To agree or disagree? An analysis of CSR ratings firms.

ABSTRACT

With the increasing use of CSR ratings firms' data to guide ethical investing and to derive findings in academic studies, there has been a rise in the number of different ratings firms. As there is little evidence more than one ratings firm is used to draw conclusions (particularly in academic studies), it is important to understand whether ratings are commensurable. This paper assesses the level of agreement between two ratings firms, Bloomberg and CSR Hub across three main CSR sub-categories and an Overall score. It uses Lin's concordance correlation coefficient and intraclass correlation coefficient on continuous ratings, and cross-tabulation and Cohen's kappa on ranked ratings within a sample of 720 US and EU companies.

For both continuous and ranked data, there is most agreement on Employees/Social, Community/Social and Overall categories and weaker agreement on Environmental and Governance categories. Firms in the German DAX are most consistently rated, as are large and medium-sized firms.

These findings propose a degree of caution for investors and academics using only one rater as the basis for their decisions/inferences. Accounting practitioners should be aware their CSR disclosures result in differing ratings and should consider which raters their key investors use.

This paper is original in the comprehensive range of methods used to analyse two ratings firms across all CSR sub-categories, in samples from both the US and EU.

Keywords: Lin's concordance correlation coefficient; Cohen's kappa; intraclass correlation coefficient; interrater agreement; ratings firms.

To agree or disagree? An analysis of CSR ratings firms.

1. Background

Corporate Social Responsibility (CSR) has been the subject of hundreds of research papers since the 1970s. One of the difficulties in such research is to derive a reliable and meaningful proxy for CSR. There are criticisms that corporate intentions expressed in annual reports and press releases do not translate into actions (so-called 'greenwashing') (Lyon and Maxwell 2011) or that there is bias in results (by asking firms their own assessments of their CSR activities) (Chatterji, Levine, & Toffel, 2008). During the 1980s, the first mainstream independent rating firm, Kinder, Lydenberg and Domini (KLD) (KLD Investments, 2014) came to the fore, providing a measure of the multiple elements of CSR (social, environmental and governance) based on publicly available information but quasi-independently from the firms being rated. Since then, a multiplicity of ratings firms (referred to subsequently as 'raters') has emerged, all purporting to assess firms on their CSR activities, to assist investors to construct their portfolios around socially responsible criteria.

Socially responsible investments (SRI) have now become virtually mainstream in much of the international investment community. $22.89 trillion was invested in SRI assets in 2016, representing 26% of professionally managed assets globally; an increase of 35% since 2014 (Global Sustainable Investment Alliance 2016). Decisions to invest or disinvest and academic hypotheses are made on the basis of the CSR ratings provided by the raters (Chatterji et al. 2008, Ioannou & Serafeim 2010). Therefore, it is important to have confidence in the construct validity of the raters' scores, in other words, that they measure what they purport to measure. Consistency of ratings approach over time is also essential to be able to assess any apparent changes in results.

This paper takes two raters, a subscription-only service, Bloomberg (with their ESG (Environmental, Social and Governance) dataset) and CSR Hub, a free internet service, and compares their ratings across a sample of 720 firms listed on international stock exchanges: the US Standard & Poor's 500, the UK FTSE 350, the German DAX, the French CAC40 and the Spanish IBEX. A cumulative assessment of all four combined European stock exchanges versus the US S&P was also evaluated to even out the sample sizes. These particular stock exchanges were selected because they had a large enough number of constituents for a reasonable sample and CSR data was available from both raters on the same firms in each stock exchange. Most studies to date on ratings firms have considered only one stock exchange (for example,

there have been studies on the rater KLD who focuses almost exclusively on US firms) or one aspect of CSR ratings, such as environmental performance (Chatterji et al., 2014; Delmas, Etzion, & Nairn-Birch, 2013). As more academic studies seek to look beyond one individual country and wish to consider the full range of CSR activities, this paper assesses the agreement of two raters on the same firms across multiple countries, and across all three strands of CSR, environmental, social and governance. This study does not address the construct validity of the scores (do the raters actually measure what they purport to measure) or the quality of the disclosures being made by the firms on which the raters reach their scoring. The aim of the paper is to ascertain whether there are significant differences in the way in which raters score the same firms (either in absolute terms but also in terms of ranking) such that the choice of rater might affect decisions taken or research inferences made.

The paper assesses the two raters using statistical approaches for both continuous and categorical data. Since the scores of both raters range from 0-100 on a continuous scale, it is appropriate to use techniques suited to continuous data; the most accepted methods are Lin's concordance correlation coefficient and the intraclass correlation coefficient (Watson and Petrie 2010). However, given the limited likelihood that two individual raters with different methodologies would totally agree on a firm's performance based on such a wide range, the scores were also ranked into tertiles, of high, medium and low-scoring to assess whether, even if the individual 0-100 scores were different, the raters classified the firms in the same category within the peer group; hence the use of two statistical methods designed for use with categorical data, Cohen's kappa and interrater agreement (cross tabs) (Banerjee, Capozzoli, McSweeney, & Sinha, 1999; Tanner & Young, 1985; Watson & Petrie, 2010).

This paper differs from previous studies in its range of different statistical methods to provide a more comprehensive assessment of rater agreement. It considers a much broader scope of firms than other studies, from both the US and EU, from an industry and firm size perspective across all three elements of the CSR spectrum (Social, Environmental and Governance), as well as an overall score. The paper includes all firms rated by both raters for whom there is a score, not just those who have been positively screened or considered best-in-class. This provides a broader assessment of rater agreements, since screens can skew results and are often based on partial, subjective criteria (Chatterji, Durand, Levine, & Touboul, 2014; Ziegler & Schröder, 2010).

One of the reasons for assessing different countries in this study is that it has been found that there is a country effect on CSR scoring, due to the fact that countries regard the responsibilities of firms in different ways (Yunwook and Soo-yeon 2010; Fisher 2017). For example, countries that expect firms to engage with CSR can drive greater engagement, which in turn can lead to greater CSR disclosures, and hence potentially higher rater scores. Increased disclosures means there is more primary data for raters to analyse, which is likely to lead to increased reputation (irrespective of the quality of the disclosures) and increased rater scores (Cho et al. 2012; Hughey and Sulkowski 2012). Increased disclosures have also been found to result in greater rater agreement in more highly rated firms (Herzel et al., 2012).

Previous studies have also found that some industries engage in more CSR activities than others, depending on whether they are visible to the public eye or have a great impact on one element of CSR (Cai, Jo, and Pan 2012) (extraction industries involved with environment issues, for example). Hence one might expect to see greater CSR engagement result in greater disclosures and hence more rater agreement, but as this effect may be limited to only one aspect of CSR, this industry impact may be less evident. Indeed, Wanderley et al. (2008) found that the influence of industry on CSR was much less apparent than country influence.

In terms of company size, most studies have found that large firms are rated higher in CSR since they have greater resources at their disposal to invest in those activities and they tend to be more visible to investors and the general public which increases their need for legitimacy (Udayasankar 2008). As stated previously, this increased disclosure may lead to greater congruence between raters as there is more primary data available on which to make CSR ratings and judgements.

This paper has found that there are greater agreements between raters on the Social aspects of CSR and the Overall category (Environmental, Social and Governance combined). There is greatest agreement across the DAX and S&P firms. For industry, the results were very varied; whilst generally there were fair levels of agreement for some industries in basic extraction and energy, this finding changed when using a ranked methodology, where Communications and Technology industries were most closely rated. For firm size however, the raters agree most on ratings for medium and large firms in comparison with smaller firms.

The rest of this paper is structured as follows: a brief review on the uses made by investors and academics of raters' outputs, followed by a review of prior literature on data validity and commeasurability of CSR raters. This is then followed by sections on rater selection, methodology and sample selection, leading on to a discussion of the results, followed by a conclusion with suggestions for future research.

2. Using raters as an assessment of CSR

The main aim of raters is to support investors in their selection of investments, taking account of an investor's personal ethos to invest in socially responsible stocks, otherwise known as socially responsible investment (SRI). Referred to as a strategy which underpins financial returns with social and/or environmental concerns in line with investors' own social, ethical, ecological and economic aspirations (Brzeszczynski & McIntosh 2014; Benson & Humphrey 2008; Renneboog et al. 2008; Delmas et al. 2013), SRI represented 26% of professionally managed assets globally, worth $22.89 trillion in 2016 (Global Sustainable Investment Alliance 2016). Initially emanating from religious institutions' desires to manage their investments in accordance with their moral and ethical beliefs (Brzeszczynski and McIntosh 2014; Schueth 2003), SRI has now developed to incorporate environmental, often political (e.g. apartheid in South Africa during the 1960-1990s) and social concerns (Schaltegger 2011).

Many academic studies have evaluated whether SRI provides a better, worse or equivalent return to investors than those conventionally invested without regard to CSR performance (for a useful meta-analysis of 85 studies and 190 experiments on SRI and financial performance, see Revelli & Viviani (2015)). Despite the expectation that returns on an SRI portfolio should be lower than a 'conventionally' invested portfolio, as it reduces the availability of some investments (e.g. so-called 'sin' stocks, such as tobacco, armaments and alcohol) (Benson and Humphrey 2008), some studies have found that returns on SRI portfolios are actually higher (Brzeszczynski and McIntosh 2014; Kempf and Osthoff 2007). Other studies posit that SRI investments should perform better than non-SRI portfolios, since by being socially responsible, SRI investments exhibit lower risk (through lower litigation costs, for example by having good employee/environmental policies), benefit from better relationships with a wider range of stakeholders (Delmas, Etzion, and Nairn-Birch 2013) or have lower cost of capital (Kang 2012). Other authors have found a neutral relationship between SRI and financial performance (Cellier & Chollet, 2015; Revelli & Viviani, 2015).

The ratings provided by raters, such as MCSI ESG STATS (formerly known as KLD), ASSET4, CSR Hub, Bloomberg, SAM, Vigeo Eiris and others, provide the basis on which many of the SRI portfolios are constructed and maintained. So in theory, if a firm were to fall in the ratings, then it may be divested from the SRI portfolio and vice versa (Ioannou and Serafeim 2010; Wilbert 2006; Cheung and Roca 2013). All raters use publicly available information on the CSR activities of firms to a greater or lesser degree. From an accounting perspective, an understanding of how the raters interpret and use the data provided to them to create the rating is key to assessing whether the rating is fair, as access to investment funds may depend upon it. Other impacts are loss of reputation, which may result in a fall in sales or loss in confidence from suppliers or employees – all of which impact on the firm's profitability. Either way, 'the business imperative behind CR (corporate responsibility) reporting is reputation' (Birkey et al. 2016, 144). Practising accountants are very cognisant of the importance of reputation to maintain stable relationships with customers, suppliers and capital providers. Hence the translation of CSR reporting into raters' scores is critical – a fall in reputation can result in investors withdrawing their financial support (particularly those motivated by SRI). Accountants need to manage the tension between traditional (profit-maximising) performance management and reporting and the stakeholder expectations about CSR reporting (which may not be so causally related to increasing shareholder wealth).

As many accountants are engaged in the performance management and reporting of the CSR activities of their firms (Huang and Watson 2015; Michelon, Pilonato, and Ricceri 2015; Gray 2010) either in standalone reports or as part of an integrated report, they need an understanding of how their disclosures might impact their firm. CSR disclosures reward firms for having good CSR performance and they can also reduce the reputational impacts of poor CSR performance (Huang and Watson 2015; Cho et al. 2012). Disclosures are perceived to be a commitment to social investment, improving firm reputation (Huang and Watson 2015). Hence many academics view CSR reporting as a symbolic legitimacy tool (Cho and Patten 2007; Roberts 1992; Gray et al. 1995; Rodrigue, Magnan, and Boulianne 2013), where organisations report on actions which they believe will improve their reputation and perception among stakeholders (Lys, Naughton, and Wang 2015). Disclosures can also be used as a communications tool to explain the organisation's substantive actions and policies to stakeholders, or to avoid more detailed enquiry and/or regulation (Cho et al. 2012; Brammer, Brooks, and Pavelin 2006). There is some evidence that disclosures can improve brand value (Brooks and Oikonomou 2018), increase firm valuation (Plumlee et al. 2015;

Clarkson 1995) and reduce the cost of equity capital, enabling firms to raise additional equity financing (Dhaliwal et al. 2011). However, given these benefits to firms and the lack of auditability around much CSR reporting, this has led to a concern that firms are deliberately using disclosures as signalling techniques (Lys, Naughton, and Wang 2015; Jahn and Brühl 2019). Equally, CSR raters may place undue reliance on firm disclosures as an indicator of actual performance, which may not be an accurate portrayal of underlying behaviour (Cho et al. 2012; Huang and Watson 2015). However, if decisions are being made on the basis of ratings which might affect funding or reputational risk (Gödker and Mertins 2017), then accountants ignore the power of disclosures at their peril.

Academics also use the raters as an independent proxy for CSR performance to make data collection simple and to compare firms across a variety of industries and in some cases, countries (Galant and Cadez 2017). In organisational studies, one of the most popular uses for the scores from the raters is to attempt to answer the perennial question – whether 'good' CSR performance translates into superior financial performance (McWilliams & Siegel, 2001; Orlitzky, Schmidt, & Rynes, 2003; Waddock & Graves, 1997; Waddock, 2003) (a few examples of the many studies on this subject). Accounting studies have examined CSR ratings and information asymmetry (Cho et al. 2013), insider trading (Gao, Lisic, and Zhang 2014), earnings quality (Kim, Park, and Wier 2012), tax avoidance and earnings quality (Watson 2015) and financial performance (Qiu, Shaukat, and Tharyan 2016; Brooks and Oikonomou 2018).

Chatteriji et al. (2009) pointed out how flawed reliance on these ratings might be: studies that found little or no correlation between CSR scores and financial performance may be as a result of understatement of the company's CSR activities in the scores, whereas those finding a positive relationship could be as a result of an overstatement of the activities of the firm in the scores. This is because each rater has their own proprietary (and not always scientific) method of converting CSR reporting to a raw score. Hence using one given rater as the sole basis for CSR data may skew results, particularly if (as found in this paper) raters even rank the firms differently (i.e. do not regard the same firms as 'high', 'medium' or 'low' scoring).

Using these raters without due regard to the validity of their assessment can undermine investment decisions on the one hand, but when used in academic study can reduce the generalisability of findings or even potentially change the findings (Chatterji et al., 2014; Kang, 2012). Whether differences between raters assessing the same firms arise through error of measurement, difference in methodology or

alternative conceptualisations of the elements of CSR, the impacts can be significant (Griffin and Mahon 1997; McWilliams and Siegel 2001; von Arx and Ziegler 2014).

One of the inherent difficulties for all raters is to be able to demonstrate that what they are measuring is a reflection of what is actually happening in the firms being rated (known as 'construct validity'), as the concepts of CSR are multi-faceted (Chatterji et al., 2014; Chatterji et al., 2008; Kang, 2012; Weber & Gladstone, 2014). Given the proliferation of raters, there is an ever growing diversity of methodologies to assess the elements of CSR, and unlike financial ratios, there is no universal definition of what constitutes social and environmental performance and how that can be adequately assessed (Chatterji et al., 2014; Delmas et al., 2013; Schafer, Beer, Zenker, & Fernandes, 2006; Yates-Smith, 2013). There is a concern that without some convergence, flawed methodologies may be relied upon and confidence may be lost in the use of raters as the basis for SRI (Delmas, Etzion, and Nairn-Birch 2013; Chatterji et al. 2014). However, assessing construct validity for a concept such as CSR which itself has not been universally defined is notoriously difficult. Raters only have limited access to data on the firms they rate, and as much of it has been released by the firms themselves, it cannot be said to unbiased or impartial. Hence raters rely on quantity of information available rather than the quality or accuracy of such information (Eccles, Ioannou, and Serafeim 2014). As each firm will release different information (depending on their own preferences or CSR strategy), and each rater will have their own (proprietary) methodology of assessing that information, standardising the measurement of construct validity is virtually impossible. Hence the aim of this paper is not to assess this aspect of rater performance, but to determine to what extent two raters (one subscription only and the other free) rate the same firms with the same scores or ranking. As raters use their methodologies on a consistent basis, this paper is not attempting to assess whether they are really measuring what firms do, but to what extent, with similar information, they would rate similarly.

3. Literature review

Despite the popular use of CSR raters in both investment communities and academia, there are only a few academic studies which have evaluated the validity of the results from raters, usually by comparing the raters' scores with some other measure of the CSR element under review (in other words, the construct validity of the measure or 'does it measure what it is aiming to measure?'). These studies generally concentrate on just one element of the CSR gamut of measures, such as social performance in the KLD

database in Sharfman's (1996) study, workforce diversity and corporate governance also in the KLD database (Kang 2012), governance within the datasets from RiskMetrics/Institutional Shareholder Services, GovernanceMetrics International and The Corporate Library in the Daines et al. study (2010), the environmental element from KLD, Trucost and Sustainable Asset Management (SAM) datasets in the Delmas et al. study (2013) or the social aspect from KLD, ASSET4, Calvert, FTSE4Good,DJSI and Innovest in the Chatterji et al. study (2014).

However, with the exception of the Delmas et al. (2013) and the Chatterji et al. (2014) studies which do compare scores across raters (although only for one aspect of CSR), there are even fewer academics who have attempted to assess rater commensurability, especially across the whole range of CSR metrics. Whilst SustainAbility undertook a large review of 108 raters in 2010 (SustainAbility 2010b), they did not directly compare results from them, choosing rather to review the methods and criteria which the raters choose to use in their assessments. The few commensurability studies will now be discussed.

Herzel et al. (2012) used log linear analysis to determine the patterns of agreement between the raters KLD and ASSET4 and found that there was more agreement on which firms were the 'leaders' at the higher-performing end of the ratings scale, rather than the 'laggards'. They excluded Governance from their analysis due to large differences in interpretation by the raters of the constituents of Governance. The authors also found that the amalgamated categories (such as Integrated and Social) exhibited greater levels of agreement than the single sub-categories.

In Delmas et al.'s (2013) study of the environmental ratings of US-based firms from three raters (KLD, Trucost and Sustainable Asset Management (SAM)), they found that environmental processes (what measures the firms being rated have in place to improve their environmental performance) and environmental outcomes (the results from the processes on the actual environmental performance of the firm) explains approximately 80% of the variance of the different data (ranking) sources.

A more recent study by Chatterji et al (2014) took firms in the following six indices (KLD Domini 400 Social Index, the Calvert Social index, the DJSI World Index (all three US-based raters), Innovest's 'Top 100 Leaders in Sustainability' (Canadian rater), the FTSE4Good Index and ASSET4 firms ranked A+ (both EU-based raters)). They tracked approximately 500 US-based firms across multiple years. These firms were positively identified by the indices as being socially responsible (i.e. positive screening) (with the

exception of ASSET 4 which does not use screens). The Chatterji et al (2014) study assessed overlaps of membership of firms in the indices as their measure of agreement. Given the different (screening) methodologies and universes of the raters (European versus US-based firms) it is perhaps not altogether surprising that the results demonstrated low convergent validity between the raters. One interesting finding was that the geographically proximate raters, i.e. US-based as a group, and European Union (EU)-based as another group, demonstrated better agreement (0.45 on a tetra choric correlation for the US-based raters and 0.53 for the EU-based raters) than overall average correlation (0.31) (Chatterji et al., 2014). Even taking data for two raters over a four-year period and adjusting for differences in theorization, their study did not find any improvement in convergence. This Chatterji et al. (2014) study only found six firms that were in all the indices, hence the comparison across the six indices was small (although they did carry out some bi-lateral reviews using only two indices). They also used different years for different raters which was not adequately explained and undermines the results since inevitably occurrences which might affect the rating of firms change year-on-year and would naturally change the ratings of firms. Another drawback in their approach was the comparison of raters which use screens (i.e. do not rate those considered to be 'sin' stocks) with those that do not, as inevitably this reduces any commensurability of data. Some of the raters use very different scales to rate firms, with some using 0-100 and others -1 for 'concerns', 0 for neutral and +1 for 'positives'; this makes comparisons of commensurability more difficult to assess. The raters used in this paper do not use screens, they use a 0-100 scale as their measurement system and the same data year was used for comparison.

4. Rater selection

The majority of academic studies on CSR have focused on US stocks (A K Chatterji, Levine, and Toffel 2008; Chatterji et al. 2014; Kang 2012; Delmas, Etzion, and Nairn-Birch 2013), however this study reviews a wider range of both US and EU stocks using two different datasets which have not previously been assessed, Bloomberg ESG and CSR Hub. Both these raters cover a much wider range of European firms.

Whilst the most popular rater in academic circles historically has been the KLD database (Huang and Watson 2015), its exclusive focus on large US companies (Harrison and Freeman 1999) excludes any extrapolation of findings to a broader consideration of CSR issues globally, particularly considering the impact of country-specific characteristics (such as legislation or institutional environments (Fisher 2017)). Since 2014, the CSR dataset from a European firm, Sustainalytics was incorporated into Bloomberg's

investor and financial database as its ESG data (Sustainalytics 2014). This enables investors and academics access to information on listed companies in all major world stock exchanges. Bloomberg has by far the largest market share of the global market data/analysis spend (33.22% in 2017, second was Thomson Reuters at 22.50%) (Burton-Taylor International Consulting 2018), hence the choice of Bloomberg in this study both for its coverage of firms but also its dominant use in the investor community. Bloomberg is also increasingly used in academic settings within university economics departments, with over 300 universities using their Professional Service (Bloomberg 2018) as part of simulated trading floor environments. Hence this has made it more accessible as a data source for academic studies for staff and students alike, which is another strong motivation for its inclusion in this study, particularly given that it has not been previously assessed for commensurability of ratings.

The CSR Hub dataset, which has also not been previously assessed for commensurability of ratings, was selected as the comparator for its coverage of a wide range of companies worldwide and for its openness, as it provides its overall and sub-category data free of charge on its website, to enable a wider variety of users to access the data, including management and the academic community. This makes it accessible to both professional and amateur investors who can use it as the basis for their investment decisions. Given this, it is of interest to both academics and investors who wish to use/already use this free data source to determine how such a database compares with a subscription-only service such as Bloomberg ESG, which may be out of financial reach for some scholars and investors.

Bloomberg:

Since 2014, Sustainalytics have provided the dataset for inclusion in the Bloomberg Professional service (Sustainalytics 2014). The ethos behind Sustainalytics' rating service is to allow their target audience of institutional investors to make more informed investment decisions, rather than suggest that those firms rated higher by Sustainalytics are 'better' than those who are rated lower (SustainAbility 2013). Therefore, the rater does not screen any firms; rather it leaves that to the investors themselves should they wish to do so in order to achieve their investment objectives.

Whilst they share their methodology and weightings with their clients, they do not publicly disclose their methodology and analysis of the raw scores. The basic sources of information to develop the ratings are publicly available data, such as financial reporting, sustainability reports, websites and press releases,

but also a scan of the environmental, social and employment issues through third party searches. The firm also contacts each firm to solicit their input (SustainAbility 2013). Updates are made annually, except for controversy indicators which are reviewed monthly. They do not screen out particular industries such as tobacco and arms, but they do include involvement in such 'controversial' industries as part of their scoring system.

The rater multiplies the raw scores from each category by a weighting factor to obtain the overall scores. This methodology is designed to compare a company within its industry against best practice to derive weighed indicator scores for each firm (Sustainalytics 2015). Their categories include: Environmental, Social, Governance and an overall total ESG disclosure score. Beneath these higher-level categories, they consider sub-categories such as: CO_2 emissions, greenhouse gas (GHG) emissions, energy consumption, water consumption, hazardous waste, environmental fines, number of employees, number of women in the workforce/management, community spending, size of the board, number of independent directors, time served on the board, board attendance and political donations. They use the Global Reporting Initiative (GRI) mapping within their methodology (SustainAbility 2013). They also review monthly ten 'controversy indicators' to monitor unforeseen events during the year. Their scoring is on a scale of 0-100.

Since the data used in this study was derived from the Bloomberg Professional Service, it will hereinafter be referred to as the Bloomberg dataset, rather than Sustainalytics dataset.

CSR Hub:

The primary objective of CSR Hub is to provide access to a global database of CSR ratings over as broad a range of companies and countries possible (CSR Hub 2015b). Its information sources (they cite more than 548 data sources (CSR Hub 2019)) are largely publicly available, although the firm itself subscribes to twelve ESG analyst datasets (e.g. MSCI, Thomson Reuters, Vigeo Eiris, Trucost) to obtain their ratings. They do not list either Bloomberg ESG data or Sustainalytics as a data source (CSR Hub 2019) so there is no influence of either source in their ratings. For each item of information, they normalise different data sources for each company in order to reduce bias and provide a more consistent rating. They rate information from original sources more highly than those who provide summarised aggregated data which may less reliable and will not provide a rating if there is only one source of data which cannot be

triangulated from other sources (CSR Hub 2015b). CSR Hub provide data on twelve sub-categories, community development and philanthropy, product, human rights and supply chain, compensation and benefits, diversity and labour rights, training, health and safety, energy and climate change, environment policy and reporting, resource management, board structure, leadership ethics and transparency and reporting (CSR Hub 2014). These sub-categories are then amalgamated into four categories, Community, Employees, Environment and Governance, plus an overall score. They normalise and aggregate data from various sources (minimum two, maximum six different sources), mapping to the Global Reporting Initiative (GRI) principles, across their categories in order to attempt to triangulate findings and then provide scores on a 0-100 rating scale (CSR Hub 2015a). They do not solicit information directly from the companies being rated in their methodology. Unlike the Bloomberg dataset, they do not alter weightings to assess any specific industry trends, but like Bloomberg, they do not screen out controversial stocks either.

The principal customers for CSR Hub are given as 25% corporate managers, 20% advisors, 40% individual activists, 5% not-for-profit organisations and 10% students and educators (SustainAbility 2013) and uniquely, the firm provides its current data on the basic ratings free of charge (although the more detailed sub-category and historic data is fee-based).

This study compares the higher-level categories from both raters, plus the overall scores. For the Environmental and Governance categories, given that both firms adopt GRI guidelines in their methodology, there appears to be reasonable similarity on the issues covered. These GRI guidelines (GRI, 2013) contain a wide range of sub-categories on which to assess firms and whilst they focus on the processes within the firms being rated (e.g. policies and programmes in evidence), they also take account of outcomes (e.g. fines paid or targets achieved), which provides a more rounded view of the behaviours of the firms being rated (Delmas, Etzion, and Nairn-Birch 2013). This commonality of approach equates to Chatterji et al.'s (2014) use of the term 'common theorization', to describe the degree to which raters share a common understanding of the definitions of the elements within CSR, and hence which elements to measure. For the Environmental and Governance categories, this common theorization appears high across Bloomberg and CSR Hub. Both raters include the same sub-categories in these two higher level categories, which Chatterji et al describe as high 'commensurability' (Chatterji et al., 2014), or when raters appear to measure the same construct, e.g. Environmental, in a similar way. For the Social category, the picture is less clear-cut. The Bloomberg coverage of Social is covered by two categories within CSR Hub,

Community and Employees. Whilst both raters measure similar things at sub-category level, since they use similar GRI-based topics (so the level of common theorization appears high), CSR Hub has chosen to partition it into two top level categories, Community and Employees, rather than an all-encompassing one of Social as in the Bloomberg dataset.

Given the different sources of data used by the two raters, it is again possible that commensurability of some elements will be low, as the Bloomberg dataset incorporates direct input from the firms being rated, whereas the CSR Hub dataset retains an 'arm's-length' approach to its ratings.

Neither rater evaluated utilises the best-in-class approach, whereby firms are rejected if they are not best-in-class (although Bloomberg does make some consideration of where a firm being rated stands in comparison to its industry peer group). This approach is used by the SAM Group and can lead to the elimination of companies who, despite having improved their CSR performance may still be rejected if it falls below that of those who have performed even better (Consolandi et al., 2009).

5. Research methodology

In medical and social science, it is often necessary to assess the reliability of a rating system by evaluating interrater agreement (for example, checking how two doctors independently assess the same patient on a medical scale of symptoms to derive an appropriate diagnosis) (Banerjee et al. 1999). If there are high measures of agreement, then there is consensus on what is being measured and hence the raters can be deemed to be interchangeable (Banerjee et al. 1999). There are various accepted methods of evaluating interrater agreement, depending on how many raters are being assessed, how many times the subjects are being rated and whether the ratings scale used is continuous or discrete.

This study assesses agreement between two raters, on the same subjects, on a single occasion using a continuous 0-100 scale. From the literature, the most appropriate statistical techniques in this scenario are Lin's concordance correlation coefficient and the intraclass correlation coefficient (Chen and Barnhart 2013). However, it would be unrealistic to assume that two independent raters using their own proprietary methodologies would score the same firms exactly the same 0-100 score. Therefore, a second assessment was carried out, ranking the scores into high, medium and low-scoring tertiles (i.e. turning the continuous data into categorical data) to assess even if the raters did not score the same firm on exactly the same score, they did consider them to be higher, medium or lower performing in comparison with

others. In this case, statistical techniques best suited to categorical data were used (cross-tabulation and Cohen's kappa). The next section discusses the methodologies used in assessment 1 (scores as continuous data) and assessment 2 (scores as categorical data).

Assessment 1: Scores as continuous data

Lin's concordance correlation coefficient (CCC)

Whilst some studies use Pearson's correlation coefficients to assess rater agreement (for example, Sharfman (1996)), this is actually not a measure of agreement (Bland and Altman 1986). Although the scores can be strongly correlated (r>0.7), they may not agree, for example in the case where one rater consistently rates higher than the other (referred to as systematic bias, where the line of best fit does not pass through the origin) or in the case where the scores plot in a scatter around the line of best fit, but on average the methods rate similarly (but not the same) (Watson and Petrie 2010). Whilst the latter is arguably of less concern in social sciences, in medical sciences, the ability to assess agreement between raters more precisely is crucial. An extension of this technique is Lin's concordance correlation coefficient (CCC) (Lin 1989). Like Pearson's correlation coefficient, it examines the linear relationship between two sets of scores (*r)*, but it also assesses the slope of the line of best fit relating to the two sets of scores, with the line passing through the origin. The closer their fit is to the line and the closer that line of best fit is to a line of 45 degrees passing through the origin, the greater the agreement between the scores (Watson and Petrie 2010; Lin 1989). If all pairs of scores are on a line of best fit of 45 degrees (i.e. perfect agreement), the coefficient would equal 1. Given this ability to test relationship and the line of best fit, it is a robust technique to assess the reliability of scores. It can also handle small sample sizes well, such as those in some of the European datasets (Watson and Petrie 2010).

It is represented by the equation:

$$r_c = \frac{2s_{xy}}{s_x^2 + s_y^2 + (\bar{x} - \bar{y})^2}$$

where r_c is the estimated concordance correlation coefficient between the *n* pairs of results (*x, y), s* is the standard deviation of the *n* pairs of results and $\bar{x}$ and $\bar{y}$ are the sample means of the two samples, (Watson and Petrie 2010; Lin 1989).

Intraclass correlation coefficient

The ICC is a particularly useful measure of interrater reliability as it measures the reliability of ratings by comparing the variability of different ratings of the same subject (here firms) to the total variation across all ratings and all subjects, plus an error sum (Scherbaum & Ferreter, 2008; Shrout & Fleiss, 1979). This measure is similar to the CCC, in that it measures the variances between the pairs of data (i.e. the assessments made by both raters of a single firm) as a proportion of the total variance of the observations. The scale varies between 0 (no agreement) to 1 (total agreement) (Watson and Petrie 2010; Rothwell 2000; Chen and Barnhart 2013) and it is expressed in the following equation:

$$r = \frac{s_a^2 - s_d^2}{s_a^2 + s_d^2 + \frac{2}{n}\left(n\bar{d}^2 - s_d^2\right)}$$

where r is the intraclass correlation coefficient, *n* is the number of pairs observed, s_a^2 is the estimated variance of *n* sums, s_d^2 is the estimated variance of the *n* differences, and $\bar{d}$ is the estimated mean of the differences (Watson and Petrie 2010).

Assessment 2: Scores as categorical data

As it is unlikely that the raw scores between the raters would agree precisely, a categorical approach was also used to test if raters agreed on which firms were high, medium or low scoring. Hence the raw scores from each rater were taken and placed into tertiles as ‘high-scoring’, ‘medium-scoring’ and ‘lower scoring’. The use of cross-tabulation and Cohen’s kappa are the most appropriate techniques for categorical data.

Cross-tabulation

This method involves creating a table with the number of times raters agree into which category a firm should be ranked. One rater’s scores are represented in the columns and the other’s in the rows. This allows us to describe the relationships between the raters (Wildemuth 2006).

Cohen’s kappa

Cohen’s kappa, first developed by Jacob Cohen in 1960 is a measure of agreement between two raters based on a nominal scale, but corrected to eliminate agreement purely by chance (Cohen 1960). Used principally in medical research, it is a development from earlier models which studied agreement between raters, but which did not take into account that certain levels of agreement between two raters are inevitable, due to chance alone (Banerjee et al., 1999). Cohen’s kappa adjusts for this chance agreement

by examining the marginal distributions of the responses made by each rater, and assuming that each rater rates independently (Banerjee et al. 1999). Cohen's kappa is expressed as:

$$k = \frac{P_o - P_e}{1 - P_e}$$

where P_o is the observed proportion of agreement and P_e is the proportion of agreement expected by chance.

6. Sample selection

Data was collected on firms from the US Standard & Poor's 500 (hereinafter referred to as S&P), the UK FTSE350 (FTSE), the French CAC40 (CAC), the Spanish IBEX (IBEX) and the German DAX (DAX) stock exchanges. This was to determine whether there are country-specific institutional environments which can affect CSR behaviours in different countries (Fisher 2017; Yunwook and Soo-yeon 2010) and which might be a cause of possible divergence between raters' scores. If certain countries expect firms to engage in CSR engagement, their scores could be higher, hence there could be more agreement between raters (as was found by Herzel et al. (2012)), due to a greater level of data available from highly CSR-active firms.

For each firm, the same data was gathered: from CSR Hub: Community, Employees, Environment, Governance and an aggregated overall score, from Bloomberg: Social, Environmental, Governance and aggregated overall score. Both Community and Employees categories were compared with Bloomberg's Social category. The relative weightings of the two categories of Community and Employees are not known and it is acknowledged that users have different perspectives on the relative weightings of different sub-categories (Graves & Waddock 1994; Ruf et al. 1998). Therefore, it would not be appropriate to average the scores for these two categories to create an artificial Social category for CSR Hub, as any arbitrary weighting could significantly undermine any findings. This will, however, mean that there is likely to be limited convergence between the Social (Bloomberg) and Community and Employees (CSR Hub) categories. It is important to note that this lack of commensurability on these particular categories does not necessarily reflect low validity of the scores or of the raters themselves, merely a differing perception of that particular aspect of CSR vis-à-vis the raters' target audiences' preferences (Chatterji et al., 2014); the two raters evaluated here having quite different target audiences as noted above.

After eliminating firms where data was unavailable across any of the CSR proxies, the final total sample of firms was 720, which was split across S&P (370 firms), FTSE (252 firms), CAC (37), IBEX (32) and DAX (29). Only firms from these European exchanges were used as there was insufficient data available on one or more of the CSR elements for a meaningful sample of companies in other European countries. All data used related to ratings from one year: 2014.

The data was analysed according to the Global Industry Classification System (GICS) used by Bloomberg to test for any industry effects (Cai, Jo, and Pan 2012). The distribution of firms by country and industry are shown in table 1. There is no single predominant industry influencing the results.

INSERT TABLE 1 HERE

Certain industries (e.g. basic materials) have more impact on the environment hence they are likely to produce more data and information on environmental issues. This data and information is the input to the raters' scorings and hence one might expect greater correlation and agreement on these aspects of ESG than others where less information is in the public domain.

Companies were also analysed by size, by grouping firms into tertiles of 'large' 'medium' and 'small' using market capitalisation (number of ordinary shares outstanding multiplied by share price) as the determinant of firm size, in common with that used by the stock exchanges themselves. This was to assess whether there was more agreement on CSR metrics in larger or smaller firms. Again, the logic behind this approach is that larger firms tend to be more visible (Udayasankar 2008) and hence provide more information on their CSR activities which might influence CSR scores and promote greater rater agreement.

7. Results

General data review

INSERT TABLE 2 ABOUT HERE

General descriptive statistics (minimum, maximum and mean scores and the standard deviation) were obtained from the absolute scores of CSR Hub and Bloomberg (see Table 2). On all categories (Overall, Social (and the equivalents in CSR Hub of Community and Employees), Environment and Governance), Bloomberg uses a wider variety of scores; in general, lower minimums and higher maximums, but universally greater standard deviations. Across those exchanges with larger sample sizes, such as All

Firms, S&P, all EU and where tertiles are used as sample groupings, the effects of these variations are smoothed, but for smaller sample groupings, such as the smaller exchanges and industry samples where there are only a few sample firms, the wider variations may create impacts on the outcomes of the statistical tests.

Assessment 1: Continuous data

Lin's concordance correlation coefficient (CCC)

INSERT TABLE 3 ABOUT HERE

Table 3 shows Lin's concordance correlation coefficients (CCC). In general, the association between the two datasets is very low (less than 0.20): given that both raters use different methodologies and possibly use different data sources (for example, Bloomberg contact firms directly as well as using publicly available data, whereas CSR Hub does not), this is not surprising.

Within the data, some further analysis is possible. One initial hypothesis was that the division of CSR Hub's social dimension into two sub-categories of Employees and Community was expected to yield poor correlations with Bloomberg's all-encompassing Social category. Of all categories assessed, this category performs the highest, and environment (which is arguably more quantitatively evaluated by firms in comparison to social issues) is the category where the raters diverge the most.

Raters agreed most closely with firms in DAX, perhaps due to the smaller sample size, although the second most agreement was within the S&P, which had the largest sample size. By industry, the closest in agreement was in Energy, followed by Basic Materials and Industrial. This does tend to support prior findings that industries which have the greatest impacts on environmental issues provide more data and hence there may be greater rater congruence (Cai, Jo, and Pan 2012). Equally, prior literature has found that larger firms provide more data which can be analysed, and hence which might provoke closer ratings; this study supports that finding.

Intraclass correlation coefficient (ICC)

These findings were supported by the intraclass correlation coefficient (ICC) findings (two-way mixed effects) when estimated at the most rigorous level of agreement – absolute agreement (as opposed to consistency). The single measures iteration of ICC produces almost identical results to CCC (and hence for brevity has not been shown here); however, when estimating average measures (which instead of

determining whether there is agreement on a single firm, assesses whether on average there is agreement), there are some slight variations as can be seen in table 4.

INSERT TABLE 4 ABOUT HERE

From this we can see that based on average measures, the S&P and DAX are the countries where the raters are most likely to agree (with CAC and FTSE being the least). In terms of industry, Basic materials and Industrial again appear to be industries where there is best agreement, but Communications rates better using average measures. Both methods agree that Technology, Consumer Cyclical and Utilities are where raters disagree the most.

Both of the Community/Social and Employee/Social categories are areas of highest agreement, and the ICC method confirms that the Environmental category is the worst. The ICC results confirmed the CCC findings that results were closest in large firms.

Assessment 2: Categorical data

Crosstab analysis

The raters' scores were placed into tertiles, representing firms being rated as 'high scoring', 'medium scoring' and 'low scoring' in each category. There is no inference by placing the firms into these categories that they are in anyway 'better' or 'worse', as the groupings are used as nominal categories based on the scorings from the raters themselves to be used to determine agreement between the raters. These categories are then used to construct a two-way (as there are two raters being compared) contingency or crosstab table. A sample table from the results from this study are displayed in table 5 below.

TABLE 5 SHOULD BE INSERTED HERE

The diagonal cells of the table represent where the ratings agree. Therefore, a simple method of measuring agreement is to see how many agreements were observed in the table, which in this example equates to 136+116+139=381. This is then divided by the number of observations (here 720) giving a percentage of agreement of 54%.

Cohen's kappa

Whilst crosstab is undoubtedly a simple method, it fails to account for which place in the table the agreement occurred (e.g. did the two raters both agree more on the higher scoring firms?) and it also does

not take into account the element of chance (Altman 1991). Cohen's kappa is used to compare a model of independence as a baseline (what you would expect to see without adjusting for chance agreement) with the sum of the differences between the observed counts on the main diagonal cells of the table with that independent model. If there is a large difference between the observed cell count and the expected cell count, then there is deemed to be a high level of agreement which goes beyond mere chance of the raters agreeing alone (Tanner & Young, 1985).

Table 6 summarises the Cohen's kappa scores from this study.

TABLE 6 SHOULD BE INSERTED HERE

If there is complete agreement between the raters, then Cohen's kappa = 1, but if there is no agreement beyond that what would be expected by chance, then kappa is less than or equal to 0. As for what is deemed 'poor', 'fair' or 'excellent' values of Cohen's kappa to demonstrate agreement of raters beyond pure chance, Cohen (1960) suggested that values ≤ 0 show no interrater agreement, 0.01–0.20 as none to slight, 0.21–0.40 as fair, 0.41–0.60 as moderate, 0.61–0.80 as substantial, and 0.81–1.00 as almost perfect agreement. Landis and Koch (1977) proposed that kappa values above 0.75 suggests excellent agreement beyond mere chance, values between 0.40 and 0.75 represent fair to good agreement beyond chance, whilst values below 0.40 demonstrate poor agreement beyond chance. Where Cohen's kappa is used in medical studies levels of agreement must be high as otherwise poor results may impact on clinical practice (McHugh 2012), whereas in more general social sciences, levels of acceptance may be lower. Whilst there is also discussion of bias in the use of kappa for marginal distributions (Banerjee et al., 1999), as the author has derived the tertile rankings from the raters' scores as nominal categories, rather than the raters themselves deciding the rankings, there is limited effect of any conscious bias of raters in this case.

Based on the suggested thresholds of agreement noted above, there is generally only fair agreement beyond that which would occur through chance whether firms are rated high, medium or low. In general, the Cohen's kappas are higher than those in the more stringent individual firm score level results from the CCC and ICC methods. This is to be expected: it is more likely that raters would agree which firms are in the highest ranking tertile, even if they do not agree on a specific single score for each firm.

In common with the CCC method, Employee/Social is the category with greatest agreement, followed by the Overall category. However, raters score Environmental scores quite consistently into similar tertile ranks unlike in the continuous data assessment. The DAX and S&P exchanges both ranked most closely (in agreement with the CCC assessment), with similarly the FTSE and CAC being where raters agree the least. This is despite the fact that Cohen's kappa results tend to be weaker across smaller sample populations (Altman, 1991; Tractenberg et al., 2010; McHugh, 2012).

Where these results differ from the CCC is across industries, with CCC and ICC methods finding better agreement for Basic Materials and Industrial sectors, whilst Cohen's kappa finds better rated agreement in the Communications and Technology sectors. All of these groups have relatively small sample sizes (see table 1), which might explain the reduced variability in these results. At the lower end of the results, Cohen's kappa ranks Consumer Cyclical, Utilities and Consumer Non-cyclical as the industries in which raters do not consistently rank the same.

Based on firm size, the 'best' level of agreement across this measure is for medium-sized firms, then large firms and finally smaller firms.

8. Conclusion

This study has analysed 720 firms from the US and EU for the agreement in ratings by two CSR raters using both a continuous scale approach and a categorical ranked approach to assess even if exact scores do not agree, that there is some convergence of opinion on whether firms are likely to be in the top, mid or lower tertile of rating. Whilst it is perhaps not surprising that there is low rater agreement using the continuous scale approach, there are only 'fair' levels of agreement using a ranked approach. This study adds to the literature on CSR rater commensurability by examining a greater number of firms across both the EU and the US, across industries and different firm sizes, using robust techniques to assess absolute and relative ratings.

One unexpected finding from this study is that irrespective of the approach used (continuous or categorical) the Social categories of CSR tend to be rated most closely than Environmental or Governance. This is despite there often being more legislation for listed companies around reporting such matters and for many of the environmental issues to be measured quantitatively (e.g. monitoring of greenhouse gases) in reports, which one might imagine would be consistently interpreted by raters. The Social dimensions on

the other hand are often more individual to the firm and hence would not appear to be so consistent in approach for raters to assess. This finding is in contrast to Delmas et al.'s (2013) decision not to study the social aspects of CSR since they were less 'quantifiable' than the environmental aspects which they chose to focus on.

Another interesting finding unique to this study (as others focused only on one stock exchange) is that both the largest dataset (US S&P) and the smallest (DAX) consistently rated closer than the other stock exchanges, although less surprising is the finding that large firms are rated more consistently than small ones (possibly due to the higher visibility of these firms and a greater level of disclosure and activities). Industry findings were mixed, with the continuous approach finding the Basic Materials and Industrial sectors more consistent, but the categorical approach finding more agreement in the Communications and Technology sectors.

Reliable results from raters assist investors and academics alike in making informed decisions or inferences, but the choice of rater could affect outcomes significantly (Chatterji et al., 2014). By using a wider variety of statistical techniques than previous studies, this paper presents a more robust analysis of the two raters, Bloomberg and CSR Hub. For students and others who wish to make use of CSR Hub's free access to CSR data may do so in the knowledge that based on this current study, its theorization and commensurability is has a fair level of similarity to that of a subscription-only provider (Bloomberg) on Employees, Community and Overall categories. However, caution should be exercised in Governance and Environmental assessments. That is not to say that they are inaccurate with CSR Hub (or indeed with Bloomberg), but the individual researcher should satisfy him/herself that the methodology and issues covered are in line with the aims of their particular study, and hence triangulation with a second CSR rater is advisable.

There is little evidence that academics (and perhaps even investors) assess firm CSR performance using more than one rater (Delmas, Etzion, and Nairn-Birch 2013), and within academia, the emphasis has been on the use of KLD. However, Chatterji et al,'s (2014) (and this study's) recommendation is that those using CSR raters should consider using more than one rater to validate findings. For investors, the recommendation is that they should consider a much broader scope of information other than just raters' scores as the basis for their investments.

There are continuing criticisms of raters on issues of credibility, transparency of methodology, appropriate focus on the issues that matter, conflicts of interest with firms being rated, the difficulty of making inter-industry comparisons and limited validity of metrics (they do not measure what they claim) (Chatterji et al., 2014; Delmas et al., 2013; SustainAbility, 2010a; Yates-Smith, 2013). However, it is likely, given that the rise in SRI which shows no sign of abating (Global Sustainable Investment Alliance 2015; Renneboog, Ter Horst, and Zhang 2008), that raters will continue to be relied on for both practitioner/investor support and for academic purposes, given their at least quasi-independence from the firms they are rating. SRI indices and raters can act as positive drivers on corporate behaviour to drive engagement in CSR and to maintain a positive rating or inclusion in a specific CSR index (Slager 2009). They can also be used to reduce information asymmetry between firms and investors (Schafer et al., 2006). However, the lack of accountability and stewardship of the raters themselves (Yates-Smith 2013) and the continuing plethora of different measures and metrics will make assessing firms on a level playing field increasingly difficult (Chatterji et al., 2014).

Over time it is likely that the merger and acquisition activity of raters that has already been seen (Schafer et al., 2006; Weber and Gladstone, 2014) will continue, as will pressures to standardise ratings categories and measurement methods (leading to greater 'common theorization' and 'commensurability' (Chatterji et al., 2014)). That there is a level of agreement on some of the CSR categories (Overall, Employees/Social and Community/Social) between the two firms in this study, particularly in comparison to the recent six rater firm study by Chatterji et al (2014), perhaps owes much to the fact that they use a form of the standardised GRI guidelines within their methodology (better theorization and commensurability). If approaches like this gain more ground (Willis 2003) and methodologies from the raters become more transparent (SustainAbility 2010a), it is likely that there will be a further reduction in the number of disparate approaches used by raters, which will increase the credibility and reliability of their ratings.

This is paper is not without limitations. The issue of appropriate sample sizes and their impact on the statistical power of findings is often not adequately addressed in many academic studies (Scherbaum and Ferreter 2008) and which is demonstrated by some of the weaker findings in this study relating to the smaller EU exchanges. Despite this, there does not appear to be any clear explanation for the relatively good levels of agreement across many of the DAX categories, as the sample size was small, and the

general descriptive statistics discussed earlier in this paper illustrated no demonstrable differences in comparison with the other smaller exchanges (CAC and IBEX).

Another limitation to this paper is that it did not attempt to assess whether Bloomberg and CSR Hub measure what actually occurs in firms (construct validity), due to the inherent difficulty in assessing this, given the range of information provided by firms and the proprietary processes the raters use to assess it. However, more research in this regard would be very useful. This study has also only assessed two CSR raters over the higher-level categories of CSR data (Overall, Social/Employees, Social/Community, Environment and Governance) and only over one year of data. Further work should be undertaken to assess levels of agreement from a greater number of raters encompassing additional CSR sub-categories in order to provide greater assurance to investors and academics that decisions taken, and inferences made are based on appropriate data. It would also be of particular use in academic studies to assess the consistency of the raters over time by studying longitudinal data over several years. This historical consistency would be of arguably less importance for investors, in particular those who invest conventionally rather than under SRI criteria. Investors often make rapid (dis)investment decisions based on very immediate events, rather than on historic trends. Only the SRI investor would wish to be assured that the firm's performance was not attributable to past unethical behaviour (Chatterji, Levine, & Toffel, 2009; Kang, 2012), and hence the historical consistency of a rater has more value.

This research has potential impacts for investment practitioners using raters to determine and manage SRI portfolios and in academic studies by providing greater appraisal of the levels of agreement (and disagreement) between ratings. For accountant-practitioners, it is important that they are aware of the differing levels of agreement between raters and should seek to identify which rater(s) are used (if any) by their own major investors. As investors may take the decision to invest or disinvest on the basis of the rating of the firm (Gödker and Mertins 2017), practitioners should be aware that their disclosures may impact their reputation but also the behaviour of the investment community (Cho et al. 2012; Huang and Watson 2015; Lys, Naughton, and Wang 2015).

References

Altman, D. G. 1991. *Practical Statistics for Medical Research*. Chapman and Hall, London.

Arx, Urs von, and Andreas Ziegler. 2014. "The Effect of Corporate Social Responsibility on Stock Performance: New Evidence for the USA and Europe." *Quantitative Finance* 14 (6). Affiliation: Center for Corporate Responsibility and Sustainability, University of Zurich, Zähringerstr. 24, 8001, Zurich, Switzerland; Affiliation: Center of Economic Research, Swiss Federal Institute of Technology (ETH) Zurich, Zürichbergstr. 18, 8032, : 977–91. https://doi.org/10.1080/14697688.2013.815796.

Banerjee, Mousumi, Michelle Capozzoli, Laura McSweeney, and Debajyoti Sinha. 1999. "Beyond Kappa: A Review of Interrater Agreement Measures." *Canadian Journal of Statistics* 27 (1): 3–23. https://doi.org/10.2307/3315487.

Benson, Karen L, and Jacquelyn E Humphrey. 2008. "Socially Responsible Investment Funds: Investor Reaction to Current and Past Returns." *Journal of Banking & Finance* 32 (9): 1850–59. https://doi.org/http://dx.doi.org/10.1016/j.jbankfin.2007.12.013.

Birkey, Rachel N., Giovanna Michelon, Dennis M. Patten, and Jomo Sankara. 2016. "Does Assurance on CSR Reporting Enhance Environmental Reputation? An Examination in the U.S. Context." *Accounting Forum* 40 (3). Elsevier Ltd: 143–52. https://doi.org/10.1016/j.accfor.2016.07.001.

Bland, J M, and D G Altman. 1986. "Statistical Methods for Assessing Agreement between Two Methods of Clinical Measurement." *Lancet* 1 (8476): 307–10. https://doi.org/10.1016/S0140-6736(86)90837-8.

Bloomberg. 2018. "Bloomberg Professional Services." Bloomberg Market Concepts. 2018. https://www.bloomberg.com/professional/expertise/universities/.

Brammer, Stephen, Chris Brooks, and Stephen Pavelin. 2006. "Corporate Social Performance and Stock Returns: UK Evidence from Disaggregate Measures." *Financial Management* 35 (3): 97–116. https://doi.org/10.1111/j.1755-053X.2006.tb00149.x.

Brooks, Chris, and Ioannis Oikonomou. 2018. "The Effects of Environmental, Social and Governance Disclosures and Performance on Firm Value: A Review of the Literature in Accounting and Finance." *British Accounting Review* 50 (1): 1–15. https://doi.org/10.1016/j.bar.2017.11.005.

Brzeszczynski, Janusz, and Graham McIntosh. 2014. "Performance of Portfolios Composed of British SRI Stocks." *Journal of Business Ethics*. Vol. 120. Springer Science & Business Media B.V. https://doi.org/10.1007/s10551-012-1541-x.

Burton-Taylor International Consulting. 2018. "Financial Market Data/Analysis Global Share & Segment Sizing." https://burton-taylor.com/wp-content/uploads/2018/03/B-T-Global-Market-Data-Analysis-5-Year-Competitor-Segment-Product-User-Institution-Analysis-2018-Information-Kit-Final-1.pdf.

Cai, Ye, Hoje Jo, and Carrie Pan. 2012. "Doing Well While Doing Bad? CSR in Controversial Industry Sectors." *Journal of Business Ethics* 108 (4): 467–80. https://doi.org/10.1007/s10551-011-1103-7.

Cellier, A., Chollet, P. 2015. "The Effects of Social Ratings on FIrm Value." *Research in International Business and Finance* In press. https://doi.org/http://dx.doi.org/10.1016/j.ribaf.2015.05.001.

Chatterji, A, R Durand, D Levine, and S Touboul. 2014. "Do Ratings of Firms Converge? Implications for Strategy Research." Vol. IRLE Worki. http://irle.berkeley.edu/workingpapers/107-14.pdf: IRLE Berkeley.

Chatterji, A K, D I Levine, and M W Toffel. 2008. "How Well Do Social Ratings Actually Measure Corporate Social Responsibility?" UC Berkeley, USA: UC Berkeley: Center for Responsible Business.

Chatterji, Aaron K., David I. Levine, and Michael W. Toffel. 2009. "How Well Do Social Ratings Actually Measure Corporate Social Responsibility?" *Journal of Economics and Management Strategy* 18 (1): 125–69. https://doi.org/10.1111/j.1530-9134.2009.00210.x.

Chen, Chia Cheng, and Huiman X. Barnhart. 2013. "Assessing Agreement with Intraclass Correlation Coefficient and Concordance Correlation Coefficient for Data with Repeated Measures." *Computational Statistics and Data Analysis* 60 (1). Elsevier B.V.: 132–45. https://doi.org/10.1016/j.csda.2012.11.004.

Cheung, Adrian (Wai Kong), and Eduardo Roca. 2013. "The Effect on Price, Liquidity and Risk When Stocks Are Added to and Deleted from a Sustainability Index: Evidence from the Asia Pacific Context." *Journal of Asian Economics* 24 (0): 51–65.

https://doi.org/http://dx.doi.org/10.1016/j.asieco.2012.08.002.

Cho, Charles H., Ronald P. Guidry, Amy M. Hageman, and Dennis M. Patten. 2012. “Do Actions Speak Louder than Words? An Empirical Investigation of Corporate Environmental Reputation.” *Accounting, Organizations and Society* 37 (1): 14–25. https://doi.org/http://dx.doi.org/10.1016/j.aos.2011.12.001.

Cho, Charles H, and Dennis M Patten. 2007. “The Role of Environmental Disclosures as Tools of Legitimacy: A Research Note.” *Accounting, Organizations and Society* 32 (7–8): 639–47. https://doi.org/http://dx.doi.org/10.1016/j.aos.2006.09.009.

Cho, Seong Y., Cheol Lee, Ray J. Pfeiffer, and Ray J Pfeiffer Jr. 2013. “Corporate Social Responsibility Performance and Information Asymmetry.” *Journal of Accounting and Public Policy* 32 (1). Elsevier Inc.: 71–83. https://doi.org/10.1016/j.jaccpubpol.2012.10.005.

Clarkson, Max E. 1995. “A Stakeholder Framework for Analyzing and Evaluating Corporate Social Performance.” *Academy of Management Review* 20 (1). Academy of Management: 92–117. https://doi.org/10.5465/AMR.1995.9503271994.

Cohen, Jacob. 1960. “A Coefficient of Agreement for Nominal Scales.” *Educational and Psychological Measurement* XX (1): 37–46. https://doi.org/10.1177/001316446002000104.

Consolandi, Costanza, Ameeta Jaiswal-Dale, Elisa Poggiani, and Alessandro Vercelli. 2009. “Global Standards and Ethical Stock Indexes: The Case of the Dow Jones Sustainability Stoxx Index.” *Journal of Business Ethics* 87 (February). Springer Science & Business Media B.V: 185–97. https://doi.org/10.1007/s10551-008-9793-1.

CSR Hub. 2014. “CSRHub Category and Subcategory Schema Community.” http://www.csrhub.com/content/csrhub-data-schema/.

———. 2015a. “CSRHub Rating Rules | CSR Ratings.” CSR Hub Website. 2015. http://www.csrhub.com/content/csrhub-rating-rules/.

———. 2015b. “The CSRHub Ratings Methodology | CSR Ratings.” 2015. http://www.csrhub.com/content/csrhub-ratings-methodology/.

———. 2019. “Data Sources CSR Ratings.” Data Sources. 2019. https://www.csrhub.com/our_data_sources/.

Daines, Robert M, Ian D Gow, and David F Larcker. 2010. “Rating the Ratings: How Good Are Commercial Governance Ratings?” *Journal of Financial Economics* 98 (3): 439–61. https://doi.org/http://dx.doi.org/10.1016/j.jfineco.2010.06.005.

Delmas, Megali A., D. Etzion, and N. Nairn-Birch. 2013. “Triangulating Environmental Performance: What Do Corporate Social Responsibility Ratings Really Capture?” *Academy of Management Perspectives* 27 (3): 255–67. https://doi.org/10.5465/amp.2012.0123.

Dhaliwal, Dan S., Oliver Zhen Li, Albert Tsang, and Yong George Yang. 2011. “Voluntary Nonfinancial Disclosure and the Cost of Equity Capital: The Initiation of Corporate Social Responsibility Reporting.” *Accounting Review* 86 (1): 59–100. https://doi.org/10.2308/accr.00000005.

Eccles, Robert G., Ioannis Ioannou, and George Serafeim. 2014. “The Impact of Corporate Sustainability on Organizational Processes and Performance.” *Management Science* 60 (11): 2835–57. https://doi.org/10.2139/ssrn.1964011.

Fisher, Victoria. 2017. “An International Analysis of CSR Rankings and a Country’s Culture.” *Senior Honors Theses* 561: 49. http://commons.emich.edu/honors/561.

Galant, Adriana, and Simon Cadez. 2017. “Corporate Social Responsibility and Financial Performance Relationship: A Review of Measurement Approaches.” *Economic Research-Ekonomska Istrazivanja* 30 (1). Routledge: 676–93. https://doi.org/10.1080/1331677X.2017.1313122.

Gao, Feng, Ling Lei Lisic, and Ivy Xiying Zhang. 2014. “Commitment to Social Good and Insider Trading.” *Journal of Accounting and Economics* 57 (2–3). Elsevier: 149–75. https://doi.org/10.1016/j.jacceco.2014.03.001.

Global Reporting Initiative. 2013. “G4 Sustainability Reporting Guidelines: Implementation Manual.” Amsterdam, The Netherlands: GRI.

Global Sustainable Investment Alliance. 2015. “Global Sustainable Investments 2012-2014.” www.ussif.org/files/publications/.

———. 2016. “Global Sustainable Investment Review 2016.”

Gödker, Katrin, and Lasse Mertins. 2017. “CSR Disclosure and Investor Behavior: A Proposed Framework and Research Agenda.” *Behavioral Research in Accounting* 30 (2): 37–53. https://doi.org/10.2308/bria-51976.

Graves, Samuel B, and Sandra A Waddock. 1994. “Institutional Owners and Corporate Social Performance.” *Academy of Management Journal* 37 (4). Academy of Management: 1034–46. https://doi.org/10.2307/256611.

Gray, Rob. 2010. “Is Accounting for Sustainability Actually Accounting for Sustainability…and How Would We Know? An Exploration of Narratives of Organisations and the Planet.” *Accounting, Organizations and Society* 35 (1): 47–62. https://doi.org/http://dx.doi.org/10.1016/j.aos.2009.04.006.

Gray, Rob, Reza Kouhy, Simon Lavers, Rob Gray, Reza Kouhy, and Simon Lavers. 1995. “Longitudinal Study of UK Disclosure.” *Accounting, Auditing & Accountability Journal* 2: 47–77.

Griffin, Jennifer J, and John F Mahon. 1997. “The Corporate Social Performance and Corporate Financial Performance Debate: Twenty-Five Years of Incomparable Research.” *Business & Society* 36 (1): 5–31. https://doi.org/10.1177/000765039703600102.

Harrison, Jeffrey S, and R E Freeman. 1999. “Stakeholders, Social Responsibility, and Performance: Empirical Evidence and Theoretical Perspectives.” *Academy of Management Journal* 42 (5). Academy of Management: 479–85. https://doi.org/10.2307/256971.

Herzel, Stefano,; Becchetti, L.; Nicolosi, M.; Fabretti, A.; Ciciretti, R.; Giovannelli, A.; Palit, I.; Stanghellini, E. 2012. “The Dissemination of CSR Information in Financial Markets.” www.sirp.se/getfile.ashx?cid=315341&cc=3&refid=76.

Huang, Xiaobei, and Luke Watson. 2015. “Corporate Social Responsibility Research in Accounting.” *Journal of Accounting Literature* 34: 1–16.

Hughey, Christopher J., and Adam J. Sulkowski. 2012. “More Disclosure = Better CSR Reputation? An Examination of CSR Reputation Leaders and Laggards in the Global Oil and Gas Industry.” *Journal of the Academy of Business and Economics* 12: 24–34.

Investments, Domini Social. 2014. “Domini Social Investments.” 2014. https://www.domini.com/domini-funds.

Ioannou, Ioannis, and George Serafeim. 2010. “The Impact of Corporate Social Responsibility on Investment Recommendations.” Harvard, US: Harvard Business School.

Jahn, Johannes, and Rolf Brühl. 2019. “Can Bad News Be Good? On the Positive and Negative Effects of Including Moderately Negative Information in CSR Disclosures.” *Journal of Business Research* 97 (January). Elsevier: 117–28. https://doi.org/10.1016/j.jbusres.2018.12.070.

Kang, J. 2012. “Effectiveness of the KLD Social Ratings as a Measure of Workforce Diversity and Corporate Governance.” *Business & Society*. https://doi.org/10.1177/0007650312461602.

Kempf, A, and P Osthoff. 2007. “The Effect of Socially Responsible Investing on Financial Performance .” *European Financial Management* 13: 908–22.

Kim, Yongtae, Myung Seok Park, and Benson Wier. 2012. “Is Earnings Quality Associated with Corporate Social Responsibility?” *The Accounting Review* 87 (3): 761–96.

Landis, J R, and G G Koch. 1977. “An Application of Hierarchical Kappa-Type Statistics in the Assessment of Majority Agreement among Multiple Observers.” *Biometrics* 33 (2): 363–74. https://doi.org/10.2307/2529786.

Lin, L.I. 1989. “A Concordance Correlation Coefficient to Evaluate Reproducibility.” *Biometrics* 45: 255–68.

Lyon, Thomas P., and John W. Maxwell. 2011. “Greenwash: Corporate Environmental Disclosure under Threat of Audit.” *Journal of Economics and Management Strategy* 20 (1): 3–41. https://doi.org/10.1111/j.1530-9134.2010.00282.x.

Lys, Thomas, James P. Naughton, and Clare Wang. 2015. “Signaling through Corporate Accountability Reporting.” *Journal of Accounting and Economics* 60 (1). Elsevier: 56–72. https://doi.org/10.1016/j.jacceco.2015.03.001.

Mahoney, Lois S, Linda Thorne, Lianna Cecil, and William LaGore. 2013. “A Research Note on Standalone

Corporate Social Responsibility Reports: Signaling or Greenwashing?" *Critical Perspectives on Accounting* 24 (4–5): 350–59. https://doi.org/http://dx.doi.org/10.1016/j.cpa.2012.09.008.

McHugh, M. 2012. "Interrater Reliability: The Kappa Statistic | Biochemia Medica." *Biochemia Medica* 22 (3): 276–82. https://doi.org/http://dx.doi.org/10.11613/BM.2012.031.

McWilliams, Abagail, and Donald Siegel. 2001. "Corporate Social Responsibility: A Theory of the Firm Perspective." *Academy of Management Review* 26 (1). Academy of Management: 117–27. https://doi.org/10.5465/AMR.2001.4011987.

Michelon, Giovanna, Silvia Pilonato, and Federica Ricceri. 2015. "CSR Reporting Practices and the Quality of Disclosure: An Empirical Analysis." *Critical Perspectives on Accounting* 33. Elsevier Ltd: 59–78. https://doi.org/10.1016/j.cpa.2014.10.003.

Orlitzky, Marc, Frank L Schmidt, and Sara L Rynes. 2003. "Corporate Social and Financial Performance: A Meta-Analysis." *Organization Studies* 24 (3): 403–41. https://doi.org/10.1177/0170840603024003910.

Plumlee, Marlene, Darrell Brown, Rachel M. Hayes, and R. Scott Marshall. 2015. "Voluntary Environmental Disclosure Quality and Firm Value: Further Evidence." *Journal of Accounting and Public Policy* 34 (4). Elsevier Inc.: 336–61. https://doi.org/10.1016/j.jaccpubpol.2015.04.004.

Qiu, Yan, Amama Shaukat, and Rajesh Tharyan. 2016. "Environmental and Social Disclosures: Link with Corporate Financial Performance." *The British Accounting Review* 48: 102–16. https://doi.org/http://dx.doi.org/10.1016/j.bar.2014.10.007.

Renneboog, Luc, Jenke Ter Horst, and Chendi Zhang. 2008. "Socially Responsible Investments: Institutional Aspects, Performance, and Investor Behavior." *Journal of Banking & Finance* 32 (9): 1723–42. https://doi.org/http://dx.doi.org/10.1016/j.jbankfin.2007.12.039.

Revelli, Christophe, and Jean-Laurent Viviani. 2015. "Financial Performance of Socially Responsible Investing (SRI): What Have We Learned? A Meta-Analysis." *Business Ethics: A European Review* 24 (2): 158–85. https://doi.org/10.1111/beer.12076.

Roberts, Robin W. 1992. "Determinants of Corporate Social Responsibility Disclosure: An Application of Stakeholder Theory." *Accounting, Organizations and Society* 17 (6): 595–612. https://doi.org/http://dx.doi.org/10.1016/0361-3682(92)90015-K.

Rodrigue, Michelle, Michel Magnan, and Emilio Boulianne. 2013. "Stakeholders' Influence on Environmental Strategy and Performance Indicators: A Managerial Perspective." *Sustainable Development, Management and Accounting: Boundary Crossing* 24 (4): 301–16. https://doi.org/http://dx.doi.org/10.1016/j.mar.2013.06.004.

Rothwell, Peter M. 2000. "Analysis of Agreement between Measurements of Continuous Variables: General Principles and Lessons from Studies of Imaging of Carotid Stenosis." *Journal of Neurology* 247 (11): 825–34. https://doi.org/10.1007/s004150070068.

Ruf, Bernadette M, Krishnamurty Muralidhar, and Karen Paul. 1998. "The Development of a Systematic, Aggregate Measure of Corporate Social Performance." *Journal of Management* 24 (1): 119–33. https://doi.org/10.1177/014920639802400101.

Schafer, H., J. Beer, J. Zenker, and P. Fernandes. 2006. "Who Is Who in Corporate Social Responsibility Rating. A Survey of Internationally Established Rating Systems That Measure Corporate Responsibility." http://www.global-ethic-now.de/gen-eng/0d_weltethos-und-wirtschaft/0d-pdf/04-verantwortung/transparenzstudie_bertelsmann.pdf.

Schaltegger, Stefan. 2011. "Sustainability as a Driver for Corporate Economic Success." *Society and Economy* 33 (1). Akadémiai Kiadó: 15–28. https://doi.org/10.1556/SocEc.33.2011.1.4.

Scherbaum, C. a., and J. M. Ferreter. 2008. "Estimating Statistical Power and Required Sample Sizes for Organizational Research Using Multilevel Modeling." *Organizational Research Methods* 12 (2): 347–67. https://doi.org/10.1177/1094428107308906.

Schueth, Steve. 2003. "Socially Responsible Investing in the United States." *Journal of Business Ethics* 43 (3). Springer Science & Business Media B.V: 189–94. http://search.ebscohost.com/login.aspx?direct=true&db=bth&AN=9699309&site=ehost-live.

Sharfman, Mark. 1996. "The Construct Validity of the Kinder, Lydenberg & Domini Social Performance Ratings Data." *Journal of Business Ethics* 15 (3). Springer Science & Business Media B.V: 287–96. http://search.ebscohost.com/login.aspx?direct=true&db=bth&AN=12134975&site=ehost-live.

Slager, Dr. Rieneke. 2009. “The FTSE4GOOD Index: Engagement and Impact.” http://www.nottingham.ac.uk/business/ICCSR/assets/The-FTSE4GOOD-index_engagement-and-impact.pdf.

SustainAbility. 2010a. “Rate the Raters: Phase One. Look Back and Current State,” no. May: 1–8. http://www.sustainability.com/library.

———. 2010b. “Rate the Raters: Phase Two. Taking Inventory of the Ratings Universe.” http://www.sustainability.com/library.

———. 2013. “Rate the Raters Phase Five Questionnaire for Raters Dow Jones Sustainability Indices February 2013.” http://www.sustainability.com/library/the-raters-response#.VhesfE2FN9A.

Sustainalytics. 2014. “Sustainalytics ESG Research on Bloomberg | Sustainalytics.” Sustainalytics Website. 2014. http://www.sustainalytics.com/sustainalytics-esg-research-now-available-bloomberg.

———. 2015. “Sustainalytics | ESG Research Methodology | Sustainalytics.” 2015. http://www.sustainalytics.com/research-methodology.

Tanner, M.A., Young, M.A. 1985. “Modeling Agreement among Raters.” *Journal of American Statistical Association* 80: 175–80.

Tanner, Martin A., and Michael A. Young. 1985. “Modeling Agreement Among Raters.” *Journal of the American Statistical Association* 80 (389): 175. https://doi.org/10.2307/2288068.

Tractenberg, Rochelle E, Futoshi Yumoto, Shelia Jin, and John C Morris. 2010. “Sample Size Requirements for Training to a Kappa Agreement Criterion on Clinical Dementia Ratings.” *Alzheimer Disease and Associated Disorders* 24 (3): 264–68. https://doi.org/10.1097/WAD.0b013e3181d489c6.

Udayasankar, Krishna. 2008. “Corporate Social Responsibility and Firm Size.” *Journal of Business Ethics* 83 (2): 167–75. https://doi.org/10.1007/s10551-007-9609-8.

Waddock, S. 2003. “Myths and Realities of Social Investing. .” *Organization and Environment* 16: 369–80.

Waddock, Sandra A, and Samuel B Graves. 1997. “The Corporate Social Performance- Financial Performance Link.” *Strategic Management Journal* 18 (4). John Wiley & Sons, Inc: 303–19. http://search.ebscohost.com/login.aspx?direct=true&db=bth&AN=12493386&site=ehost-live.

Wanderley, Lilian Soares Outtes, Rafael Lucian, Francisca Farache, and José Milton De Sousa Filho. 2008. “CSR Information Disclosure on the Web: A Context-Based Approach Analysing the Influence of Country of Origin and Industry Sector.” *Journal of Business Ethics* 82 (2): 369–78. https://doi.org/10.1007/s10551-008-9892-z.

Watson, Luke. 2015. “Corporate Social Responsibility, Tax Avoidance, and Earnings Performance.” *The Journal of the American Taxation Association* 37 (2): 1–21. https://doi.org/10.2308/atax-51022.

Watson, P.F., and A. Petrie. 2010. “Method Agreement Analysis: A Review of Correct Methodology.” *Theriogenology* 73: 1167–79.

Weber, J, and J Gladstone. 2014. “Rethinking the Corporate Financial–Social Performance Relationship: Examining the Complex, Multistakeholder Notion of Corporate Social Performance.” *Business and Society Review* 119 (3): 297–336.

Wilbert, C. 2006. “Campaign to Stop Killer Coke | Breaking News Archives | 2006 | Social Responsibility of Coca-Cola Questioned; Giant Retirement Fund Decides to Sell Shares.” 2006. http://killercoke.org/article_ajc060719.php.

Wildemuth, Barbara M. 2006. *Applications of Social Research Methods to Questions in Information and Library Science*. 2nd ed. Santa Barbara, California: Libraries Unlimited.

Willis, Alan. 2003. “The Role of the Global Reporting Initiative’s Sustainability Reporting Guidelines in the Social Screening of Investments.” *Journal of Business Ethics* 43 (3). Springer Science & Business Media B.V: 137–233. http://search.ebscohost.com/login.aspx?direct=true&db=bth&AN=9699334&site=ehost-live.

Yates-Smith, Chris. 2013. “Socially Responsible Investment: Good Corporate Citizenship or Hidden Portfolio Risk?” *Law & Financial Markets Review* 7 (2). Hart Publishing Ltd: 112–17. https://doi.org/10.5235/17521440.7.2.112.

Yunwook, K., and K. Soo-yeon. 2010. “The Influence of Cultural Values on Perceptions of Corporate Social

Responsibility: Application of Hofstede's Dimensions to Korean Public Relations Practitioners." *Journal of Business Ethics* 91 (4): 485–500.

Ziegler, Andreas, and Michael Schröder. 2010. "What Determines the Inclusion in a Sustainability Stock Index?: A Panel Data Analysis for European Firms." *Special Section: Coevolutionary Ecological Economics: Theory and Applications* 69 (4): 848–56. https://doi.org/http://dx.doi.org/10.1016/j.ecolecon.2009.10.009.

TABLES

TABLE 1: Distribution of firms by industry and bourse

Industry	Cac40	Dax	FTSE350	Ibex	SP500	Total
Basic Materials	2	4	20	2	20	48
Communications	4	1	15	2	17	39
Consumer, Cyclical	7	6	47	2	47	109
Consumer, Non-cyclical	5	6	51	2	70	134
Energy	2		13	2	34	51
Financial	5	5	56	7	61	134
Industrial	8	3	36	8	55	110
Technology	1	2	7	2	39	51
Utilities	3	2	7	5	27	44
Total	**37**	**29**	**252**	**32**	**370**	**720**

TABLE 2: Summary statistics of all firms across Environmental, Social and Governance categories

	Overall		Social	Community	Employee	Environment		Governance	
All firms	Bloomberg	CSR hub	Bloomberg	CSR Hub	CSR Hub	Bloomberg	CSR hub	Bloomberg	CSR hub
Minimum	12.8	41.0	3.3	37.0	36.0	1.4	29.0	17.9	35.0
Maximum	72.6	75.0	75.4	75.0	89.0	76.0	80.0	85.7	73.0
Mean	35.3	61.0	35.6	57.2	64.0	25.7	64.1	56.6	57.2
Std dev.	13.2	6.6	16.2	7.8	9.2	17.3	7.4	7.7	7.7

TABLE 3: Lin's concordance correlation coefficients on all firms (continuous data)

	Overall	Community /Social	Employee/ Social	Environment	Governance
All firms	0.108	0.149	0.149	0.066	0.066
Panel 1: by country					
S&P	0.086	0.079	0.090	0.049	0.196
FTSE	0.047	0.048	0.072	0.041	0.019
DAX	0.138	0.288	0.153	0.060	0.268
IBEX	0.044	0.200	0.166	- 0.016	0.011
CAC	0.059	0.105	0.047	0.054	0.104
European	0.072	0.114	0.129	0.065	- 0.044
Panel 2: by industry					
Basic materials	0.143	0.205	0.105	0.080	0.161
Communications	0.120	0.174	0.206	0.070	0.013
Consumer cyclical	0.092	0.134	0.144	0.059	0.062
Consumer non-cyclical	0.095	0.140	0.123	0.070	0.134
Energy	0.256	0.250	0.209	0.161	0.352
Financial	0.097	0.132	0.143	0.048	0.135
Industrial	0.099	0.162	0.141	0.059	0.169
Technology	0.072	0.119	0.126	0.041	0.089
Utilities	0.091	0.173	0.109	0.055	0.066
Panel 3: by firm size (tertiles)					
Large	0.167	0.198	0.168	0.096	0.330
Medium	0.129	0.165	0.190	0.070	0.227
Small	0.052	0.077	0.074	0.037	0.053

TABLE 4: Intraclass correlation coefficient (ICC) (continuous data (average measures))

Scale data	Overall	Community /Social	Employee /Social	Environment	Governance
All firms	0.195	0.260	0.260	0.124	0.227
Panel A: By country					
S&P	0.158	0.166	0.940	0.125	0.254
FTSE	0.091	0.093	0.134	0.078	0.037
DAX	0.249	0.455	0.272	0.170	0.431
IBEX	0.086	0.340	0.292	0.330	0.023
CAC	0.115	0.194	0.093	0.105	0.193
EU	0.134	0.204	0.229	0.123	0.167
Panel B: by industry					
Basic materials	0.254	0.345	0.193	0.150	0.282
Communications	0.218	0.302	0.347	0.134	0.026
Consumer cyclical	0.169	0.238	0.254	0.112	0.112
Consumer non-cyclical	0.175	0.247	0.220	0.131	0.238
Energy	0.175	0.247	0.220	0.131	0.238
Financial	0.178	0.235	0.251	0.092	0.240
Industrial	0.181	0.280	0.249	0.113	0.291
Technology	0.136	0.216	0.227	0.079	0.167
Utilities	0.170	0.299	0.200	0.106	0.127
Panel C: by firm size (tertile)					
Large	0.287	0.331	0.288	0.176	0.497
Medium	0.230	0.285	0.320	0.131	0.371
Small	0.099	0.143	0.138	0.073	0.102

TABLE 5: Crosstab of Overall category

		Bloomberg overall			
		Low	Medium	High	Total
CSR Hub overall	Low	136	42	22	200
	Medium	77	116	81	274
	High	26	81	139	246
Total		239	239	242	720

TABLE 6: Summary table of Cohen's kappa scores

	Overall	Community/ Social	Employee/ Social	Environment	Governance
All firms	.315***	.234***	.331***	.238***	.075**
Panel A: By bourse					
S&P	.278***	.159***	.181***	.204***	.011
FTSE	.274***	.066***	.175***	.238***	-.007
DAX	.379**	.171	.118	.274*	.219*
IBEX	.250*	.109	.155	.201	.098
CAC	.065	.148	.309*	.025	-.027
EU	.250***	.117**	.213***	.245***	.001
Panel B: by industry					
Basic materials	.322***	.239***	.335***	.234**	.071*
Communications	.331**	.301*	.386**	.413***	.047
Consumer cyclical	.202**	.302***	.342***	.074	-.003
Consumer non-cyclical	.261***	.236***	.249***	.206**	.068
Energy	.343***	.180**	.365***	.171*	.176*
Financial	.383***	.267***	.415***	.214***	.105*
Industrial	.429***	.198**	.333***	.359***	.089
Technology	.303***	.199**	.353***	.235**	.056
Utilities	.293***	.222**	.155	.313**	.024
Panel C: by firm size (tertile)					
Large	.337***	.297***	.336***	.231***	.180***
Medium	.346***	.342***	.435***	.286***	.073
Small	.250***	.060	.179***	.220***	.046

***Significant at p<.001, ** at p<.005, * at p<0.1 (2 tailed)