

***Breaking Down The Wall Of Codes:
Evaluating Non-Financial Performance Measurement***

Aaron Chatterji and David Levine, Haas School of Business

I. Introduction

In the shadow of recent corporate scandals, measures of the social performance of businesses have become increasingly popular. These measures cover everything from environmental performance and the treatment of workers, to corporate governance and charitable giving. Concurrently, appeals for corporate reform and better business ethics have come from policy makers, shareholders, ordinary citizens, and business leaders themselves. As a result, the proliferation of measures has left managers and other stakeholders with many options to gauge a firm's social responsibility, but no true consensus on which measures work best and what is actually being measured.

Such nonfinancial performance metrics 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 or instead taking potentially unprofitable risks and depreciating hard-to-measure assets such as employee skill or customer loyalty. As such, non-financial performance metrics can be part of a "balanced scorecard" (Kaplan and Norton 1996) 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, reporting such metrics can affect the profit-maximizing choice for businesses.

Still, the use of metrics that are not reliable, valid, or comparable can lead to outcomes that actually harm corporate social performance and overall welfare.¹ Even if many of the metrics are sensible, the proliferation of overlapping metrics on a single topic burdens managers and is costly to shareholders.

Unfortunately, all of the metrics 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. What person looking for a new shirt, for example, can distinguish which of the certifications from Worker Rights Consortium, Worldwide Responsible Apparel Production, The Clean Clothes Campaign, or Fair Labor Association match their desires to avoid products made in sweatshops? Importantly, these are only a subset of the apparel standards discussed below. Thus,

¹ 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 all attempts to measure it give the same answer. If that emission is not costly to human or environmental health – but stakeholders are misled by the metric – then it is not a valid measure of environmental performance.

more measurement is not always desirable. How then can managers make sense of the numerous non-financial performance measures that are used today and maximize long-term shareholder value? Finally, what can interested parties do to reduce the response burden on managers and yield more reliable, valid, and comparable metrics?

As we will discuss below, many of the existing metrics have severe problems regarding reliability, validity, and comparability. Thus, some common sense reforms and adoption of best practices could vastly improve measurement. In the remainder of this paper, we will first discuss what non-financial performance metrics have traditionally aimed to measure and evaluate the impact of the proliferation of non-financial performance measures. We will then outline the characteristics of an ideal non-financial performance metric. Next, we will briefly review the proliferation of codes of conduct in the apparel industry and the various methodologies of socially responsible investing firms. We will then classify the existing space of metrics across a few useful categories. Finally, we will offer some practical suggestions that will help managers to reform the measurement of non-financial performance and improve the quality of existing metrics.

II. The Wall of Codes

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 such as those listed above.

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 imposes big costs on consumers and other stakeholders. Not only must they pay the (passed on) costs of compliance, but with so many labor standards they cannot identify which ones are valid measures of true social responsibility.

In addition to the codes required by customers, an even larger set of measures are voluntary. 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*, additional surveys from *Fortune* and *The Financial Times*, and 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.

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 measurements with high-sounding names, give themselves passing marks, and deceive customers or other stakeholders. {Philip Morris USA has a long list of recent honors on their website for example} Second, the proliferation confuses many consumers and socially responsible investors so that they reduce the importance they give to the metrics with high validity.

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 (E.U.) recently passed a law banning the use of lead (among other hazardous substances) in electronic equipment. Unfortunately, a popular lead substitute often used comes from coral reefs. While lead is clearly hazardous, it is not clear if the alternative is environmentally less harmful.²

III. Overview of Non-Financial Performance Measurement

The goal of non-financial performance measurement is presumably to align managerial incentives with long-term shareholder value and to better align shareholder value creation with social value creation. What characteristics should an ideal measure have? Ideal non-financial performance measures are reliable, comparable, and valid.

1. Reliability

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

² Thank you to Mike Toffel for this example

that when two respondents at the same organization describe the workplace, their responses are only moderately correlated.³ 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. 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.⁴ Measurement of human rights, equal opportunity, or environmental hazards are at least as challenging as describing whether an employment test is present, yet many metrics rely on single respondents from companies.

At their worst, unreliable measures can be used by poor performers to indicate improvement in key areas or can be misinterpreted by stakeholders in assessing a company. Thus, ensuring the reliability of non-financial performance metrics should be a priority.

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.

2. Validity

Validity – whether the measure identifies performance that is important to society -- is a 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. We could reasonably expect that this measure would be reliable (different attempts to measure the minority representation on the board would usually come up with the same answer) and allow for comparison across companies and sectors. But a deeper question, which goes to the heart of the concept of validity, is whether or not 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, total emissions are often publicized while total toxicity, arguably more important, is not. Total toxicity would be a valid measure of the true externalities associated with a firm’s production process, while emissions may not be.⁵ We could

³ Gerhart, B., Wright, P., McMahan, G., & Snell, S. (in press). 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:

⁴ (<http://www.ilr.cornell.edu/depts/cahrs/downloads/pdfs/workingpapers/WP00-21.pdf>).

⁵ Thanks to Jason Scorse for this point

make a similar critique of many measures used in identifying “social responsibility”; for example, Dow Jones Sustainability Indexes uses the size of the corporate board as an 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 non-financial performance measurement.

In addition, because very few non-financial performance metrics capture externalities related to suppliers and the supply chain, validity is an even more serious concern. 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. The struggle at Nike and other apparel producers has been to link their brand image to the behavior at their suppliers. That logic has not spread to other sectors. Thus, firms can reduce their reported emissions level by selling their high-emissions plants or shifting those plants to low-emissions products while buying the intermediate inputs from outside suppliers with even higher emissions.

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 perspective. Consider the application of U.S. safety standards for well-understood hazards to developing nation. Most economists for example, would not view this extension of policy as desirable 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. 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 can approve.)

Validity can also depend on context. Many non-financial performance metrics measure water use in terms of meters cubed (e.g. Dow Jones), but in an environment where water is renewable and priced at the market price, 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. No existing metrics distinguish whether water is initially misallocated.

More generally, to construct an ideal non-financial performance measurement system, there needs to be consensus on what the most significant externalities are and how they relate to social welfare. As the examples given above show, such 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.

IV. The Proliferation of Non-Financial Performance Metrics

In discussing the sources of non-financial performance metrics, we can choose from global standards and codes like the United Nations Global Compact and the International Financial Corporation Equator Principles, or examine media surveys like *Fortune's Most Admired Company* list and the *Financial Times' Most Respected Companies* list, or environmental management standards like ISO 14000 or social reporting tools such as the Global Reporting Initiative. In this paper, for the sake of brevity, we focus on two interesting areas where non-financial performance measures have proliferated- The Apparel Industry and Socially Responsible Investing Firms. In this section, we attempt to understand why metrics and codes have continued to proliferate and speculate as to what the costs of this trend may be for firms and consumers.

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 (June, 1996)

"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.
(From Reclaiming America, Randy Shaw, UC Press)*

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. 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.” (HBS Case Study 9-700-047)

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.” (Shaw, 1999) 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. (HBS Case Study 9-700-047; Shaw, 1999)

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. (quoted in Brownell, In Progress)

Kathie Lee Gifford was the popular star of the morning show “Live with Regis and Kathie Lee” and an unlikely labor activist. Still, when it was discovered that her clothing line was manufactured in Honduras by 13 year old girls working long hours for low wages, she transformed almost overnight into an advocate for the broader social responsibility of enterprise, telling her story on Larry King Live. 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 comprised of industry representatives (both apparel manufacturers and importers), unions, and non-governmental organizations and aimed to develop an industry wide code of conduct and monitoring plan.

(Eliot and Freeman, 2003)

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.

(Eliot and Freeman, 2003) (<http://www.newecon.org/DouglasCodesofConduct.html>, Last accessed March 26th, 2005)

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, 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.

(Eliot and Freeman, 2003)

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 began 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 Worker Rights Consortium.

(<http://www.prospect.org/webfeatures/2001/06/gourevitch-a-06-29.html>. Last accessed March 26th, 2005)

The differences between Workers Rights Consortium and the Fair Labor Association are predictable based on the reasons for the original split of Apparel Industry Partnership. 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 Consortium 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 announced 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)

(<http://www.prospect.org/webfeatures/2001/06/gourevitch-a-06-29.html>. Last accessed March 26th, 2005)
(O'Rourke, 2002)

While critics on the left felt 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 Worldwide Responsible Apparel Production (WRAP). The code of conduct prescribed by Worldwide Responsible Apparel Production is considerably weaker than that of the Fair Labor Association, especially on the central issue of wages. For example, 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 the 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.

(Eliot and Freeman, 2003)

Other Codes of Conduct and Monitoring Systems

Other codes of conducts and monitoring systems have arisen in recent years. SA 8000, started in 1998, is modeled after the success of ISO 9000 and ISO 14000 which are produced by the International Standards Organization. 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, 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. (O’Rourke, 2002)

(<http://www.cepaa.org/SA8000/SA8000.htm>. Last accessed March 26th, 2005)

(<http://www.sweatshops-retail.org/nrf%20website/initiatives.htm>. Last accessed March 26th, 2005)

The Fair Wear Foundation, started in Holland in 1999, also utilizes International Labor Organization Standards, and is comprised 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 European wide 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.

(<http://www.mvo-platform.nl/mvotekst/FWF%20Principles%20and%20Polocies.pdf>. Last accessed March 26th, 2005)

(Eliot and Freeman, 2003)

The Ethical Trading Initiative (ETI), is a U.K. based organization of apparel companies, unions, and non-governmental organizations, that 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 Fair Wear Foundation and Ethical Trading Initiative define this as a “living wage” but are not as specific about to calculate this wage as the Worker Rights Consortium is) Prominent corporate members are the Gap, Levi Strauss, and The Body Shop. The Ethical Trade Initiative does not certify companies or auditors, and not does disclose specific information about companies. Currently, they are seeking to identify best practices to improve codes of conduct governing labor practices.

(http://www.ethicaltrade.org/Z/lib/base/code_en.shtml. Last accessed March 26th, 2005)

Summary

This brief history of the evolution of apparel codes of conduct is not comprehensive, but 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 standards discussed in this section have significant overlap on standards 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.

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 outlined above (and many more not listed) 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

Apparel industry codes of conduct are not the only area where non-financial performance metrics have proliferated. Consider now 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. While we are uncertain of the “correct” weight to put on environmental issues, all three weighting schemes cannot be correct, unless each firm is appealing to different investors.

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. Once again, if each of these firms purports to measure the same construct, why are their major areas of divergence among the criteria that they use?

Thus, in terms of weighting various components of non-financial performance, three notable firms have quite different methodologies, which results in unreliable comparisons of social performance between companies.

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 rewards 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 are 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 sometimes ethical rules are 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; thus, in these indices. While DJSI began purely examining relative performance, 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 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, which is considered adequate in a poor nation in the 1990s might not be best practice by 2005. Similarly, the Environmental Defense Fund's Scorecard reports the plants with the highest emissions in each state.⁶ 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; (2001).

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. To their credit, recently KLD 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.

Finally, one-size-fits all standards are rarely sensible. FTSE4Good social responsible investing metric require 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. The Dow Jones survey response rate was 25% in 2001 for example, and only respondents were included in the index. The advantage of the relative performance bar discussed above 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 U.K. 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

⁶ Environmental Defense Fund, *Scorecard: The Pollution Information Site*, [<http://www.scorecard.org/>], last accessed March 25, 2005.

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 sub-scores 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, allow membership in the UN Global Compact, SA 8000, 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. (On the unclear validity of ISO 14000 as a measure of good environmental performance, see Toffel 2005).

V. Defining the Dimensions of Non-Financial Performance Measurement

The previous discussion outlined some of the dimensions of non-financial performance. As the universe of non-financial performance metrics widens, it is important to isolate the key features of various metrics. In this section, classify non-financial performance metrics across a few major criteria. Many of the criteria can be directly applied to the metrics and codes described above and this system of classification can help understand present and future metrics. We follow with some specific recommendations for improvement.

Our basic scheme is outlined in Box 1.

BOX 1: Dimensions of a Social Performance Metric

1. What broad area is this metric concerned with?
 - a. Environment
 - b. Workplace
 - c. Community
 - d. Corporate governance
 - e. Hybrid of the above

2. What kind of organization created the metric?
 - a. Socially responsible investment firm
 - b. Government
 - c. Media
 - d. Standards Organization
 - e. Customer

3. What level of monitoring is required?
 - a. None
 - b. Self-report
 - c. Auditor/external monitor

4. Does this metric evaluate an internal process, outcome, or both?

5. Level of Analysis
 - a. Firm
 - b. Plant
 - c. Product

6. Level of Monitoring the entire value chain (suppliers and lifecycle environmental costs)

7. Level of Transparency of Methodology

8. Reliability and Comparability of Measures

9. Validity of Measures

There are several ways to classify the metrics we consider. Our framework identifies several important attributes of each metric to facilitate discussion, rather than providing a definitive classification of CSR metrics.

The **domain** of a metric is the broad area that it operates in. The domain of ISO 14000, for example, is environmental performance. Many of the apparel standards discussed above emphasize the workplace domain, although most also discuss environmental

issues. Some metrics, like those proposed by socially responsible investing firms, are hybrid standards covering more than one domain.

We are also concerned with classifying the organizations that create the metrics. SRI firms, government organizations, and standards organizations all produce non-financial performance measures to suit their own organizational goals. One of the first steps in understanding a metric is understanding **who created it. (The source)**

Metrics also vary along **the quality of monitoring**. Some non-financial performance frameworks like the U.N. Global Compact, a global agreement outlining its own code of conduct, have no monitoring at all, while others recommend self-reporting. Some of the apparel standards have internal and external monitors to ensure compliance. Even among external monitors, there is variation in whether auditor visits are pre-announced, carried out by independent auditors, and other dimensions that can affect the validity of the findings. The quality of monitoring can provide some insight into how effective a particular metric, standard, or code may be in influencing firm behavior.

The distinction between measuring a **process and/or an outcome** is also critical to understanding non-financial performance measurement. Many systems, like the ISO 14000 standard for environmental management systems examine the processes of an organization such as whether they monitor emissions and try to reduce them. Other measures, like the EPA's Toxic Release Inventory, focus on the outcome, in this case the amount of toxins released.

The level of analysis is also a key component of any metric. Some systems focus on the firm, while others, like ISO14001 management system for environmental performance, can be applied at the plant level. Organizations like the socially responsible investment rating firm KLD go a step further, measuring product quality and safety. Among apparel standards, the level of analysis is an important distinction. The Fair Labor Association (FLA) certifies companies (brands) and their suppliers while The Worldwide Responsible Apparel Production (WRAP) system certifies individual plants (as does SA 8000), and the Worker Rights Consortium does not provide any certification but only exposes violations.

(O'Rourke,2002)

(<http://www.newecon.org/DouglasCodesofConduct.html>. Last accessed March 26th, 2005)

An important extension of non-financial performance metrics is **supplier monitoring**. To ascertain the social responsibility of any organization, we must also examine its suppliers. While few measurement system currently account for this, there has been some movement towards increased supplier monitoring. Even less common is lifecycle measurement that includes whether products can be recycled or reused, although such metrics are important in understanding the environmental effects of a product or organization.

Transparency is another important distinguishing characteristic between metrics. A few of the metrics we studied explained exactly what was being measured and how the measurement process was carried out. Unfortunately, many other sources of non-financial

performance metrics leave their methodology vague, which adds to mounting confusion over new metrics, imposing costs of consumers.

Finally, we discussed the **reliability, comparability and validity** of metrics above. Metrics vary widely along these dimensions, although few have demonstrated they meet these criteria.

VI. Recommendations and Best Practices

There are several recommendations and best practices that would improve the state of the non-financial performance metrics today. While some improvements may take time and money to implement, we strongly believe that the long run benefits of more reliable, comparable, and valid metrics are immense. In addition, our list is far from comprehensive, as there are several different ways to conceptualize and implement a new generation of non-financial performance metrics. The following recommendations will 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.

We have divided our recommendations into 4 categories: Improving the Measurement Process, Improving Transparency, Reducing Compliance Costs, and Ensuring Data Quality.

A. 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. Of course, there always the possibility that alternate metrics could be worse, because existing metrics at least meet some basic criteria. The measurement organization needs to be precise about what it seeks to measure, reward, and punish, and explain exactly how it goes about it.

Next, SRI firms should create diversification at the portfolio weights, not the social scores. In addition, by releasing sub-scores organizations can allow independent users create their own weights.

Finally, 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.

B. Improving Transparency

In general, metrics should be much more transparent. Dow Jones is a good example of showing weights on items, a detailed description for each topic, 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 website. Industry-specific criteria and measures should be used. In addition, codes of conduct (like those in the apparel industry) should specify how they calculate a living wage and what assumptions underlie their model.

It would be extremely 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. It is crucial to include firms that were not selected, to provide a reasonable counterfactual to the components of the index.

Finally, the source of the weights in SRI 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.

C. 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 environmental 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 of replying to later requests for information.

The advantage of standard certifications is 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. 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.

In addition, there should be more coordination among data collectors. KLD and the Sustainable Investment Research International Group, a consortium 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 and increase response rates and data quality.

In the case of apparel, given how much overlap there is between the various codes of conduct, there exist significant opportunities for cooperation on agreed-upon concerns

like 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 system 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.

Finally, measurement organizations should provide websites with secure online questionnaire for companies to complete. By making the process as painless as possible, response rates will rise.

D. 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 EMS standard. Even magazine rankings of most admired firms will include a survey of employees or other stakeholders. 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.⁷

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.

VII. A Call to Action

Given the recommendations offered above, managers can play a vital role in reforming non-financial performance metrics. While some may argue that non-financial performance metrics are not a core concern for managers, the proliferation of these measures and increasing stakeholder concern suggest otherwise. Metrics that are reliable, valid, and comparable would allow firms to manage long-term shareholder value and stakeholders to more effectively monitor the social responsibility of firms. To spur this

⁷ Dow Jones claims its data are audited, meaning that its handling of the unaudited data is checked by PriceWaterhouseCoopers. This claim is sufficiently confusing we can only assume it is meant to mislead stakeholders into believing the companies’ responses are audited, which is not true.

process, managers can take a few steps. First, managers should respond only to surveys that use reliable, comparable, and valid measures. Organizations involved in non-financial performance measurement should work together with managers to ensure that their metrics meet this standard. Second, internal metrics of non-financial performance can be designed to meet the above criteria and consistently re-evaluated to ensure that the metrics are meeting organizational and stakeholder goals.

Because the organizations that measure non-financial performance are dependent on managerial cooperation, managers have a unique opportunity to influence the development of new metrics. Managers collectively – presumably through trade associations -- need to interact with organizations interested in measurement and interested in the 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). Coordination is expensive in terms of time and hassle. At the same time, the expense of measuring the social performance right are vastly lower than the costs we currently pay to measure it wrong.

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 suffer. So managers and measurement/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: more informative measures that are reliable, comparable, and valid.

References

Brownell, Josiah (In progress). “*Hey Wal-Mart, Don't You Know!? Sweatshop's Labor Gotta....Stay?*”; Working Paper

Burns, Jennifer L. 2000, “*Hitting the Wall: Nike and International Labor Practices*”; Harvard Business School Case Study 9-700-047.

Douglas, William A. 2001, “*Who's Who in Codes of Conduct?*”; <http://www.newecon.org/DouglasCodesofConduct.html> Last accessed on March 26th, 2005.

Gerhart, Barry. 1999. *Human Resource Management and Firm Performance: Measurement Issues and their Effect on Causal and Policy Inferences*. Research in Personnel and Human Resources Management 17: 31-51.

Gerhart B, Wright P, McMahan G, Snell S. 2000, *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: 803-834.

Gourevitch, Alexander. 2001. <http://www.prospect.org/webfeatures/2001/06/gourevitch-a-06-29.html>, Last accessed March 26th, 2005. The American Prospect

Eliot, Kimberly Ann and Richard Freeman. 2003. "Can Labor Standards Improve Under Globalization?"; Institute for International Economics.

Sabel, Charles, Dara O'Rourke, and Archon Fung. 2000, "Ratcheting Labor Standards: How Open Competition Can Save Ethical Sourcing," in (ed.) R. Thamootheram, Visions of Ethical Sourcing, London: Financial Times Prentice Hall.

Kaplan, Robert S. and David P. Norton, 1996. *The Balanced Scorecard: Translating Strategy Into Action*; Harvard Business School Publishing.

O'Rourke, Dara, 2003. *Outsourcing Regulation: Analyzing Non-Governmental Systems of Labor Standards and Monitoring*; Policy Studies Journal.

Shaw, Randy, 1999. *Reclaiming America: Nike, Clean Air, and The New National Activism*; UC Press.

Toffel, Michael W. 2005. *Resolving Information Asymmetries in Supply Chains: The Role of Certified Voluntary Programs*. Working Paper. Berkeley, CA: University of California Haas School of Business.

Fair Labor Association, <http://www.fairlabor.org/all/code/index.html> Last accessed March 26th, 2005

Fair Wear Foundation,
[http://www.mvo-platform.nl/mvotekst/FWF Principles and Polocies.pdf](http://www.mvo-platform.nl/mvotekst/FWF%20Principles%20and%20Policies.pdf). Last accessed March 26th, 2005

Measuring Sustainability 2004, Class Project for MBA 251 at Kenan Flagler Business School
<http://www.kenan-flagler.unc.edu/assets/documents/sustainabilityIndexSummaries.pdf>.
Last accessed March 26th, 2005

Social Accountability International, <http://www.cepaa.org/SA8000/SA8000.htm>. Last accessed March 26th, 2005

The Global Justice Report, http://www.bc.edu/bc_org/avp/cas/soc/Justice/resources/nosweat5.htm.
Last accessed March 26th, 2005

The National Retail Federation, <http://www.sweatshops-retail.org/nrf%20website/initiatives.htm>. Last accessed March 26th, 2005