dependencies), and for stream order and ecoregion (fit to account for the points of randomization in the WSA and NRSA design). Table 4 presents bootstrapped confidence intervals for each parameter; the level of significance specified in the table refers to the maximum bootstrapped interval at which a parameter is “significant,” that is, does not contain 0. This is the optimal way to test hypotheses related to linear mixed model effects, since residual degrees of freedom are uncertain for a multilevel model[18] (Bates et al., 2014; Bolker et al., 2009).

Before assessing the variables of interest, it is important to consider the consistency of parameter estimates for known sources of environmental degradation included as control variables in each model. These models are able to identify established causes of water

[17]In the data, there are 291 observations associated with an active group. As the number of years required to be considered part of the treatment group increases, the treatment group sample size declines to 270 (two or more years), 260 (3), 233 (4), 219 (5), 193 (6), 178 (7), and 159 (8).

[18]The multilevel model is a compromise between a complete pooling (no fixed effects) and no-pooling (fixed effects) model, where the precise amount of pooling differs for each group. Thus, it is unclear what the correct degrees of freedom used to calculate the *t* or *F* statistic and test a given parameter should be, since the appropriate degrees of freedom presumably differ across each group of observations.

quality changes, such as road density and agricultural land use. For instance, the results in Table 4 show that an increase in agricultural land usage within a watershed has a significant positive effect on phosphorus and nitrogen levels (i.e., increased pollution levels); this speaks to the face validity of this modeling approach.[19] Note that the estimated effect of an active group in Table 4 is not directly comparable to the ATE estimates in Table 3, because each model in Table 4 has an additional interaction term that acts on the treatment variable. The regression-based ATE estimates will likely differ in any case given that the AIPW estimator uses a nonparametric differencing approach.[20]

Group Responsibility

The interaction term "WG × Management" in Table 4 represents the predicted difference in each outcome metric between groups that have actual management responsibilities and those that do not (e.g., groups that serve as coordinative bodies or that engage in stakeholder outreach and education). As described in the specification of the model above, group characteristics are interacted with group presence because a characteristic only has meaning in the context of an active group (for instance, a group must be active to have any type of responsibility, management, or otherwise). Using interaction terms ensures that the group characteristic coefficients can be interpreted as representing the predicted difference between groups with and without said characteristic. Table 4 suggests that a group with management responsibilities has a significant negative impact on phosphorus levels and a significant positive impact on benthic community health.

Using the same method of interpretation applied to the ATE estimates above (multiplying the parameter estimate by two times the standard deviation of the variable and then exponentiating the result to get a multiplicative effect), a group with management responsibilities is associated with a 37 percent (SD = 1.52) lower phosphorus level and a 27 percent higher benthic index score (SD = 0.74). While the sign of the coefficient for the estimated effect on nitrogen level, riparian cover, and in-stream habitat complexity is in the hypothesized direction (reduced pollution, improved habitat condition), these effects are all insignificant; the estimated effect on turbidity is insignificant and not in the hypothesized direction. These results provide limited support for the hypothesis that collaborative groups with management responsibilities have a relatively greater impact on water quality. Table 4 shows that the difference in phosphorus level between the two group types is very similar to that of the predicted difference (in terms of both sign and significance) associated with a two standard deviation increase in county median income. Similarly, the difference associated with management groups with regards to benthic community health is similar in magnitude to the change associated with a two standard deviation increase in agricultural land usage.

[19]Phosphorus, nitrogen are strongly linked to agricultural land use (Tong & Chen, 2002). Similarly, in Table 4 road density is positively related to stream turbidity, phosphorus level, and nitrogen level, and negatively related to benthic condition. This fits with established ecological findings; for instance, roads increase erosion and sediment yield, thereby increasing stream turbidity (Forman, 1998; Montgomery, 1994), and water runoff from roads carries heavy-metal pollutants that can harm benthic communities (Forman, 1998; Horner & Mar, 1983).

[20]A potential complicating factor in testing group characteristics is that correlation between characteristics might hinder simultaneous estimation (i.e., multicollinearity). I tested for this possibility by fitting a distinct model for each group characteristic, and comparing these isolated estimates to the parameter estimates from the unrestricted model including all group characteristics; parameters from the restricted models (one characteristic each) were almost identical to those in the unrestricted model. Thus, I present only the unrestricted model results.

Stakeholder Representation

The number of stakeholder types in a group is considered as a continuous variable (standardized in the same way as the continuous covariates) in the "WG × Stakeholders" interaction term. While the parameter estimates predict a small increase in pollution (phosphorus, nitrogen, and turbidity) as the number of stakeholder types in a watershed group increase, none of these parameters are statistically significant. As the number of stakeholder types is mean centered and standardized, this means that there is not a great deal of difference between an "average group" involving local governmental representatives and four or five additional stakeholder types (the mean number of additional types is 4.3) and either a limited group involving only local governmental representatives (e.g., local city and county officials) or a diverse group involving all coded stakeholder types. This does not corroborate the theory that collaborative institutions are made more effective by incorporating a broader range of perspectives (Burby, 2003; Innes & Booher, 1999; Margerum, 2011; Wondolleck & Yaffee, 2000), but it also does not evidence that broader involvement dilutes policy actions (Coglianese, 1997, 1999; Koontz et al., 2004).

Group Formalization

Table 4 also tests two aspects of group formalization: (1) whether or not a group has a dedicated coordinator, and (2) the level of goal specification a group codifies. The "WG × Coordinator" interaction term represents the predicted difference between a group that has a coordinator and a group that does not. Only one coefficient is significant (turbidity, which is predicted to decrease by 33 percent [SD = 1.65]), but five of six are of a sign suggesting that groups with a coordinator achieve greater environmental gains. Given that it is fairly accepted that coordinators serve a valuable purpose, it is very interesting that these results do evidence a stronger, more substantive difference between groups that have a coordinator and those that do not. One potential source of variation not captured available in these data is the work level of the coordinator. In some cases, a group coordinator works on a part time—or even largely volunteer—basis, or serves as coordinator as part of her broader job description at a government agency. Other groups have a coordinator who works full time in support of the group. Presumably, better data that are able to codify coordinator effort level would more carefully test the benefit of having a full-time, dedicated coordinator.

The "WG × Goals/Objectives" term in Table 4 compares groups that have either itemized goals or objectives to the reference category, groups that publish only a mission statement. While several of the results are insignificant and of almost zero magnitude (phosphorus level and in-stream habitat condition), it is interesting to note that groups with itemized goals and objectives are associated with much higher (40 percent, SD = 1.29) nitrogen levels and reduced benthic community health (20 percent, SD = 0.74). While the lack of significance and a consistent pattern among the remaining outcome metrics makes it difficult to draw an overarching conclusion, these results suggest that itemizing purposes and goals does not necessarily make a group any more impactful.

DISCUSSION

The results of this analysis suggest that collaborative watershed groups achieve water quality and in-stream habitat gains. Watersheds with a collaborative group that has been active at least four years are estimated to have significantly lower levels of phosphorus content, nitrogen content, and turbidity. These watersheds are also estimated to have significantly greater in-stream habitat complexity (e.g., woody debris and aquatic plants). There is no significant estimated difference with regards to benthic community health or riparian cover (but the sign of each parameter is positive, meaning that all six ATE estimates are of a sign suggesting environmental improvement). As discussed above, considering the conceptual linkage between the actions of a collaborative watershed group and site-specific measurables lends explanatory context to these results. Water content metrics are obviously subject to proximate inputs, but they also more broadly reflect land use and environmental behavior throughout the watershed. For instance, if group actions help several farms mitigate fertilizer runoff, such activities would be reflected in a water quality sample taken downstream. Similarly, while in-stream habitat conditions are of course subject to onsite actions such as channelization, upstream land usage such as logging and development (or conversely restoration actions) can result in flow changes and floods that alter downstream habitat (Crispin et al., 1993; Wang et al., 1997).

Given that benthic community health is related to upstream activities such as logging (Harding et al., 1998), it is interesting that no significant effect is found on this metric. This result is likely attributable to the fact that benthic communities continue to demonstrate the effects of land use actions decades after the activity has ceased (Harding et al., 1998; Zhang et al., 2009, actually refer to stream biodiversity as “the ghost of past land use”). Accordingly, linking collaborative group presence to changes in benthic community health likely requires a longer time horizon. Lastly, inability to identify a link between collaborative groups and riparian cover is likely attributable to the fact that this metric least reflects aggregate watershed management and restoration actions and most reflects proximate actions by whichever entity owns that piece of property.

Comparing different types of groups, this analysis identifies a distinction between collaborative groups given management responsibilities (e.g., lead management entity for an Endangered Species Act recovery plan) and those tasked solely with coordination or planning. Two of the six parameters are significant and five of six are of the hypothesized sign, providing limited evidence that groups given management responsibilities stand apart as more effective. These results suggest that the additional costs of collaborative management (as opposed to coordination or planning), such as increased power sharing (Margerum, 2011), time and resource commitment (Hill & Lynn, 2003; Sabatier et al., 2005), and investment in the process (Margerum, 2011; Sabatier et al., 2005; Wondolleck & Yaffee, 2000), do result in increased environmental benefits as well.

Little differentiation is found with regards to stakeholder representation. A possible reason for the lack of significance associated with stakeholder representation is that this analysis focuses specifically on government-sponsored collaborative watershed groups. For instance, in the face of existing theory and evidence (Burby, 2003; Dryzek, 1997; Gregory et al.,

2001; Innes & Booher, 1999; Margerum, 2011; Smith, 2004; Wondolleck & Yaffee, 2000), it would seem unlikely that incorporating additional perspectives and interests into the policy process does not matter at all. However, it is very plausible that the role of government-sponsored watershed groups and scope of their activities are fairly constrained, such that there is ultimately little variation in outcomes regardless of inputs. For instance, even watershed groups that engage in management actions do not have broad rulemaking and enforcement authority, but rather manage land use or similar issues (that are otherwise typically the responsibility of local governments [Koontz et al., 2004]). Watershed groups cannot pass new laws or implement a market-based water quality trading system. Thus, groups are necessarily limited by their legislative and regulatory environment.

Group impacts are shown to differ somewhat by group formalization, but not necessarily in the hypothesized direction. The results suggest that groups with itemized goals and objectives actually perform worse with regards to phosphorus level and benthic community health. It is not readily clear why this is the case, but given the prevailing wisdom that increased specification helps groups to resolve conflict and better assess efficacy (Margerum, 2011; Mattessich et al., 2001; Susskind & Cruikshank, 1987), the lack of association between goal specificity and improved environmental outcomes is noteworthy. It is plausible that regardless of goal specificity, the goals or objectives that end up being prioritized are those that closely dovetail with existing regulatory mandates, and thus the nominal goals of the group do not track closely with empirical actions. While this does not explain the significant results opposite of the hypothesized direction, it does perhaps explain why goals and objectives are not linked to environmental improvements. Finally, while the presence of a group coordinator or facilitator is linked only to a significant decrease in turbidity, as discussed above the lack of conclusive results in this regard is likely due to the fact that these data combine coordinators of various types and capacities. Resource limitations and a lack of data availability prevent ascertaining the effort level of a group coordinator.

A prominent distinction that emerges from these results is the contrast between the significant ATE estimates associated directly with collaborative group presence and the inability of the group characteristics tested to “account” for the predicted difference between a watershed with an active collaborative governance institution and a watershed without such an entity. One potential reason is that the variables tested might not be the variables that drive group effectiveness (as measured by environmental impact). A notable omission, of course, is group funding levels. While one would presume that differential effectiveness associated with funding discrepancies is somewhat of a given, this relationship is worth testing to posit whether public agencies devoting funds to collaborative endeavors are getting any “bang” for their “buck.” Of course, resource munificence alone cannot be the sole driver of group effectiveness. For instance, the findings above conclusively identify a significant benefit associated with having a group coordinator.

These results also highlight the essential role of qualitative research in understanding the role and function of collaborative governance. It is unlikely that large-N statistical analysis alone can definitively answer these questions. For instance, the extensive case studies conducted by Margerum (2011) speak to contextual variables and localized drivers of group

efficacy that do not necessarily emerge in a larger-N cross-sectional analysis. Given that the uncertainty and complexity of environmental systems makes it difficult to parse effects of collaboration amidst other influences (Koontz & Thomas, 2006; Rapp, 2008), process tracing and the use of program logic models (Bickman, 1987; Margerum, 2011; McLaughlin & Jordan, 1999) might serve as an evaluatory complement to the systematic analysis of outcomes conducted in this project. Most importantly, these works highlight the idiosyncratic nature of local environmental management; what works well in one context might not work well in another, and this is very difficult to tease out in a large-N statistical analysis.

CONCLUSION

It is easy to lose sight of the fact that collaborative governance requires the expenditure of time and effort by public actors, and that these resources could be applied elsewhere. In other words, collaboration is not just a "concept… [but potentially] a way of solving problems… and achieving results" (Margerum, 2011, p. 306). However, much of what we currently know about the environmental impacts of government-supported collaborative institutions is based upon evidence from small-N case studies or studies that use subjective measures (e.g., stakeholder perceptions or quality of policy outputs) as proxies for environmental outcomes. Previous research (e.g., Biddle & Koontz, 2014; Hoornbeek et al., 2012; Koontz, 2003; Leach & Sabatier, 2005; Lubell, 2004a; Ulibarri, 2015,) has shown that collaborative governance has a positive effect both on intermediate outputs and perceived policy or program effectiveness. This analysis uses a unique data set and a rigorous analytical approach to build upon these works by conducting one of the first large-N statistical analyses that systematically tests the relationship between collaborative governance and environmental outcomes. Most importantly, the use of objective outcome data (water quality and habitat metrics) across a large geographic scale represents a major advancement.

Simply put, these results evidence that collaborative governance institutions (in this case, collaborative watershed groups) do improve ecological outcomes. It is also important to note that I find no indication across any of the six outcomes metrics that collaborative governance engenders worse environmental outcomes. This demonstrates that the lowest common denominator effect (Coglianese, 1999) is less of a concern than might be thought, and that fears of collaboration leading to more talk and less action might be somewhat unfounded. Despite the rigorous matching approach employed, it remains possible that the ATE estimates are biased upwards due to unobserved factors that are positively related both to the presence of a group and to water quality outcomes. However, since matching is based upon research evidence regarding the factors that drive collaborative group formation (Lubell et al., 2002), it is likely that this approach successfully reduces omitted variable bias. Further, as discussed, these estimates represent a significant step forward even if some omitted variable bias remains (and given the infeasibility of randomized assignment in this context, it is unlikely that experimental data will be available in the future).

Due to the inconclusive findings with regards to group characteristics, this work does not shed a great deal of light on the question of how collaborative watershed management

improves environmental outcomes. The lack of definitive results associated with group characteristics generally accepted as beneficial (e.g., presence of a coordinator, goal specificity) is particularly interesting. One suggestion emerging from the literature is that collaborative watershed governance in general represents a shift in focus toward nonpoint water quality problems that are not suitably addressed by state and Federal regulatory authorities (Hardy & Koontz, 2008; Hoornbeek et al., 2012; Koontz et al., 2004). This shift in focus, however operationalized, might matter more than the specific details of the group itself at a macro level. Group characteristics are not likely irrelevant, but perhaps are more of a contextual issue rather than the basis for collaborative group impact.

Future research directions include the addition of a third wave of data from the 2013 to 2014 NRSA (to be released in 2016 following laboratory analysis of samples), as well as additional data concerning group budgets, procedures, and activities. Also, while I control for land and resource use in this analysis, future research will take advantage of larger samples to allow for a better understanding of how collaborative groups can be more or less effective (in terms of producing desired outcomes) in specific contexts. This ongoing work is important, as perhaps the central takeaway of this analysis is that we as policy scholars and practitioners need to think more deeply about why we believe that collaborative groups are an effective tool for achieving public policy goals.

APPENDIX A: AIPW SPECIFICATION AND PROPENSITY SCORE BALANCE

The ATE is estimated using the AIPW via:

$$\widehat{ATE}_{AIPW} = \frac{1}{n}\sum_{i=1}^{n}\left\{\left(\frac{X_i Y_i}{\hat{\pi}(Z_i)} - \frac{(1-X_i)Y_i}{1-\hat{\pi}(Z_i)}\right) - \frac{(X_i - \hat{\pi}(Z_i)}{\hat{\pi}(Z_i)(1-\hat{\pi}(Z_i))} * \left[(1-\hat{\pi}(Z_i)\widehat{\mathbb{E}}(Y_i|X_i=1,Z_i) + \hat{\pi}(Z_i\widehat{\mathbb{E}}(Y_i|X_i=0,Z_i)\right]\right\} \quad \text{(A.1)}$$

where $\hat{\pi}(Z_i)$ is the estimated propensity score given the set of control variables Z for site i, Y_i is the observed outcome, and X_i is the treatment variable for site i. Equation (A1) builds upon the basic inverse propensity weight (IPW) estimator by adjusting for a weighted average of the two regression estimators.[21] Glynn and Quinn (2010, p. 41) show that $\widehat{ATE}_{AIPW}$ is a consistent estimator for ATE when either (1) the propensity score model is correctly specified; or (2) the two outcome regression models are correctly specified (see also Scharfstein et al., 1999). This means that the estimate is "doubly robust" (Bang & Robins, 2005; Glynn & Quinn, 2010) to uncertainty about both the selection process and the outcome model. Since the empirical processes that drive the existence of collaborative watershed groups and water quality conditions are both complex, this is a significant advantage. Using the $\widehat{ATE}_{AIPW}$, I estimate the effect of an active collaborative group for each of the outcome metrics used in this analysis.[22]

[21]The adjustment term is $[(1-\hat{\pi}(Z_i)\widehat{\mathbb{E}}(Y_i|X_i=1,Z_i) + \hat{\pi}(Z_i\widehat{\mathbb{E}}(Y_i|X_i=0,Z_i)]$ such that $\frac{1}{n}\sum_{i=1}^{n}\{(\frac{X_i Y_i}{\hat{\pi}(Z_i)} - \frac{(1-X_i)Y_i}{1-\hat{\pi}(Z_i)}) - \frac{(X_i-\hat{\pi}(Z_i)}{\hat{\pi}(Z_i)(1-\hat{\pi}(Z_i))}\}$ is the basic $\widehat{ATE}_{IPW}$ estimator.

While the AIPW estimator has many advantages, it also can exhibit an extremely high variance when weights applied to observations are very small or very large. This occurs when estimated propensity scores are close to 0 or 1 (since weights are derived from the inverse propensity scores). Pohlmeier et al. (2013) thus recommend a shrinkage method that stabilizes the treatment estimators by shrinking the propensity scores ($\hat{p}_i = \Pr[\mathbf{Z_i} = 1 | X_i]$) toward the unconditional mean treatment value (see also Busso et al., 2014; Frolich, 2004):

$$\hat{p}_i^2 = \left(1 - \lambda_i(n)\hat{p}_i + \lambda_i(n)\hat{D} \right. \tag{A.2}$$

where $\hat{D}$ is the mean treatment value and $\lambda_i(n)$ the "tuning" parameter used to shrink the scores, is specified as $\lambda_i(n) = 1/\sqrt{(n)}$ (Pohlmeier et al., 2013). This method reduces weight variance, thereby reducing variance in the ATE estimators as well (Pohlmeier et al., 2013).[23] This procedure results in ATE estimates nearly identical to those generated without tuning, but serves to greatly reduce the estimate standard errors.

APPENDIX B: EXAMINING PROPENSITY SCORE BALANCE

To examine covariate balance, Table B1 presents average values for the treatment and control groups for each variable included in the propensity score model. The third column presents the *P*-value resulting from a standard two-sample *t*-test comparing the mean values. Three substantive differences that do emerge are that (1) control observations have, on average, a county population density twice as high as treatment observations; (2) control observations have, on average, about four more active NPDES permits within their HUC8 watershed than do treatment observations (since the variance of population densities among watersheds is quite high, the *t*-test fails to identify a statistical significant difference between the treatment and control groups even though the average difference is quite high in substantive terms); and (3) watersheds with an active group have, on average, about 25 percent agricultural land, compared to 20 percent in the control group.

This is consistent with the predicted role of collaborative groups as primarily targeting nonpoint source pollution (Hardy & Koontz, 2008; Hoornbeek et al., 2012; Koontz et al., 2004; addressed in the Discussion and Conclusion sections); watersheds with a higher level of NPDES permits are presumably those for which point source pollution is a more significant issue, whereas watersheds with fewer NPDES permits (as well as lower population density and a higher percentage of agricultural land) are likely those for which nonpoint source pollution is a more significant driver of water quality. It is important to remember, however, that the reason we examine the degree to which the treatment and control groups are balanced on observables is due to concern that the two groups might differ in ways that are unobserved as well (differences in observables can obviously be

[22]Discussed in Data section, six outcome metrics are used in order to comprehensively assess environmental condition: nitrogen content, phosphorus content, turbidity, benthic community health index, in-stream habitat complexity, and riparian cover. Thus, six ATE estimates are generated.

[23]Other common techniques applied to very low or very high weights are (1) to discard all weights above and below specified cutoffs; and (2) to truncate all weights above and below specified cutoffs. The former method discards potentially valuable information and sacrifices efficiency, and given the relatively small sample size for this analysis is not the best option. The latter method essentially shrinks only extreme parameters; the advantage of the shrinking method used in this analysis is that it applies a consistent procedure to all observations.

controlled for in the model). In this case, it is highly plausible that the observables do sufficiently account for variation between the treatment and control groups, since land cover, development, and point source pollution permits provide a comprehensive reflection of local watershed characteristics. This is likely why coarser metrics such as voting records (tested but left out of the final models) are insignificant and do not improve model fit, since the aforementioned covariates do a better job of capturing local heterogeneity.

Since the AIPW estimator relies on observed covariates to estimate selection probability, it is also important that the distributions of each covariate are relatively balanced between the treatment and control groups; otherwise, the selection model lacks common support. Figure B1 shows that while there are more control observations overall, covariate frequency distributions for the treatment and control groups are highly similar.

While Table B1 shows that the mean covariate values differ somewhat between the treatment and control groups, Figure B1demonstrates that the overall distributions for each covariate overlap nicely. While there are more control observations, the range and relative frequency of observations are highly similar between the two groups. Figure B2 shows a similar distribution for the actual propensity scores. As might be expected, the distribution of propensity scores for the control group is skewed slightly lower than the distribution of propensity scores for the treatment group; nonetheless, the overall distribution for each spans the same range. Also, note that there are very few scores close to 0 or 1; this is in part due to the application of the shrinkage method, and helps ensure stable estimation within the weighting process.

Table B1.

Comparison of covariates between treatment and control groups.

	Group active ≤ 4 years	No group or active <4 years	*P*-value
Percent developed	0.020	0.030	0.007**
Percent agricultural	0.250	0.200	0.010*
Watershed area	4,772.560	4,488.880	0.102
Pop. density	21.660	44.070	0.178
Median income	45,264.620	47,677.090	0.032*
NPDES permits	6.810	10.820	0.009**
NPDES enforce ratio	0.690	0.760	0.651

Note:

** $P < 0.01$

* $P < 0.05$.

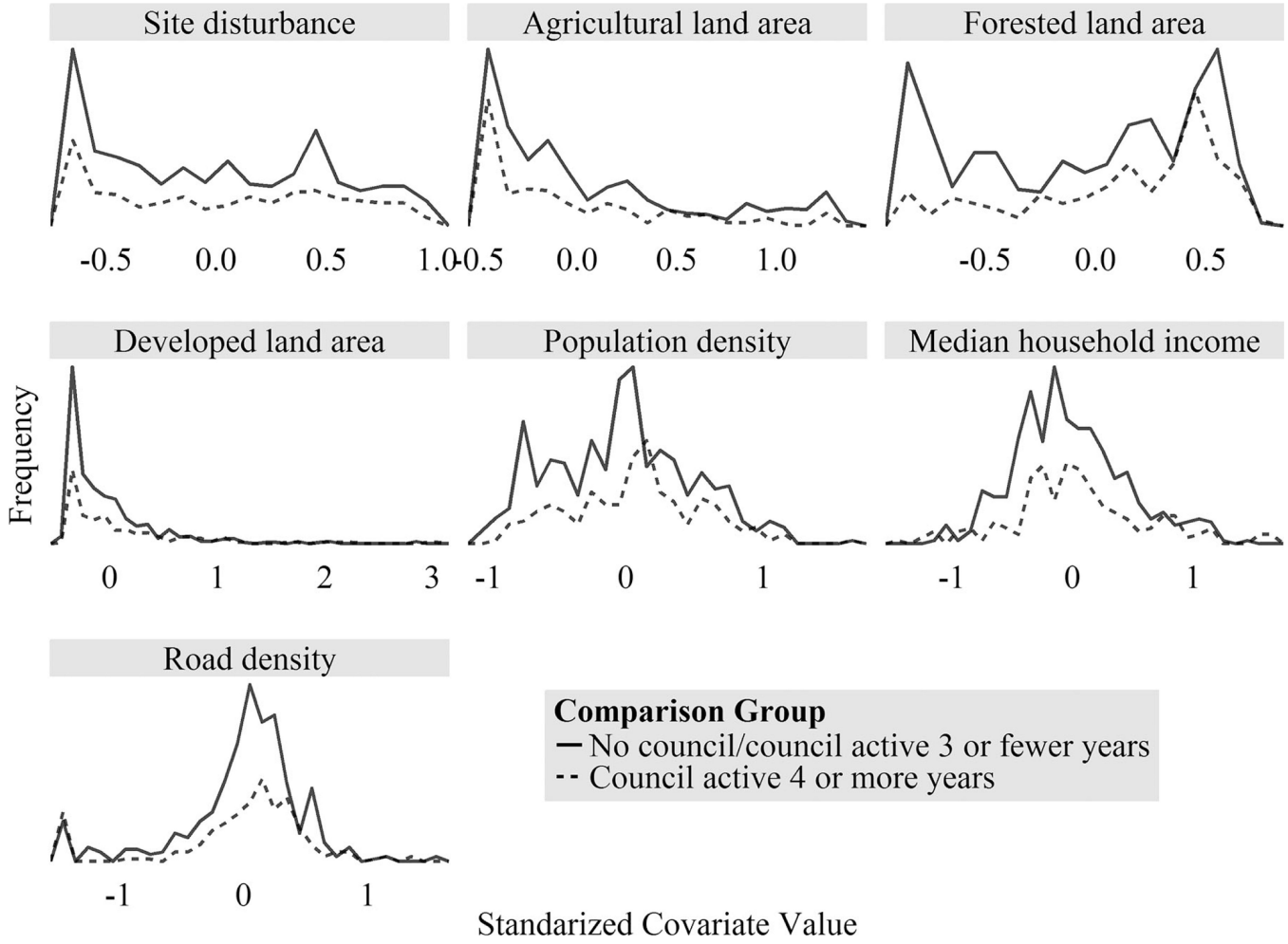

Figure B1.
Covariate Distributions, Treatment, and Control Groups.

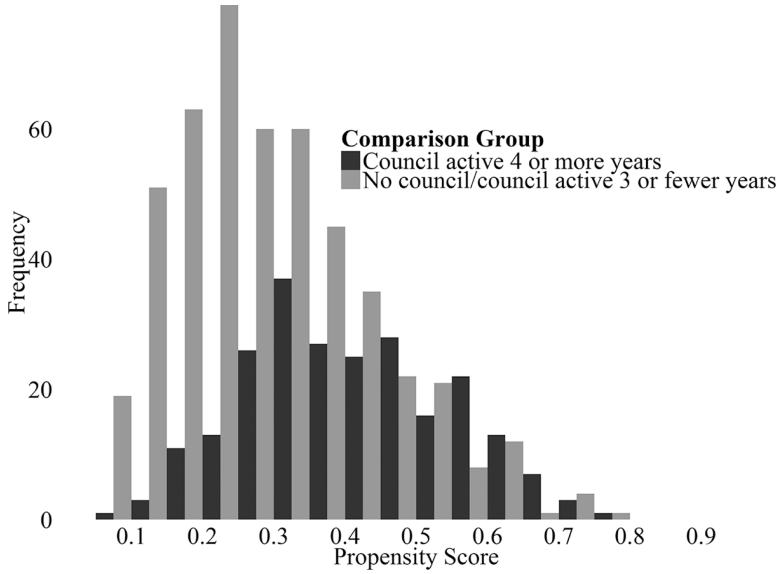

Figure B2.
P-score Distributions, Treatment, and Control Groups.

APPENDIX C: EXAMPLE CODING PROTOCOL

Note: This process is applied iteratively across available documents. Since many groups have distinct documents that reference bylaws, membership, and funding, for instance, a group might initially receive a "0" for each category of stakeholder representation when coding the bylaws document; these variables will then be updated to reflect new data subsequently produced by analyzing the membership roster.

Q1: Is this textual source an (1) official group Web site; (2) annual group report; (3) group bylaw or charter document; (4) piece of authorizing legislation?

If no → disregard. If yes → proceed to Question 2.

Q2: Does the textual source contain language that addresses a group's purpose?

If no → proceed to Question 5. If yes → proceed to Question 3.

Q3: Does the text speaking to a group's purpose present an itemized set of purposes?

If no → code Objective Formalization as "MISSION STATEMENT." If yes → proceed to Question 4.

Q4: Does the itemized set of purposes contain specific, measurable points of reference (e.g., "reduce total nitrogen level" instead of "improve water quality").

If no → code Objective Formalization as "GOALS." If yes → code Objective Formalization as "OBJECTIVES."

Q5: Does the textual source contain language describing or listing group membership?

If no → proceed to Question 14. If yes → proceed to Question 6.

Q6 to Q13: Does text describing group membership list a tribe (or business, Federal agency, etc.) as a member of the group?

If no → code Tribal Representation as 0, proceed to next question. If yes → code "Tribal Representation" as 1, proceed to next question—for Q6 to 12, proceed to next stakeholder type; for Q13, proceed to Q14.

Q14: Does textual source contain specific reference to a group coordinator or facilitator?

If no → code COORDINATOR as 0, proceed to Q21. If yes → code COORDINATOR as 1, proceed to Q21.

Q15: Does textual source identify year in which group was formed?

If no → proceed to Q22. If yes → code FORMATION YEAR as specified year.

Q16 to Q23: Does textual source contain language reference to group actions or responsibilities related to EDUCATION (e.g., group "runs environmental education programs in local schools")?

If no → proceed to next question. If yes → code GROUP ACTIVITY as "education."

Q17: Outreach (e.g., group "reaches out to local farmers").

Q18: Coordination (e.g., group "provides forum where agencies can share information").

Q19: Monitoring (e.g., group "conducts ongoing monitoring of stream pollutants").

Q20: Projects (e.g., group is "conducting restoration on local creek").

Q21: Planning (e.g., group is "charged with developing comprehensive action plan").

Q22: Management (e.g., group is "lead local entity for water improvement program").

Q23: Permitting (e.g., group is "administers land use permits within the watershed") If GROUP ACTIVITY is "Management" or "Permitting" code GROUP RESPONSIBILITY as "1" otherwise code GROUP RESPONSIBILITY as "0."

ACKNOWLEDGMENTS

The author thanks Craig Thomas, Ryan Scott, Grant Blume, Ann Bostrom, and Winfield Wilson for feedback on early drafts, and three anonymous reviewers for their careful and constructive reviews. Partial support for this research came from a Eunice Kennedy Shriver National Institute of Child Health and Human Development research infrastructure grant, R24 HD042828, to the Center for Studies in Demography & Ecology at the University of Washington.

REFERENCES

Agranoff R, & McGuire M (2003). Inside the matrix: Integrating the paradigms of intergovernmental and network management. International Journal of Public Administration, 26, 1401–1422.

Alexander ER (1993). Interorganizational coordination: Theory and practice. Journal of Planning Literature, 7, 328–343.

Alexander ER (1995). How organizations act together: Interorganizational coordination in theory and practice. New York, NY: Routledge.

Altheide DL, & Schneider CJ (2012). Qualitative media analysis. Qualitative research methods (2nd ed., Vol. 38). Thousand Oaks, CA: Sage.

Altheide D, Coyle M, DeVriese K, & Schneider C (2008). Emergent qualitative document analysis. In Hesse-Biber SN & Leavy P (Eds.), Handbook of emergent methods (pp. 127–151). New York, NY: Guilford Press.

Anderson LT (1995). Guidelines for preparing urban plans. Chicago, IL: Planners Press, American Planning Association.

Anderson SE, Hodges HE, & Anderson TL (2013). Technical management in an age of openness: The political, public, and environmental forest ranger. Journal of Policy Analysis and Management, 32, 554–573.

Ansell C, & Gash A (2008). Collaborative governance in theory and practice. Journal of Public Administration Research and Theory, 18, 543–571.

Bang H, & Robins JM (2005). Doubly robust estimation in missing data and causal inference models. Biometrics, 61, 962–973. [PubMed: 16401269]

Bates D, Maechler M, Bolker B, & Walker S (2014). lme4: Linear mixed-effects models using Eigen and S4. ARXIV e-print retrieved from https://github.com/lme4/lme4/.

Beierle TC (2002). The quality of stakeholder-based decisions. Risk Analysis, 22, 739–749. [PubMed: 12224747]

Bickman L (1987). The functions of program theory. New Directions for Program Evaluation, 33, 5–18.

Biddle JC, & Koontz TM (2014). Goal specificity: A proxy measure for improvements in environmental outcomes in collaborative governance. Journal of Environmental Management, 145, 268–276. [PubMed: 25083592]

Bidwell RD, & Ryan CM (2006). Collaborative partnership design: The implications of organizational affiliation for watershed partnerships. Society and Natural Resources, 19, 827–843.

Bingham LB, & O'Leary R (2008). Big ideas in collaborative public management. Armonk, NY: M. E. Sharpe.

Black DA, & Smith JA (2004). How robust is the evidence on the effects of college quality? Evidence from matching. Journal of Econometrics, 121, 99–124.

Bolker BM, Brooks ME, Clark CJ, Geange SW, Poulsen JR, Stevens MHH, & White JS (2009). Generalized linear mixed models: A practical guide for ecology and evolution. Trends in Ecology & Evolution, 24, 127–135. [PubMed: 19185386]

Burby RJ (2003). Making plans that matter: Citizen involvement and government action. Journal of the American Planning Association, 69, 33–49.

Busso M, DiNardo J, & McCrary J (2014). New evidence on the finite sample properties of propensity score reweighting and matching estimators. Review of Economics and Statistics, 96, 885–897.

Cameron AC, & Trivedi PK (2005). Microeconometrics: Methods and applications. New York, NY: Cambridge University Press.

Carlson C (1999). Convening. In Susskind L, McKearnen S, & Thomas-Lamar J (Eds.), The consensus building handbook: A comprehensive guide to reaching agreement (pp. 169–198). Thousand Oaks, CA: Sage.

Carr G, Bloschl G, & Loucks DP (2012). Evaluating participation in water resource management: A review. Water Resources Research, 48, 11.

Cheng AS, & Daniels SE (2005). Getting to we: Examining the relationship between geographic scale and ingroup emergence in collaborative watershed planning. Human Ecology Review, 12, 30–43.

Coglianese C (1997). Assessing consensus: The promise and performance of negotiated rulemaking. Duke Law Journal, 46, 1255–1349.

Coglianese C (1999). The limits of consensus: The environmental protection system in transition: Toward a more desirable future. Environment: Science and Policy for Sustainable Development, 41, 28–33.

Collins K, Blackmore C, Morris D, & Watson D (2007). A systemic approach to managing multiple perspectives and stakeholding in water catchments: Some findings from three UK case studies. Environmental Science & Policy, 10, 564–574.

Crispin V, House R, & Roberts D (1993). Changes in instream habitat, large woody debris, and salmon habitat after the restructuring of a coastal Oregon stream. North American Journal of Fisheries Management, 13, 96–102.

Curtis A, & Byron I (2002). Understanding the social drivers of catchment management in the Wimmera region. Bathurst, New South Wales: Johnstone Centre for Research in Natural Resources and Society.

Donahue JD, & Zeckhauser R (2011). Collaborative governance: Private roles for public goals in turbulent times. Princeton, NJ: Princeton University Press.

Dryzek JS (1997). The politics of the Earth: Environmental discourses. New York, NY: Oxford University Press.

Emerson K, Nabatchi T, & Balogh S (2012). An integrative framework for collaborative governance. Journal of Public Administration Research and Theory, 22, 1–29.

EPA (U.S. Environmental Protection Agency). (2013). National rivers and streams assessment 2008–2009: A collaborative report. Washington, DC: U.S. EPA Office of Research and Development, EPA/841/D-12/001.

Ferguson C (2004). Governance of collaborations: A case study. Administration in Social Work, 28, 7–28.

Forman RTT (1998). Roads and their major ecological effects. Annual Review of Ecology and Systematics, 29, 207–231.

Frame TM, Gunton T, & Day JC (2004). The role of collaboration in environmental management: An evaluation of land and resource planning in British Columbia. Journal of Environmental Planning and Management, 47, 59–82.

Freedman DA, & Berk RA (2008). Weighting regressions by propensity scores. Evaluation Review, 32, 392–409. [PubMed: 18591709]

Frolich M (2004). Finite-sample properties of propensity-score matching and weighting estimators. Review of Economics and Statistics, 86, 77–90.

Fuller BW (2009). Surprising cooperation despite apparently irreconcilable differences: Agricultural water use efficiency and CALFED. Environmental Science and Policy, 12, 663–673.

Funk MJ, Westreich D, Wiesen C, Sturmer T, Brookhart MA, & Davidian M. (2011). Doubly robust estimation of causal effects. American Journal of Epidemiology, 173, 761–767. [PubMed: 21385832]

Gelman A (2006). Multilevel (hierarchical) modeling: What it can and cannot do. Technometrics, 48, 432–435.

Gelman A (2008). Scaling regression inputs by dividing by two standard deviations. Statistics in Medicine, 27, 2865–2873. [PubMed: 17960576]

Gelman A, & Hill J (2006). Data analysis using regression and multilevel/hierarchical models. New York, NY: Cambridge University Press.

Gelman A, Hill J, & Yajima M (2012). Why we (usually) don't have to worry about multiple comparisons. Journal of Research on Educational Effectiveness, 5, 189–211.

Gerlak AK, Lubell M, & Heikkila T (2012). The promise and performance of collaborative governance. In Kamieniecki S & Kraft ME (Eds.), Oxford handbook of U.S. Environmental Policy (pp. 413–434). Oxford, U. K.: Oxford University Press.

Gigone D, & Hastie R (1993). The common knowledge effect: Information sharing and group judgment. Journal of Personality and Social Psychology, 65, 959–974.

Glynn AN, & Quinn KM (2010). An introduction to the augmented inverse propensity weighted estimator. Political Analysis, 18, 36–56.

Grafton RQ, & Hussey K (2011). Water resources planning and management. New York, NY: Cambridge University Press.

Gray B (1989). Collaborating: Finding common ground for multiparty problems. San Francisco, CA: Jossey-Bass Publishers.

Greenland S, Maclure M, Schlesselman JJ, Poole C, & Morgenstern H (1991). Standardized regression coefficients: A further critique and review of some alternatives. Epidemiology, 2, 387–392. [PubMed: 1742392]

Gregory R, McDaniels T, & Fields D (2001). Decision aiding, not dispute resolution: Creating insights through structured environmental decisions. Journal of Policy Analysis and Management, 20, 415–432.

Guba EG, & Lincoln TS (1994). Competing paradigms in qualitative research. In Denzin NK & Lincoln YS, (Eds.), Handbook of qualitative research (pp. 105–117). Thousand Oaks, CA: Sage.

Hall TE, & O'Toole LJ (2000). Structures for policy implementation: An analysis of national legislation, 1965–1966 and 1993–1994. Administration & Society, 31, 667–686.

Hall TE, & O'Toole LJ (2004). Shaping formal networks through the regulatory process. Administration & Society, 36, 186–207.

Harding JS, Benfield EF, Bolstad PV, Helfman GS, & Jones EBD III. (1998). Stream biodiversity: The ghost of land use past. Proceedings of the National Academy of Sciences, 95, 14843–14847.

Hardy SD, & Koontz TM (2008). Reducing nonpoint source pollution through collaboration: Policies and programs across the U.S. states. Environmental Management, 41, 301–310. [PubMed: 17999107]

Hill C, & Lynn L (2003). Producing human services: Why do agencies collaborate? Public Management Review, 5, 63–81.

Hoch C (2000). The practice of local government planning. Washington, DC: Published for the ICMA Training Institute by the International City/County Management Association.

Hoornbeek J, Hansen E, Rinquist E, & Carlson R (2012). Implementing water pollution policy in the United States: Total maximum daily loads and collaborative watershed management. Society & Natural Resources, 26, 420–436.

Horner RR, & Mar BW (1983). Guide for assessing water-quality impacts of highway operations and maintenance. Transportation research record (Vol. 948, pp. 31–40). Washin gton, DC: National Research Council.

Huxham C (2003). Theorizing collaboration practice. Public Management Review, 5, 401–423.

Huxham C, & Vangen S (2000). Leadership in the shaping and implementation of collaboration agendas: How things happen in a (not quite) joined-up world. Academy of Management Journal, 43, 1159–1175.

Huxham C, & Vangen S (2005). Managing to collaborate: The theory and practice of collaborative advantage. New York, NY: Routledge.

Imperial MT (2005). Using collaboration as a governance strategy: Lessons from six watershed management programs. Administration & Society, 37, 281–320.

Innes JE (1996). Planning through consensus building: A new view of the comprehensive planning ideal. Journal of the American Planning Association, 62, 460–472.

Innes JE (1998). Information in communicative planning. Journal of the American Planning Association, 64, 52–63.

Innes JE, & Booher DE (1999). Consensus building and complex adaptive systems. Journal of the American Planning Association, 65, 412–423.

Innes JE, & Booher DE (2004). Reframing public participation: Strategies for the 21st century. Planning Theory and Practice, 5, 419–436.

Innes JE, & Booher DE (2010). Planning with complexity: An introduction to collaborative rationality for public policy. New York, NY: Routledge.

King G, & Zeng L (2006). The dangers of extreme counterfactuals. Political Analysis, 14, 131–159.

Koontz TM (2003). The farmer, the planner, and the local citizen in the dell: How collaborative groups plan for farmland preservation. Landscape and Urban Planning, 66, 19–34.

Koontz TM, & Thomas CW (2006). What do we know and need to know about the environmental outcomes of collaborative management? Public Administration Review, 66, 111–121.

Koontz TM, Steelman TA, Carmin J, Korfmacher KS, Moseley C, & Thomas CW (2004). Collaborative environmental management: What roles for government? Washington, DC: RFF Press.

Langbein LI (2000). Regulatory negotiation versus conventional rule making: Claims, counterclaims, and empirical evidence. Journal of Public Administration Research and Theory, 10, 599–632.

Layzer JA (2008). Natural experiments: Ecosystem-based management and the environment. Cambridge, MA: MIT Press.

Leach WD (2006). Collaborative public management and democracy: Evidence from western watershed partnerships. Public Administration Review, 66, 100–110.

Leach WD, & Sabatier PA (2005). Are trust and social capital the keys to success? Watershed partnerships in California and Washington. In Sabatier PA (Ed.), Swimming upstream: Collaborative approaches to watershed management (pp. 233–258). Cambridge, MA: MIT Press.

Leach WD, Pelkey NW, & Sabatier PA (2002). Stakeholder partnerships as collaborative policymaking: Evaluation criteria applied to watershed management in California and Washington. Journal of Policy Analysis and Management, 21, 645–670.

Leach WD, Weible CM, Vince SR, Siddiki SN, & Calanni JC (2013). Fostering learning through collaboration: Knowledge acquisition and belief change in marine aquaculture partnerships. Journal of Public Administration Research and Theory, 24, 591–622.

Levy JM (2013). Contemporary urban planning. Upper Saddle River, NJ: Pearson Education.

Lubell M (2004a). Resolving conflict and building cooperation in the National Estuary Program. Environmental Management, 33, 677–691. [PubMed: 14727073]

Lubell M (2004b). Collaborative environmental institutions: All talk and no action? Journal of Policy Analysis and Management, 23, 549–573.

Lubell M (2004c). Collaborative watershed management: A view from the grassroots. Policy Studies Journal, 32, 341–361.

Lubell M (2005). Political institutions and conservation by local governments. Urban Affairs Review, 40, 706–729.

Lubell M, Schneider M, Scholz JT, & Mete M (2002). Watershed partnerships and the emergence of collective action institutions. American Journal of Political Science, 46, 148–163.

Mandell MP (1999). The impact of collaborative efforts. Review of Policy Research, 16, 4–17.

Mandell MP (2001). Collaboration through network structures for community building efforts. National Civic Review, 90, 279–288.

Margerum RD (2002). Collaborative planning: Building consensus and building a distinct model for practice. Journal of Planning Education and Research, 21, 237–253.

Margerum RD (2007). Overcoming locally based collaboration constraints. Society & Natural Resources, 20, 135–152.

Margerum RD (2008). A typology of collaboration efforts in environmental management. Environmental Management, 41, 487–500. [PubMed: 18228089]

Margerum RD (2011). Beyond consensus: Improving collaborative planning and management. Cambridge, MA: MIT Press.

Margerum RD, & Born SM (2000). A co-ordination diagnostic for improving integrated environmental management. Journal of Environmental Planning and Management, 43, 5–21.

Margerum RD, & Holland R (2001). South East Queensland 2001: Has it helped improve environmental planning? Australian Planner, 38, 142–150.

Mattessich PW, Murray-Close M, & Monsey BR (2001). Collaboration: What makes it work. St. Paul, MN: Amherst H. Wilder Foundation.

Mazmanian DA, & Kraft ME (2009). Toward sustainable communities: Transition and transformations in environmental policy. Cambridge, MA: MIT Press.

McGuire M (2006). Collaborative public management: Assessing what we know and how we know it. Public Administration Review, 66, 33–43.

McGuire M, & Silvia C (2010). The effect of problem severity, managerial and organizational capacity, and agency structure on intergovernmental collaboration: Evidence from local emergency management. Public Administration Review, 70, 279–288.

McLaughlin JA, & Jordan GB (1999). Logic models: A tool for telling your programs performance story. Evaluation and Program Planning, 22, 65–72.

Meier KJ (2005). Managerial networking: Issues of measurement and research design. Administration & Society, 37, 523–541.

Montgomery DR (1994). Road surface drainage, channel initiation, and slope instability. Water Resources Research, 30, 1925–1932.

Moore E, & Koontz T (2003). A typology of collaborative watershed groups: Citizen- based, agency-based, and mixed partnerships. Society & Natural Resources, 16, 451–460.

Moreland RL, Levine JM, & Cini MA (1993). Group socialization: The role of commitment. In Hogg MA & Abrams D (Eds.), Group motivation: Social psychological perspectives (pp. 105–129). London: Harvester Wheatsheaf.

Newig J, & Fritsch O (2009a). Environmental governance: Participatory, multilevel, and effective? Environmental Policy and Governance, 19, 197–214.

Newig J, & Fritsch O (2009b). More input-better output: Does citizen involvement improve environmental governance? In Bluhdorn I (Ed.), Search of legitimacy. Policy making in Europe and the challenge of complexity (pp. 205–224). Opladen, Germany: Farmington Hills.

O'Leary R, & Bingham LB (2009). The collaborative public manager: New ideas for the twenty-first century. Washington, DC: Georgetown University Press.

Parker KB, Margerum RD, Dedrick DC, & Dedrick JP (2010). Sustaining watershed collaboratives: The issue of coordinator-board relationships. Society & Natural Resources, 23, 469–484.

Platt J (1981). Evidence and proof in documentary research: Some specific problems of documentary research. Sociological Review, 29, 31–52.

Pohlmeier W, Seiberlich R, & Uysal D (2013). A simple and successful method to shrink the weight. Retrieved from SSRN: http://ssrn.com/abstract=2238335.

Poole C (1991). Multiple comparisons? No problem! Epidemiology, 2, 241–243. [PubMed: 1912038]

Provan KG, & Milward HB (1995). A preliminary theory of interorganizational network effectiveness: A comparative study of four community mental health systems. Administrative Science Quarterly, 40, 1–33.

Rapp V (2008). Northwest Forest Plan, the first 10 years (1994–2003): First-decade results of the Northwest Forest Plan. Technical Report. Portland, OR: United State Forest Service: Pacific Northwest Research Station.

Raudenbush SW, & Bryk AS (2001). Hierarchical linear models: Applications and data analysis methods (2nd ed.). Thousand Oaks, CA: Sage.

Robins JM, & Rotnitzky A (1995). Semiparametric efficiency in multivariate regression models with missing data. Journal of the American Statistical Association 90, 122–129.

Robins JM, Rotnitzky A, & Zhao LP (1994). Estimation of regression coefficients when some regressors are not always observed. Journal of the American Statistical Association 89, 846–866.

Rosenbaum PR, & Rubin DB (1983). The central role of the propensity score in observational studies for causal effects. Biometrika, 70, 41–55.

Sabatier PA (2005). Swimming upstream: Collaborative approaches to watershed management. Cambridge, MA: MIT Press.

Sabatier PA, & Shaw LK (2009). Are collaborative watershed management groups democratic? An analysis of California and Washington partnerships. Journal of Soil and Water Conservation, 64, 61A–64A.

Sabatier PA, Leach WD, Lubell M, & Pelkey N (2005). Theoretical frameworks explaining partnership success. In Sabatier P (Ed.), Swimming upstream: Collaborative approaches to watershed management (pp. 173–200). Cambridge, MA: MIT Press.

Scharfstein DO, Rotnitzky A, & Robins JM (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models. Journal of the American Statistical Association, 94, 1096–1120.

Schively C (2007). A quantitative analysis of consensus building in local environmental review. Journal of Planning Education and Research, 27, 82–98.

Schneider M, Scholz J, Lubell M, Mindruta D, & Edwardson M (2003). Building consensual institutions: Networks and the National Estuary Program. American Journal of Political Science, 47, 143–158.

Smith SR (2004). Government and nonprofits in the modern age: Is independence possible? In Frumkin P & Imber JB (Eds.), Search of the nonprofit sector (pp. 3–18). New Brunswick, NJ: Transaction Publishers.

Susskind LE (1996). Dealing with an angry public: The mutual gains approach to resolving disputes. New York, NY: Free Press.

Susskind LE, & Cruikshank J (1987). Breaking the impasse: Consensual approaches to resolving public disputes. New York, NY: Basic Books.

Susskind LE, Thomas-Lamar J, & McKearnen S (1999). The consensus building handbook: A comprehensive guide to reaching agreement. Thousand Oaks, CA: Sage.

Sweeney BW, Bott TL, Jackson JK, Kaplan LA, Newbold JD, Standley LJ, Hession WC, & Horwitz RJ (2004). Riparian deforestation, stream narrowing, and loss of stream ecosystem services. Proceedings of the National Academy of Sciences of the United States of America, 101, 14132–14137. [PubMed: 15381768]

Thomas CW, & Koontz TM (2011). Research designs for evaluating the impact of community-based management on natural resource conservation. Journal of Natural Resources Policy Research, 3, 97–111.

Tong S, & Chen W (2002). Modeling the relationship between land use and surface water quality. Journal of Environmental Management, 66, 377–393. [PubMed: 12503494]

Ulibarri N (2015). Tracing process to performance of collaborative governance: A comparative case study of federal hydropower licensing. Policy Studies Journal, DOI: 10.1111/psj.12096

Vangen S, & Huxham C (2003). Nurturing collaborative relations: Building trust in interor-ganizational collaboration. Journal of Applied Behavioral Science, 39, 5–31.

Wang L, Lyons J, Kanehl P, & Gatti R (1997). Influences of watershed land use on habitat quality and biotic integrity in Wisconsin streams. Fisheries, 22, 6–12.

Weible C, Sabatier PA, & Lubell M (2004). A comparison of a collaborative and top-down approach to the use of science in policy: Establishing marine protected areas in California. Policy Studies Journal, 32, 187–207.

Wondolleck JM, & Yaffee SL (2000). Making collaboration work: Lessons from innovation in natural resource management. Washington, DC: Island Press.

Wood DJ, & Gray B (1991). Toward a comprehensive theory of collaboration. Journal of Applied Behavioral Science, 27, 139.

Yaffee SL, Phillips AR, Frentz IC, Hardy PW, Maleki SM, & Thorpe BE (1996). Ecosystem management in the United States: An assessment of current experience. Washington, DC: Island Press.

Zhang Y, Richardson JS, & Pinto X (2009). Catchment-scale effects of forestry practices on benthic invertebrate communities in Pacific coastal streams. Journal of Applied Ecology, 46, 1292–1303.

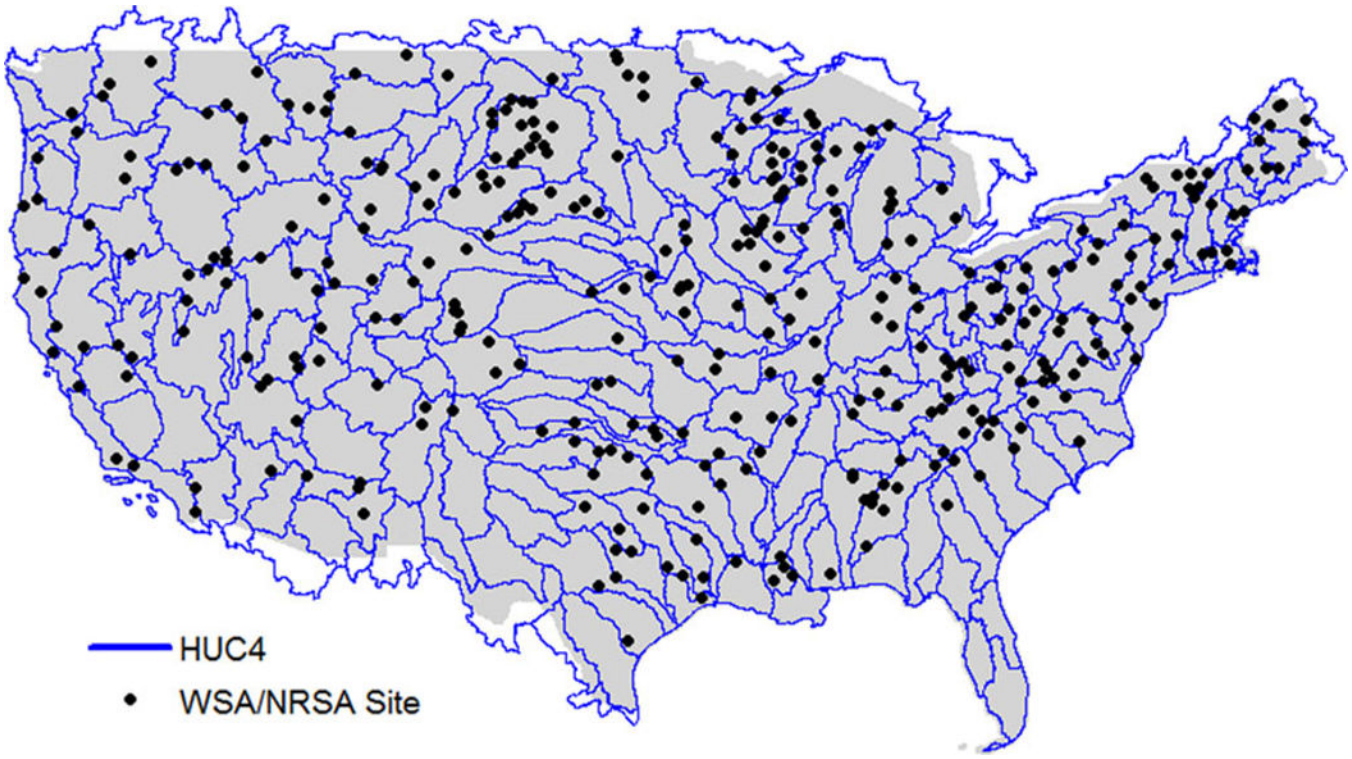

Figure 1.
Sites Sampled in WSA and NRSA.

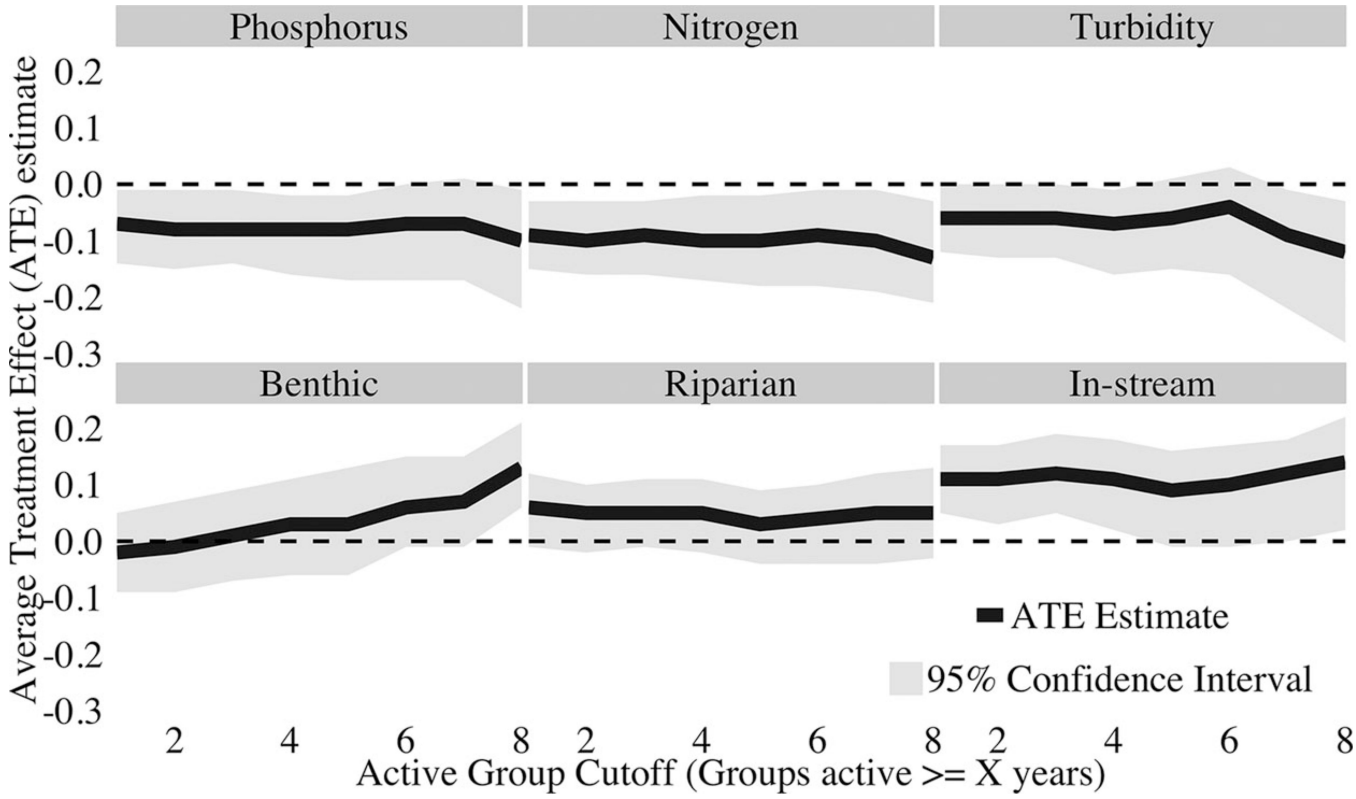

Figure 2.
ATE Estimates with Varying Cutoff.

Table 1.

Outcome metrics (unstandardized).

	Mean	SD	Units	Details
Phosphorus	131.68	402.49	µg/L	Total nitrogen content
Nitrogen	1,216.69	2,206.71	µg/L	Total phosphorus content
Turbidity	289.08	166.67	NTU	Turbidity level
Benthic health	351.53	206.20	Index score (0 to 100)	Benthic multimetric index
Riparian cover	298.19	183.53	Sum areal prop.	Ground + mid + canopy cover
In-stream habitat	197.37	133.28	Sum areal prop.	All natural cover types

NTU, nephelometric turbidity unit.

Table 2.

Group variables.

Variable	Levels	*n*	Percent
Has coordinator	None	54	23.2
	Dedicated coordinator	179	76.8
	All	233	100.0
Goal Formalization	0	138	59.2
	1	95	40.8
	All	233	100.0
Group Responsibility	0	129	55.4
	1	104	44.6
	All	233	100.0
Total stakeholder types	0 (local government only)	12	5.2
	1	13	5.6
	2	15	6.4
	3	25	10.7
	4	50	21.5
	5	64	27.5
	6	35	15.0
	7	14	6.0
	8	5	2.1
	All	233	100.0

Table 3.

Average treatment effect (ATE) (y < 4).

	Phosphorus	Nitrogen	Turbidity	Benthic	Riparian	In stream
ATE	−0.08*	−0.10*	−0.07*	0.03	0.05	0.11*
	[−0.16; −0.02]	[−0.17; −0.02]	[−0.16; −0.01]	[−0.06; 0.11]	[−0.02; 0.11]	[0.02; 0.18]

Note:

* 0 is outside the confidence interval.

Table 4.

Multilevel model results.

	Phosphorus	Nitrogen	Turbidity	Benthic	Riparian	In stream
Site disturbance	0.10*** (0.04, 0.17)	0.10*** (0.05, 0.16)	0.001 (−0.07, 0.07)	−0.03 (−0.10, 0.05)	−0.14*** (−0.21, −0.07)	−0.02 (−0.09, 0.05)
Percent agricultural	0.25*** (0.13, 0.37)	0.24*** (0.15, 0.36)	0.02 (−0.09, 0.17)	−0.12** (−0.25, −0.01)	−0.04 (−0.16, 0.08)	−0.02 (−0.15, 0.11)
Percent forest	−0.14** (−0.28, 0.004)	−0.23*** (−0.30, −0.06)	−0.18** (−0.34, −0.05)	0.21*** (0.06, 0.34)	0.19*** (0.06, 0.32)	0.05 (−0.09, 0.20)
Percent developed	0.01 (−0.12, 0.13)	0.01 (−0.09, 0.10)	−0.05 (−0.26, 0.01)	0.16** (0.03, 0.29)	0.05 (−0.06, 0.17)	0.02 (−0.11, 0.14)
Pop. density	0.11* (−0.02, 0.24)	0.18*** (0.07, 0.27)	0.13* (−0.01, 0.27)	−0.20*** (−0.35, −0.06)	−0.01 (−0.14, 0.12)	−0.05 (−0.18, 0.09)
Med. income	−0.14*** (−0.25, −0.05)	−0.07 (−0.13, 0.03)	−0.13*** (−0.23, −0.02)	0.07 (−0.04, 0.17)	0.01 (−0.09, 0.11)	0.06 (−0.04, 0.15)
Road density	0.12*** (0.05, 0.20)	0.12*** (0.06, 0.19)	0.14*** (0.05, 0.21)	−0.09** (−0.18, 0.000)	−0.04 (−0.12, 0.04)	−0.07 (−0.15, 0.01)
Watershed group (WG)	−0.001 (−0.13, 0.13)	−0.09 (−0.20, 0.03)	0.01 (−0.12, 0.16)	−0.01 (−0.16, 0.14)	0.11 (−0.03, 0.24)	0.06 (−0.09, 0.21)
WG × goals/objectives	−0.01 (−0.12, 0.10)	0.13*** (0.02, 0.21)	−0.05 (−0.16, 0.07)	−0.15** (−0.28, −0.03)	−0.01 (−0.15, 0.09)	0.01 (−0.12, 0.13)
WG × coordinator	−0.04 (−0.17, 0.09)	−0.03 (−0.14, 0.09)	−0.12* (−0.27, 0.02)	0.04 (−0.11, 0.19)	−0.10 (−0.23, 0.04)	0.02 (−0.11, 0.16)
WG × management	−0.15*** (−0.26, −0.03)	−0.05 (−0.15, 0.05)	0.09 (−0.04, 0.19)	0.16** (0.02, 0.29)	0.04 (−0.09, 0.15)	0.05 (−0.07, 0.18)
WG × stakeholders	0.08 (−0.03, 0.19)	0.06 (−0.03, 0.17)	0.05 (−0.06, 0.18)	0.06 (−0.07, 0.19)	−0.05 (−0.17, 0.07)	−0.03 (−0.15, 0.08)
BIC	837.28	609.63	919.89	1.157.54	901.90	1.036.49

Note:

*** $P < 0.01$

** $P < 0.05$

* $P < 0.1$.

P-values refer to bootstrapped confidence intervals that do not contain 0.

Models also include propensity scores and random effects for year, HUC4, state, ecoregion, and stream order