Breaking Down the Wall of Codes:

EVALUATING NON-FINANCIAL PERFORMANCE MEASUREMENT

Aaron Chatterji
David Levine

You can walk into some third-world factories and face a literal "wall of codes." Dozens of codes of conduct are pasted higher than anyone can read, assuring visitors of the social responsibility of the enterprise as defined by dozens of customers such as Levi-Strauss, Gap, and Nike and a host of certifying organizations with names as helpful as SA 8000, ISO 14000, and FLA.

The cost of this wall of codes is clear for managers: They must fill out endless forms—repeatedly explaining their safety procedures, overtime rules, and so forth. In addition, they must host endless visits from compliance auditors. Less obviously, the wall of codes is costly for consumers and other stakeholders who care about the social performance of businesses. Not only must they pay the (passed on) costs of compliance, but with so many standards they cannot always identify which standards and codes are valid measures of true social responsibility.

In addition to the codes required or preferred by customers, managers at major U.S. employers receive literally thousands of pages of surveys each year on their social, environmental, governance, and ethics policies. For example, a typical firm may be invited to fill out forms detailing its treatment of women on surveys from *WorkingMother* and *Working Women*, from *Fortune* and *The Financial Times*, and from socially responsible investment firms such as FTSE4Good, Dow Jones Sustainability Indexes (DJSI), and KLD. A similar list exists on environmental issues, with plant managers choosing among ISO 14000, Energy Star, The Coalition for Environmentally Responsible Economies (CERES), and many others.

We appreciate comments from Mary Gentile, Jason Scorse, Mike Toffel, and participants at the 2005 Haas School of Business conference on Social Metrics.

What are the costs associated with the proliferation of non-financial performance metrics and associated surveys? When managers face too many surveys and measurement systems, the natural response is to ignore them. Consequently, many metrics suffer from non-response bias or incomplete survey responses that make it difficult to generalize findings. Each additional survey raises the cost of compliance and reduces the impact of each existing metric. Besides serving as an indication that non-financial performance metrics are yet to be ingrained, high non-response rates can also significantly bias results if the group of respondents differs significantly from the non-respondents. To be blunt, because many surveys have response rates below 20 percent, we cannot be sure about the accuracy of many non-financial performance metrics.

More generally, the introduction of each additional performance metric dilutes the importance of all that preceded it, so more measurement is not always advisable. A proliferation of measures can also benefit poor performers in two ways. First, poor performers can design their own metrics with high-sounding names, give themselves passing marks, and deceive customers or other stakeholders. Philip Morris USA, for example, has a long list of recent honors on their web site. Second, the proliferation confuses many consumers and socially responsible investors so that they reduce the importance they give to the metrics with high validity. For example, The American Forest and Paper Association's Sustainable Forest Initiative has been criticized as being a weak, industry-backed standard that dilutes more strict certifications.[1]

Even worse, when metrics do not measure what is socially important (that is, they are not valid measures of social performance), increased measurement can decrease social performance. For example, a metric that consumers do care about that encourages companies to spend millions of dollars cleaning up minor environmental hazards might reduce spending on much more serious hazards that are not measured. The European Union (EU) recently passed a law banning the use of lead (among other hazardous substances) in electronic equipment. While lead is clearly hazardous, it is not clear if the alternatives, lead-free solder alloys, are environmentally less harmful.[2] Thus, this restriction could actually provide incentives for firms to invest in more hazardous production processes using lead substitutes.

A similar example can be found in child labor restrictions. While all of the major codes of conduct in apparel prohibit child labor, it is implicitly assumed that children not working in factories will attend school. Unfortunately, the alternative can sometimes be working in a more dangerous industry, such as prostitution.[3] Thus, the unintended consequences of metrics can *decrease* overall welfare.[4] This pitfall is a challenge for many non-financial performance metrics.

Why Measure Social Performance?

Measures of corporate social performance are supposed to solve two problems. First, they can help top managers, boards of directors, and other stakeholders understand if operational managers are building valuable long-term

relationships and assets. Exclusive reliance on short-term financial metrics provides incentives to take potentially unprofitable risks and to depreciate hard-to-measure assets such as employee skill or customer loyalty.[5] As such, non-financial performance metrics can be part of a "balanced scorecard"[6] that can help to build long-term shareholder value. Second, such metrics can help customers, communities, regulators, and potential employees judge the social performance of enterprises. If some of these stakeholders favor socially responsible businesses and have power to reward it, reporting such metrics can increase the level of social performance that maximizes profits. For example, if pollution is publicized, some plants may voluntarily reduce their emissions to avoid public disapproval.

As noted above, metrics that are not reliable, valid, or comparable can lead to outcomes that harm corporate social performance and overall welfare.[7] Even if many of the metrics are sensible, the proliferation of overlapping metrics on a single topic burdens managers and is costly to shareholders and consumers.

Unfortunately, some of the metrics currently used are not sensible. Poor performers have incentives to invent and adopt unreliable, invalid, and non-comparable standards, because stakeholders will find it difficult to differentiate which standards are valid. For example, when looking for a new shirt, few consumers can distinguish whether certifications from Worker Rights Consortium, Worldwide Responsible Apparel Production, The Clean Clothes Campaign, or Fair Labor Association best match their desire to avoid products made in sweatshops. In fact, each additional certification and corresponding acronym can actually decrease overall welfare, even while increasing the amount of measurement (and resulting costs).

Overview of Non-Financial Performance Measurement

The goal of non-financial performance measurement is to align managerial incentives with long-term shareholder value and to better align shareholder value creation with social value creation.

Metrics, Ratings, and Codes

At an abstract level, a measurement system often starts with a system designer who formulates a standard or rating system, which is accompanied in most cases by an assessment. The assessment can be done by the system designer or by an outside organization that the system designer certifies. The product of that assessment leads to a quantitative rating or certification and is utilized by various stakeholders to gain insight on the non-financial performance of an enterprise.

This description matches the formal structure in financial reporting in the United States. The system designer is the Securities and Exchange Commission (SEC) in cooperation with the Public Company Accounting Oversight Board (PCAOB) and the Financial Accounting Standards Board (FASB), the criteria is

the Generally Accepted Accounting Principles (GAAP), and the SEC and FASB have a complex procedure to approve auditing firms and auditors.

In the realm of nonfinancial performance, socially responsible investing firms develop their ratings criteria (one example of a system) and assess the firms according to those criteria. The ratings are then used to construct socially responsible portfolios for the end user. Organizations such as the Fair Labor Association develop their code of conduct (another example of a system) and leave the assessment to third-party certification groups. If certified, firms and factories can use the certification to signal their social performance to retailers, governments, and consumers. The United Nations developed the Global Compact, but because there is no assessment mechanism (just corporate self-reports), the Global Compact has been widely criticized.[9]

The non-financial performance metrics have another difference from the U.S. accounting standards in that standards are not derived from a legislated authority. Instead, there are usually complex negotiations among companies, NGOs, and other stakeholders in determining who will administer codes, the content of codes, and how to measure compliance. In addition, there is a dimension of market competition, where some codes are adopted more widely, often due to pressure from companies themselves, consumers, major customers, financial markets, or other stakeholders.

Within the realm of nonfinancial metrics, there are differences between compliance with a given code of conduct and a numerical rating of social responsibility. Importantly, our analysis and critiques apply to both codes of conduct and to ratings. Thus, we use the term "metric" to include all types of measures, standards, and codes relating to non-financial performance.

Reliability and Comparability

A measure is reliable if it provides the same answer when applied more than one time. While this seems intuitive, for many non-financial performance metrics, it is not a given. If a questionnaire is filled out at different times, by different people, in different divisions of the same firm, the answers can vary widely. In addition, because many non-financial performance surveys cover a wide range of topics, it is unlikely that one individual in an organization will have all the necessary information at their disposal. Thus, in many cases the quality of survey responses depends on organizational efforts to coordinate information from many different sources.

Importantly, Gerhart and his colleagues find that when two respondents at the same organization describe the workplace, their responses are only moderately correlated.[10] For example, even on a simple question such as "Applicants for this job take formal tests (paper and pencil or work sample) before being hired" the correlation of responses was only .38—an unimpressive figure. Simply put, these findings suggest that relying on a single rater could be unwise. Moreover, in a follow up study they found that when a manager responds about the workplace practices in the organization, that response is only weakly related to the practices that employees report.[11] Measurement of human rights, equal

opportunity, or environmental hazards are significantly more challenging as describing whether an employment test is present, yet many metrics rely on single respondents from each organization.

A related issue is the comparability of a particular metric across different firms and over time. Many environmental performance measures suffer from lack of comparability, which hinders improvements because top performers are difficult to discern. How should emissions of toxic materials be compared across industries for example? If comparable measures were used, researchers could easily compare firms across several social responsibility metrics or track a single firm's performance over time. These types of analyses would help us to identify key issues in corporate social responsibility.

Validity

Validity—whether the measure identifies performance that is important to society—is more difficult to assess than reliability (which just identifies whether the measure comes out the same each time it is used). A measure may be reliable but not accurately measure an outcome that matters to stakeholders. For example, a reliable metric could be the number of minorities on the firm's board of directors. This metric is reliable, in that different attempts to measure the minority representation on the board probably come up with the similar answers. This metric is also comparable across companies and sectors. A deeper question, which goes to the heart of the concept of validity, is whether this metric really tells us anything about whether minority employees at a particular firm face equal opportunities? It would be possible for a firm to have minority board members and still not treat their minority employees fairly.

With regards to the Environmental Protection Agency's Toxic Release Inventory, for example, pounds of emissions released by each facility are often publicized. Total toxicity (an index that emphasizes the truly hazardous emissions) would be a more valid measure of the harmful externalities associated with a firm's production process than pounds of emissions.

We could make a similar critique of many measures used in identifying "social responsibility." For example, the Dow Jones Sustainability Indexes use the size of the corporate board as one indication of good corporate governance. A poorly designed measure with low validity may not provide stakeholders with the information that they desire about a particular social performance category.

Unfortunately, the metrics that are easiest to report are not always the most informative. As a result, it is easy to imagine a situation where a firm reports superior environmental performance based on available measures, while it causes environmental damage in ways that are difficult to monitor. This issue presents a serious challenge to measuring non-financial performance.

Validity is also reduced because very few non-financial performance metrics capture the social performance of suppliers and the supply chain. For example, if Nike is surveyed about the working conditions in the facilities it owns versus those who supply them, the answers (and social implications) are very different—Nike owns no manufacturing plants. Labor activists' struggle against

Nike and other apparel producers has been to link their brand image to the behavior of their suppliers.[12]

Ignoring suppliers' social performance allows firms to reduce their reported emissions level by selling their high-emissions plants, which would not necessarily improve overall welfare. In the worst case, ignoring suppliers means that monitoring the social performance of an enterprise can worsen outcomes. For example, focusing on emissions by company-owned plants can provide incentives for companies to stop producing some products and instead to import them from nations with weak environmental laws and enforcement—increasing global pollution.

The concept of validity also depends on stakeholders' values and their beliefs about labor markets. Consider the application to developing nations of U.S. safety standards for well-understood workplace hazards. Most economists would view this policy as undesirable because it would increase the cost of operating plants in poor nations. As costs rise, employment and incomes fall for poor workers. If workers have full information about the safety risks they face, it is plausible that they accept the risk only because they receive above-average wages in their (poor) nation. The stricter safety regulations might cost them their jobs or, at a minimum, lower their pay by more than they value the lower risk. In short, if poor workers are compensated for a safety risk through wages or benefits, then imposing U.S.-level safety standards helps nobody.

At the same time, while economic theory suggests that workers are compensated for known risks, there is only mixed evidence of this effect in the United States and other industrialized nations.[13] Moreover, when health hazards are hard to observe (such as long-term cancer risks) it is unlikely that workers will be compensated for any risk; thus, imposing higher regulations are particularly appealing—even to economists—in these cases. The key point is that U.S.-level standards make more sense for hard-to-observe health hazards than for hazards that workers can take into account when accepting a job. Nevertheless, few standards distinguish these types of risk. (Some standards such as SA 8000 have emphasized workers' right to know about the hazards they work with—the sort of standard even economists approve.)

Validity can also depend on the context. Many non-financial performance metrics measure water use in terms of meters cubed (e.g., the Dow Jones Sustainability Index), but in regions where water is renewable and priced at its social value, water conservation is not an important social goal. Society is not better off if this "environmental" metric induces firms to spend $100 to save water that is only worth $50 to society. At the same time, in most of the world water is not priced near its social value, especially when water is being pumped out of an aquifer that is slow to refill. In that case, it makes sense to encourage firms to spend $100 to save water worth $200 to society. Unfortunately, few of the existing metrics distinguish whether water is initially misallocated.

Finally, to construct an ideal system for measuring non-financial performance, there needs to be an understanding of what the most significant externalities are in different contexts. As the examples given above show, such a

FIGURE 1. Can you match the characteristics to the code?
(Hint: Some codes match more than one characteristic.)

Code of Conduct	Characteristics
1. FLA	a. No prison labor
2. WRC	b. Must pay a living wage
3. WRAP	c. Must have independent monitors
4. SA 8000	d. Guaranteed right to organize unions
5. ETI	e. Monitors should make announced visits
6. Fair Wear Foundation	f. Factory reports are made public
7. Wal-Mart's Code of Conduct	g. Certify only facilities not brands

Suggested Answer Key (open to interpretation):
1. a, c, d, f 2. a, b, c, d, f 3. a, g 4. a, c, d, f, g 5. a, b 6. a, b 7. a

consensus must be based on a nuanced understanding of where markets are currently failing—but this understanding does not form the basis of current metrics.

The Proliferation of Non-Financial Performance Metrics

The apparel industry and socially responsible investment firms are two sectors that are illustrative of the general issues relating to a proliferation of codes. They are sectors that have already garnered considerable public attention. It is important to note that these two cases are not intended to formally test hypotheses related to corporate social performance and existing metrics. Rather, they are used to highlight the salient points of our analysis.

Apparel Industry Codes of Conduct and Monitoring Systems

> "I'm not really aware of that. My job with Nike is to endorse the product. Their job is to be up on that."—Michael Jordan, when asked by *Time Magazine* about Nike's alleged exploitation of its workers[14]

> "I don't think that it's something that, when we get the various sneakers, that we think of who made them. Maybe its ignorance on our part, but it's a very honest ignorance."—Jim Calhoun, coach of the men's basketball team at the University of Connecticut, when asked about the allegations against Nike and its foreign production facilities[15]

The current proliferation of codes of conduct in the apparel industry has grown out of fundamental disagreements over how much workers should be paid, what kinds of safety standards they should have, who should monitor the factories, and whether or not workers should be guaranteed the right to organize. Unfortunately, the proliferation of codes has resulted in considerable confusion for consumers (see Figure 1). While considerable differences remain between codes of conduct from the Fair Labor Association, the Worker Rights

Consortium, Worldwide Responsible Apparel Production, and others, it is striking that the codes largely overlap on key issues that were still controversial when Jordan made his comment. Indeed, Nike was not "up on" the working conditions of its foreign suppliers, and even if they were, the company argued early on that they should not be held responsible because, in the words of their regional spokesperson in Asia, "we don't make shoes."[16]

To understand how the boundaries of Nike's corporate responsibility broadened from its own direct operations to also include that of its foreign suppliers, we must go back to the stories Jeff Ballinger, Kathie Lee Gifford, and others who played prominent roles in the evolution of industry codes of conduct in apparel. In doing so, we can identify the major differences between existing codes and assess the impact of the proliferation of non-financial performance monitoring and measurement in this industry.

Jeff Ballinger and Kathie Lee Gifford

Jeff Ballinger was a labor activist at the AFL-CIO's branch in Indonesia in the late 1980s. During a softball game with Nike employees, Ballinger mentioned that he was working to protect workers at supplier factories for American firms. A Nike employee then quipped, "I am your worst nightmare."[17] Little did the employee know that Ballinger would instead soon become Nike's worst nightmare. Ballinger quickly became aware of Nike suppliers' suspect practices in Indonesia and decided to make Nike the focal point of a broader war on sweatshops and economic injustice. Between 1988-1992, Ballinger began to compile data on Nike's production facilities in the country, focusing on workers' pay. When Ballinger reported his findings that Nike production facility workers were working long hours for meager pay (even by Indonesian standards), Nike initially responded with a weak code of conduct (consisting of 7 broad principles and requiring compliance with local laws), all the while maintaining that the firm could not be held responsible for the actions of their independent suppliers. While it did include progressive environmental and non-discrimination clauses, the critical component of a fair wage was calculated by Indonesian standards, in accordance with the daily minimum caloric intake for an individual, not his family. In addition, there were no guarantees that the Indonesian government would actually enforce the law.[18]

Ballinger persisted, publishing an article on Nike in *Harper's* magazine in 1992 that drew attention to Nike's labor practices. However, given Nike's immense advertising budget and public relations skill, it is unclear if Ballinger could have ever succeeded in his cause without the unlikely help of a television talk show host named Kathie Lee Gifford.

> "I felt like I was being [,] of all people, being kicked in the teeth for trying to help kids."—a tearful Kathy Lee Gifford, commenting of the allegations that her clothing line was being produced in sweatshops using child labor[19]

Kathie Lee Gifford was the popular star of the morning show "Live with Regis and Kathie Lee" and an unlikely target for a media scandal. Still, when it

was discovered that her clothing line was manufactured in Honduras by 13-year-old girls working long hours for low wages, Gifford's reputation was threatened and intense media attention was focused on labor conditions in the developing world. The Gifford story, although unrelated to Nike, was a powerful flashpoint for the mainstream media, allowing them to focus on corporate labor practices in general and the mounting critiques against Nike specifically.

At this point, the Clinton administration, led by Secretary of Labor Robert Reich, formed the Apparel Industry Partnership (AIP) in August 1996. Reich had been waging a domestic campaign against sweatshops, called No Sweat, and the mainstream attention of the Gifford scandal added new momentum to the international component of his efforts. The original Apparel Industry Partnership was composed of industry representatives (both apparel manufacturers and importers), unions, and non-governmental organizations, and it aimed to develop an industrywide code of conduct and monitoring plan.[20]

Immediately upon inception, some criticized the presence of industry representatives in the organizations, believing that they would block meaningful reform and use weak standards to tout their corporate citizenship. The debate over the proper role of industry in designing codes of conduct and monitoring schemes is at the heart of many of the differences between standards today.

The Fair Labor Association and its Discontents

In November 1998, when the Apparel Industry Partnership announced the formation of the Fair Labor Association (FLA) to enact the provisions of the code of conduct and facilitate monitoring of compliance, the two major union members—Union of Needletrades, Industrial, and Textile Employees (UNITE) and the Retail, Wholesale, and Department Store Union—left the group. In addition, the Interfaith Center for Corporate Responsibility (ICCR), a religious group that uses shareholders' resolutions to try to affect corporate practices, also decided to withdraw. Two apparel firms also left the partnership, presumably because the monitoring component was too strict.[21]

The split of the Apparel Industry Partnership was across lines that still divide the different codes of conduct today. The FLA system called for paying workers the prevailing wage in the industry and instituted a monitoring scheme where firms could select and pay their own monitors. Its critics wanted a "living wage" and a stronger and more independent monitoring system. In addition, some detractors felt that the Fair Labor Association lacked strong language on the right to organize in nations where union activity is illegal.

The original corporations involved in the Fair Labor Association— Nike, Reebok, Phillips-Van Heusen, Liz Claiborne, Adidas-Salomon, Eddie Bauer, GEAR for Sports, Patagonia, and Polo Ralph Lauren—were joined by Nordstrom but lost Levi Strauss in 2002. The major accomplishments thus far appear to be the accreditation of several external monitors and the affiliation with 170 universities to institute a limited version of the FLA monitoring scheme with their suppliers of university logo apparel.[22]

For similar reasons that unions and NGOs had been dissatisfied with the Fair Labor Association, student activists began the United Students Against Sweatshops (USAS) in 1998. The student movement had begun in response to increasing awareness that university logo apparel was being manufactured abroad in factories that shared much in common with Nike's production facilities. The student movement was also bolstered by union support. United Students Against Sweatshops was also dissatisfied with the Fair Labor Association, so they formed the Worker Rights Consortium (WRC) in 2000. The Worker Rights Consortium was composed of union representatives, university representatives, and students. No industry representatives were involved in the formation of the Worker Rights Consortium.[23]

The differences between the Workers Rights Consortium and the Fair Labor Association are predictable based on the reasons for the original split of the Apparel Industry Partnership. The Worker Rights Consortium calls for a "living wage" defined as take home pay for a workweek of 48 hours or less that provides for basic needs and a 10% reserve for emergencies. Fair Labor Association only requires a wage in accordance with local law or industry standard, whichever is higher. Worker Rights Consortium also calls for independent monitors that make unannounced visits with full public disclosure of investigations, while Fair Labor Association certifies external monitors (from which the company can choose from and will pay for) and makes only report summaries public. Under the Fair Labor Association, only 10 percent of a firm's factories must be monitored yearly. (In 2002, the Fair Labor Association responded to some of these critiques and modified its monitoring systems to allow for unannounced visits and greater public disclosure. In addition, companies no longer choose and pay the monitors).[24]

While critics on the left felt that the Fair Labor Association standards were too weak, many clothing makers felt they were too costly. Thus, the American Apparel and Footwear Association devised its own system in 2000 for monitoring and certifying individual factories, called the Worldwide Responsible Apparel Production (WRAP). The code of conduct prescribed by the Worldwide Responsible Apparel Production is considerably weaker than that of the Fair Labor Association, especially on the central issue of wages. For example, the Worldwide Responsible Apparel Production requires only that companies pay the local minimum wage. In addition, while the Fair Labor Association code calls for 1 day off of out of seven, the Worldwide Responsible Apparel Production code allows employers to suspend this rule during busy times. Lastly, the Worldwide Responsible Apparel Production code does much less to publicize the results of their certification than the Fair Labor Association or the Worker Rights Consortium. In sum, the Worldwide Responsible Apparel Production code and monitoring system is the weakest and most industry-friendly of the major codes.[25]

Other Codes of Conduct and Monitoring Systems

Other codes of conducts and monitoring systems have arisen in recent years. The SA 8000, started in 1998, is modeled after the success of ISO 9000

and ISO 14000, which are produced by the International Standards Organization. The SA 8000 only accredits facilities through external monitors, and unlike Fair Labor Association and Worldwide Responsible Apparel Production, it does not examine the results of the audits. In terms of wages, the SA 8000 code calls for a wage that is legal and in accordance with industry standards, and satisfies "the basic need of the workers and their families." This standard is somewhat stronger than the Fair Labor Association code. In addition, the SA 8000 has stronger language on the right to organize, calling on firms to "facilitate parallel means of independent and free association and bargaining" where those freedoms do not exist.[26]

The Fair Wear Foundation, started in Holland in 1999, also utilizes International Labor Organization Standards, and is composed of business representatives, union members, and non-governmental organizations. The Fair Wear Foundation was instituted by the Dutch chapter of the Clean Clothes Campaign, a Europeanwide organization. The Fair Wear Foundation follows the SA 8000 code with regards to wages, calling for a wage to "meet the basic needs of workers and their families and to provide some discretionary income." However, Fair Wear certifies companies rather than facilities.[27]

The Ethical Trading Initiative (ETI) is a UK-based organization of apparel companies, unions, and non-governmental organizations, which also relies on the International Labor Organization Conventions. The language on wages in the Ethical Trading Initiative is nearly identical to the Fair Wear Foundation in its discussion of "basic needs" and "some discretionary income." (Both the Fair Wear Foundation and the Ethical Trading Initiative define this as a "living wage" but are not as specific about how to calculate this wage as is the Worker Rights Consortium.) Prominent corporate members are the Gap, Levi Strauss, and The Body Shop. The Ethical Trade Initiative does not certify companies or auditors, and it does not disclose specific information about companies. Currently, they are seeking to identify best practices to improve codes of conduct governing labor practices.[28]

These codes are for the apparel industry, but most products are not produced by a single "industry." Clothing makers routinely produce clothing for dolls, which can put them under the standards of that industry. Textile factories for synthetic fibers are often closer to being chemical plants than anything in the apparel industry, and different standards apply there. The result can be an endless mish-mash of standards that confuse even those working full-time in the sector.

At the same time, some industries have done a better job at coordinating than has apparel. Both major electronics manufacturers and toy makers have created common base standards for their industries.[29] Specific manufacturers can add additional requirements, but this common floor reduces the cost of inspections and reporting for the entire value chain.

Summary

This brief history of the evolution of apparel codes of conduct is not comprehensive, but it illustrates the salient points. The major differentiation between various standards is the language on wages, which ranges from the specified "living wage" in the Worker Rights Consortium and SA8000 codes to the local wage in the Worldwide Responsible Apparel Production code. Other key differences involve the independence and character of the monitor, where the Worker Rights Consortium requires independent monitors with unannounced visits, while the Fair Labor Association allows companies to choose from a list of monitors. Finally, differences in the length of the workweek, overtime pay, and the right to organize are also apparent.

On the other hand, all of these standards have significant overlap regarding child labor, forced labor, and physical abuse of workers. In fact, each of the codes implicitly acknowledges some responsibility by the company for the working conditions of its suppliers, which, as the case of Nike elucidated, was not always a given.

Another instructive aspect of the apparel story is that the largest and most well known firms (such as Nike in apparel) are often targeted by advocates of responsible business. While this strategy makes sense in terms of attracting media attention, it may not always target the worst performers in the industry and it may reduce incentives for leading firms to improve their non-financial performance if they expect to be criticized regardless of what they do.[30]

In the case of apparel, the proliferation of codes of conduct is related to the differences among the major stakeholders—management, unions, and non-governmental organizations—over the appropriate requirements for wages, working conditions and rights, and monitoring. On the one hand, more choices allow consumers to choose the standard that satisfies them best. On the other hand, the numerous acronyms can shield poor performers and obscure the achievements of industry leaders. In that sense, the proliferation of apparel codes, despite noble intentions, could end up imposing serious costs on consumers.

Socially Responsible Investing Firms and Their Methodologies

In another area where non-financial performance metrics have proliferated, there are three major socially responsible investing (SRI) indices: KLD's Domini 400, Dow Jones Sustainability Indexes (DJSI), and the FTSE4Good (Produced jointly by *The Financial Times* and the London Stock Exchange). Each receives prominent media attention, and socially responsible investing is a huge business, with over $2.2 trillion in assets (or one out of every nine dollars invested) in professionally managed portfolios that use socially responsible investing strategies.

Despite the prominence of these indices, their measurement systems and the resulting selection and ranking of companies differ considerably.

Weighting Systems

Each of the three firms weight various aspects of financial and non-financial performance differently. To see some of the divergences, consider the importance given to environmental issues (as opposed to social or governance or workplace issues). Dow Jones typically puts a third of its points on environmental issues but raises that share in environmentally sensitive industries. KLD puts only 20% of its explicit points on environmental issues. It then uses these explicit scores as only one of the inputs in deciding who should be included in their index; the ultimate decision is made subjectively by a committee. FTSE4Good includes environmental performance as part of its criteria but does not assign it a weight.

KLD has detailed product safety criteria, while FTSE4Good has human rights as 1 of its 3 main criteria. Meanwhile, only Dow Jones Sustainability Index explicitly considers "Economic" criteria. In terms of weighting various components of non-financial performance, these three notable firms have quite different methodologies, which results in unreliable comparisons of social performance between companies. There might be good reasons to have different weighting systems to use as benchmarks against a user's own portfolio, but the implicit values that drive these differences in measurement are not elucidated by SRI firms.

A deeper question surrounds what the ideal weighting system would be in any case. Equally dividing points between each of the categories may be convenient, but the best solution might be to let investors decide how much weight to put on each category when creating their portfolio. The user can then decide how to weight different aspects of performance to suit their own goals, rather than having the measurement firm decide for them.

Relative vs. Absolute Bars of Performance

Another interesting difference between the socially responsible investing firms' methodologies involves the concept of relative vs. absolute bars of performance. Most measures of social responsibility reward an absolute level of performance: you are either certified by the Fair Labor Association or by the ISO 14000 auditor of your environmental management system standard or not. FTSE4Good follows this practice, certifying any company above its bar.

Other social performance metrics have a fixed number of "winners" who gain certification. KLD, for example, picks the top half of the S&P 500 and 150 other firms to create its Domini 400 list of socially responsible firms. (Media awards for "top 100" or "Most Admired" also have this practice of rewarding a fixed number of best performers).

The Dow Jones Sustainability Indexes chooses the top 10% of each industry. Other standards such as "Most Admired" lists and the KLD Domini 400 reward performance relative to all firms, even those in different industries. The Domini 400 has the added features that some sectors such as tobacco are completely ruled out; this element of the screening is, therefore, an absolute standard. FTSE4Good will only include firms that meet all of its criteria.

Until recently, Dow Jones selected the top 10 percent of companies in 64 sectors, which is a relative performance bar, because companies are competing with others in the same industry.

The choice of relative versus absolute bars of performance should be influenced by the overall goal of the index. Under an absolute bar of performance, where companies are either deemed socially responsible or not according to some fixed criteria, firms from the mining industry would have a much more difficult time than software firms getting into a socially responsible index. Would an absolute bar of performance then discourage entire industries from improving their non-financial performance? A relative bar of performance would allow mining firms to compete amongst each other and perhaps provide better incentives for improved non-financial performance. On the other hand, would a socially responsible investing portfolio properly include the "best" tobacco company in its index? These questions must be addressed when deciding between absolute and relative bars of performance.

The choice of relative versus absolute standards is rarely defended by those designing the standards. In fact, absolute standards, standards relative to an industry, and standards relative to all firms all answer different ethical questions and provide different incentives to high- and low-performing firms. There is not a single "correct" principle; as usual, a nuanced approach is needed.

For example, a downside of relative standards is that ethical rules are often more absolute: most socially minded investors believe there is no "most ethical" tobacco company whose profits are cleansed of unnecessary deaths. Thus, KLD and FTSE4Good have absolute screens that rule out tobacco and firearms companies, while DJSI originally did not. DJSI now purely examines relative performance and recently the index began excluding industry groups where the top performer did not score at least 25 percent of the maximum score.

For items where difficulty varies across industries, managers' incentives are maximized with relative incentives. It makes no sense to punish a utility for more carbon dioxide emissions than a consulting firm; instead, each should be competing against peers to improve environmental performance. If the utility is competing for a top score against a consulting firm, the utility can never win; thus, they do not have any incentive to reduce emissions.

At the same time, if socially responsible investors reduced the cost of capital for favored firms (a doubtful point), most such investors would prefer that low-polluting sectors of the economy grew more rapidly. As such, it *does* make sense to shift funds to lower-polluting sectors—as happens with an absolute standard, but not a standard comparing firms with their industry peers. The case is clearer for socially conscious consumers, who can shift product demand to sectors with lower absolute levels of harm.

Relative standards have two additional advantages, one technical and one substantive. The technical advantage is that "best in industry" makes it easy to rate firms that do not respond to a survey—assume they are not in the top slice that wins the certification. While some deserving firms are denied certifications they deserve, this procedure makes life easier for the certification-granting

agency. FTSE4Good apparently follows this procedure, although their materials are not clear on this point.

An important substantive feature of relative standards is that they ratchet up over time as the average level of social performance improves. A level of health and safety, for example, that was considered adequate in 1990, might not be best practice in 2005. Similarly, the Environmental Defense Fund's Scorecard reports the plants with the highest emissions in each state.[31] Such a standard always puts pressure on at least some plants to improve performance. This pressure creates the "ratcheting labor standards" whose beneficial properties are described by Sabel et al.[32]

In addition, KLD uses S&P membership as another variable strongly influencing inclusion in the Domini 400. This decision rule is unrelated to the social performance of enterprises and ensures the index has less social responsibility than is needed to achieve diversification.

Finally, one-size-fits-all standards are rarely sensible. FTSE4Good social responsible investing metric requires more reporting and more proactive policies from enterprises in sectors likely to have larger problems in an arena. For example, environmental compliance is far more important in a refinery than in an apartment building.

Data Collection

FTSE4Good and Dow Jones use surveys to collect data, while KLD found that survey response rates were too low. For example, the Dow Jones survey response rate was 25% in 2001, and only respondents were included in the index. The advantage of the relative performance bar is that DJSI might not be too far off assuming respondents are in the top 10 percent of their industry—although we have no evidence on the social responsibility of non-responders. FTSE4Good claims a high response rate for their survey in the UK, but we have been unable to find a response rate for U.S. firms (and FTSE4Good's reply to our question did not include this information)**.** The use of surveys can introduce significant non-response bias, which may reduce the reliability and generalizability of the results.

Surveys can certainly improve the quality of information collected on each firm, but with the proliferation of socially responsible investing firms and methodologies, surveys will likely continue to have low response rates in the future. One way to increase survey response rates would be to coordinate research efforts among socially responsible investing firms. KLD has taken the lead in this effort by co-founding the Sustainable Investment Research International Group (SIRI), a joint effort with 9 other socially responsible investing firms to improve global research efforts. Organizations of this type hold considerable promise in reducing compliance costs for companies and improving the quality of information obtained from survey respondents.

On the other hand, relying on media reports has severe flaws as well, as firm may be implicitly rewarded for superior public relations strategy rather than non-financial performance.

Transparency

The three methodologies also differ greatly in their accessibility and transparency. Generally, the methodologies are not well explained and difficult to understand even after a detailed reading. It is certain that Dow Jones and FTSE4Good rely more on quantitative data than KLD. This method should produce higher reliability, but if the measures and weights are poorly chosen, the measures may still have low validity. KLD uses largely qualitative and subjective measures, which make it difficult to produce comparable and reliable metrics.

While each socially responsible investing firm was responsive to questions we posed in trying to understand their measurement, the precise scores and subscores for each company in the index (much less companies not included in the index but ranked anyway) were difficult or impossible to obtain. This lack of transparency is disappointing and makes it difficult to understand what these socially responsible firms are measuring and whether we can make reasonable comparisons between them.

Other Concerns

The use of standard certifications in the scoring process also varies by firm. FTSE4Good, for example, allows membership in the UN Global Compact, SA 8000 accreditation, or support for the OECD Guidelines for Multinational Enterprises to meet the policy component of their human rights criteria. FTSE4Good also views ISO 14001 certification as equivalent to compliance with all six indicators for its environmental management criteria. Dow Jones uses these standard certifications and memberships in global standards organizations in similar ways. This overlap is a positive first step. Ideally, all socially responsible investing firms would build on common certifications. In doing so, the cost of compliance could be significantly decreased. At the same time, if the building-block certifications were not valid, the overall measures would remain problematic.[33]

Recommendations and Best Practices

There are several recommendations and best practices that would improve the state of non-financial performance metrics today. While some improvements may take time and money to implement, the long-run benefits of more reliable, comparable, and valid metrics are immense. The following recommendations can help top managers more efficiently monitor their organizations to improve long-term shareholder value and allow other stakeholders to make better-informed judgments about the social responsibility of an enterprise.

Improving the Measurement Process

First, non-financial performance metrics need to be integrated into an economic and philosophical structure that reflects that measurement organization's goals. The measurement organization needs to be precise about what it

seeks to measure, reward, and punish, and it must explain exactly how it goes about it.

Next, there should be greater consensus over what exactly should be measured and how it will be accounted for. If each firm is measuring diversity using a different method, comparability among metrics is futile. Differences due to different *values* provide valuable variety for the users of metrics. Unfortunately, many differences across metrics today are due to arbitrary choices, not thoughtful weighing of fundamental values.

Finally, as noted above, supply chains for most products span industry lines. Thus, standards must improve their comparability not only within industries, but also across them. In the computer industry, the saying is: "Standards are wonderful, there are always so many to choose from." Workplace standards should be created from a common set of measures. Different workplaces can choose different level of achievements on a single dimension, but the various levels can all be expressed in a common language. This will improve the coordination of standards within and between industries while reducing the response burden on companies.

Improving Transparency

In general, the process of measuring social performance should be much more transparent. The Dow Jones World Index provides a good example of showing the relative importance given to different topics, providing a detailed description for each item, and precisely outlining the criteria for inclusion. Whenever possible, survey instruments and the relative weights for various items and sub-scores should be detailed on a web site. Industry-specific criteria and measures should be used. In addition, schemes that use concepts such as a "living wage" should explain how they calculate that figure.

It would helpful if the measurement organizations provided detailed case studies of one or more firms and explained exactly how the information was collected and scored, and why the firm was accepted or denied for inclusion into the index.

Finally, the source of the weights in socially responsible investment indices appears arbitrary. The origins of these weights should be clearly explained. Perhaps a sensitivity analysis would help in explaining this. In addition, it would be useful to provide sub-scores so users could optimize with their own weights. If the different weights were well understood by data users, they can capture different preferences and values among the measurement organizations. In the current system, when weights and values are not very visible to users, then the variation just adds noise to the measurement process.

Reducing Compliance Costs

The proliferation of codes means that organizations must provide overlapping information many times a year. Certifying organizations should allow both standard certifications to meet criteria and also permit alternative evidence. For example, an organization might show either ISO 14000 certification of its envi-

The Global Reporting Initiative (GRI)

The Global Reporting Initiative has created a set of standards so that if companies choose to report social performance metrics, there is a common language to do that reporting. (See their web site <www.globalreporting.org/>, last accessed April 4, 2005.) These guidelines have several strengths, such as the multiple stakeholders who help design them, the presence (under development) of supplements appropriate to specific industries, and the increasing acceptance of the GRI guidelines. For example, the GRI guidelines call for information on air pollution emissions broken down by major pollutant (e.g., SOx, NOx), number of workplace injuries, a description of the policy prohibiting child labor, and hundreds of other quantitative and qualitative metrics.

ronmental management system or answer a longer questionnaire on the features of its environmental management system. Such a norm would enhance incentives for firms to achieve the building-block certifications, as they would lower the cost to replying to later requests for information.

The advantage of standard certifications is that some are becoming widely accepted and reduce measurement costs. For example, firms could voluntarily use the well-known Global Reporting Initiative (GRI) guidelines to post reports on the web first (see sidebar). Measurement organizations would then only ask what was not covered by the Global Reporting Initiative. If the Global Reporting Initiative were to promptly adopt machine-readable standards (perhaps using XML), organizations with surveys could automatically fill in many of the blanks using information that firms voluntarily posted in standard GRI formats. Firms could begin this process immediately by releasing their GRI reports in machine-readable format.

If an NGO were to collect all GRI-formatted reports into a single database, additional features could be added. Activist groups could upload their own machine-readable qualitative and quantitative ratings for each firm. It would then be straightforward for other groups to create tools that downloaded data from the database and created decision aids for socially conscious consumers, investors, potential employees, and other stakeholders. Such tools could prioritize items based on the stakeholders' values; for example, one data user might want to weight environmental scores twice as heavily as workplace issues, while another data user may only want to avoid tobacco and alcohol firms. Alternatively, data users could use decision weights chosen by groups they respect; for example, one data user might use the decision weights provided by Amnesty International while another might use weights chosen by the Catholic Church.

An additional set of groups might create tools to utilize the datasets. One tool might use a cell phone to read bar codes and provide information to consumers directly at the point of sale.[34] Another tool might use the information to help socially minded investors design optimal portfolios.

In addition, there should be more coordination among data collectors. KLD and the Sustainable Investment Research International Group, a consor-

tium of 10 socially responsible investing firms, have started to circulate a common questionnaire on social performance. This is a good example of coordination of research efforts across firms that can lower compliance costs, increase response rates, and improve data quality.

In the case of apparel, given how much overlap there is between the various codes of conduct, there exists significant opportunities for cooperation on agreed upon concerns such as child labor, forced labor, and abuse, which would allow further specialization of codes along the issues of a living wage and the right to organize. In fact, by grouping the non-controversial issues into one standard, it would be easier to differentiate the very worst performers who fall below the minimum bar of performance.

All measurement systems should follow the example of FTSE4Good in examining the most dangerous industries in more detail. There is no reason to detail the environmental performance of medical device firms as much as energy companies.

The development of good internal metrics can help to lower compliance costs as well. If internal metrics are valid, reliable, and comparable, then external compliance will be easier to satisfy and measurement will likely improve.[35]

Finally, measurement organizations should provide web sites with secure online questionnaires for companies to complete. By making the process as painless as possible, response rates will rise. SRI World Report's OneReport system is one example of how better data collection could be organized.

Improving Data Quality

When important decisions are at stake, measurement organizations should not rely solely on management to describe reality. Most of the apparel standards include outside auditing, as does the ISO 14000 standard for environmental management systems. Even some magazine rankings of most admired firms include a survey of employees or other stakeholders (e.g., *The Financial Times*' "World's Most Respected Companies" also polled fund managers in 2000). SRI firms, in contrast rely on press reports coupled with management surveys with no independent confirmation of data quality, leading to concerns about the validity of self-reports. Ensuring data quality is a top priority and the auditing process needs to be reexamined and explained clearly to stakeholders.[36] In particular, greater efforts need to be made to communicate the relevant information to consumers at the point of sale.[37] Consumers should be provided with simple and clear ratings, perhaps in similar detail to the nutritional value packaging on food, which will allow them to make an informed decision in a short amount of time.

Finally, there should be a formal system to encourage continuous improvement and more research to validate the metrics. The major stakeholders in social responsibility should be funding ongoing research to examine which metrics are valid measures of the social performance they claim to measure. For example, are babies in poor nations healthier when they live downstream of factories with "good" certifications than living downstream of other factories?

Are safety certifications improving safety in certified workplaces? The list goes on. Thus far, the socially responsible investing firms have shown willingness to update their methodologies over time, but there is no systematic effort to validate their measures.

Improving SRI with Modern Finance

Although economics and finance have developed sophisticated tools to optimize risk-adjusted returns subject to constraints, the most famous socially responsible indexes (the KLD, DJSI, and FTSE4Good) do not use these tools; instead, they merely include or exclude stocks. To better match the performance of diversified portfolios, they either give preference to S&P 500 members (KLD) or pick the top stocks of industries—even industries with low social performance.

Modern portfolio theory explains how to do better, in terms of higher social performance for any level of risk and expected return. The basic intuition is that stocks should be included in the optimal socially conscious portfolio if they add diversification or if they are socially responsible. If they add both, then they are particularly highly weighted in the portfolio.

For example, consider two firms that have equal social scores: the first has a close substitute in terms of diversification (that is, its return is highly positively correlated with one or more other socially responsible stocks); while the second is uncorrelated with the rest of the portfolio. Optimal portfolio theory tells us to hold more of the second stock than of the first (unless both are below some threshold level of ethical acceptability, in which case both might be excluded).

Fortunately, several social responsible funds have moved away from simple lists of stocks and integrated both risk-return characteristics and social factors into their portfolio selection. KLD recently partnered with Barclays and introduced a social fund that combines modern portfolio theory with KLD social performance metrics to achieve the highest possible risk-adjusted returns while maintaining social factors in choosing the portfolio (iShares KLD Select Social Index Fund). Ideally, all of the indices should move away from inclusion versus exclusion and provide continuous ratings that can be used to create optimized portfolios. More social investors should combine their social concerns with modern portfolio theory to ensure that they hold the least increment in risk for any given level of social concern.[38]

A Call to Action

Managers can play a vital role in reforming non-financial performance metrics. The proliferation of these measures of nonfinancial performance and increasing stakeholder attention to them provide incentives for an increasing number of managers to engage those defining and measuring the social performance of enterprises. Metrics that are reliable, valid, and comparable are in the

interest of responsible corporations, shareholders, and the variety of stakeholders interested in corporate social responsibility.

Because the organizations that measure non-financial performance are largely dependent on managerial cooperation, managers have sufficient leverage to rationalize the metrics. Managers collectively—in many cases through trade associations—need to interact with organizations interested in measurement and measures (e.g., NGOs). These groups must work together to understand where various numbers come from, which are meaningful, and which do not connect to the goals of the NGOs or other stakeholders (such as customers or socially responsible investors). With this information, managers should, thus, respond only to surveys that use reliable, comparable, and valid measures. Managers can then design internal metrics of non-financial performance to meet the above criteria. Managers and stakeholders can work together to evaluate the validity of metrics, ensuring they meet organizational and stakeholder goals. Managers should focus on surveys that are well designed so as to be meaningful, easy to fill out, and have a high response rate.[39]

Determining the appropriate role of business in developing metrics is difficult. On one hand, business participation is essential for metrics to be widely accepted and continuously improved. However, as in the apparel codes example, participation of business can also reduce the legitimacy (and sometimes the validity) of the resulting standards. Thus, corporations can add the most value through partnerships with NGOs and other stakeholders, rather than leading initiatives themselves. In addition, businesses can lead in designing a common language that reduces reporting costs while describing performance.

Coordination is expensive in terms of time and hassle. At the same time, the expense of correctly measuring social performance is vastly lower than the costs both organizations and society currently pay to measure it incorrectly. Furthermore, cooperation need not only be on the organizational level. In many cases, one individual inside one organization can have a large effect. In some cases, we find that enterprising and socially conscious individuals within large organizations are leading the improvement of metrics, through their personal networks with other stakeholders. These personal relationships can help to form the bedrock of trust necessary for coordination.[40]

As the saying goes, "What gets measured gets managed." When well-meaning groups measure the wrong thing or too many things, then managerial effort is being wasted and both financial and social performance suffers. So managers and the measurement and standards groups need to come to a consensus on what to measure and how to do it. The results should combine better measurement with less paperwork to provide more information that matters and that is reliable, comparable, and valid.

Many organizations are pleased to be judged on their social performance. Stakeholders and managers must work together to ensure responsible organizations receive appropriate credit without endless forms and hassle.

Notes

1. The Forest Service Employees for Environmental Ethics, <www.fseee.org/forestmag/0102yuska.shtml>, last accessed May 17, 2004.
2. We thank Mike Toffel for this example. Also see Julie Schoenung, M., Oladele A. Ogunseitanm, Jean-Daniel M. Saphores, and Andrew Shapiro, "Adopting Lead-Free Electronics: Policy Differences and Knowledge Gaps," *Journal of Industrial Ecology*, 8/4 (2004): 59-85.
3. David Vogel, *The Market For Virtue: The Potential and Limits of Corporate Social Responsibility* (Washington, D.C.: Brookings Institution Press, 2005).
4. Thoughtful versions couple restrictions on child labor with policies to promote education and nutrition, as The Nike Foundation and its partners do in Bangladesh. Nike Foundation Website, <www.nike.com/nikebiz/nikefoundation/home.jhtml>, last accessed September 2, 2005.
5. Managerial incentives based on the firm's stock price is a common example.
6. Robert S. Kaplan and David P. Norton, *The Balanced Scorecard: Translating Strategy Into Action* (Boston, MA: Harvard Business School Publishing, 1996).
7. By "reliable" we mean that another test would give the same answer; by "valid" we mean that the answer given correctly describes the underlying dimension of social responsibility. For example, a measurement of emissions might be reliable in that repeated measures yield the same answer. If that emission is not costly to human or environmental health, then that metric is not a valid measure of environmental performance.
8. Due to lack of space, we ignore an array of other measurements of the social performance of enterprises ranging from global standards and codes (such as the United Nations Global Compact and the International Financial Corporation Equator Principles) for very large project loans; media surveys such as those for *Fortune's Most Admired Company* list and the *Financial Times' Most Respected Companies* list; environmental management standards such as ISO 14000; and social reporting tools such as the Global Reporting Initiative. For a brief description of many of these standards, see <www.newecon.org/DouglasCodesofConduct.html>.
9. *BusinessWeek*, July 12, 2004.
10. B. Gerhart, P. Wright, G. McMahan, and S. Snell, "Measurement Error in Research on Human Resources and Firm Performance: How Much Error Is There and How Does It Influence Effect Size Estimates?" *Personnel Psychology*, 53 (in press).
11. <www.ilr.cornell.edu/depts/cahrs/downloads/pdfs/workingpapers/WP00-21.pdf>.
12. Randy Shaw, *Reclaiming America: Nike, Clean Air, and The New National Activism* (Berkeley, CA: University of California Press, 1999).
13. Charles Brown, "Equalizing Differences in Labor Markets," *Quarterly Journal of Economics*, 85 (1980).
14. *Time Magazine* (June 1996).
15. Shaw, op. cit.
16. Harvard Business School Case Study 9-700-047
17. Shaw, op. cit.
18. Harvard Business School Case Study 9-700-047; Shaw, op. cit.
19. Quoted in Josiah Brownell, "Hey Wal-Mart, Don't You Know!? Sweatshop's Labor Gotta. . . . Stay?" working paper.
20. Kimberly Ann Eliot and Richard Freeman, "Can Labor Standards Improve Under Globalization?" Institute for International Economics, 2003.
21. Eliot and Freeman, op. cit.; William A. Douglas, "Who's Who in Codes of Conduct?" <www.newecon.org/DouglasCodesofConduct.html>, last accessed March 26, 2005.
22. Eliot and Freeman, op. cit.
23. Alexander Gourevitch, "No Justice, No Contract: The Worker Rights Consortium Leads the Fight Against Sweatshops," *The American Prospect*, June 29, 2001, <www.prospect.org/webfeatures/2001/06/gourevitch-a-06-29.html>, last accessed March 26, 2005.
24. Gourevitch, op. cit.; Dara O'Rourke, "Outsourcing Regulation: Analyzing Non-Governmental Systems of Labor Standards and Monitoring," *Policy Studies Journal*, 31/1 (2003): 1-29.
25. Eliot and Freeman, op. cit.
26. O'Rourke, op. cit.; Social Accountability International web site, <www.cepaa.org/SA8000/SA8000.htm>, last accessed March 26, 2005; National Retail

Federation web site, <www.sweatshops-retail.org/nrf%20website/initiatives.htm>, last accessed March 26, 2005.

27. The MVO Platform web site, <www.mvo-platform.nl/mvotekst/FWF%20Principles%20and%20Polocies.pdf>, last accessed March 26, 2005; Eliot and Freeman, op. cit.
28. The Ethical Trade Initiative web site, <www.ethicaltrade.org/Z/lib/base/code_en.shtml>, last accessed March 26, 2005.
29. For electronics, see the *Electronic Industry Code of Conduct*, 2004, <www-03.ibm.com/procurement/proweb.nsf/objectdocswebview/fileelectronic+industry+supply+code+of+conduct/$file/electronic+industry+supplier+code+of+conduct_oct18.04.pdf>, last accessed April 12, 2005. For toys, see International Council of Toy Industries (ICTI), *Code of Business Practices*, <www.toy-tia.org/Content/NavigationMenu/Library/Publications_Resources1/ICTI_Code/ICTI_Code_of_Business_Practices.htm>, last accessed April 12, 2005.
30. Thanks to Mike Toffel for this insight.
31. Environmental Defense Fund, *Scorecard: The Pollution Information Site*, <www.scorecard.org/>, last accessed March 25, 2005.
32. Charles Sabel, Dara O'Rourke, and Archon Fung, "Ratcheting Labor Standards: How Open Competition Can Save Ethical Sourcing," in R. Thamotheram, ed., *Visions of Ethical Sourcing* (London: Financial Times-Prentice Hall, 2000).
33. On the unclear validity of ISO 14000 as a measure of good environmental performance, see Michael W. Toffel, "Resolving Information Asymmetries in Supply Chains: The Role of Certified Voluntary Programs," working paper, Haas School of Business, University of California, Berkeley, CA, 2005.
34. Dara O'Rourke, "The Social Responsibility Barcode: Providing Information to Consumers on the Environmental and Social Impacts of Products," unpublished manuscript, University of California, Berkeley, 2005.
35. We thank Mary Gentile for this point.
36. Dow Jones claims its data are audited, meaning that its handling of the unaudited data is checked by PricewaterhouseCoopers. This claim is sufficiently confusing that cynics might believe it is meant to mislead stakeholders into believing the quite different claim that *companies'* responses are audited.
37. We thank Ellen Konar for this point.
38. If social performance metrics help align managers' interests with long-term shareholder value creation—for example, by reducing incentives to reduce investments in long-term relations—socially responsible portfolios can achieve excess returns. It still remains important to use modern portfolio theory to reduce risk.
39. Dillman(2000) provides useful techniques for increasing survey response rates. Don A. Dillman, Mail and Internet Surveys: The Tailored Design Method (New York, NY: Wiley, 2000).
40. We thank Mary Gentile for this point.