Evaluating the Impact of SA 8000 Certification

Michael J. Hiscox, Claire Schwartz, and Michael W. Toffel

May 20, 2008

---

1. Introduction: The need for evaluation

The Social Accountability 8000 Standard (SA 8000), along with other types of certification standards and corporate codes of conduct, represents a new form of voluntary "self-governance" of working conditions in the private sector, initiated and implemented by companies, labor unions, and non-governmental activist groups cooperating together. There is an ongoing debate about whether this type of governance represents real and substantial progress or mere symbolism. Advocates promote SA 8000 and similar codes as a necessary tool to improve workplace conditions, especially in nations that lack robust enforcement of regulatory standards.

Many detractors worry that codes place too much emphasis on process rather than performance, and note that to be effective such codes require scrupulous monitoring by a reliable and credible third-party organizations. While SA 8000 has more stringent and specific requirements than many alternative codes, critics have questioned the qualifications and training of the auditors SAI authorizes to monitor compliance with the code.[1] Similar concerns have been voiced about auditors monitoring compliance with alternative codes, including those administered by the Fair Labor Association (FLA) and the Worldwide Responsible Apparel Production (WRAP).[2] The manager of one auditing organization authorized to certify factories to SA 8000 in China stated in 2000 that "Right now, in labour-intensive industries in southern China, the SA 8000 standard cannot be enforced effectively… The factories always find a way around the auditors."[3] Some critics worry that effective monitoring is not possible using commercial (for profit) auditors, like the ones authorized to audit SA 8000 thus far.[4] There are real concerns that SA 8000 and other types of codes might be adding to the costs of doing business without fundamentally improving social or environmental outcomes.

To date, the debates about the impact of private governance of working conditions have been dominated by philosophical and political discourse because the SA 8000 and similar codes have yet to be subjected to systematic evaluation (see O'Rourke 2003; Esbenshade 2004; Vogel 2005). Very little empirical evidence is available to indicate whether those companies that have adopted such codes offer significantly better working environments in terms of safety, health, freedom of association, and fair pay practices. Almost no systematic evidence exists to indicate whether independent organizations, such as SAI, have been able to establish effective monitoring programs that ensure compliance with their codes, or whether they are simply being used as political cover for businesses hoping to avoid further scrutiny from activists and negative publicity. In addition, we have no rigorous evidence on whether adopting SA 8000 or similar

---

[1] Labour Rights in China. 1999. *No Illusions: Against the Global Cosmetic SA 8000*. Hong Kong: Asia Monitor Resource Center.
[2] For example, *see* Dara O'Rourke. 2000. *Monitoring the Monitors: A Critique of PricewaterhouseCooper's Labor Monitoring*. Working Paper. http://nature.berkeley.edu/orourke/
[3] Jennifer Ehrlich. 2000. Sweatshop swindlers. *South China Morning Post* (Hong Kong), December 18, p. 15.
[4] Maquila Solidarity Network. 2001. Memo: Codes Update No. 8 (August). http://en.maquilasolidarity.org/en/node/531

codes has any positive or negative impacts upon staff turnover and absenteeism, product defect rates, sales growth, order size, and other measures of business performance.

We are aware of no evaluations that have sought to examine whether and how adopting SA 8000, specifically, affects workplace practices or business outcomes. Until the results of rigorous evaluations of SA 8000 are published by independent researchers, preferably in peer-reviewed academic journals, the debates about the effectiveness of the code will continue. New research that meets the highest methodological standards is critical for resolving these debates. In addition, rigorous evaluation studies can highlight areas on which to focus efforts to improve existing standards and procedures, perhaps revealing that SA 8000 certification has been more effective in dealing with some types of issues than others, for example, or that it has been more effective in some industries or countries than in others.

In this chapter, we review prior evaluations of other private codes governing workplace conditions, including ETI's Base Code, Nike's code of conduct and Fair Trade. These have taken the form of focused studies examining only producers who have adopted a specific code, and comparative studies that compare code adopters to non-adopters. We then discuss several principles and techniques well known as best practices for evaluation studies in other domains. We believe there is a critical need to incorporate these principles and techniques in future evaluations of codes like SA 8000 to bolster their robustness and enable researchers to make clear causal inferences, distinguishing the performance differences between code adopters and non-adopters before adoption ("selection effects") and the performance differences that emerge after adoption ("treatment effects").

2. Evaluation studies: A review of existing research[5]

Have other voluntary private codes governing workplace conditions like SA8000 really improved labor standards in the places they have been adopted? What impact do these codes have on business performance? To date, the research aimed at answering these questions has assumed two forms. First, a variety of *focused studies* have examined evidence from a set of producers (companies, factories, and farms) who have adopted a particular voluntary code or standard, aiming to evaluate compliance with that code and to describe the processes through which the adoption of the code might have affected practices and outcomes. Some of these studies examine quantitative data on a large number of code-adopting producers, while others provided more finely-detailed, qualitative descriptions of the ways in which a code can affect producers. Second, a smaller set of *comparison studies* have attempted to measure the impact of a particular code or standard by examining differences in statistical measures of standards and outcomes gathered from producers who have adopted the code and from producers who have not adopted the code. Both of these methods involve non-experimental research designs that place serious limits on the degree to which they can be used to make causal inferences about the impact of private governance.

2.1 Focused studies: producers who have adopted a particular code

Barrientos and Smith (2007) provide one of the most extensive studies of producers who have adopted a voluntary code, examining British companies participating in the Ethical Trading Initiative (ETI). The study aims to evaluate the impact of the ETI's Base Code on labor standards (e.g., health and safety standards and compliance with minimum wage laws) and rights of workers to organize and bargain collectively with employers. The ETI Base Code formally prohibits forced labor, child labor, discrimination, harsh or inhuman treatment, and excessive hours of work, and guarantees safe and

[5] While outside the scope of this chapter, for reviews of studies evaluating codes, standards, and government voluntary programs regarding environmental issues, see Bennear and Coglianese (2005), Coglianese, Nash, and Borck (2008), Darnall and Sides (2008), and King and Toffel (2008).

hygienic conditions, regular employment, a living wage, and freedom of association and the right to collective bargaining.

The Barrientos and Smith study combines a survey of 29 ETI member companies with more detailed case studies of 23 supplier facilities located in Britain, South Africa, India, Vietnam, and Costa Rica. The researchers interviewed managers at the 29 member companies, reviewed annual reports, and asked each company to complete a survey by email. In each supplier case study, researchers conducted interviews with workers, managers, trade union officials, and representatives of other non-governmental groups and government agencies and asked each group about the impact of the ETI code. The authors report that greater compliance with the ETI code among suppliers is associated with more proactive management approaches to code implementation among buyers and the leverage held by such buyers (in terms of the percentage of supplier output they account for and the duration of the supply relationship). The interviews with workers indicated that, in general, the ETI code had led to some improvements in health and safety standards, minimum wage compliance, benefits, and working hours, but little or no improvement in terms of freedom of association and rights to collective bargaining and protection against various types of discrimination. The authors note that the mechanisms for monitoring compliance with the code often rely upon firms that specialize in technical or financial audits and are better suited to identifying the visible aspects of code compliance, such as health and safety measures, than less visible aspects relating to workers' rights and discrimination. Barrientos and Smith argue that this weakness is compounded by the fact that auditors often collect information primarily from management, without significantly engaging workers.

Locke, Qin, and Brause (2007) examine the workplace conditions of Nike's suppliers, all of whom have agreed to comply with Nike's Code of Conduct, which "directs them to respect the rights of their employees, and to provide them with a safe and healthy work environment."[6] The authors analyze quantitative data on the working conditions among some 800 suppliers in 51 countries, gathered by Nike's own internal auditing system. The data reveal higher audit scores at suppliers that were visited more frequently by Nike production specialists and that were located in countries with stronger regulations and legal institutions. Better working conditions were also associated with smaller plant size and with more formal partnership ties with Nike. Working conditions were generally worse in facilities in Asia compared with those located elsewhere. The audit data suggest that, in general, working conditions among all Nike suppliers have improved only slightly over time. The authors also note that data from Nike's separate "compliance rating program" (which assigns simple letter grades to suppliers based on their overall compliance with health, safety, labor, and environmental standards) indicate that working conditions in over 80 percent of the firm's suppliers have either remained the same or fallen over time.

A follow-up study by Locke and Romis (2007) examined evidence from detailed case studies of two of Nike's supplier facilities in Mexico.[7] The authors chose two facilities that were very similar in terms of their size, location, product line, and place in Nike's supply chain, but were noticeably different in terms of working conditions (wage levels, work hours, and employee satisfaction). After visiting the facilities and conducting numerous interviews with managers, workers, and representatives of non-governmental groups, the authors attribute this divergence in work conditions to differences in work organization and human resource management. Specifically, they argue that the introduction of lean manufacturing techniques in one factory had led to larger investments in worker training and greater work autonomy, raising productivity and improving working conditions at the same time.

[6] Nike Inc. Nike Responsibility Governance. http://www.nikebiz.com/responsibility/cr_governance.html (accessed March 22, 2008)

[7] See also Locke, Kochan, Romis, and Qin (2007).

In addition to these studies of corporate codes of conduct in the manufacturing sector, several scholars have examined evidence on the impact of Fair Trade certification among selected sets of agricultural producers. One prominent example is the study of nine Fair Trade certified coffee cooperatives in Costa Rica conducted by Ronchi (2002), based upon interviews with managers and farmers in each of the cooperatives and evidence from documents (e.g., accounting reports) generated by the cooperatives and the national Costa Rican coffee growers' consortium, Coocafé. The study reported that these cooperatives had benefited from participation in the Fair Trade system in financial terms, receiving a higher and more stable price for their coffee, and also in non-financial ways, enjoying more support for building organizational capacities. Similar conclusions were reported by a team of researchers who interviewed farmers in seven Fair Trade certified coffee grower cooperatives in Mexico, Guatemala, and El Salvador (Murray, Raynolds, and Taylor 2003, 2005; Raynolds, Murray, and Taylor 2004). Among the non-financial benefits of participation in the Fair Trade system, the study suggested that increased social stability and greater access to technical training and education were particular important for these coffee farmers.

All these studies are similar in that they focus on one set of producers (factories or farms) that has adopted a particular voluntary code or standard. They provide richly detailed descriptions of the methods by which specific codes have been implemented and the mechanisms by which codes can succeed or fail in improving working and living conditions. However, these studies are extremely limited in terms of their ability to measure the impact of a particular code on outcomes because they lack control or comparison groups – producers who have *not* adopted the code in question, but who are otherwise virtually identical to the producers who have adopted the code, and whose experience can be compared with that of the code adopters.

2.2 Comparison studies: producers that have adopted a code vs. others that have not

Weil and Mallo (2007) examine the impact of private monitoring of labor standards by manufacturers in the American garment industry. It is not a study of the impact of a voluntary code or standard, but it does assess one aspect of private governance in the form of private monitoring. The program examined in the study is a novel combination of a government regulation (and enforcement power) and private monitoring. The Wage and Hour Division at the U.S. Department of Labor is the government agency in charge of enforcing workplace standards regulations, as enacted in the Fair Labor Standards Act (FLSA). When the Division discovers a violation of the FLSA at a contractor doing assembly work for an apparel manufacturer, it has the power to embargo sales of goods from that contractor. Faced with such an embargo, manufacturers enter into agreements with the Division to monitor their contractors, remediate violations, and notify the Division of non-compliance. Weil and Mallo study the effect of these monitoring agreements on compliance with minimum wage laws based on data from four Division surveys of around 70 apparel contractors (randomly selected from all manufacturing and contractor firms appearing on the California and New York registration lists between 1998 and 2001). Except among the New York contractors surveyed in 2001, the results indicate that contractors subject to monitoring by manufacturers had substantially higher compliance with minimum wage laws when compared with contractors not subject to such monitoring.

Nelson, Martin and Ewert (2007) have examined the impact of several different codes of conduct in the South African wine and Kenyan cut flowers industries. In South Africa the authors compared 5 wine companies that had adopted the code of practice of the Wine and Agricultural Ethical Trade Association (which is broadly similar to the ETI code) with 15 companies that had not adopted the code. In Kenya, they compared 6 farms that had adopted any of a variety of codes that address labor practices and environmental standards (including those of the Kenya Flower Council, the Flower Label Programme, and Fair Trade) with 6 farms that had not adopted any such code. The study finds that, compared to workers in the non-adopting enterprises, workers in code-adopting companies and farms experienced

better material conditions (e.g., wages, working hours, and housing quality) and better social conditions (e.g., daycare facilities, access to HIV/AIDS education and medical care). Overall, the study reports that the adoption of codes of conduct is associated with better working conditions in both industries, although the effects are not always large and are less pronounced for casual workers.

Perhaps the most ambitious comparison study aimed at measuring the impact of Fair Trade certification has been provided by Arnould, Plastina, and Ball (2006). The study was based on a survey of over 1,200 certified and non-certified coffee farmers in Peru, Nicaragua, and Guatemala (see also Plastina and Arnould 2007). In each of the three countries the researchers surveyed a random sample of farmers from within a sample of certified coffee growing cooperatives, and then surveyed a random sample of non-certified coffee farmers in areas adjacent to those cooperatives. The certified and non-certified groups were then compared. The study reports that the Fair Trade certified farmers generally received higher prices and sold larger quantities of coffee than non-certified farmers, and also appeared to enjoy a slightly higher material quality of life (in terms of access to water, medical care, cement floors, etc.) and higher levels of self reported well being. Certified farmers also appeared somewhat better off than their counterparts in terms of education and health levels within their families. Becchetti and Constantino (2006) used a similar comparison approach in a smaller study, conducting a survey of 120 Fair Trade certified and non-certified fruit farmers in Kenya. They report that certified farmers appear to have greater satisfaction with prices and incomes, greater crop diversification, and higher food consumption and dietary quality than non-certified farmers.

These types of the comparison studies provide a first step toward assessing the impact of particular voluntary codes (and private monitoring programs). However, as discussed further below, they too suffer from severe methodological limitations, the most important of which is *selection bias*. To make causal inferences, it is not sufficient to simply compare working conditions and outcomes between existing producers already participating in the code in question with producers who are not participating in the code. Since one group is participating in the code and one group is not, these two groups are likely to differ in many observed and unobserved ways that could explain any difference in outcomes between them.

3. Evaluation design: general principals for impact studies

The remedy for the methodological problems encountered in standard focused and comparison studies of voluntary codes lies in better research design. Robust evaluations should incorporate three fundamental research designs principles: (1) examine performance over time; (2) compare the performance of participants to a very similar set of non-participants; and (3) address selection bias using randomized trials wherever possible. We describe general guidelines for designing impact studies below.

3.1 Examine performance over time

Studies that simply compare the performance of producers who have adopted a particular code to that of non-adopters at a particular moment in time may well find significant differences between the groups in terms of performance. Such cross-sectional studies have several attractive features, including the ability to survey or interview managers and staff at these companies in a compressed timeframe. However, these types of studies have a fundamental design weakness when it comes to evaluation: one cannot learn whether performance differences between adopters and non-adopters actually arose due to the adoption of the code, or whether the differences observed were already evident even before the producers decided whether or not to adopt the code. For example, some producers that adopt SA 8000 may do so because they already have implemented strong work conditions, find it relatively easy to conform to the standards

requirements, and thus need to make few subsequent changes to their operations. On the other hand, some producers adopting SA 8000 may do so to drive improvements in their operations, using the standards as benchmarks to achieve. It is possible that in both of these cases, a year after adoption, adopters may outperform non-adopters. However, research designs that involve comparing the performance of adopters to non-adopters only in the post-adoption timeframe are unable to distinguish between these two very different scenarios (referred to as "selection effects" and "treatment effects", respectively).

Ideally, evaluations should gather data from all organizations in the sample before a substantial number of them adopt the code and again afterwards. Data should also be gathered from non-adopters at the same times (the importance of including non-adopters in any evaluation study is described in the next section).

Pre-implementation data. Gathering baseline data is an essential part of evaluations. Researchers should collect data on as many important producer characteristics as possible, and on all of the outcome or performance indicators that will be used in the evaluation. The important issue here is that these data are collected *before* the code is adopted by a substantial number of the producers in the sample, making it possible to distinguish between selection and treatment effects in the analysis.

Post-implementation data. The follow-up survey should be conducted well after the period during which producers adopt the code and should measure all the same producer characteristics and outcomes measured in the baseline survey. By gathering the same types of data in the baseline and the follow-up surveys, both across- and within-group differences can be examined, making the causal inferences more robust.

3.2 The importance of examining non-adopters

Examining only the performance of participants in a code seldom produces convincing evidence of the impact of adopting the code. If one only examines companies and suppliers subject to the ETI Base code, for example, without considering how these firms compare with counterparts who have *not* adopted the code, it is impossible to make any valid causal inferences about the impact of the code. How do we know what would have happened among the ETI companies and firms if they had not been part of the ETI system? Even if one can show that there has been very little improvement over time in workplace standards among suppliers subject to Nike's code of conduct, to switch to another example, if we do not know what occurred among similar suppliers who were *not* subject to the Nike code it is extremely difficult to say anything about the impact of the code.

Imagine a study that tracked the performance of organizations that adopted SA 8000 adopters in 2003, and found their performance was steady from 2001 to 2003 and then improved dramatically from 2003 to 2005. While it would be temping to conclude that SA 8000 adoption was responsible for this improvement, such a conclusion would be unwarranted. After all, non-participants may also have experienced improvements since 2003 due to factors having nothing to do with SA 8000 adoption, such as economic cycles and inflation, changes in the labor market, new regulations, changes in regulatory enforcement, or the availability of new technologies.

To be much more confident that performance changes are associated with the code being studied, researchers need to compare the performance of code adopters to that of non-adopters over time. That said, the inclusion of non-adopters and temporal data in an evaluation are necessary—but not sufficient—conditions to generate convincing causal inferences, as explained in the next section.

3.3 Overcoming selection bias

To make causal inferences, as noted above, it is not sufficient to simply compare the performance of adopters to the performance of an arbitrary group of non-adopters, as these two groups are likely to differ in many ways that could explain any difference in outcomes between them. To examine the causal effect of Fair Trade certification on farmer income, for example, it is not sufficient to simply compare the incomes of certified and non-certified farmers because the two groups are likely to differ in terms of many other characteristics—such as farming skills, innovativeness, ambition, risk acceptance, etc.—that could help explain why they have decided to join the Fair Trade program or not, and could also independently explain any difference in incomes between them.

This point can be illustrated with a simple example. Assume coffee farmers are of two types: low skill and high skill. Assume that the high skilled farmers achieve higher incomes. If the high skilled farmers tend to select into Fair Trade certification, then a simple comparison would find that certified farmers are better off than non-certified farmers, even if certification itself had zero impact on incomes. The difference in incomes could be entirely driven by the difference in skills between the two groups of farmers. Here, the selection bias can lead to a very wrong inference, with all the difference in income falsely attributed to the Fair Trade program and not to the true cause (farming skills). It need not be skill differences, of course, but any of a large set of characteristics, many of them extremely difficult to observe and measure.

It is worth noting that selection bias can undermine attempts to measure the impact of a program even if participation in the program is not voluntary. In the case of the private monitoring program in the US apparel industry, for instance, manufacturers and contractors did not voluntarily select themselves into the program: entering into a monitoring agreement (covering all its contractors) was mandatory for a manufacturer sourcing from a contractor found to be in violation of exiting labor laws by the Department of Labor. Selection bias still makes any evaluation of the impact the program very difficult. The manufacturers that had been caught sourcing from a contractor that violates labor laws and compelled to sign a monitoring agreement are actually likely to differ in many ways from counterparts who have not been in this position. Some of these differences may reflect managers' attitudes about the importance of labor standards, or expectations that the firm will be targeted for government inspections in the future. These kinds of differences between the two groups might have large effects on compliance with labor standards and it is very difficult to account for all such differences.

Below, we describe three alternative research design approaches that can provide ways to overcome selection bias in evaluations of codes such as SA 8000: (1) randomization; (2) matching to establish quasi-control groups; and (3) using instrumental variables. The first two approaches provide alternative methods that attempt to compare code adopters to a group of non-adopters that are as similar as possible in every way except for the fact that they did not adopt the code. The third approach, most commonly used in economics, involves identifying a measure that meets a unique set of criteria for eliminating selection bias effects in statistical analysis when randomization and matching designs are impossible.

Randomization. The ideal research design for evaluations would incorporate the use of randomized trials to create "treatment" and "control" groups. This is the critical methodological principle guiding the best new research evaluating policy programs and interventions associated with development and poverty alleviation.[8] Why is randomizing so helpful? Random assignments of individuals or organizations to the intervention—such as randomly assigning producers to adopt a code of conduct—will create groups that are essentially identical on all observed and unobserved characteristics. This approach completely

[8] For example, all research sponsored by the Massachusetts Institute of Technology (MIT) Poverty Action Lab incorporates randomized trials (see: http://www.povertyactionlab.com).

overcomes concerns that organizations that voluntarily sought the intervention (e.g., adopted the code) differ in important ways from those that did not. With random assignment, any difference in outcomes can be directly attributed to the intervention of interest.

There are creative ways to apply randomization even in situations in which it may seem impractical. In particular, if the actual treatment cannot be randomized, randomized encouragements to get treated can often be used. For research on the adoption of a specific voluntary code, it may be impossible to randomly assign code adoption to producers, but it may be possible to randomize some form of encouragement to adopt a code such as SA 8000 (e.g., training seminars, consultations, etc.).

There are also ways to apply randomization in cases in which it may not seem fair or ethical to randomly assign potentially beneficial treatments among needy subjects. This may be an issue when designing a study with encouragements to assist small farmers or firms in adopting a code or standard (e.g., Fair Trade). One simple approach in these types of cases involves randomization in the *order* in which the encouragements are administered among farmers (i.e., creating early and late treatment program groups), with all receiving the same assistance over time.

While experiments with random assignment represent the "gold standard" of evaluation design, they are not feasible when seeking to evaluate the impact of codes on organizations that have already voluntarily adopted them. Furthermore, random assignment is often only implemented after the program administrator is convinced of its merits by a researcher, which is not always possible due to a conflict of interest that sometimes arises. While program evaluators are interested in understanding the conditions under which the program is effective, program administrators are often most interested in deploying the code as rapidly and broadly as possible. Depriving some potential adopters from adopting—or even intentionally delaying some applicants—in order to randomly assign them to a control group may sometimes be directly at odds with the goal of rapid diffusion.

Beyond questions of feasibility, experiments sometimes require substantial patience. Even in cases where researchers successfully convince program administrators to engage in random assignment, researchers may need to wait months or years before post-assignment performance data becomes available. Indeed, the World Bank notes that "Randomized evaluation designs, involving the collection of information on project and control groups at two or more points in time, provide the most rigorous statistical analysis of project impacts and the contribution of other factors. But in practice it is rarely possible to use these designs for reasons of cost, time, methodological or ethical constraints."[9] Fortunately, two other approaches can yield robust evaluations, although each one requires important assumptions and has its own limitations. We turn to those now.

Matched control groups. Recall that the purpose of randomization is to create a group of organizations that adopted a particular code that possess the same characteristics as a group of organizations that did not adopt the code, so that the performance of both of these groups can be tracked over time – ideally before and after the former group actually adopts the code. As already alluded to, the challenges to implementing randomization are often large, and sometimes insurmountable. Fortunately, researchers can alternatively identify a "matched" group of non-adopters with very similar characteristics as the adopters. This technique requires two things of the researcher: (1) to have a thorough understanding of the factors that lead organizations to voluntarily adopt the code; and (2) to have access to data on many—ideally all —of these factors. Matching is a widely used approach in evaluation conducted by academics. Researchers can sometimes identify one or more non-adopters with characteristics that are virtually identical to each adopter ("exact matching"). Even when exact matches are not available, researchers can apply statistical

[9] World Bank. 2004 *Monitoring and Evaluation: Some Tools, Methods and Approaches.* Washington DC: World Bank Operations Evaluation Department.

techniques (e.g., "propensity score matching") to create groups of participants and non-participants that, as a whole, are very similar.[10]

Compared to randomization, developing matched control groups offers some significant advantages to researchers. It alleviates the need for working with the program administer to encourage randomization, and rather than having to wait long periods for the implementation of randomized trials, in many cases the researcher can make immediate use of post-adoption data. However, relative drawbacks include the need to understand a wide array of factors that encourage adoption, and the need to access data on those factors in the year(s) of adoption for both adopters and non-adopters. Furthermore, matching methods require the assumption that, besides those used by the researcher to create the matched groups, all *other* factors are randomly distributed across both the adopters and matched non-adopters.

Instrumental variable approaches. A third approach to avoiding selection bias is the "instrumental variables" technique. While this approach avoids the challenges of convincing program administrators to randomize assignment and waiting long periods for trials to be conducted, or of having to gather data on a wide array of factors that influence adoption, the challenges of this third technique are no less formidable. This approach requires identifying a variable with particularly unusual relationships to the other variables in the study: this special variable has to be correlated with adoption *and* have no direct influence on the performance variables once all other available factors are controlled for.[11] In effect, it has to approximate the features of random assignment.

It is often very difficult to identify appropriate instrumental variables. But we can illustrate the application to the study of the impact of voluntary codes with a hypothetical example. Imagine that the organization that oversees a particular code persuaded several major airline and hotel companies to distribute promotional fliers to their customers on a particular day to mark some relevant celebration or event (e.g., the 60th anniversary of the Universal Declaration of Human Rights). Suppose that managers of firms who read these fliers about the code are substantially more likely to adopt the code than counterparts who do not see the fliers. In this case we could use whether managers of firms traveled on these airlines or stayed in these hotels on that particular day as an instrumental variable, modeling adoption of the code in a way that allows us to weed out selection bias effects when conducting an analysis of performance measures among a sample of firms.

4. Conclusion

While the research design criteria described in this chapter are not easy to implement, they are nonetheless standard practice in other domains of program evaluation. The number of organizations becoming certified under SA 8000 and similar codes continues to rise, while additional codes governing working conditions continue to emerge. At the same time, a fierce debate is raging about whether these codes represent substantive efforts to improve working conditions or merely symbolic efforts that allow organizations to score marketing points and counteract stakeholder pressure by merely filing some paperwork. Many critics have also raised questions about the skills and incentives of those charged with monitoring and certifying organizations as being in compliance with all the terms of these codes. Until more rigorous evaluations are conducted, these debates will continue unresolved. We believe evaluations

[10] For more on matching techniques, Toffel (2006) provides a practical example of propensity score matching in his evaluation of the effects of the ISO 14001 Environmental Management System standard, and Smith and Todd (2005) provide excellent practical advice. For the conceptual and theoretical basis of matching techniques, see Heckman, Ichimura, and Todd (1998) and Rosenbaum & Rubin (1985).

[11] Technically, the instrumental variable has to be correlated with the adoption variable, but uncorrelated with the error term in the regression that predicts performance. The researcher can easily test the former, but the latter is an assumption that cannot be empirical proven (or disproven) and instead relies on the researcher to make a compelling argument about why the assumption is plausible.

designed with the features described in this chapter will help introduce systematic evidence to these important debates. This could help identify which particular codes are best able to distinguish those organizations with superior working conditions. Just as importantly, such evaluations may shed light on which elements of the codes are most effective and which types of monitoring systems represent best practices, and which areas are most in need of improvement.

References:

Arnould, Eric J., Alejandro Plastina, and Dwayne Ball, 2006. Market Disintermediation and Producer Value Capture: The Case of Fair Trade Coffee in Nicaragua, Peru and Guatemala. Paper prepared for presentation at the Conference, "Product and Market Development for Subsistence Marketplaces: Consumption and Entrepreneurship Beyond Literacy and Resource Barriers," University of Illinois at Chicago, August 2-4, 2006. (http://www.uic.edu/depts/oee/submarkets/program.htm)

Barrientos, Stephanie and Sally Smith. 2006. *The ETI Code of Labour Practice: Do Workers Really Benefit?* Institute of Development Studies, University of Sussex.

Becchetti, Leonardo and Marco Constantino. 2006. The Effects of Fair Trade on Marginalised Producers: An Impact Analysis on Kenyan Farmers. Working Paper ECINEQ WP 2006-41. European Center for the Study of Income Inequality. www.ecineq.org/milano/WP/ECINEQ2006-41.pdf

Bennear, Lori S. and Cary Coglianese. 2005. Measuring Progress: Program Evaluation of Environmental Policies. *Environment* 47(2):22-39.

Coglianese, Cary, Jennifer Nash, and Jonathan Borck. 2008. Evaluating the Social Effects of Performance-Based Environmental Programs. Paper prepared for presentation at the Conference, "Dialogue on Performance-Based Environmental Programs: Better Ways to Measure and Communicate Results," Harvard University Kennedy School of Government, March 11, 2008. www.mswg.org/documents/PathtoWashington/Coglianese_Nash_Borck_Measuring%20Social%20EffectsFNL.pdf

Darnall, Nicole and Stephen Sides. 2008. Assessing the Performance of Voluntary Environmental Programs: Does Certification Matter? *The Policy Studies Journal* 36(1): 95-117.

Esbenshade, Jill. 2004. *Monitoring Sweatshops: Workers, Consumers and the Global Apparel Industry*. Philadelphia: Temple University Press.

Heckman, James J., Hidehiko Ichimura, and Petra Todd. 1998. Matching as an econometric evaluation estimator. *The Review of Economic Studies* 65(2): 261-294.

King, Andrew A. and Michael W. Toffel. Forthcoming. Self-regulatory Institutions for Solving Environmental Problems: Perspectives and Contributions from the Management Literature. In *New Perspectives in Governance for Sustainable Development*, Magali Delmas and Oran Young (Eds.). Cambridge University Press.

Locke, Richard M., Thomas Kochan, Monica Romis, and Fei Qin. 2007. Beyond Corporate Codes of Conduct: Work Organization and Labour Standards at Nike's Suppliers. *International Labour Review* 146(1-2): 21-40.

Locke, Richard M. and Monica Romis. 2007. Improving Work Conditions in a Global Supply Chain. *MIT Sloan Management Review* 48(2): 54-62.

Locke, Richard M., Fei Qin, and Alberto Brause. 2007. Does Monitoring Improve Labor Standards? Lessons from Nike. *Industrial and Labor Relations Review* 61(1): 3-31.

Murray, Douglas L., Laura T. Raynolds, and Peter L. Taylor. 2003. One Cup at a Time: Poverty Alleviation and Fair Trade in Latin America. Manuscript prepared for the Ford Foundation. http://www.colostate.edu/Depts/Sociology/FairTradeResearchGroup/

Murray, Douglas L., Laura T. Raynolds, and Peter L. Taylor. 2006. The Future of Fair Trade Coffee: Dilemmas Facing Latin America's Small-Scale Producers. *Development in Practice* 16 (2): 179-192.

Nelson, Valerie, Adrienne Martin, and Joachim Ewert. 2007. The Impacts of Codes of Practice on Worker Livelihoods: Empirical Evidence from the South African Wine and Kenyan Cut Flower Industries. *The Journal of Corporate Citizenship* 28: 61-72.

O'Rourke, Dara. 2003. Outsourcing Regulation: Analyzing Nongovernmental Systems of Labor Standards and Monitoring. *Policy Studies Journal* 31(1): 1–30.

Plastina, Alejandro and Eric J. Arnould. 2007: Fair Trade Impacts on Educational Attainment and Health: A Three Country Comparison. White Paper 07-2001, Norton School University of Arizona.

Raynolds, Laura T., Douglas Murray, and Peter L. Taylor. 2004. Fair Trade Coffee: Building Producer Capacity via Global Networks. *Journal of International Development* 16(8):1109-1121.

Ronchi, Loraine. 2002. The Impact of Fair Trade on Producers and Their Organizations: A Case Study with Coocafé in Costa Rica. PRUS Working Paper No. 11. Poverty Research Unit at Sussex. University of Sussex. Brighton. http://www.sussex.ac.uk/Units/PRU/wps/wp11.pdf

Rosenbaum, Paul R. and Donald B. Rubin. 1985. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician* 39(1): 33-38.

Smith, Jeffrey A. and Petra E. Todd. 2005. Does matching overcome LaLonde's critique of nonexperimental estimators? *Journal of Econometrics* 125(1-2): 305-353.

Toffel, Michael W. 2006. Resolving Information Asymmetries in Markets: The Role of Certified Management Programs. Working Paper No. 07-023, Harvard Business School. http://www.hbs.edu/research/pdf/07-023.pdf