Item Response Models to measure Corporate Social Responsibility

Nicolosi Marco[*]
Dipartimento di Economia, Università di Perugia, Perugia, Italy

Stefano Grassi[†]
School of Economics, University of Kent, U.K., and
CREATES, Aahrus University, Denmark

Elena Stanghellini[‡]
Dipartimento di Economia, Università di Perugia, Perugia, Italy.

March 13, 2014

Abstract

Corporate Social Responsibility (CSR) is a multidimensional concept that involves several aspects, ranging from Environment, to Social and Governance. Companies aiming to comply with CSR standards have to face challenges that vary from one aspect to the other and from one industry to the other. Latent variable models may be usefully employed to provide a unidimensional measure of the grade of compliance of a firm with CSR standards that is both understandable and theoretically solid. A methodology based on Item Response Theory has been implemented on the multidimensional sustainability rating as expressed by KLD dataset from 1991 to 2007. Results suggest that companies in the industry Oil & Gas together with firms in Industrials, Basic Materials and Telecommunications have a higher difficulty to meet the CSR standards. Criteria based on Human rights, Environment, Community and Product quality have a large capacity to select the best performing firms, as they are very discriminant, while Governance does not exhibit similar behavior. A stock selection based on the ranking of the firms according to the proposed CSR measure supports the hypothesis of a positive relationship between CSR and financial performance.

JEL Classification: C40; G11; G14

Keywords: Item Response Theory; Latent Variable Models; Portfolio management; Ranking; Socially Responsible Investment.

[*] Address for Correspondence: Via A. Pascoli 20, 06123, Perugia, Italia. E-mail: marco.nicolosi@unipg.it

[†] E-mail:S.Grassi@kent.ac.uk

[‡] E-Mail: elena.stanghellini@stat.unipg.it

1 Introduction

Socially Responsible Investments (SRI) have become a widespread practice in most industrialized countries, as the share of investors aware of the threats posed by companies that violate environmental and social standards is growing. Estimates of the Social Investment Forum (US-SIF) indicate that at the end of 2011 about one out of every nine dollars under professional management in the US was invested along this line, with an increase of 22% over the period end 2009 - end 2011 (US-SIF, 2012).

Responsible investors or portfolio managers with a SRI mandate aim to integrate Corporate Social Responsibility (CSR) criteria with the usual risk/return trade-off. One common strategy is to consider only a few of these aspects, for example including in the investment set only stocks that are high ranked in "Environment" or excluding stocks operating in controversial fields, such as tobacco, alcohol or weapon production. Different strategies have different impacts on the portfolio's financial performance. A fundamental question for investors, which is at the center of a vast debate, is whether there is a negative or a positive association between financial performance and the various SRI strategies; see Section 2 for a review.

We argue that CSR has a multidimensional nature covering various dimensions related to Environmental, Social and Governance issues. A unidimensional quantitative syntheses of all the available information is a desirable requirement. Indeed, rating agencies, such as Asset4 or Sustainalytics, form a linear combination of different indicators to provide the users with a single score that measures the overall CSR compliance of a company. A survey of 43 publications, made by Chen and Delmas (2011), found that only 8 did not use aggregating methods. Out of the remaining 35, 9 used linear aggregating methods with unequal weights, while 26 used simple average, a procedure that is often criticized as it gives all aspects the same weight, see Hopkins (2005).

Linear weighting schemes with unequal weights are more theoretically grounded. However, the issue is on how to choose weights. Ruf, Muralidhar, and Paul (1998) assign weights that take into account the preferences of different stakeholders, while Waddock and Graves (1997) developed a system based on experts' opinions. However, a linear combination of the CSR scores with subjective weighting of the CSR aspects is prone to criticism, as the resulting overall CSR performance of a company depends strongly on the choice which is often considered arbitrary (Bird, Hall, Momente, and Reggiani, 2007). Moreover, weights can be no longer adequate when applied to a different dataset or to the same dataset but at different time points (Rowley and Berman, 2000). An approach based on Data Envelopment Analysis, which is independent of any subjective a priori weight specification and can be implemented on different datasets, was first introduced by Bendheim, Waddock, and Graves (1998) and later extended by Chen and

Delmas (2011).

In this paper we suggest that latent variable models (Skrondal and Rabe-Hesket, 2004) may provide a tool to capture the firm's grade of compliance with CSR standards, by postulating a model that links the observable measurements of CSR aspects to a unidimensional trait, which is latent, and that we call "CSR ability". We use Item Response Models (see De Boeck and Wilson, 2004) to extract the one-dimensional latent variable. Item Response Models weight differently and in a non linear way the different dimensions of CSR. The influence of each aspect on the total ability depends on its capacity to discriminate well behaving companies from the others. Furthermore, as compliance with CSR standards may imply additional costs that vary across industries, any analysis that does not explicitly take this issue into account is confounded (Benson, Brailsford, and Humphrey, 2006). The proposed model allows an industry effect on CSR ability.

The Item Response model has been implemented yearly from 1991 to 2007 on the KLD rating system,[1] that measures the CSR performances of firms in the US with respect to seven dimensions capturing Corporate Governance, Environment and Social issues. Results show that in general the CSR ability is most influenced by criteria based on Human rights, Environment, Community and Product quality. Companies in Oil&Gas as well as in Industrials, Basic Materials and Telecommunications have a higher difficulty to get high CSR ratings. Therefore, given a certain pattern of CSR ratings, a company in one of the difficult industries will receive an ability score higher than a company in another industry.

The CSR ability is used to rank companies and construct high and low ranked (equally weighted, value weighted and mean-variance optimal) portfolios. The proposed measure is used as a criterion for stock selection to investigate the impact that the integration of CSR criteria in the investment strategies has on the portfolios' financial performance, as measured by the Jensen's α in the Carhart's model. A comparison between SR and conventional funds has been done for example in Statman (2000) and Bauer, Koedijk, and Otten (2005), where no significative differences were found between their financial performances, and in Nofsinger and Varma (2013), where it was found that SR funds outperformed conventional funds during the global financial crisis. However, a comparison between investment funds can be altered by managerial skills and expenses that vary between funds. On the contrary, following Kempf and Osthoff (2007) and Statman and Glushkov (2009), we compare the performance of hand-made strategies, or indexes, constructed on the basis of the CSR, avoiding in such a way this confounding effect. We find that the high ranked strategies outperform the

[1] It is theoretically possible to estimate the model simultaneously to all years, however, it would lead to serious problems of non-identifiability and instability as the ranked companies vary from one year to the other.

low ranked ones in terms of the Jensen's α. Similar conclusions are in Kempf and Osthoff (2007) and Statman and Glushkov (2009). For the sake of comparison, portfolios are formed also on the basis of 4 other known rankings used in literature, see Section 4.2. Results show that the proposed ranking outperforms, in terms of statistical significance and robustness, three of the four rankings analyzed for comparison.

There is not a clear evidence in literature that the CSR's impact on financial performance is positive. Rather, as it is widely reviewed in Section 2, results vary depending on the period, the CSR aspects and the measure of financial performance/risk considered. Our proposal is mainly methodological and aims to foster the use of theoretically robust methods, based on latent variable models, to describe the phenomenon and to adequately measure the ability of a firm to comply with CSR standards, which is a latent construct. Since the class of latent variable models is rather large, they can be implemented on other rating databases, with different measurement schemes, leading to aggregated measures that do not depend on a priori subjective choices of weights.

The rest of the paper is organized as follows: Section 2 provides a brief review of the literature on SRI while in Section 3 an introduction of latent variable models is given. In Section 4 a description of the CSR data is provided and ways to aggregate positive and negative ratings of KLD are reviewed. The Item Response model is described in Section 5 while in Section 6 a qualitative discussion of the main features of the proposed measure is given. Section 7 details the evolution over time as results from the model estimates. In Section 8 results on low and high ranked portfolios are illustrated and in Section 9 some conclusions are presented.

2 Socially Responsible Investments

The main concern of both Socially Responsible (SR) investors and conventional ones is financial profitability. Therefore the issue of whether there is a positive, negative or neutral association between CSR practise and financial performance is a crucial question. CSR implies costs and benefits for a company, and their combination affects the stock price. However the costs/benefits ratio is not the only factor to consider. For example, values-based investors, who accept a lower performance to obtain non-financial utility, may affect the stock price negatively, by excluding from their investment set the stocks of not SR companies. Therefore, behavior of investors has to be taken into account.

A positive association may exist whenever benefits outweigh the costs, by creating intangible capital or, viceversa, not performing Socially Responsible (SR) actions may reduce the overall productivity of the firm. Under this scenario, if the investors underestimate (overestimate) the CSR benefits (costs), then the risk-adjusted returns of the stocks that are high ranked in CSR will be higher than those of the low ranked stocks. Along this line is the *errors-in-expectations hypothesis* (see Derwall, Koedijk, and Horst, 2011), according to which a slow reaction to recognize the impact of CSR practices on future cash flows explains why SR stocks have higher risk-adjusted returns.

Empirical studies who find a positive relation are, among others, Becchetti and Ciciretti (2011) who find that companies with good ratings in Product quality or in Governance reacted better to the Lehman Brothers' default, and Nofsinger and Varma (2013), who document that SRI funds outperformed conventional funds during the global financial crisis and, as such, might be considered as an insurance against the crisis. The Employment conditions as well as the Community relations have been found to have a positive relationship with stock returns by Kempf and Osthoff (2007), and Statman and Glushkov (2009). The Community dimension has been found to be positively associated with higher financial performance also by Brammer, Brooks, and Pavelin (2006) and Manescu (2011). Kempf and Osthoff (2007), and Statman and Glushkov (2009) analyzed also the impact of different combinations of the CSR aspects on financial performance. In particular, Kempf and Osthoff (2007) found that the highest abnormal returns were reached by a portfolio of stocks with the highest average rating of all the considered CSR characteristics and using a best-in-class approach to correct for the industry effects. Statman and Glushkov (2009) constructed portfolios by taking the best and worst companies ranked by an industry-adjusted score. They found that the superior performance of stocks with the higher CSR scores is particular evident in the long-short portfolio of the "top-overall" and "bottom-overall" companies, where a "top(bottom)-overall" company is one in the top(bottom) third of companies by two or more CSR characteristics and not in the bottom(top) third by any others.

A negative association is bound to be there whenever the costs of company's actions to comply with CSR standards outweigh the profits or, vice versa, if the savings from not performing SR behavior are larger than the implied losses, and the investors overestimate(underestimate) benefits(costs). Under this scenario, the low ranked stocks outperform the high ranked ones in terms of risk adjusted returns. As noticed in Manescu (2011), a negative association may be consistent also with an additional risk premium that the low ranked stocks carry as they are exposed to a non-sustainability risk factor that accounts for environmental social or litigation risk. However the author didn't find any strong evidence that the difference in risk adjusted returns can be due to compensation for risk. Another mechanism that may induce low ranked stocks to have higher performance is described in Heinkel, Kraus, and Zechner (2001). They propose an equilibrium model that shows how the exclusion of polluting companies from the investment set may push down their price, thereby producing higher expected returns, as soon as the percentage of SR investors is sufficiently high.

Empirical evidence of a negative association is in Hong and Kacperczyk (2009), who find that the stocks of companies involved in controversial activities have higher performance, and in Becchetti and Ciciretti (2009), who find that SR stocks are less remunerative but also less risky than the others. Further findings of a negative association between environmental or employment indicators and financial returns are in Brammer et al. (2006). A weak negative effect of Human rights and Product quality on financial performance was observed in the more recent period by Manescu (2011).

Finally, a neutral association may arise either if SR actions imply no additional costs and benefits, or they do imply costs and benefits but their effects compensate each other. Most studies compared the performances of ethical and conventional mutual funds and found no statistically significant differences between their risk-adjusted returns. Among them, we cite Hamilton, Jo, and Statman (1993), Statman (2000), and Bauer et al. (2005) where also evidence is provided that ethical funds tend to be less exposed to market variability of returns and more growth-oriented than conventional ones.

3 Latent variable models

In many educational and psychological measurement situations there is an underlying variable of interest. This variable is often something that is intuitively understood, such as intelligence or attitude, but it cannot be measured directly as can height or weight, for example, since it is a concept rather than a physical dimension. This is what psychometricians refer to as an unobservable, or latent, trait. Such a variable is easily described via its attributes, which altogether constitute partial and imperfect measurements. A review on latent variable models can be found in Skrondal and Rabe-Hesket (2004).

A primary goal of educational and psychological measurement is the determination of how much of such a latent trait a person possesses. Since most of the research has dealt with variables such as scholastic, mathematical or language skills, the generic term “ability” is used in this context. For this purpose, it is necessary to have a scale of measurement, a ruler having a given metric. This can be for example a set of questions, or items, with binary answers taking value 1 if the answer is correct and 0 otherwise. Models to deal with such a data are called Item Response models. Later extensions, that we used here, include the possibility to deal with data that are categorical ordered responses.

Items may possess different capacity to discriminate among people. For example, all subjects tend to give correct answers to a trivial item or wrong answers to a difficult one. In both cases, this item is not discriminant and therefore the ability of a subject has to be less influenced by the answers to that item. Furthermore, different background variables, such as gender or social class, have to be taken into consideration, as their effect may be in the direction of making an item more complex to some subjects than for others. If so, a correct answer provided by a subject of one group may require a higher level of ability than a correct answer provided by a subject of another group. It then follows that (a) the responses are not permutable and (b) the same pattern of responses has a different influence on the ability depending on the personal features of the respondent.

An objective way to extract the ability of subjects has to give different weights to each item according to its capacity to discriminate among subjects and has to take the effect of background variables into account. In our context, we want to extract the ability of a firm to fulfill sustainability standards. Corporate Social Responsibility has a complex structure, usually measured by several dimensions. If we substitute a generic person with a generic firm, and we consider each dimension as an item of a measurement model, after introducing covariates to take into account the effects of industries, latent variable models can be successfully applied to measure the socially responsible performance of a firm.

4 Sustainability scores

4.1 The KLD dataset

KLD Research and Analytics[2] provides the longest time series of ESG information. From 1991 to 2000, KLD covered approximately 650 companies belonging to the Domini 400 Social index[3] and/or to the S&P500 index. In 2001 KLD expanded its coverage to include the largest 1000 US companies by market capitalization. Since 2003 KLD has provided ratings for the largest 3000 US firms. KLD has used the names and, since 1995, the CUSIP codes in order to identify the companies.

Companies are rated according to seven dimensions: Governance, Community, Diversity, Employee relations, Environment, Human rights, Product quality. Scores were assigned on the basis of the company's corporate social responsibility reports and public information, and after a direct engagement. KLD released ratings yearly. The scores on the performances of a certain year were published at the beginning of the following year. The focus here is on members of either the S&P500 index or the MSCI KLD 400 Social index[4].

[2] KLD Research and Analytics was acquired by RiskMetrics at the end of 2009. We here focus on data released prior to the merging.

[3] The Domini 400 Social index is now called MSCI KLD 400 Social index.

[4] Financial data were downloaded from Datastream that uses ISIN codes to identify the companies. Matching the financial data with the KLD data, identified through the names

For each dimension, KLD considers different qualitative binary indicators taking values 0 or 1. There are two types of indicators: "strength" and "concern". A score 1 in a strength is "positive", meaning that the company has a proactive behavior in complying with the standards; on the other hand, a score 1 in a concern indicator has to be considered as "negative", indicating a weakness of the company to satisfy the standards. Ambiguity arises as a 0 in one indicator can result from either neutral performance or lack of rating. However, as noticed by Manescu (2011), membership of S&P500 index or MSCI KLD 400 Social index minimizes the presence of companies for which a score 0 is due to the latter. In addition, KLD also provides negative ratings on controversial business issues such as Alcohol, Gambling, Firearms, Military, Nuclear Power and Tobacco, for negative screening that consists in excluding from the investment set the stocks of companies raising a concern in one of those items. As the present analysis is not focused on negative screening, association with controversial activities is not considered. For a complete description of the indicators accounted by KLD see Becchetti and Ciciretti (2009).

Table 1 shows some descriptive statistics of the KLD dataset for years 1991 and 2007, corresponding to the first and the last year in our dataset. For each dimension, the number of indicators as well as the maximum, the mean and standard deviation of the distribution of the sum of the corresponding binary indicators are reported for strengths and concerns separately. The total number of indicators has varied over the years, from 30 strengths and 24 concerns in 1991 to 40 strengths and 34 concerns in 2007. Also, the number of strengths and concerns is different from one aspect to the other. These features hinder comparisons across years and across different CSR dimensions. Moreover, distributions are mainly centered around the lowest values with small tails. In Figure 1, the histograms of Environment and Corporate Governance strengths and concerns are presented, that show skewness of the univariate distributions. This pattern repeats in most of the data and gives rise to sparsity of the contingency table obtained, for each year, by the cross-classifications of companies according to all dimensions.

4.2 Aggregating strengths and concerns

Different ways to aggregate strengths and concerns to form the firm's overall score have been presented in the literature and here we review them, as they will form the benchmark for our proposed measure. An extensive commentary is in Manescu (2011). For each year and for each company p in the investment set, let s_{ik}^{p} and c_{ik}^{p} be the k-th indicator of, in order, strengths and concerns in dimension $i = 1, \dots, 7$. Then the simplest method (see e.g. Becchetti and Ciciretti, 2011 or Statman and Glushkov, 2009) is to take the

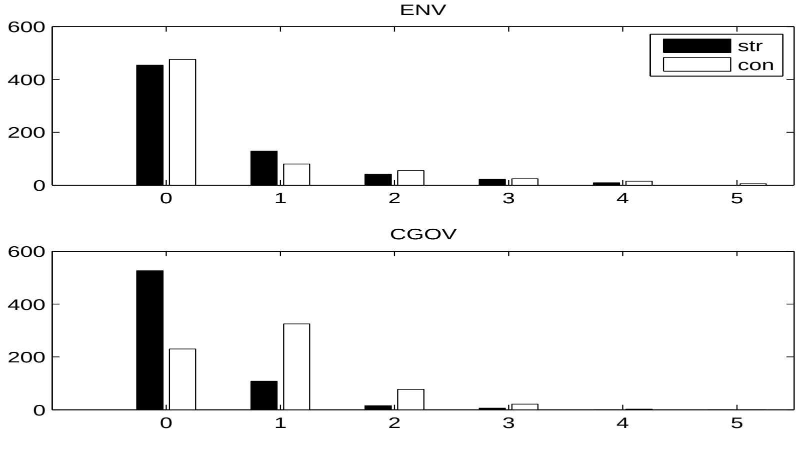
or the CUSIP codes, produced a loss of about 5% in the total market capitalization of our investment universe as some companies of the KLD dataset were not identifiable.

Table 1: Descriptive Statistics for KLD dataset

	strengths				concerns			
	#indicators	max	mean	sd	#indicators	max	mean	sd
Dataset 2007								
COM	7	5	0.41	0.81	4	3	0.23	0.47
CGOV	6	3	0.24	0.53	7	4	0.84	0.77
DIV	8	7	1.50	1.51	3	2	0.26	0.48
EMP	6	5	0.71	0.89	5	4	0.80	0.88
ENV	6	4	0.48	0.86	7	5	0.53	1.03
HUM	3	1	0.02	0.14	4	3	0.14	0.40
PRO	4	2	0.15	0.38	4	4	0.62	0.89
Dataset 1991								
COM	4	3	0.36	0.62	4	1	0.04	0.20
CGOV	3	1	0.02	0.13	2	1	0.04	0.20
DIV	7	3	0.26	0.57	2	1	0.03	0.17
EMP	6	3	0.24	0.49	4	2	0.12	0.36
ENV	6	2	0.24	0.47	6	5	0.37	0.78
HUM	-	-	-	-	2	2	0.15	0.40
PRO	4	2	0.15	0.38	4	2	0.20	0.45

The table shows some descriptive statistics of the KLD dataset in years 2007 and 1991. For any given dimension, the number of indicators as well as the maximum, the mean and standard deviation of the distribution of the sum of the corresponding binary indicators are reported for strengths and concerns separately.

Figure 1: KLD histograms (year 2007)



Histograms of the strengths and concerns for the Environment and Corporate Governance dimensions in year 2007.

sum of the strengths net the sum of concerns:

$$ESG^1_{pi} = \sum_{k=1}^{n_i} s^k_{pi} - \sum_{k=1}^{m_i} c^k_{pi} \tag{1}$$

where n_i and m_i are, in order, the number of strength and concern indicators for dimension i in the year considered. Since these numbers vary from one year to another and from one dimension to another, the first limitation of this method is that it prevents comparisons across years and across dimensions. The other criticism is related to the industry effect: if in a particular industry a given concern is not possible, then firms in that industry would receive inflated scores. The distortion induced by the industry effect can be counterbalanced by a best-in-class approach, that in fact is applied both by Kempf and Osthoff (2007) and Manescu (2011).

The first drawback of the previous measure is avoided by the measure in Kempf and Osthoff (2007). The authors propose first to convert the weaknesses into strengths by taking the complements to 1, i.e. by building $\bar{c}^k_{pi} = 1 - c^k_{pi}$. By doing so, for a particular firm p: if a weakness k is present than $\bar{c}^k_{pi} = 0$ and it becomes 1 otherwise. Then, they form the sum of the overall indicators, divided by the total number of the indicators:

$$ESG^2_{pi} = \frac{\sum_{k=1}^{n_i} s^k_{pi} + \sum_{k=1}^{m_i} \bar{c}^k_{pi}}{n_i + m_i} \tag{2}$$

which ranges from 0 to 1. The industry effect, however, still remains and again, without adjustment, some firms would receive inflated scores, depending on their industry. Another criticism is that the same weight is given to "not having a weakness", that in its essence is passive behavior, and to "having a strength", which in its essence is a proactive behavior.

Another measure has been proposed by Manescu (2011) and used in Herzel, Nicolosi, and Stărică (2012). This is the difference between the average value over the strengths and the average value over the concerns:

$$ESG^3_{pi} = \frac{\sum_{k=1}^{n_i} s^k_{pi}}{n_i} - \frac{\sum_{k=1}^{m_i} c^k_{pi}}{m_i}, \tag{3}$$

that ranges from -1 to +1. Averaging makes the measure comparable across dimensions and years, but it is prone to another criticism: firms with a high absolute number of strengths, which is however relatively small to the total n_i, and with a low number of weaknesses, relatively high to the total m_i, may get a negative score. This may happen when the numbers m_i and n_i are far apart.

For any of these measures, the best-in-class adjusted score of a firm p in a given industry I is defined as

$$\widehat{ESG}^j_{pi} = ESG^j_{pi} - < ESG^j_i >_I\ , \tag{4}$$

and its overall score is then formed by averaging the best-in-class adjusted scores $\widehat{ESG}_{pi}^{j}$ along the seven dimensions i:

$$\widehat{ESG}_{p}^{j} = \frac{1}{7}\sum_{i}^{7} \widehat{ESG}_{pi}^{j}, \tag{5}$$

where $j = 1, 2, 3$ labels the method used and $< ESG_i^j >_I$ is the average score in item i for stocks in the industry I.

Another proposal to rank companies without aggregating their ratings is described in Statman and Glushkov (2009). In their analysis, the authors do not aggregate over CSR dimensions, but define a top-overall company as one that is in the top third of companies for at least two dimensions but not in the bottom third by any dimension according to $\widehat{ESG}_{pi}^{1}$. In an analogous way, a bottom-overall company is one that is in the bottom third of companies by two or more characteristics, but not in the top third by any others.

4.3 Categorical ordinal responses

Item Response model requires as input categorically ordered responses[5], that will be denoted by Y_{pi} where p labels the company and i the KLD characteristic. Although such variables can be constructed using any of the three measures described in Section 4.2, we choose ESG^3 defined in Equation (3), as comparability across dimensions and years is needed and as the negative meaning of concerns is retained. Moreover, in our sample, firms with a negative value when the total strengths is greater then the total concerns are extremely rare (about 2% of the cases).

The contingency table obtained by the cross-classifications of companies according to ESG^3 in all dimensions i is sparse.[6] Sparsity reduces the efficiency and reliability of the estimates and sometimes leads even to a non-identifiability of the parameters. Moreover, specification of the model requires that each item ranges over the same categories. For these reasons it is necessary to reduce the dimensionality, hence we argued as follows. Firms with negative ESG^3 value on item i have raised relatively more concerns than strengths on that item, and this information is more important than the value of ESG^3 itself, so we decided to group them together. The same happens for firms with positive ESG^3 value. Firms with zero ESG^3 value are considered neutral in terms of their performance on item i. We decided therefore to form a three category response variable Y_{pi}, as follows:

[5] Extension to unordered responses, not used here, exist, see De Boeck and Wilson, 2004, Ch. 3.

[6] As already explained in Section 4.1, sparsity of joint distribution is a feature of the KLD dataset and it is not due to the particular aggregation scheme.

(i) $Y_{pi} = -1$ if $ESG^3_{pi} < 0$

(ii) $Y_{pi} = 0$ if $ESG^3_{pi} = 0$

(iii) $Y_{pi} = 1$ if $ESG^3_{pi} > 0$

This is not the unique choice we could adopt, and other strategies that led to an ordinal response with more categories have been considered. However, the contingency table of the data was still sparse for many of the years considered, and computational problems were encountered when estimating the model. The solution here proposed turned out to be the best, as the model performed well in all years considered. The solution here proposed accounts only part of the information given by the dataset but it earns in efficiency of the estimates.

5 Polytomous Item Response Models

We here introduce the class of latent variable models used for our purposes, i.e. the polytomous Item Response models (see De Boeck and Wilson, 2004). The starting point is the set of categorically ordered responses Y_{pi} built as in the previous section. For each firm p, we modeled these responses as expressions of a latent dimension $\tilde{\eta}_p$, measuring the firm's social responsibility. The variable $\tilde{\eta}_p$ is supposed to be a stochastic variable whose prior distribution is normal with zero mean and an unknown variance.

Item Response models give an expression of the probability of an individual p to have a score Y_{pi} in the item i not lower than category $u, u \in \{-1, 0, 1\}$ conditionally to his/her latent ability $\tilde{\eta}_p$. Mathematically this is

$$P(Y_{pi} \geq u \mid \tilde{\eta}_p) = G(f(\tilde{\eta}_p)), \tag{6}$$

where G^{-1} is usually referred to as the link function, connecting the categorical inputs Y_{pi} to the continuous distributed latent variable $\tilde{\eta}_p$. In this work, the Probit, that is the inverse of the standard normal cumulative distribution function, is used for the link function.

Function f is a linear function of $\tilde{\eta}_p$ that has to be specified in order to set the model. This function accounts for the capability of discrimination of an item and for the difficulty that companies in different industries may have to comply with different aspects of CSR. According to the Industrial Classification Benchmark (ICB), the following industries have been considered: Basic Materials, Consumer Goods, Consumer Services, Financials, Healthcare, Industrials, Oil&Gas, Technology, Telecommunications, Utilities.

The following specification, corresponding to the Rasch model, is used for f:

$$f(\tilde{\eta}_p) = k_u + b_{iu} + \sum_r \beta_r D^r_p + \tilde{\eta}_p. \tag{7}$$

Coefficients k_u are thresholds that differentiate between categories $u, u \in \{-1, 0, 1\}$. Such thresholds are modified by the b_{iu} parameters, depending both on item i and category u. Notice that $k_{-1} = \infty$ and $b_{i,-1} = 0$ by construction, as $P(Y_{pi} \geq -1 \mid \tilde{\eta}_p) = 1$. The binary indicator D_p^r takes value 1 if firm p belongs to industry r, where label r runs over the industries. Therefore β_r captures the industry effect. High values indicate that firms in industry r have higher difficulty in getting high scores than the others. We define the overall difficulty for companies in industry r to have a score higher than u in item i as the cutoff value $\Lambda_{i,r,u}$ of the ability $\tilde{\eta}$ for which the corresponding probability (6) is higher than 0.5:

$$\Lambda_{i,r,u} = -\left(k_u + b_{i,u} + \beta_r\right) \qquad u = 0, 1. \tag{8}$$

The cutoff relative to the single category u is not really important in itself. What is relevant is the difference:

$$\Delta\Lambda_i = \Lambda_{i,r,1} - \Lambda_{i,r,0} = k_0 + b_{i,0} - (k_1 + b_{i,1}), \tag{9}$$

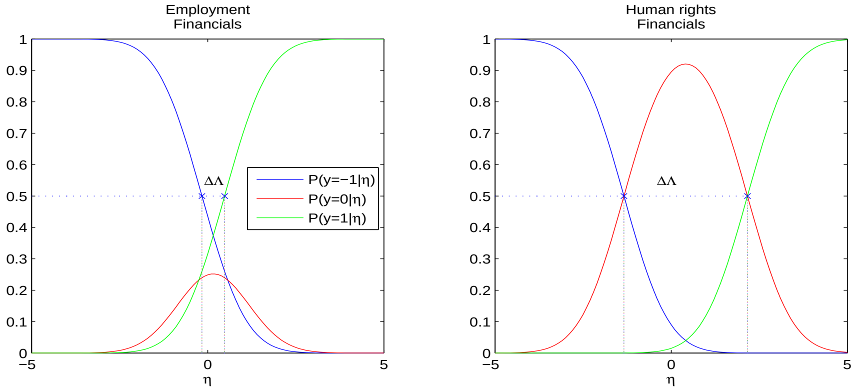
that may be interpreted as the capability of discrimination of an item i. More discriminant items have higher values of this difference: the more the difference, the higher the probability that an individual with a large (small) value of the latent ability has a high (low) score. We call such a difference the "discriminatory coefficient". Let us note that in a dichotomous model the only way to allow for different discrimination power of the items is to add an item-dependent loading factor in front of the latent variable. As a matter of fact, a large loading in a given item means that the further apart individuals fall on the latent dimension the greater their differences in giving a positive response to that item. In that case the discrimination power of an item increases if the steepness of the probabilities defined in Equation (6) for that item increases. Differently, in a polytomous model, the discrimination of an item may vary also if the distance of the probabilities of getting different scores in that item varies, even though its loading is equal to 1.

The discrimination coefficient defined in Equation (9) does not depend on the industry, as the model disentangles the item effect from the industry effect, that instead is captured by β_r. A more sophisticated model (with different specification of function f), not supported by our data, could account also for a mixed term that makes the discriminatory coefficient of the single item depend on the industry.[7]

The model is estimated by maximization of the marginal likelihood. Identification constraints impose that, for industry Consumer Services, $\beta_r =$

[7] Statistical measures based on AIC criteria show that this additional feature is unnecessary for the data at hand. For example in 2005 the AIC for the richer model is 8580 to be compared with the AIC of the proposed model that is 7982. Results are similar also for the other years.

Figure 2: Estimated probabilities for different CSR characteristics

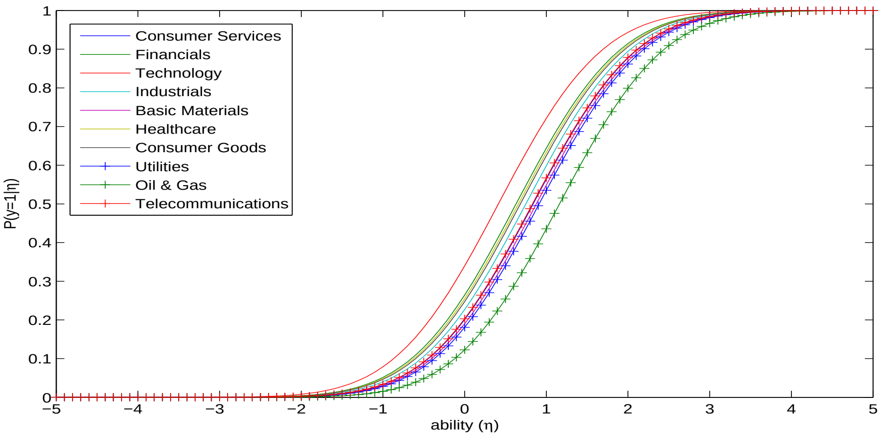


Estimated probability of a score $u, u \in \{-1, 0, 1\}$ as a function of the ability for item (left panel) Employment and (right panel) Human rights for companies in industry Financials (year 2007)

0 and, for item Community, $b_{iu} = 0$ for all categories u. Once the model is estimated, following the Bayes' formula, the posterior distribution of the latent variable $\tilde{\eta}_p$ is computed. The expected value of this posterior distribution is taken as the CSR ability for company p, that we indicate with η_p. This is a one dimensional variable that provides a synthesis of the multidimensional nature of CSR behavior. It is built in a way that takes the industry effects and the different discriminant power of each item into account. The largest contribution to the ability is given from the most discriminant items. Moreover, given a set of scores, a company will have a higher ability if it belongs to an industry r with a higher parameter β_r.

To give a hint of the implications of the model, in Figure 2 the estimated probability of a score u with respect to the ability for item Employment (left panel) and Human rights (right panel) for companies in the industry Financials in year 2007 is presented. The first one is a non discriminating item with a small value of $\Delta\Lambda$, as for a wide range of ability around zero there is a rather substantial probability of each category u; on the contrary the second one is a discriminating item, with a higher value of $\Delta\Lambda$, as low values of the ability give a high probability of $u = -1$, values around zero of the ability give a high probability of $u = 0$ and high values of the ability give a high probability of $u = 1$. Looking at the effect of industries, in Figure 3 the probability of a response $Y = 1$ in item Environment for different industries for year 2007 is reported as a function of the ability. Curves are shifted according to the β_r coefficients: companies belonging to industries with probability shifted on the right need a high level of ability to have $Y = 1$ on Environment. The figure shows that, in that year firms in Oil&Gas have the greatest difficulty to present such a score, while firms in industry Technology have the lowest difficulty. Since the industry effect is disentangled from the item effect, this pattern repeats over other items.

Figure 3: Estimated Probabilities for different industries



Estimated probability of a score $Y = 1$ in item Environment for different industries (year 2007), as a function of the ability

Furthermore, since β_r does not depend on categories, it repeats also over categories.

6 A discussion on the ability measure

To better understand the properties of the η measure, Table 2 (upper panel) reports the number of strengths and concerns and the measures $\widehat{ESG}^1$, $\widehat{ESG}^2$, and $\widehat{ESG}^3$ in each single CSR characteristics of International Business Machines Corp. in 2004, a company belonging to industry Technology. The averages of the $\widehat{ESG}$ scores give the aggregated measures according to the different methods. The first method gives an aggregated score of 0.44 while the second and third measures rate the stock 0.04 and 0.03 respectively. All these three rankings pose the stock among the top 25% of stocks in the same industry. The last column shows the ordinal responses used as input of the Item Response model and the relative estimated value for η, that is 0.06, in the third quartile of the industry distribution. The η measure is not a linear combination of the ordinal responses. Rather, it is a non linear function of the single scores that is most influenced by high discriminant dimensions that, in 2004, according to the model estimates are Human rights, Community, Environment and Product quality. Moreover the way the scores are aggregated depends on the industry. This is the main difference of the η-ranking from the other ranking proposed, where the best-in-class industry adjusted scores for each item are averaged with the same weights, independently from the industry.

To fully comprehend how these two features are incorporated by the

Table 2: The role of different patterns of scores for ranking

International Business Machines Corporation, year 2004, Technology

	str	*con*	$\widehat{ESG}^1$	$\widehat{ESG}^2$	$\widehat{ESG}^3$	*ordinal*
COM	2	1	0.77	0.08	0.05	1
CGOV	1	2	-0.22	-0.03	-0.15	-1
DIV	5	0	3.71	0.34	0.52	1
EMP	3	2	0.62	0.06	0.06	1
ENV	0	2	-1.99	0.17	-0.29	-1
HUM	0	1	-0.74	-0.11	-0.19	-1
PRO	1	0	0.90	0.11	0.23	1
aggregated measure			*ranking*-1	*ranking*-2	*ranking*-3	η-*ranking*
			0.44	0.04	0.03	0.06
quartile			IV	IV	IV	III

Simulation 1: permutating the scores, same industry (Technology)

	str	*con*	$\widehat{ESG}^1$	$\widehat{ESG}^2$	$\widehat{ESG}^3$	*ordinal*
COM	1	2	-0.22	-0.03	-0.15	-1
CGOV	2	1	0.77	0.08	0.05	1
DIV	5	0	3.71	0.34	0.52	1
EMP	3	2	0.62	0.06	0.06	1
ENV	0	2	-1.99	0.17	-0.29	-1
HUM	0	1	-0.74	-0.11	-0.19	-1
PRO	1	0	0.90	0.11	0.23	1
aggregated measure			*ranking*-1	*ranking*-2	*ranking*-3	η-*ranking*
			0.44	0.04	0.03	0.03
quartile			IV	IV	IV	III

Simulation 2: same scores, different industry (Oil&Gas)

	str	*con*	$\widehat{ESG}^1$	$\widehat{ESG}^2$	$\widehat{ESG}^3$	*ordinal*
COM	2	1	1.03	0.10	0.12	1
CGOV	1	2	-0.33	-0.05	-0.18	-1
DIV	5	0	4.79	0.44	0.66	1
EMP	3	2	1.15	0.10	0.16	1
ENV	0	2	-0.64	-0.05	-0.11	-1
HUM	0	1	-0.55	-0.08	-0.14	-1
PRO	1	0	1.24	0.16	0.31	1
aggregated measure			*ranking*-1	*ranking*-2	*ranking*-3	η-*ranking*
			0.96	0.09	0.12	0.19
quartile			IV	IV	IV	IV

The table shows the number of strengths and concerns, and the measures $\widehat{ESG}^1$, $\widehat{ESG}^2$, and $\widehat{ESG}^3$ of the single CSR characteristics, as well as the aggregated measures for International Business Machines Corp. in 2004 (upper panel) in Technology and two hypothetical cases (lower panels). Ordinal responses are reported in the last column, together with the ability measure η_p. The row "*quartile*" indicates the quartile within the industry to which the company belongs, according to each ranking. Intermediate panel shows $\widehat{ESG}$ measures for a hypothetical company in the same industry but with permutated scores. Lowest panel shows the case of another hypothetical company with the same scores but in industry Oil&Gas.

ability measure, the intermediate and lower panels show the $\widehat{ESG}$ measures, the ordinal responses and η for two hypothetical companies: the first one in the same industry as International Business Machines Corp. but with permutated scores (intermediate panel); the second with the same scores but in a different industry, i.e. Oil&Gas (lower panel). Their ability is calculated using the estimated coefficients.

In the first case, the strengths and concerns relative to Community are permuted with those of Corporate Governance while all the other scores are unchanged. Straightforward averages give the same aggregated ratings as for International Business Machines Corp., because the averages do not depend on the patterns. On the other hand, the value of η decreases to 0.03 because the hypothetical company has now a negative score in Community that is a highly discriminant dimension, while has a positive score in Corporate Governance that has a low discriminant power. The influence of industry is highlighted by the lower panel. The fact that the hypothetical company belongs to a more difficult industry, Oil&Gas, makes the ability necessary to have the same partial ratings increasing to 0.19. The result is in the forth quartile of the Oil&Gas distribution.Notice that the ability measure does not depend on any subjective choice of the relative importance to assign to the different CSR characteristics.

Table 3 shows the average (over the period 1991-2007) number of stocks per industry in the top 25% (upper panel) and bottom 25% (lower panel) according to the η measure as well as the other three measures $\widehat{ESG}^j$ with $j = 1, 2, 3$. The average numbers of stocks that are high(low) ranked for both the *η-ranking* and one of the other three measures are also shown. In the last two columns the table reports the average number of stocks in the top(bottom)-overall sets defined in Statman and Glushkov (2009) and their intersection with the high(low) ranked sets according to η. Most of the stocks that are high(low) ranked according to η are also high(low) ranked according to the other measures. For example, among the 12 stocks in industry Basic Materials that are high ranked according to η, 11 out of them are also high ranked according to $\widehat{ESG}^1$, that instead counts 16 stocks. Considering the same industry, only 6 stocks are in the top-overall set and 5 of them are also high ranked according to η. The same behavior is noticed for the other industries. This means that the top(bottom)-overall procedure selects a fewer number of stocks, and most of them are in the top(bottom) 25% of the η distribution.

We finally remark that the number of high(low) ranked companies can be different according to the different measures considered. This is due to the fact that for discrete distributions quantiles are not well defined. This problem is overcome by using the η measure that instead has a continuous distribution.

Table 3: Portfolios' size per industry

Average number of high ranked stocks according to

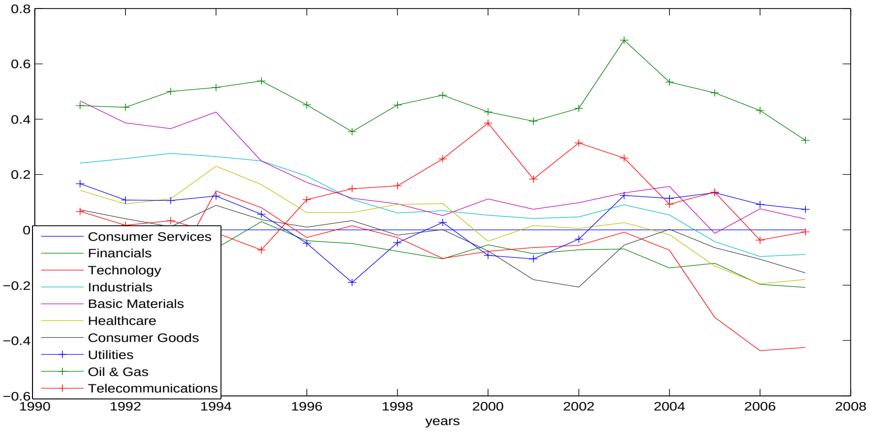
	η	$\widehat{ESG}^1$	η and $\widehat{ESG}^1$	$\widehat{ESG}^2$	η and $\widehat{ESG}^2$	$\widehat{ESG}^3$	η and $\widehat{ESG}^3$	top overall	η and top overall
Basic Mat.	12	16	11	12	10	12	10	6	5
Cons. Goods	22	26	19	23	17	22	18	14	13
Cons. Services	25	33	22	25	20	25	22	17	15
Financials	21	28	18	21	15	22	16	10	8
Healthcare	12	15	10	12	9	13	9	7	5
Industrials	30	39	29	30	25	31	26	18	17
Oil&Gas	8	10	7	8	7	8	7	3	3
Technology	15	18	12	15	12	15	12	11	9
Telecom.	4	4	3	4	3	4	3	1	1
Utilities	12	16	11	12	10	13	10	7	6

Average number of low ranked stocks according to

	η	$\widehat{ESG}^1$	η and $\widehat{ESG}^1$	$\widehat{ESG}^2$	η and $\widehat{ESG}^2$	$\widehat{ESG}^3$	η and $\widehat{ESG}^3$	bottom overall	η and bottom overall
Basic Mat.	12	15	11	12	10	13	10	7	6
Cons. Goods	22	28	20	22	19	22	18	10	9
Cons. Services	25	34	23	27	22	27	21	11	10
Financials	21	27	18	22	18	22	18	8	7
Healthcare	12	19	11	13	10	13	10	5	4
Industrials	30	37	28	31	26	31	25	16	15
Oil&Gas	8	10	7	8	6	8	7	5	4
Technology	15	20	13	16	12	16	13	4	4
Telecom.	4	5	3	4	3	4	3	2	2
Utilities	12	16	11	13	10	12	10	7	6

Average (over the period 1991-2007) number of stocks per industry in the top 25% (upper panel) and bottom 25% (lower panel) of the stocks according to the η measure as well as the other 3 measures $\widehat{ESG}^j$ with $j = 1, 2, 3$. The average numbers of stocks that are high(low) ranked for both the *η-ranking* and one of the other 3 measures are also shown. The last two columns show the average number of stocks in the top(bottom)-overall sets defined in Statman and Glushkov (2009) and their intersection with the high(low) ranked sets according to η.

Figure 4: Evolution over time of the industry effects



Evolution over time of the industry effects as measured by β_r coefficients, where r labels the industries

7 Evolution over time of CSR aspects

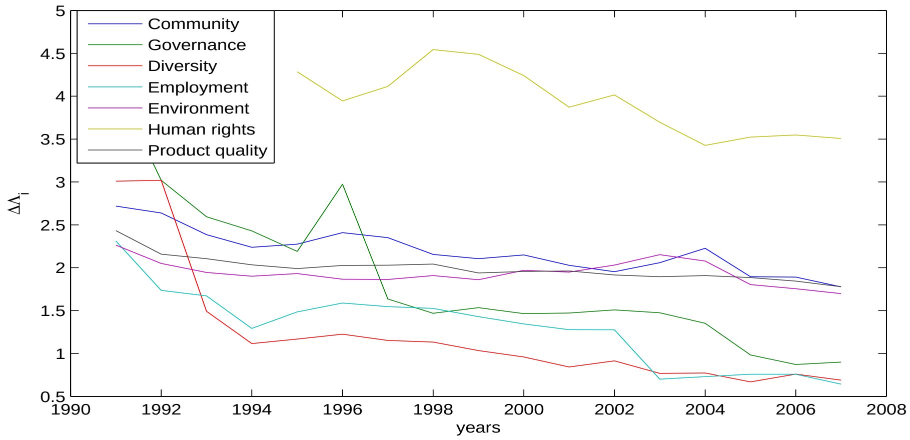
Figure 4 shows the evolution of industry coefficients β_r over time. Industry Oil&Gas consistently presents the highest difficulty to meet CSR standards. Industrials, Basic Materials and Telecommunications show a higher difficulty than the reference industry, Consumer Services, contrary to what exhibited by the industry Financials. Figure 5 shows the evolution over time of the discriminatory coefficients for each item i. The picture exhibits a rather stable behavior, especially after 1997. Human rights, Environment, Community and Product quality are rather discriminant items.

After 1997, Governance was not a very discriminant dimension, as it was very difficult to score 1 but also to score 0, and therefore the CSR ability is not strongly influenced by the Governance. This is in line with most of the literature regarding the effectiveness of the governance indices in predicting corporate performances, see for example Bhagat, Bolton, and Romano (2008). For almost all years, Employment and Diversity were the easiest items both for score 0 and 1 and they were not discriminant. This implies that the ability extracted from our model is less influenced by these items.

8 Socially responsible stock selection

One of the most straightforward application of rankings is stock selection. Rankings are formed according to the η measure (η-*ranking*), as well as according to $\widehat{ESG}^j$, $j = 1, 2, 3$ (*ranking-j*). Moreover, the top-overall, bottom-overall ordering of Statman and Glushkov (2009) is also considered

Figure 5: Evolution over time of the discriminatory power of different items



Evolution over time of the discriminatory coefficients $\Delta\Lambda_i$ for all the items i

and it will be referred to as the "*ranking*-4". On the basis of such rankings, high and low ranked portfolios are formed, as explained in Section 8.1. The aim of this section is to investigate the relationship between the portfolios' CSR performance, as measured by the η-*ranking*, as well as by the other rankings considered for comparison, and the portfolios' financial performance, as measured by the Jensen's α in the Carhart's model (Carhart, 1997).

8.1 Portfolio construction

Best(worst)-in-class portfolios are formed by taking the top(bottom) 25% of companies in each industry for all rankings but *ranking*-4, for which portfolios are formed by companies in the top(bottom)-overall sets. This guarantees that portfolios are well diversified across industries. Together with the High (H) and Low (L) ranked portfolios, the High minus Low ($H-L$) difference portfolios are also considered.

Concerning the portfolio weighting scheme, Equally Weighted (EW), Value Weighted (VW) and mean-variance optimal portfolios are formed. In the first and second scheme, portfolio weights are re-balanced at the beginning of each month, while, for computational reasons, in the latter case they are re-balanced at the beginning of each year. Notice that while the EW and VW portfolios do not depend on any financial data, the optimal portfolios mix the ex-ante mean-variance optimality with the CSR information. An optimal portfolio, in the mean-variance sense, is a portfolio that, given a certain level of expected return R, minimizes its variance keeping an expected return at least equal to R. A no short-selling constraint is imposed in the optimization. Input data for optimal portfolio allocations, namely the vector of expected returns μ and the covariance matrix of the returns Σ, were estimated similarly to Herzel et al. (2012), using the Carhart model (Carhart, 1997) to compute Σ and making a market neutral forecasting assumption for μ. In the construction of the optimal portfolios we set R as the level of the market expected return. That is an intermediate level that makes the optimal portfolios over the different subsets of allocation preserve a certain degree of diversification.

8.2 Carhart's model

To measure the financial performances of portfolios, the Carhart's model (Carhart, 1997) is implemented and the Jensen's α's are extracted.

The Carhart's model takes styles into account, as differences in investment style can be very relevant in explaining the performance differences between ethical and conventional funds (see e.g. Kurtz, 1997, or Guerard, 1997). Such model is used for example by Kempf and Osthoff (2007) or by Statman and Glushkov (2009), who constructed their own indexes, and by Bauer et al. (2005), who compared the performance of ethical and conventional funds in different regions.

The Carhart model explains the portfolio excess return $R_{j,t} - RF_t$ over the risk-free rate RF_t of portfolio j at month t in terms of 4 risk factors:

$$\begin{aligned} R_{j,t} - RF_t = &\alpha_j + \beta_{j,1}\,(R_t^M - RF_t) + \beta_{j,2}\,SMB_t \\ &+ \beta_{j,3}\,HML_t + \beta_{j,4}\,Mom_t + \epsilon_{j,t}. \end{aligned} \tag{10}$$

The factor $R_t^M - RF_t$ is the excess return of the market at time t, SMB_t is the return at time t of the small cap portfolio minus the large cap portfolio, HML_t is the return at time t of the value stocks' portfolio minus the growth stocks' portfolio, and Mom_t is the return at time t corresponding to the momentum factor[8]. The intercept, Jensen's α, gives the portfolio extra return that cannot be obtained through the leverage to the risk factors. The terms $\epsilon_{j,t}$ are the idiosyncratic errors and the β's are the factor loadings of portfolios over the risk factors.

8.3 Portfolio risk-adjusted performances

The results for the Carhart's model are summarized in Table 4 for the EW, VW and optimal portfolios constructed on the basis of the *η-ranking*. The table shows the ordinary least square estimates of the coefficients of the regressions together with their significance at levels 1%, 5% and 10%, indicated respectively with ***, ** and *. The levels of significance are computed taking a Newey-West correction into account. The Jensen's α is annualized and expressed as a percentage. The adjusted R^2 is also shown in the last column. First we comment on the α's. We are mainly interested in comparing

[8] The time series of the risk factors were downloaded from the K. R. French's web site.

Table 4: Carhart's model estimate for Equally Weighted (EW) Value Weighted (VW) and optimal best-in-class portfolios

		$\alpha(\%)$	Mkt	SMB	HML	Mom	R^2_{adj}
EW	$H\eta$	1.78	0.88***	0.19***	0.40***	−0.17***	0.80
	$L\eta$	0.95	1.01***	0.18***	0.46***	−0.28***	0.84
	$H\eta - L\eta$	0.82	−0.13***	0.01	-0.07*	0.10***	0.31
VW	$H\eta$	3.59***	0.80***	-0.07	−0.17***	−0.16***	0.81
	$L\eta$	0.11	0.88***	−0.15***	0.19***	−0.07**	0.79
	$H\eta - L\eta$	3.49***	−0.08**	0.08	−0.36***	−0.08***	0.31
optimal	$H\eta$	2.65**	0.82***	0.03	0.15***	−0.18***	0.81
	$L\eta$	0.66	0.91***	-0.04	0.24***	−0.21***	0.82
	$H\eta - L\eta$	1.99**	−0.09***	0.07*	−0.09***	0.03	0.13

Carhart's model estimate for equally weighted (EW), value weighted (VW) and optimal best in class portfolios from January 1992 to December 2008. The table shows the OLS estimates of the coefficients. The α's are annualized and expressed as a percentage. The significance at the 1%, 5% or 10% level is indicated respectively with ***, ** and *. The p-values, not shown in the table, are computed by taking a Newey-West correction into account. The rows labeled with $H\eta$, $L\eta$ report the results for portfolios constructed respectively over the highest and lowest ranked companies per industry according to the η-ordering. The rows labeled with $H\eta - L\eta$ show the results for the long-short portfolios.

the risk-adjusted performances of high ranked portfolios with the low ranked ones. Therefore we focus on the long-short positions. The α coefficients of high minus low portfolios are positive, and significant in the case of VW and optimal portfolios with respectively a value of 3.49% and 1.99%. The positive performances of the long/short portfolios are mainly driven by the high ranked portfolios that in fact have α's positive, and significant for the VW and optimal weightings, with respectively 3.59% and 2.65%. These results agree with the hypothesis of a positive relationship between the portfolios' CSR performance and their financial performance.

Concerning the other factors in the Carhart's model, we observe from Table 4, that the market and the book to market ratio loadings of the high minus low portfolios are always negative and statistically significant. This means that the high ranked portfolios are usually less exposed to the market factor and are invested more in growth stocks (low book to market ratio) than value stocks (high book to market) compared to the low ranked portfolios. These findings are in line with those in Bauer et al. (2005), in Kempf and Osthoff (2007), and in Statman and Glushkov (2009). Interestingly, we notice that the high ranked EW portfolios are more tilted towards stocks with high momentum than the low ranked ones (in line with results for top-overall minus bottom-overall strategies in Statman and Glushkov, 2009), while the high ranked VW portfolios are more invested in stocks with a negative momentum with respect to the low ranked ones.

To check for robustness over the cut-off, portfolios have been constructed also by considering the median value of the η distribution within each industry as the cutoff. Results are in agreement with those already shown and we do not detail them here[9]. In general, when a coefficient is significantly different from zero for a portfolio built with respect to the median values, then that result (in terms of either p-value or magnitude) is even strengthened in a portfolio built with respect to the quartiles. Moreover, in some cases we observed non significant coefficients of the portfolios formed with respect to the median values that became significant in the other case.

To check for robustness of the findings on α's against the estimation window, the overall period is further divided into two subperiods: January 1992 - December 2000 and January 2001 - December 2008. Table 5 shows the annualized α's estimated over the two subperiods, and their significance. For comparison, Table 5 reports also the results, for the portfolios that are high and low ranked according to the other orderings. The table shows that the *η-ranking* gives robust α's over the different subperiods, but significance may decrease due to the shorter periods considered. Concerning the other rankings, it is remarkable that the risk-adjusted returns, estimated on the whole period, of the long-short strategies based on *ranking*-1, *ranking*-2, and *ranking*-3 are never significant. Moreover, their sign changes from positive in the first subperiod to negative in the second one, even though in the most of the cases results are not significant. *Ranking*-2 corresponds to the aggregation scheme (combination1) in Kempf and Osthoff (2007). Their main findings concern a positive and significant α over the period 1992-2004 of the long-short position for *VW* portfolios with 10% cut-off (further checked with 25% cut-off). The analogous parameter here is not significant. This may be due to the different period and different weighting scheme (weights in their paper are formed at the beginning of each year and kept constant until the end of the year).

Contrary to *ranking*-1, *ranking*-2, and *ranking*-3, and similarly to *η-ranking*, *ranking*-4 provides robust results over the two subperiods. In particular, the risk adjusted returns estimated on the overall period is positive, and significant for *EW* and *VW* weighting schemes. This ranking corresponds to the top-overall bottom-overall approach of Statman and Glushkov (2009), and our results are quite similar to theirs, although on a different investment set.

9 Conclusions

Corporate Social Responsibility has a multidimensional nature. Companies with a policy of responsibility have to comply with standards related to

[9] Those results are available upon request.

Table 5: Comparison between rankings. Jensen's alphas in the Carhart model for EW, VW, and optimal best in class portfolios in different subperiods

	1992-2008			1992-1999			2000-2008		
	EW	VW	optimal	EW	VW	optimal	EW	VW	optimal
$H\eta$	1.78	3.59***	2.65**	2.11*	5.26***	3.34***	2.13	1.48	2.09
$L\eta$	0.95	0.11	0.66	0.96	1.86	1.05	0.60	-1.69	0.36
$H\eta - L\eta$	0.82	3.49***	1.99**	1.15	3.40*	2.30*	1.53	3.16*	1.73
$H1$	2.08	2.56**	2.27**	2.89**	4.05***	3.54***	1.33	-0.18	0.52
$L1$	1.76	1.81	2.38**	1.18	2.04	1.66	2.03	1.63	3.20*
$H1 - L1$	0.32	0.75	-0.12	1.71*	2.01	1.88*	-0.70	-1.81	-2.68
$H2$	1.54	1.93	1.61	2.72**	3.78**	3.21**	0.81	-0.44	0.22
$L2$	1.76	1.81	1.98**	1.24	1.91	1.47	1.66	1.38	2.50
$H2 - L2$	-0.22	0.12	-0.37	1.48	1.87	1.74	-0.85	-1.82	-2.28
$H3$	1.60	2.22*	1.71	2.33*	4.00***	2.95**	1.14	-0.38	0.26
$L3$	1.17	0.58	1.39	0.80	1.18	1.20	1.38	0.05	1.94
$H3 - L3$	0.43	1.63	0.31	1.54	2.82	1.75	-0.24	-0.43	-1.67
$H4$	2.88**	4.34***	2.59*	3.27**	4.82**	3.18*	3.06	2.48	1.57
$L4$	-0.63	-1.82	1.03	-0.60	0.27	-0.55	-1.08	-2.72	2.19
$H4 - L4$	3.51**	6.17***	1.56	3.87**	4.55	3.73**	4.14	5.19*	-0.63

Comparison of the α's for EW, VW, and optimal best in class portfolios constructed according to different rankings. The table shows the OLS estimates of the alphas in the whole period 1992-2008 as well as in the two subperiods 1992-1999 and 2000-2008. The α's are annualized and expressed as a percentage. The significance at the 1%, 5% or 10% level is indicated respectively with ***, ** and *. The p-values, not shown in the table, are computed by taking a Newey-West correction into account. The rows labeled with H and L report the results for portfolios constructed respectively over the highest and lowest ranked companies per industry according to every orderings. The rows labeled with $H\eta - L\eta$ show the results for the long-short portfolios according to the different orderings.

various aspects, that present different challenges. Therefore, when measuring the level of CSR compliance of a firm, different aspects have to be given different weights. Furthermore, Industry membership can also be highly relevant. We implemented an Item Response model that provides a univariate measure of the CSR compliance of a firm, in such a way that the different aspects of CSR are weighted according to their capacity to discriminate socially responsible companies from the others and that the Industry effect is explicitly accounted for. We called such a measure "CSR ability".

The model has been applied to the KLD ratings of companies belonging to the S&P500 index and/or to the KLD 400 Domini Social index, covering the period form 1991 to 2007. Findings indicate that firms in the industry Oil&Gas have the higher difficulty to comply with the CSR standards, followed by the ones in Industrials, Basic Materials and Telecommunications. It means that, given the same pattern of CSR ratings, the model assigns a higher CSR ability to a company in Oil&Gas than to a company in Financials. Human rights, Environment, Community and Product quality are the most discriminant items and therefore they influence the CSR ability more than the other KLD aspects. On the contrary, Governance is not a discriminant item.

According to the CSR ability measure, high and low ranked portfolios were formed and their risk adjusted returns compared. Results show that portfolios of stocks that are high ranked according to the proposed measure outperformed portfolios of low ranked stocks in terms of risk adjusted returns. The results for the long/short strategies are statistically significant for the value weighted and optimal portfolios, and are robust against the estimation period. Let us note that what we actually find is an association between CSR and financial performance. However the analysis performed does not give any information on the causality direction. Reverse causality may also explain results since companies which are better performing may afford corporate social responsibility practices. For comparison reasons, different investment strategies were formed also on the basis of other three CSR aggregated measures appeared in the literature, as well as the top(bottom)-overall approach proposed by Statman and Glushkov (2009). Results on the risk adjusted performance of the long/short position are statistically significant and robust only for the last case.

In conclusion, investigations based on the proposed measure support the hypothesis of a positive relationship between compliance with CSR and financial performances. Weights of the different dimensions to form an aggregated measure of CSR are determined by the Item Response model itself and do not have to be chosen a priori.

Acknowledgements

The authors gratefully acknowledge financial support from the Swedish Foundation for Strategic Environmental Research (MISTRA) through the Sustainable Investment Research Platform (SIRP).

References

Bauer, R., K. Koedijk, and R. Otten (2005). International evidence on ethical mutual fund performance and investment style. *Journal of Banking & Finance 29*, 1751–1767.

Becchetti, L. and R. Ciciretti (2009). Corporate social responsibility and stock market performance. *Applied Financial Economics 19*, 1283–1293.

Becchetti, L. and R. Ciciretti (2011). Stock market reaction to the global financial crisis: testing for the lehman brothers' event. *Giornale degli Economisti e Annali di Economia 70*, 3–58.

Bendheim, C. L., S. Waddock, and S. Graves (1998). Determining best practice in corporate stakeholder relations using data envelopment analysis. *Business Society 37*, 305–338.

Benson, K. L., T. J. Brailsford, and J. E. Humphrey (2006). Do socially responsible fund managers really invest differently? *Journal of Business Ethics 65*, 337–357.

Bhagat, S., B. Bolton, and R. Romano (2008). The promise and peril of corparate governance indices. *Columbia law review 108*, 1803–1882.

Bird, R., A. D. Hall, F. Momente, and F. Reggiani (2007). What corporate social responsibility activities are valued by the market? *Journal of Business Ethics 76*, 189–206.

Brammer, S., C. Brooks, and S. Pavelin (2006). Corporate social performance and stock returns: Uk evidence from disaggregated measures. *Financial Management 35*, 97–116.

Carhart, M. (1997). On persistence in mutual fund performance. *Journal of Finance 52*, 57–82.

Chen, C. M. and M. Delmas (2011). Measuring corporate social performance: an efficiency perspective. *Production and Operations Management 20*, 789–804.

De Boeck, P. and M. Wilson (2004). *Explanatory Item Response Models: A Generalized Linear and Nonlinear Approach.* New York, Springer.

Derwall, J., K. Koedijk, and J. T. Horst (2011). A tale of values-driven and profit-seeking social investors. *Journal of Banking and Finance 35*, 2137–2147.

Guerard, J. (1997). Is there a cost to being socially responsible in investing? *The Journal of Investing 6*, 11–18.

Hamilton, S., H. Jo, and M. Statman (1993). Doing well while doing good? the investment performance of socially responsible mutual funds. *Financial Analysts Journal 49*, 62–66.

Heinkel, R., A. Kraus, and J. Zechner (2001). The effect of green investment on corporate behavior. *Journal of Quantitative Analysis 4*, 421–449.

Herzel, S., M. Nicolosi, and C. Stărică (2012). The cost of sustainability in optimal portfolio decisions. *The European Journal of Finance 18*, 333–349.

Hong, H. and M. Kacperczyk (2009). The price of sin: The effects of social norms on markets. *Journal of Financial Economics 1*, 15–36.

Hopkins, M. (2005). Measurement of corporate social responsibility. *International Journal of Management and Decision Making 6*, 213–231.

Kempf, A. and P. Osthoff (2007). The effect of socially responsible investing on portfolio performance. *European Financial Management 13*, 908–922.

Kurtz, L. (1997). No effect, or no net effects? studies on socially responsible investing. *The Journal of Investing 6*, 37–49.

Manescu, C. (2011). Stock returns in relation to environmental, social and governance performance: Mispricing or compensation for risk? *Sustainable Development 19*, 95–118.

Nofsinger, J. and A. Varma (2013). Socially responsible funds and market crises. *Journal of Banking & Finance http://dx.doi.org/10.1016/j.jbankfin.2013.12.016.*

Rowley, T. and S. Berman (2000). A brand new brand of corporate social performance. *Business Society 39*, 397–418.

Ruf, B. M., K. Muralidhar, and K. Paul (1998). The development of a systematic, aggregate measure of corporate social performance. *Journal of Management 24*, 119–133.

Skrondal, A. and S. Rabe-Hesket (2004). *Generalized latent variable modeling: multilevel, longitudinal, and structural equation models.* Boca Raton, Chapman and Hall/CRC.

Statman, M. (2000). Socially responsible mutual funds. *Financial Analysts Journal 56*, 30–39.

Statman, M. and D. Glushkov (2009). The wages of social responsibility. *Financial Analysts Journal 65*, 33–46.

US-SIF (2012). *2012 Report on Sustainable and Responsible Investing Trends in the United States*.

Waddock, S. and S. Graves (1997). The corporate social performance-financial performance link. *Strategic Management Journal 18*, 303–319.