ESG Confusion and Stock Returns: Tackling the Problem of Noise*

Florian Berg†, Julian F. Koelbel‡, Anna Pavlova§, and Roberto Rigobon¶

November 19, 2021

Abstract

How strongly does ESG (environmental, social and governance) performance affect stock returns? Answering this question is difficult because existing measures of performance, ESG ratings, are noisy. To tackle the bias, we propose a noise-correction procedure, in which we *instrument* ESG ratings with ratings of other ESG rating agencies, as in the classical errors-in-variables problem. The corrected estimates demonstrate that the effect of ESG performance on stock returns is stronger than previously estimated; the standard regression estimates of ESG ratings' impact on stock returns are biased downward by about 60%. Our dataset includes scores of eight ESG rating agencies for firms located in North America, Europe, and Japan. We determine which agencies' scores are valid instruments (not all of them are) and estimate the noise-to-signal ratio for each ESG rating agency (some of which are very large). Overall, our results suggest that it is advantageous to rely on several complementary ratings. In our sample, stocks with higher ESG performance have higher expected returns. Our model provides several explanations for this finding.

*Florian Berg and Roberto Rigobon are grateful to Massachusetts Pension Reserves Investment Management Board, AQR Capital Management (Applied Quantitative Research), and MFS Investment Management — members of the Aggregate Confusion Project Council – for their generous support of this research. We are also extremely grateful to the ESG rating agencies that provided their data to this project. We thank seminar participants at London Business School, MIT Sloan, University of Amsterdam, University of Mannheim, and University of Michigan for excellent feedback. Julian serves as a member of the RepRisk Academic Advisory Council. All remaining errors are ours. Correspondence to: Roberto Rigobon, Sloan School of Management, MIT, 50 Memorial Drive, E62-520, Cambridge, MA 02142-1347, aggregateconfusion@mit.edu, tel: (617) 258 8374.

†MIT Sloan School of Management; fberg@mit.edu

‡University of Zurich and MIT Sloan School of Management; julian.koelbel@uzh.ch

§London Business School and Centre for Economic Policy Research (CEPR); apavlova@london.edu.

¶MIT Sloan School of Management and National Bureau of Economic Research (NBER); rigobon@mit.edu

1 Introduction

ESG (environmental, social, and governance) investing has taken the asset management industry by storm. In the U.S., assets under management of ESG funds increased by 42% between 2018 and 2020, while in Europe flows into ESG funds doubled in 2020 alone.[1] The US Forum for Sustainable and Responsible Investing classified assets of $17.1 trillion as ESG investments at the end of 2020, which accounts for 33% of total U.S. assets under management.[2] This unprecedented demand for assets with superior ESG performance should boost their prices. While theoretical studies suggest such an effect[3], the empirical evidence is mixed.

We believe that one of the key reasons confounding the relationship between ESG attributes and stock returns is that available measures of ESG performance are noisy. These measures are provided by ESG rating agencies,[4] and there is significant disagreement in their ESG assessments.[5] ESG ratings play a crucial role in measuring a firm's ESG attributes, guiding the investment of ESG funds, and thus linking investor preferences for ESG to portfolio choices. Our goal is to disentangle signal from noise in ESG ratings and to uncover the true impact of ESG performance on expected stock returns.

In this paper, we propose a simple model that establishes a relationship between ESG performance and stock returns and show that the noisier the measurement of ESG performance, the lower the sensitivity of stock returns to ESG performance. We show that the latter result implies that regression estimates of the relationship between stock returns and noisy measures of ESG performance would be biased towards zero; the noisier the measurement, the larger the bias. To tackle the bias, we develop a noise-correction procedure based on an instrumental variable approach. Specifically, when measuring the effects of ESG on stock returns, we instrument a given ESG rating agency's score with other rating agencies' scores as in the classical errors-in-variables problem. We conduct this analysis for the eight largest ESG rating agencies. We find that on average the standard regression estimates of ESG ratings' impact on stock returns are biased downward by more than 60%. We determine which agencies' scores are valid instruments (not all of them are) and estimate the

[1] See https://www.morningstar.com/content/dam/marketing/emea/uk/European_ESG_Fund_Landscape_2020.pdf, accessed April 8, 2021.

[2] See https://www.ussif.org/fastfacts, accessed August 8, 2021. This is an upper bound on ESG assets, as most of these assets are self-reported and come from outside the mutual fund sector. Mutual funds' ESG assets under management, covered by Morningstar, are significantly lower.

[3] Heinkel, Kraus and Zechner (2001); Pastor, Stambaugh and Taylor (2021*b*); Fama and French (2007)

[4] The ESG rating agencies in our dataset include Truvalue Labs (owned by FactSet), RepRisk (independent), MSCI IVA (owned by MSCI), Sustainalytics (owned by Morningstar), Refinitiv (formerly known as Asset4), Vigeo-Eiris (owned by Moody's), ISS ESG (majority stake owned by Deutsche Boerse), and SP Global CSA (formerly known as RobecoSAM).

[5] See Berg, Kölbel and Rigobon (2020) for a comprehensive study of ratings disagreement.

noise-to-signal ratio for each ESG rating agency (some of which are very large).

We start with a simple model that justifies our empirical strategy. There are two types of investors: traditional and ESG-conscious. The former care only about a firm's cash flow, whereas the latter care additionally about ESG performance of their portfolio holdings. The ESG attribute they care about is non-pecuniary and is uncorrelated with the firm's cash flow. We assume further that information about this ESG attribute is contained in a noisy ESG signal, provided by a rating agency. We derive stock prices in closed form and show that stocks with better ESG performance have higher prices and lower expected returns. We demonstrate that the noisier the ESG signal, the lower its effect on stock prices. This dampening effect is similar to the attenuation bias arising in OLS regressions because of a measurement error in a regressor. The regression of interest in our case is the regression of stock returns on ESG scores, which are noisy signals of true (unobserved) ESG performance.

We use an instrumental variable approach to tackle the measurement error problem and correct the attenuation bias. Specifically, we propose to instrument a rating of one agency by the ratings of other agencies for the same attribute. Standard regressions then need to be replaced by two-stage least squares (2SLS) regressions. For each region and each rater in our sample, we run the Hausman specification tests and show that the majority of OLS specifications are rejected in favour of 2SLS ones, confirming that the standard OLS estimates are biased.

We document that the effect of ESG performance on stock returns is much stronger—the coefficients on average more than double—when we replace the standard OLS regression by 2SLS. This result is consistent with the prediction of the theory that the bias we see in the standard regressions is indeed an attenuation bias. This could well be the reason why many studies do not find an effect of ESG performance on stock returns.

In our empirical analysis, we first show that each ESG rating agency score is well predicted be a combination of other rating agencies' scores. This however, is not sufficient to guarantee that ESG scores of every rater are valid instruments in our analysis. To test for instrument validity, we conduct overidenifying restrictions tests. We have a total of 8 ESG rating agencies in our sample, which gives us multiple overidentifying restrictions we can test. Based on these tests, we develop a pruning procedure to determine which instruments are to be included in and which ones should be excluded from the estimation. We find that many but not all of ESG raters pass the overidentifying restrictions test. The failure of the overidentifying restrictions test means that a candidate instrument is "endogenous," which could happen, for example, if ESG scores are backfilled retroactively by a provider, if scores of one ESG rater are influenced by another, if there are non-linearities in the way rating agencies aggregate their measurement of individual attributes into the overall ESG scores, or if measurement errors

are correlated across rating agencies. We get especially many rejections of the overidentifying restrictions test for ESG ratings of companies in North America,suggesting that some of the above possibilities are borne out in the data. In contrast, for European firms, all ESG raters' scores pass the test. We find that the instruments that are rejected least often are the ESG ratings of Refinitiv, RepRisk, S&P Global, and Truvalue Labs.

The size of the attenuation bias is a measure of noise in the standard (OLS) regression estimates. This bias varies across regions in our sample and across rating agencies. We find that, on average, the noise-to-signal ratio is about 60%. Our results show that the ratings with the smallest proportion of noise are those from MSCI, Sustainalytics, and ISS.

It is important to highlight that all raters' scores are valuable. Disregarding scores of some raters amounts to throwing away valuable information about the imperfectly measured ESG attributes. The set of valid instruments (other rating agencies' scores) identified by our procedure never includes fewer than four raters and, in many cases, includes all rating agencies. By combining information from several complementary ratings, one can get a more precise estimate of the impact of ESG performance on stock returns. It is also important to mention that our procedure does not produce or estimate a less noisy ESG score. Strictly speaking, our procedure only solves the problem of noise in estimating the relationship between ESG performance and stock returns. The impact of noise on the relationship between ESG performance and other variables (e.g., accounting measures of performance) might be identified with a similar procedure but it is not addressed here.

We now turn to the economic interpretation of the *positive* estimates of ESG performance on stock returns that we document in our sample. Our model illustrates the following two channels through which ESG ratings affect stock returns. First, ESG-conscious investors bid up prices of stocks with higher ESG scores and reduce the cost of capital of these firms. So higher ESG scores should be associated with lower expected returns. Second, (unexpected) inflows of new investors into stocks with high ESG scores results in higher returns on these stocks during the period over which inflows are observed.[6] The two channels work in opposite directions. In our estimation, we find strong evidence that the second channel dominates in our sample. That is, firms with higher ESG scores have higher stock returns. We predict that this effect would reverse, once the inflows into ESG funds stabilize and become easier to forecast.[7]

We then explore what exactly our measure of noise captures. Our estimation is done

[6]There is yet a third channel through which ESG scores affect expected stock returns, via altering cash flow risk. At the time of writing of this paper, the literature has not converged on the sign of this effect. We abstract away from it in this paper and focus on the effects of ESG due entirely to the emergence of investor clientele with a preference for ESG investing.

[7]In a recent paper, Pastor, Stambaugh and Taylor (2021*a*) advocate a similar view.

at the level of aggregate ESG scores, the key measure of ESG performance provided by ESG rating agencies. This is the main measure that fund managers take into account when constructing and benchmarking their portfolios and it therefore has the most direct link to stock returns. However, an ESG score is a complicated aggregate measure. It is a weighted average of many indicators (e.g., CO2 emissions, labor practices, etc), with weights that differ across rating agencies. We can decompose the overall noise in a rating into the (i) noise in measurement of individual indicators and (ii) noise in weights. The former is a pure measurement error. The latter reflects the fact that the weights that the representative investor puts on different attributes are unknown, and ESG rating agencies attempt to replicate these weights. This weighting, however, is imperfect, and so weight discrepancies across raters also contribute to the errors-in-variables problem that we are addressing. We show that, under certain conditions, both types of measurement errors can be corrected with our procedure. Moreover, the overidentifying restrictions tests are able to weed out instruments (rating agencies' scores) that do not satisfy the required conditions.

We demonstrate the robustness of our empirical results to the alternative econometric procedure that selects instruments to be included in the estimation. There is no established procedure (as far as the authors know) for determining the optimal set of instruments in the presence of many instruments, especially when some of them may be invalid. In our first (pruning) procedure, we chose the instruments by starting from the largest possible set and pruning instruments one at a time until the model passes the Hansen test of overidentifying restrictions. The main problem with this procedure is that the Hansen test has not being designed for this sequential search. For robustness, we implement a different approach. Although it has other weaknesses, the approach does not rely on a sequential search based on the Hansen overidentifying restrictions test. Instead, it uses a very simple Lasso selection. The resulting estimates mirror those from our main procedure. It is reassuring that there is a strong correspondence between the two sets of results.

Finally, we present a simulation that compares our main noise-correction procedure to commonly used alternative approaches such as constructing a simple average of ESG scores as a regressor and the principal component analysis. We highlight a special case in which these alternative approaches are effective. In the general case, however, our procedure performs significantly better than the alternatives. The simulation also reveals that the overidentifying restrictions test that we use to weed out invalid instruments has significant power in our setting.

Our paper is related to several strands of literature. First, it is related to asset pricing models that incorporate ESG investors who push up asset prices of green firms and lower their cost of capital (Heinkel, Kraus and Zechner, 2001; Friedman and Heinle, 2016; Oehmke

and Opp, 2019; Broccardo, Hart and Zingales, 2020; Landier and Lovo, 2020; Kashyap et al., 2021; Pastor, Stambaugh and Taylor, 2021*b*). In a closely related paper Avramov et al. (2021) look at uncertainty about the ESG profile in equilibrium asset pricing.

Second, there are many studies that have explored the link between ESG performance and stock returns empirically. The evidence is not conclusive; studies report both higher stock returns for ESG performers (Edmans, 2011; Khan, Serafeim and Yoon, 2016; Lins, Servaes and Tamayo, 2017; Albuquerque, Koskinen and Zhang, 2019) as well as lower stock returns (Chava, 2014; El Ghoul et al., 2011; Bolton and Kacperczyk, 2020). Our paper does not resolve this issue, but it adds an important point, namely, that in whichever direction the relationship is, noisy measurement will tend to attenuate the effect.

Third, our paper is related to empirical studies in finance and accounting that have explored the relationship between corporate governance and stock returns (Gompers, Ishii and Metrick, 2003; Bauer, Guenster and Otten, 2004; Adams and Ferreira, 2009; Bebchuk, Cohen and Ferrell, 2009). This literature has addressed the problem of how corporate governance ought to be measured by suggesting several alternative ways of measuring (e.g. Larcker, Richardson and Tuna, 2007; Daines, Gow and Larcker, 2010; Larcker, Reiss and Xiao, 2015). Our instrumental variable approach offers an innovative way to examine the relationship between corporate governance and stock returns when there are competing ways of measuring it.

Finally, our paper is related to the emerging literature on ESG rating divergence (Chatterji et al., 2016; Berg, Kölbel and Rigobon, 2020; Christensen, Hail and Leuz, 2021; Christensen, Serafeim and Sikochi, 2021). A closely related working paper is Gibson, Krueger and Schmidt (2021), who find that ESG rating divergence at the stock level increases stock returns. Our paper shows that ESG rating divergence diminishes the fundamental effect that ESG ratings have on stock returns.

2 Model

In this section, we present a simple model with traditional and ESG investors. Our focus is on an ESG signal, measured with noise. We will show that such measurement error leads to bias in a standard regression analysis. The noisier the ESG signal, the larger the bias.

2.1 The Economic Environment

We consider a two-period model, with $t = 0, 1$. Investment opportunities are represented by a risky stock of a single firm and a risk less bond, with the risk free rate normalized to zero.[8] The stock is a claim to the cash flow $D \sim N(\overline{D}, \sigma_D^2)$ per share, with D realized in period 1. The stock is in fixed supply of $\overline{\theta}$ shares and the riskless bond is in infinite net supply. We denote the stock price in period t by S_t, where $S_1 = D$.

There is a measure λ of ESG investors and $1 - \lambda$ of traditional investors. Both types of agents invest their funds into the stock and the bond. The ESG and traditional investors' portfolio allocation to the stock is θ^i, where $i = ESG, T$, respectively. The period-1 wealth of the investors W_1^i is then $W_0^i + \theta^i(D - S_0)$, where W_0^i is their initial wealth, $i = ESG, T$. ESG investors derive a non-pecuniary benefit Y per share from holding the stock, with $Y \sim N(\overline{Y}, \sigma_Y^2)$ independent from D. Their utility is exponential, $U(W_1, Y) = -exp(-\gamma(W_1 + \theta^{ESG}Y))$.[9] We think of Y as an ESG externality, generated by the firm, which ESG investors internalize. The traditional investors have utility $U(W_1) = -exp(-\gamma W_1)$ and do not internalize any ESG externalities. Investors' initial endowments are in terms of shares of the stock and bond and they choose their portfolios to maximize their expected utilities.

In period 0, investors receive noisy signals, s_D and s_Y, about cash flows and ESG benefit, D and Y, respectively,

$$s_D = D + \epsilon_D, \tag{1}$$

$$s_Y = Y + \epsilon_Y, \tag{2}$$

where $\epsilon_i \sim N(0, \sigma_{\epsilon_i}^2),\ i = D, Y$ are independent of each other and independent of D and Y.

2.2 Equilibrium

To solve for equilibrium, we first need to solve the inference problem of the investors. Exploiting the joint normality of random variables in our economy, we arrive at the following lemma (all proofs are in the Appendix).

[8] It is straightforward to extend the model to multiple risky stocks.

[9] Our approach to modeling ESG investors is similar to Pastor, Stambaugh and Taylor (2021*b*) and Friedman and Heinle (2016).

Lemma 1 *The mean and variance of D, conditional on signal s_D, are given by*

$$E(D|s_D) = \overline{D} + \beta(s_D - \overline{D}) = \overline{D} + \frac{\sigma_D^2}{\sigma_D^2 + \sigma_{\epsilon_D}^2}(s_D - \overline{D}), \tag{3}$$

$$Var(D|s_D) = \sigma_{\nu_D}^2 = \frac{\sigma_D^2 \sigma_{\epsilon_D}^2}{\sigma_D^2 + \sigma_{\epsilon_D}^2}. \tag{4}$$

The mean and variance of Y, conditional on signal s_Y, are as follows:

$$E(Y|s_Y) = \overline{Y} + \beta(s_Y - \overline{Y}) = \overline{Y} + \frac{\sigma_Y^2}{\sigma_Y^2 + \sigma_{\epsilon_Y}^2}(s_Y - \overline{Y}), \tag{5}$$

$$Var(Y|s_Y) = \sigma_{\eta_Y}^2 = \frac{\sigma_Y^2 \sigma_{\epsilon_Y}^2}{\sigma_Y^2 + \sigma_{\epsilon_Y}^2}. \tag{6}$$

We are now able to solve for optimal portfolios of ESG and traditional investors. These portfolios are given by

Lemma 2 (Portfolio Choice) *The investors portfolio demands are*

$$\theta^T = \frac{1}{\gamma} \frac{E(D|s_D) - S_0}{Var(D|s_D)}, \tag{7}$$

$$\theta^{ESG} = \frac{1}{\gamma} \frac{E(D|s_D) + E(Y|s_Y) - S_0}{Var(D|s_D) + Var(Y|s_Y)}. \tag{8}$$

The traditional investors hold the standard mean-variance portfolio, which optimally trades off risk (the denominator) and expected return (the numerator). In contrast, ESG investors account for ESG characteristics in their portfolio choice. The higher the stock's expected ESG benefit Y, the more shares of it ESG investors are willing to include in their portfolio. However, since ESG investors are risk-averse, the perceived risk of the stock is higher for them relative to traditional investors. This additional risk is driven by the noise in ESG ratings—the higher this noise, the less of the stock ESG investors are willing to hold (see the denominator of the portfolio demand in (8)).

The market clearing condition requires that investors' demand for the stock equals its supply, i.e.,

$$\lambda \theta^{ESG} + (1 - \lambda)\theta^T = \overline{\theta}. \tag{9}$$

To solve for the equilibrium stock price, we substitute the optimal portfolios from Lemma 2 into the market clearing condition (9). We report the resulting period-0 stock price in the following proposition.

Proposition 1 (Asset Prices) *The period-0 stock price is given by*

$$S_0 = \overline{D} + \frac{\sigma_D^2}{\sigma_D^2 + \sigma_{\epsilon_D}^2}(s_D - \overline{D}) + A\lambda \frac{\sigma_D^2 \sigma_{\epsilon_D}^2}{\sigma_D^2 + \sigma_{\epsilon_D}^2}\left[\overline{Y} + \frac{\sigma_Y^2}{\sigma_Y^2 + \sigma_{\epsilon_Y}^2}(s_Y - \overline{Y})\right] \tag{10}$$

$$- A\gamma\overline{\theta}\frac{\sigma_D^2 \sigma_{\epsilon_D}^2}{\sigma_D^2 + \sigma_{\epsilon_D}^2}\left[\frac{\sigma_D^2 \sigma_{\epsilon_D}^2}{\sigma_D^2 + \sigma_{\epsilon_D}^2} + \frac{\sigma_Y^2 \sigma_{\epsilon_Y}^2}{\sigma_Y^2 + \sigma_{\epsilon_Y}^2}\right], \tag{11}$$

where $A = \left[\frac{\sigma_D^2 \sigma_{\epsilon_D}^2}{\sigma_D^2 + \sigma_{\epsilon_D}^2} + (1-\lambda)\frac{\sigma_Y^2 \sigma_{\epsilon_Y}^2}{\sigma_Y^2 + \sigma_{\epsilon_Y}^2}\right]^{-1}$.

The ESG attribute Y does not affect fundamentals (i.e., the firm's cash flow D). However, it affects asset prices because there is a group of investors that care about it. A positive signal s_Y about the ESG attribute Y boosts the stock price. Y can be interpreted as the true ESG attribute, and s_Y as what the ESG rating agencies measure — their scores.

Suppose that the stock is a green stock, which appeals to ESG investors, i.e., $\overline{Y}$ is positive and sufficiently high. Then, relative to an economy with no ESG investors, the stock price will be higher, reflecting the additional benefit to ESG investors from holding a green stock. The mass of ESG investors λ is another important parameter. The higher the mass of ESG investors, the higher the stock price.

Let us now examine a *realized* per-share return on the stock in period 0:

$$S_0 - S_{-1} = \overline{D} - S_{-1} + \frac{\sigma_D^2}{\sigma_D^2 + \sigma_{\epsilon_D}^2}(s_D - \overline{D}) + A\lambda \frac{\sigma_D^2 \sigma_{\epsilon_D}^2}{\sigma_D^2 + \sigma_{\epsilon_D}^2}\left[\overline{Y} + \frac{\sigma_Y^2}{\sigma_Y^2 + \sigma_{\epsilon_Y}^2}(s_Y - \overline{Y})\right]$$

$$- A\gamma\overline{\theta}\frac{\sigma_D^2 \sigma_{\epsilon_D}^2}{\sigma_D^2 + \sigma_{\epsilon_D}^2}\left[\frac{\sigma_D^2 \sigma_{\epsilon_D}^2}{\sigma_D^2 + \sigma_{\epsilon_D}^2} + \frac{\sigma_Y^2 \sigma_{\epsilon_Y}^2}{\sigma_Y^2 + \sigma_{\epsilon_Y}^2}\right]. \tag{12}$$

We think about the constant S_{-1} as the value of the stock one period before ESG investors (unexpectedly) arrived in the market. The realized returns on the stock depends on the magnitude of (unanticipated) ESG investor inflows, captured by λ, with the inflows boosting returns of green stocks. In contrast, stock returns of brown firms (firms with a sufficiently negative ESG benefit Y) fall.

We now present the expression for the *expected* per-share return on the stock.

$$E(D) - S_0 = -\frac{\sigma_D^2}{\sigma_D^2 + \sigma_{\epsilon_D}^2}(s_D - \overline{D}) - A\lambda \frac{\sigma_D^2 \sigma_{\epsilon_D}^2}{\sigma_D^2 + \sigma_{\epsilon_D}^2}\left[\overline{Y} + \frac{\sigma_Y^2}{\sigma_Y^2 + \sigma_{\epsilon_Y}^2}(s_Y - \overline{Y})\right]$$

$$+ A\gamma\overline{\theta}\frac{\sigma_D^2 \sigma_{\epsilon_D}^2}{\sigma_D^2 + \sigma_{\epsilon_D}^2}\left[\frac{\sigma_D^2 \sigma_{\epsilon_D}^2}{\sigma_D^2 + \sigma_{\epsilon_D}^2} + \frac{\sigma_Y^2 \sigma_{\epsilon_Y}^2}{\sigma_Y^2 + \sigma_{\epsilon_Y}^2}\right]. \tag{13}$$

The higher the ESG signal s_Y, the *lower* the expected return on the stock. This is

because in our model the firm's cash flow D is fixed and therefore the more ESG investors push up the stock price in response to a high ESG signal, the lower the stock's expected return going forward. That is, the effects of ESG on stock prices in our model manifest themselves entirely through the cost of capital channel.[10]

Both the expected and realized returns also depend on the noise in the ESG signal, σ_{ϵ_Y}. The noisier the signal s_Y, the lower its passthrough to stock returns. In the limit of $\sigma_{\epsilon_Y} \to \infty$, there will be no effect of s_Y of stock returns. These observations will become important in our empirical analysis, which uses data on ESG signals, measured with noise. The noise in the reported ESG signals is apparent from the discrepancies in measuring the same signal by different ESG ratings providers, and in our empirical work we treat this as a classical errors-in-variables problem.

In sum, our model delivers the following testable predictions.

Prediction 1: The higher the ESG signal (or rating) s_Y, the higher the stock price and the lower the stock's expected return.

Prediction 2: (Unanticipated) inflows of ESG investors (an increase in the mass λ) lead to higher returns on green (high Y) stocks.

Prediction 3: The noisier the ESG signal s_Y, the lower its impact on stock returns.

3 ESG Discrepancies and Expected Stock Returns

In this section, we describe our data and develop an instrumental variable approach that disentangles signal from noise in ESG ratings with regard to stock returns.

3.1 Data

The ESG rating agencies included in our dataset are shown in Table 1. The analysis is performed over the time period from 2014 to 2020, the starting point being determined by Sustainalytics' data, which only starts in 2014. All ratings are organized in a way that the higher the scores the better the ESG performance, i.e. we flipped the signs of the RepRisk and Sustainalytics scores, which are designed to measure risks.

Some ratings have changed their methodology during the study period. For example,

[10] We abstract away from the cash flow risk channel by assuming that the stock's future cash flow D is uncorrelated with the ESG characteristic Y. However, it is entirely possible that firms with low ESG performance are riskier than their greener counterparts (e.g., regulation risk) and therefore their expected returns are higher.

Sustainalytics changed its methodology in December 2018, and MSCI updated its methodology in 2017. More problematically, some raters may have retro-actively changed their scores, which has been shown in the case of Refinitiv (Berg, Fabisik and Sautner, 2021). Also Sustainalytics has provided us with a dataset that is simulated backwards from 2018, based on their new methodology. It is possible that ESG rating data that are not point-in-time have a potential look-ahead bias. We address this issue as part of our methodology.

Table 1. Rater Overview. This table shows the names of the data providers, alternative names or ownership, and the exact name of the scores used in the analysis.

Rater Name	Owner/Alternative Name	Score Name
ISS ESG	Majority stake owned by Deutsche Boerse	Numeric ESG Overall Rating
MSCI	MSCI Inc.	IVA Industry Weighted score
Refinitiv	Formerly known as Asset4	TRESG Score
RepRisk	RepRisk AG	Reputation Risk Index (RRI)
Sustainalytics	Owned by Morningstar	ESG Risk Rating
S&P Global CSA	Formerly owned by RobecoSAM	ESG Score
Truvalue Labs	Owned by FactSet	Insight Score
Vigeo-Eiris	Owned by Moody's ESG Solutions	Global Score

Financial data comes from Thomson Reuters Worldscope. We downloaded monthly return data, the price-to-book ratio, the firm's beta, the firm's total assets, the firm's ebit-over-total-assets ratio, and the firm's size. We use the return data to calculate the 12 month momentum as well as 12 month volatility for each stock.

Table 2 shows the correlations between ESG scores. Correlations range from -0.57 between the pair of Refinitiv and RepRisk in Europe to 0.71 between the pair Vigeo-Eiris and Refinitiv in Japan. RepRisk scores stand out for being correlated negatively with most other scores, suggesting that this rater employs a methodology that is markedly distinct from that of others. The average pair-wise correlations are fairly similar across the three regions.

3.1.1 ESG Ratings Methodologies and Data Sources

ESG rating agencies offer a commercial service to investors by providing third-party assessments of firms' ESG performance.[11] ESG stands for environmental, social, and governance performance of companies but serves as an umbrella term for a large number of more specific attributes, such as carbon emissions, water consumption, labor relations, respect for human rights, etc.[12]

[11] ESG ratings are in some ways comparable to credit ratings, yet an important difference is that investors usually pay for access to the ratings, rather than companies paying for the assessment.

[12] See Berg, Kölbel and Rigobon (2020) for a comprehensive list.

Table 2. Correlations between ESG Rating Agencies' ESG Scores. This table shows the correlations of all three subsamples: North America, Europe, and Japan. We use MSCI's IVA Industry Weighted score, Sustainalytics' ESG Risk Ratings, Refinitiv's TRESG score, RepRisk's Reputation Risk Index (RRI), Truvalue Labs' Insight Score, Vigeo-Eiris's Global score, S&P Global's ESG score, and ISS's Numeric ESG Overall Rating. We flipped the sign for Sustainalytics and RepRisk so that a higher value corresponds to a better ESG performance.

	MSCI	ISS	RepRisk	TVL	Vigeo-Eiris	SP Global	Refinitiv
North America							
MSCI	1						
ISS	0.42	1					
RepRisk	-0.09	-0.33	1				
TVL	0.23	0.14	0.10	1			
Vigeo-Eiris	0.44	0.68	-0.39	0.12	1		
SP Global	0.39	0.57	-0.40	0.09	0.67	1	
Refinitiv	0.40	0.63	-0.45	0.10	0.70	0.65	1
Sustainalytics	0.24	0.14	0.06	-0.03	0.12	0.12	0.21
Europe							
MSCI	1						
ISS	0.46	1					
RepRisk	0.01	-0.27	1				
TVL	0.20	0.13	0.23	1			
Vigeo-Eiris	0.41	0.68	-0.44	0.08	1		
SP Global	0.29	0.54	-0.52	-0.02	0.68	1	
Refinitiv	0.31	0.59	-0.57	-0.04	0.69	0.68	1
Sustainalytics	0.42	0.27	0.26	0.09	0.22	0.13	0.08
Japan							
MSCI	1						
ISS	0.38	1					
RepRisk	0.02	-0.22	1				
TVL	0.10	0.06	0.12	1			
Vigeo-Eiris	0.42	0.62	-0.29	0.10	1		
SP Global	0.36	0.57	-0.35	0.04	0.65	1	
Refinitiv	0.34	0.57	-0.36	0.09	0.71	0.64	1
Sustainalytics	0.25	0.25	0.08	0.05	0.27	0.34	0.28

Different ESG raters provide diverging ESG ratings, as prior research has shown (Chatterji et al., 2016; Berg, Kölbel and Rigobon, 2020) and as the correlations in Table 2 confirm. The reason is that each ESG rating is generated by a unique methodology. Specifically, these methodologies differ due to different ways of aggregating ESG attributes, and due to different ways of measuring ESG attributes (Berg, Kölbel and Rigobon, 2020).

Regarding aggregation, ESG rating agencies decide which attributes should be evaluated as part of their scoring procedure, and how important they are relative to each other. There is, as of now, no generally accepted standard of what constitutes ESG. The list of relevant attributes usually includes attributes such as green house emissions, product safety, or labor practices, but can also include less obvious attributes such as electromagnetic radiation, management of systemic risks, or whether top management has monetary incentives to meet ESG targets. The weight of these attributes can also differ, and in many cases weights are industry specific and according to a proprietary methodology. ESG raters attempt to aggregate ESG attributes in a way that is consistent with what a representative ESG investor cares about. Different assumptions about the preferences of the representative ESG investor lead to different rating outcomes and is therefore one important source of noise in ESG ratings.

Regarding measurement, there is only a limited amount of standardized and publicly available data about companies' ESG performance. As a result, ESG rating agencies need to resort to a variety of data sources to support their assessment. Table 3 gives an overview of the data sources that go into ESG ratings. Mainly, data stems from five distinct sources, namely from CSR reports, regulatory filings, the media, questionnaires that rating agencies send to companies, and modelled data.[13] These sources differ along important dimensions, namely whether the information is available to the public or not, who reports the information (the company itself or a third-party observer) and whether disclosure is mandatory or voluntary.

Given that ESG raters choose which of these five data sources to use explains why measurement also contributes to the noisiness of ESG ratings. For example, a firm's labor practices could be measured on the basis of information in a CSR report, on the basis of complaints made to a regulator, or on the basis of media reports about the firm. Each of these approaches would yield different assessments. Truvalue Labs and RepRisk are two providers that rely mainly on media reporting, and it is clear from Table 2 that this results in markedly different assessments compared to the other raters, which rely on a blend of data sources.

[13]For example, several rating agencies model carbon emissions to make up for missing reported data on carbon emissions, which is voluntary.

As a result, the way that ESG ratings are produced implies that each of them offers a noisy measurement of some underlying "true" ESG performance, which itself remains unobservable.[14]

Table 3. Overview of ESG Ratings and Data Sources

Source	**CSR Reports**	**Regulatory Filings**	**Media**	**Questionnaires**	**Modelled Data**
Availability	Public	Public	Public	Private	Private
Reporting Source	Self-reported	Self-reported	Third-party	Self-reported	Third-party
Disclosure	Voluntary	Mandatory	Involuntary	Voluntary	Involuntary

3.2 Errors-in-Variables and Stock Returns

In this section, we examine the impact of ESG measurement errors on the relationship between ESG performance and stock returns. Our departure from the rapidly growing literature that examines the impact of ESG on firm performance is that we do not assume that ESG attributes are measured accurately by the ratings. As in Section 2, investors observe a noisy signal of an ESG attribute.

The key equation describing the relationship between ESG measures and expected stock returns in our model is of equation (13).[15] To emphasize the effects of ESG on stock expected returns, we rewrite it here as follows:

$$E(S_1) - S_0 = c_0 - c_{impact} \cdot c_{noise} \cdot (s_Y - \overline{Y}) + c_X \cdot X + \eta, \tag{14}$$

where X is a vector of observable cash flow-relevant firm characteristics that will be used as controls, and the coefficients c_{impact} and c_{noise} are given by

$$c_{impact} = A\lambda \frac{\sigma_D^2 \sigma_{\epsilon_D}^2}{\sigma_D^2 + \sigma_{\epsilon_D}^2},$$
$$c_{noise} = \frac{\sigma_Y^2}{\sigma_Y^2 + \sigma_{\epsilon_Y}^2},$$

[14] At an individual indicator level (e.g., CO2 emissions), "true" means precisely the actual CO2 emissions that occurred. The noise is the difference between what the rating agency uses for CO2 emissions and the actual one. For ESG ratings, which are themselves weighted averages of indicators, "true" means that the indicators are measurement-error free and that the weights assigned to the indicators coincide with the weights that the representative ESG investor assigns to individual ESG attributes — which are related to their preferences. Therefore, for the ESG ratings there are two sources of noise, the indicator measurement and the weight differences. We define these quantities in Section A.2.

[15] To establish a relationship between ESG scores and stock returns, some papers in the related literature adopt an alternative regression specification: a regression of contemporaneous (time-t) stock returns on ESG scores, as in equation (12). The bias-correction methodology developed in this section applies to both equations (12) and (13).

and where η corresponds to the terms that depend on the signal for the firm's cash flows (s_d) after controlling for observable firm characteristics (X). c_0 captures the constant terms in equation (13). The noise-to-signal ratio in the ESG signal is

$$\kappa = \frac{\sigma^2_{\epsilon_Y}}{\sigma^2_Y}. \tag{15}$$

In equation (14), the total effect of the ESG ratings on expected stock returns is decomposed into two terms: (i) the *impact* component c_{impact} that captures how the stock returns would have been affected by the improvement in the ESG rating in the absence of noise[16] in the signal and (ii) the *noise* component, c_{noise}, that dampens down the impact effect. Notice that c_{noise} is always smaller than one, and it is decreasing in σ_{ϵ_Y}, the noise of the signal. In other words, the standard cross-sectional regression of expected stock returns on the ESG signal would not recover the true impact component c_{impact}. Because of the noise in the signal, it will produce an estimate that is biased towards zero. This is a manifestation of the "attenuation bias," arising in the context of the errors-in-variables problem.

The asset pricing literature typically tackles the measurement error problem by combining individual assets into portfolios and performing analysis on portfolios. Recently, a number of papers argued that the particular method of portfolio grouping can dramatically influence the results. Chordia, Goyal, and Shanken (2015) suggest using individual stocks, as opposed to portfolios, in asset-pricing tests. In our analysis, the noise in measurement of ESG scores also affects the construction of portfolios, sorted by ESG scores. The noise in the signal leads to errors in the classification of the ESG-sorted portfolios, reducing differences across the portfolios. To develop the intuition why this occurs, take the noise to the extreme. In the limit, when the noise is infinite, the top ESG stocks and the bottom ESG stocks are just a random draw from the stock universe. Hence, after controlling for the standard characteristics and factor loadings, their expected returns should be identical.

Table 2 suggests that the measurement error problem is particularly acute for ESG ratings, which is why we propose a correction to the estimation procedure that has been used in the literature to determine the impact of ESG on expected stock returns. We run our analysis on individual stocks, not ESG-sorted portfolios, but correct the attenuation bias using instrumental variables. In order to do so, we cast our estimation procedure as the classical errors-in-variables problem.

Equation (14) is equivalent to the following structural model, to be estimated on the

[16]Strictly speaking, the population aggregate constant A also depends on the noise in the signal σ_{ϵ_Y}. This constant also decreases in the presence of noise in the signal. To facilitate the mapping of our model to the classical errors-in-variables problem, we ignore this dependence here. The model predicts that the attenuation bias is stronger than in our econometric specification.

cross-section of stocks

$$E(\Delta S_{t+1}) = a_0 + c_{impact} \cdot Y_t + c_X \cdot X_t + \eta_t, \tag{16}$$

where a_0 is the stock's average expected return, Y_t is the (unobserved at time t) true non-pecuniary ESG attribute, c_{impact} is the "true" impact of the ESG attribute on returns, X_t is a vector of observable firm characteristics related to firm cash flows, and η_t is the signal about the firm's cash flow (s_d) after controlling for observable firm characteristics. $E(\Delta S_{t+1})$ denotes the expected per-share return on the stock in period $[t, t+1]$. We estimate this model at an individual stock level, and so all variables in (16) and the equations that follow have another subscript, indexing stocks, which we suppress for expositional reasons.

We do not observe the true ESG attribute Y, but we observe multiple noisy measures of it. An ESG rating agency i measures the ESG attribute Y_t with a score $s_{i,t}$, and these scores differ across rating agencies. We model the relationship between the true attribute Y and the raters' scores as follows[17]:

$$\begin{aligned} s_{1,t} &= Y_t + \epsilon_{Y_{1,t}}, \\ s_{2,t} &= Y_t + \epsilon_{Y_{2,t}}, \\ &\vdots \\ s_{n,t} &= Y_t + \epsilon_{Y_{n,t}}. \end{aligned} \tag{17}$$

In Appendix A.2, we discuss the exact mapping between noise in (aggregate) ESG ratings, coming potentially from (i) discrepancies in measurement of each individual indicator and (ii) discrepancies in weights on each indicator, and the errors $\epsilon_{Y_{i,t}}$ in the above equations.

The regression equivalent of equation (16) is

$$\Delta S_{t+1} = a + b \cdot s_{i,t} + c_X \cdot X_t + \eta_t + \nu_t. \tag{18}$$

We make several assumptions. First, the error terms are additive and orthogonal to Y, as in the classical error-in-variables problem:

$$E[\epsilon_{Y_{i,t}} | Y_t] = 0. \tag{19}$$

Second, our errors-in-variables correction strategy is to instrument (noisy) scores of a

[17] We follow the assumption in the model in which the rating scores are just a noisy version of the true ESG fundamental. All the results on the attenuation bias go through if we assume that the rating scores are a linear transformation of the fundamental: $s_{i,t} = C_i + c_i \cdot \left(Y_t + \epsilon_{Y_{i,t}}\right)$. See discussion around equation (31).

given rater by scores of other ESG raters. The rating agencies' scores are correlated through Y_t, and we expect them to predict each other. This is the *relevance* assumption. We verify it in Section 4.

The third assumption is that the error terms $\epsilon_{Y,i,t}$ are as in the classical error-in-variables problem. In particular, the errors are independent of the stock cash-flow innovations η_t (not captured by our firm-level controls) and other stock return relevant innovations ν_t:

$$E[\epsilon_{Y_{i,t}} \cdot \eta_t] = 0 \quad \text{and} \quad E[\epsilon_{Y_{i,t}} \cdot \nu_t] = 0. \tag{20}$$

These assumptions are the *exclusion restriction.* We will use the rating agencies' scores as instruments for each other, and the exclusion restriction guarantees that our instruments are exogenous.

Finally, we assume that all errors $(\epsilon_{Y,i,t})$ are independent across rating agencies:

$$E[\epsilon_{Y_{i,t}} \cdot \epsilon_{Y_{j,t}}] = 0. \tag{21}$$

This is the *independence* assumption.

At a first glance, this set of assumptions seems to be strong or implausible. However, they can be tested using the overidentifying restrictions tests, which are possible in our setting because we have multiple instruments for the same variable. In other words, if any of these assumptions is violated, then an overidentifying restrictions test should diagnose this and reject the model. We discuss this in detail in Section 4.1.

An assumption unrelated to the instrumentation is that $E[Y_t \cdot \nu_t] = 0$. This condition means that ν is a pure expectation error.

Together, assumptions (19), (20) and (21) guarantee that ESG agencies' ratings are valid instruments for each other. We have effectively assumed the errors in measurement are white noise. We are now able to recover the true impact of the unobserved ESG attribute Y on expected stock returns.

As discussed before, the OLS estimate of the coefficient b in the regression of stock returns against ESG scores of rating agency i, is a downward biased estimate of the coefficient c_{impact} that we intend to measure:

$$\hat{b}_{OLS} = c_{impact} \cdot \frac{\sigma_Y^2}{\sigma_Y^2 + \sigma_{\epsilon_{Y,i,t}^2}}. \tag{22}$$

With other rating agencies' ESG scores used as instruments, we can formally test whether the OLS estimator is biased and correct the bias. Proposition 2 summarizes the above discussion as well as our assumptions. Define the vector of instruments $z_{i,t}$ as the ratings

from all rating agencies excluding rating agency i

$$z_{i,t} = \{s_{j,t}, \forall j \neq i\}. \tag{23}$$

Proposition 2 (Ratings as Instrumental Variables) *Suppose that $s_{i,t}$ is the noisy measure of a true ESG attribute Y_t from rating agency i, given by*

$$s_{i,t} = Y_t + \epsilon_{Y,i,t},$$

where the error terms for each rating are independent of the ESG attribute, uncorrelated with each other and orthogonal to the firms' cash flow characteristics (equations (19)*,* (20)*, and* (21)*), i.e.,*

$$\begin{aligned}
&E[\epsilon_{Y_{i,t}} \cdot \epsilon_{Y_{j,t}}] = 0, \quad \forall i \neq j,\\
&E[\epsilon_{Y,i,t} \cdot \eta_t] = 0, \quad \forall i \in [1,n],\\
&E[\epsilon_{Y,i,t} \cdot \nu_t] = 0, \quad \forall i \in [1,n],\\
&E[\epsilon_{Y_{i,t}} | Y_t] = 0, \quad \forall i \in [1,n]
\end{aligned}$$

The OLS regression estimating the impact of ESG performance on stock returns produces a downward biased estimate, given by

$$\hat{b}_{OLS} = c_{impact} \cdot \underbrace{\frac{\sigma_Y^2}{\sigma_Y^2 + \sigma_{\epsilon_{Y,i,t}}^2}}_{c_{noise}<1}.$$

The true parameter can be recovered by two-stage least squares using the other rating agencies' scores for the same ESG measure, $z_{i,t}$, as instruments. The 2SLS estimator is consistent.

$$\begin{aligned}
\hat{b}_{2SLS} &= (z's)^{-1}(z'\Delta S_t),\\
\operatorname{plim} \hat{b}_{2SLS} &= c_{impact}.
\end{aligned}$$

Two important points need to be highlighted. First, our methodology does not construct a noiseless ESG measure. In other words, we do not recover the true attribute. Second, we are reducing or eliminating the impact that the noise has on a regression coefficient — in this case the stock return coefficient. Again, this does not mean that the noise has been uncovered, but that we are able to estimate the counterfactual effect — what would have been the impact of the ESG performance on stock returns if the measure is without noise.

To test whether the OLS coefficients are biased, we implement the Hausman specification test. Specifically, we first compute the absolute value of the difference of the estimates and the variance of that difference

$$\delta = |\hat{b}_{2SLS} - \hat{b}_{OLS}|, \tag{24}$$

$$\sigma_\delta^2 = \sigma_{b_{2SLS}}^2 - \sigma_{b_{OLS}}^2, \tag{25}$$

where $\sigma_{b_{OLS}}^2$ and $\sigma_{b_{2SLS}}^2$ are the variances of the OLS and 2SLS estimates respectively. The variance of the difference of the coefficients is the difference of the variances (equation (25)) because under the null hypothesis OLS is BLUE (Best Linear Unbiased Estimator). We then test whether the difference in equation (24) is statistically different from zero.

Under the null hypothesis of no errors-in-variables, the OLS estimator is consistent and efficient, while the 2SLS estimator is consistent but inefficient. This means that the OLS estimate reaches the lowest variance among linear estimators. However, under the alternative hypothesis that there are errors-in-variables, the OLS estimate is biased (inconsistent) while the 2SLS estimator is consistent. So, if the difference in (24) is statistically different from zero, this implies that the signal is noisy.

Furthermore, notice that the ratio of the two coefficients readily provides an estimate of the noise in the signal, which we denote by κ:

$$\frac{\hat{b}_{2SLS}}{\hat{b}_{OLS}} = \frac{c_{impact}}{c_{impact} \cdot \frac{\sigma_Y^2}{\sigma_Y^2 + \sigma_{\epsilon_{Y,i,t}}^2}} = \frac{\sigma_Y^2 + \sigma_{\epsilon_{Y,i,t}}^2}{\sigma_Y^2} = 1 + \kappa. \tag{26}$$

We will construct these ratios for every rater in our sample and compare the estimated noise-to-signal ratios contained in their ratings.[18]

What should we expect when we implement the instrumental variables procedure to correct for errors-in-variables in our application? First, the 2SLS coefficients should be larger in absolute value than their OLS counterparts, and the increase in the estimates is associated with noise in the rating. Second, the variances of the 2SLS estimates should be larger than the variances of their OLS counterparts.

[18] As we mentioned before the coefficient A is also decreasing in the noise. Therefore, the estimate of noise from equation (26) is an upper bound. The reason we take this approach is because the decomposition through A requires several unobservables to perform the decomposition: The share of investors that are ESG interested, and the ratio of the volatility of the true ESG attribute relative to the dividend process.

3.3 Overidentifying Restrictions Tests and Noise-Correction (Pruning) Procedure

One major advantage of our setting is that we have more instruments than we need for our two-stage least squares estimation, i.e., our model is overidentified. Having multiple instruments for a single variable Y in (16) allows us to run tests of overidentifying restrictions (the Hansen J test) and determine the validity of the instruments. Instruments that fail the overidentifying restrictions tests should be excluded from our estimator.

Our formal procedure for identifying the set of rating agencies' whose ESG scores serve as valid instruments in the 2SLS test (18) is as follows. First, we pick a rating agency whose scores we would like to instrument and use all remaining rating agencies' scores as instruments. Second, we estimate the specification (18) by 2SLS and run the Hansen J-test of overidentifying restrictions. If the model passes the J-test, then all included instruments are valid; otherwise, we exclude instruments, one at a time, until the model passes the J-test. We report the included and excluded instruments. Finally, we repeat this process for each rating agency in our sample.

The above pruning procedure identifies the maximum number of valid instruments for each rating agency i's, $i = 1, \ldots, 8$, scores and provides the valid 2SLS estimate of the effect of ESG ratings of agency i on stock returns. We run the procedure separately for firms located in North America, Europe, and Japan.

4 Empirical Results

In this section, we fully specify the regressions we are running. We start with standard regressions of stock returns on ESG measures and contrast them to the two-stage least squares regressions, which tackle the measurement error problem in ESG scores.

Define the stock j's monthly excess return in month $t+1$ as follows:

$$r_{j,t+1} = \Delta S_{j,t+1}/S_{j,t} - r_{f_{t+1}}, \tag{27}$$

where r_{f_t} is the risk free rate.[19] Both returns and the risk free rate are measured in the local currency. The empirical counterpart of equation (13), which we implement on a panel of

[19] In models with CARA preferences and normally distributed signals, it is convenient to work with per-share stock returns, ΔS_{t+1}, with $t = 0$. In our empirical analysis, however, we use per-dollar returns, $r \equiv (S_{t+1} - S_t)/S_t$, as in the empirical literature. We acknowledge this inconsistency, but we still prefer to keep our theoretical results in terms of per-share returns, for expositional clarity.

individual firms, is

$$r_{k,t+1} = a_{OLS} + b_{OLS} \cdot s_{k,i,t} + c_{OLS} \cdot X_{k,t} + \mu_t + \nu_{k,t}, \tag{28}$$

where $s_{k,i,t}$ denotes the ESG rating of firm k, by rater i, in month t, X_k is the vector of stock-level controls (log size, book-to-market, EBIT over total assets, market beta, volatility, and momentum). X_k also includes industry and country fixed effects.[20] Finally, μ_t is the monthly time fixed effect, and $\nu_{k,t}$ is the error term, which is mean zero and independent of the regressors.

As argued in section 3.2, the coefficient $\hat{b}_{OLS}$ is biased towards zero. We compare this estimate with those obtained from a two-stage least squares procedure. The first stage regression uses corresponding scores of other rating agencies as instruments for a given rater's score and includes the same controls as in (28):

$$s_{k,i,t} = a_1 + b_1 \cdot z_{k,i,t} + c_1 \cdot X_{k,t} + \mu_t + \varepsilon_{k,t}, \tag{29}$$

where $s_{k,i,t}$ is the ESG rating of rater i for firm k in month t. Denote $\hat{s}_{k,i,t}$ as the fitted value from estimating equation (29). Then the second stage regression is

$$r_{k,t+1} = a_{2SLS} + b_{2SLS} \cdot \hat{s}_{k,i,t} + c_{2SLS} \cdot X_{k,t} + \mu_t + \nu_{k,t}. \tag{30}$$

Provided that our assumptions are satisfied, we expect that $|\hat{b}_{2SLS}| > |\hat{b}_{OLS}|$.

Table 4 reports the results of the OLS and 2SLS regressions of expected stock returns on ESG scores of a given rater. The estimate of interest is the estimated effect of ESG performance on expected stock returns.

The first result to highlight is that the majority of coefficients are quite precisely estimated. In the OLS regressions, 11 out of 24 coefficients are statistically significant at the 5% level (with the t-statistic higher than 1.96), while in the 2SLS regressions, 15 out of 24 are significant at the 5% level. This is particularly interesting given the small time series we have. Interestingly, all but one of the coefficients are positive. The average OLS coefficient is 0.127, which means that an increase in the rating of one standard deviation increases the return by 0.127 percent. The interpretation of the positive coefficients is that an improvement in the ESG score is associated with a positive expected return. In this paper, we are

[20] We do not use firm fixed effects because the frequency of most ESG ratings changes is annual. As a result we do not have enough within-firm variation to support a firm fixed effects estimation. For the same reason, we cannot cluster standard errors, however we correct them for autocorrelation and heteroscedasticity using the Newey-West method. Future research should cluster standard errors by industry when the time series of ESG ratings is long enough.

Table 4. Empirical results: OLS vs. 2SLS estimates. This table reports estimates of b_{OLS} from the OLS regression $r_{k,t+1} = a_{OLS} + b_{OLS} \cdot s_{k,i,t} + c_{OLS} \cdot X_{k,t} + \mu_t + \nu_{k,t}$ and estimates of b_{2SLS} from our second-stage regression $r_{k,t+1} = a_{2SLS} + b_{2SLS} \cdot \hat{s}_{k,i,t} + c_{2SLS} \cdot X_{k,t} + \mu_t + \nu_{k,t}$, where $r_{k,t+1}$ denotes excess return of stock k in month $t+1$, measured in local currency, $s_{k,i,t}$ is the ESG rating of firm k by rater i in month t, X_k is the vector of stock-level controls (log size (in local currency), book-to-market, EBIT over total assets, market beta, volatility, momentum, and industry and country fixed effects) and $\hat{s}_{k,i,t}$ denotes the fitted value of the ESG rating of firm k by rater i in month t from the first-stage regression $s_{k,i,t} = a_1 + b_1 \cdot z_{k,i,t} + c_1 \cdot X_{k,t} + \mu_t + \varepsilon_{k,t}$. All reported coefficients and standard errors are multiplied by 100. The regressions are run for each rater, whose names are reported in the left column. TVL stands for Truvalue Labs. The left column also reports regions in which rated firms are located. Pruned IV refers to the 2SLS estimate of the two-stage regression with the largest set of instruments that pass the Hansen J test (passed if p-value > 0.05 in the last column). Standard errors are computed using the Newey-West procedure. T-statistics > 1.96 are highlighted in bold.

		OLS			Pruned IV			
		Coeffs	StdErr	t-stat	Coeffs	StdErr	t-stat	χ^2(p-value)
North America	MSCI	0.1860	0.0505	**3.69**	0.2047	0.1060	1.93	0.094
	ISS	0.1933	0.0683	**2.83**	0.0716	0.0913	0.78	0.169
	RepRisk	0.0813	0.0646	1.26	0.0206	0.2376	0.09	0.154
	TVL	0.1329	0.0693	1.92	0.6899	0.2638	**2.62**	0.057
	Vigeo-Eiris	0.0453	0.0786	0.58	0.4159	0.1159	**3.59**	0.057
	SP Global	0.0104	0.0554	0.19	0.1997	0.0879	**2.27**	0.061
	Refinitiv	0.0730	0.0602	1.21	0.0383	0.0847	0.45	0.180
	Sustainalytics	0.1529	0.0645	**2.37**	0.1787	0.1751	1.02	0.216
Europe	MSCI	0.2070	0.0797	**2.60**	0.2830	0.1504	1.88	0.825
	ISS	0.2055	0.0953	**2.16**	0.2474	0.1285	1.93	0.491
	RepRisk	0.0175	0.0962	0.18	0.0612	0.2566	0.24	0.064
	TVL	0.0326	0.0723	0.45	0.6081	0.2930	**2.08**	0.371
	Vigeo-Eiris	0.1499	0.0843	1.78	0.2614	0.1193	**2.19**	0.455
	SP Global	0.1477	0.0789	1.87	0.2613	0.1240	**2.11**	0.419
	Refinitiv	0.0757	0.1082	0.70	0.3527	0.1460	**2.42**	0.729
	Sustainalytics	0.1070	0.0936	1.14	0.5281	0.2047	**2.58**	0.660
Japan	MSCI	0.1434	0.0709	**2.02**	0.6340	0.1622	**3.91**	0.524
	ISS	0.2702	0.0896	**3.02**	0.5285	0.1617	**3.27**	0.183
	RepRisk	0.1588	0.0766	**2.07**	-0.5444	0.2849	1.91	0.137
	TVL	0.0014	0.0775	0.02	1.2021	0.3260	**3.69**	0.159
	Vigeo-Eiris	0.1000	0.0845	1.18	0.3621	0.1073	**3.38**	0.182
	SP Global	0.1629	0.0649	**2.51**	0.2602	0.0927	**2.81**	0.082
	Refinitiv	0.1910	0.0760	**2.51**	0.4343	0.1320	**3.29**	0.485
	Sustainalytics	0.1969	0.0711	**2.77**	0.5822	0.1551	**3.75**	0.161

concerned primarily with the estimation problem, but at the end of this subsection we come back to this result and present our explanation for the sign of the coefficients.

The second, and most important result, is that we detect attenuation bias in the majority of our OLS regressions. Notice that the 2SLS coefficients are almost always larger than the OLS coefficients (in absolute terms). For 20 out of 24 estimates presented in Table 4, the 2SLS coefficient is larger than the OLS coefficient, only 3 decrease, and 1 switches signs (moves from positive to negative). The average coefficient increases from 0.127 in the OLS estimation to 0.328 in the 2SLS estimation. The corrected estimates demonstrate that the effect of ESG performance on stock returns is stronger than previously estimated. The OLS regression estimates of ESG ratings' impact on stock returns are biased downward by about 60%, i.e., after the 2SLS correction, the size of the effect is more than twice as large. It is

worth discussing what is the economic interpretation of the 2SLS coefficient and the reason why the attenuation bias occurs. There are two channels which lead to an increase in a score of a rating agency. One is an improvement in the true ESG performance (Y). An increase in Y has stock return implications, and the 2SLS coefficient estimates that effect. The second one is that a change is explained by noise. However, the noise in measurement should have zero impact on stock returns. The OLS coefficient reflect the combination of both the changes due to Y and the noise, and hence it is biased toward zero.

Third, notice that the coefficients are different across regions and raters. It is worth exploring why this is the case. For OLS estimates, there are two possible reasons which explain their divergence across regions and rating agencies. The first and simplest explanation is that the magnitude of the noise in each rating agency is different. This implies that even if the true coefficients were to be identical the attenuation bias would be larger for those rating agencies that are noisier (Proposition 2). However, if this were the only explanation, then the IV estimates should be identical across rating agencies, and they are not. The second explanation is due to the different normalizations that are performed by the rating agencies — i.e., the scores of the rating agencies are a linear transformation of the true ESG performance. Formally,

$$\begin{aligned} s_{1,t} &= C_1 + c_1 \cdot (Y_t + \epsilon_{Y_{1,t}}), \\ s_{2,t} &= C_2 + c_2 \cdot (Y_t + \epsilon_{Y_{2,t}}), \\ &\vdots \\ s_{n,t} &= C_n + c_n \cdot (Y_t + \epsilon_{Y_{n,t}}). \end{aligned} \tag{31}$$

In this case, the OLS and 2SLS coefficients on rating agency i will be scaled by $1/c_i$.[21] This implies that every region/rating agency has a different coefficient.

The rating agencies perform many normalizations in the data: (i) they transform individual indicators from self-reporting, surveys, and news outlets into 1-100 numeric values; (ii) Those values are further normalized when "best-in-class" considerations are included;[22]; (iii) finally, the indicators are aggregated in a single score which further normalizes the data. These are three of the typical steps that most rating agencies perform that could explain why the scores are not an unbiased version of the true ESG performance. It is important

[21] To understand the intuition, we assume zero noise for simplicity, we use the structural equation (16), in which the true coefficient should be c_{impact}. We replicate part of that equation here: $\Delta S_{t+1} = a_0 + c_{impact} \cdot Y_t$. Solving for Y_t from $s_{i,t} = C_i + c_i \cdot (Y_t + \epsilon_{Y_{i,t}})$ and substituting it in (16), we arrive at the following the reduced-form regression: $\Delta S_{t+1} = a_0 - \frac{C_i}{c_i} + \frac{c_{impact}}{c_i} \cdot s_{i,t} - \epsilon_{Y_{i,t}}$. The coefficient of interest is $\frac{c_{impact}}{c_i}$.

[22] "Best-in-class" refers to the practice of comparing firms within a given industry. The CO2 emissions, for instance, instead of being the total emissions the ranking is based on firms that are the top ten polluters within the industry. This procedure implies an industry specific normalization.

to mention that assumptions (19), (20), and (21) imply that the instruments are valid even when the scores are linear transformations of the true ESG performance Y. The attenuation bias is also unaffected, given that the normalization affects OLS and 2SLS in the same way.

Figure 1. Differences between OLS and 2SLS estimates. This figure plots the difference between 2SLS and OLS estimates reported in Table 4. Note that the coefficients in Table 4 have been multiplied by 100.

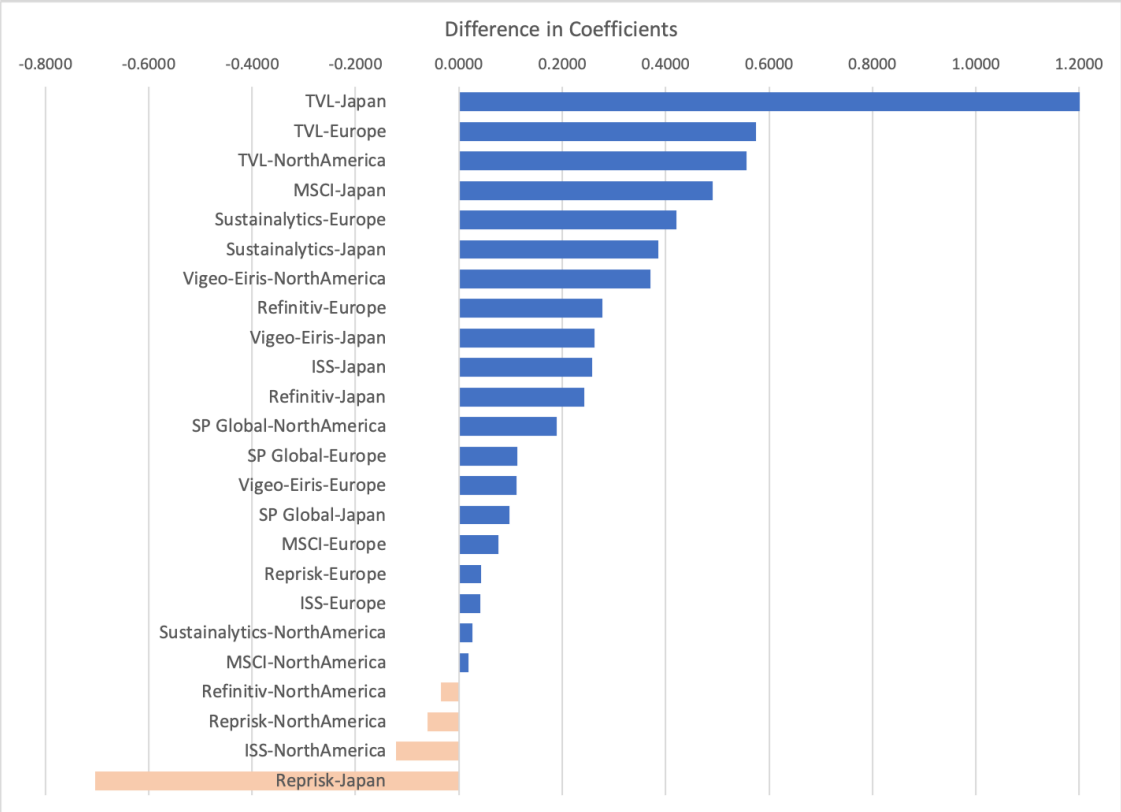


Figure 1 plots the difference between the 2SLS and OLS coefficients to illustrate the attenuation bias. It reveals that the largest increases in the coefficients are all from Truvalue Labs, indicating that in this case, the estimation benefits substantially from instrumentation. The extent to which a rating's coefficient increases is related to the implied noise of the rating, which we discuss in detail in the following section. Only in four cases the coefficients decrease when switching from OLS to 2SLS estimation. Note that two of those four cases are associated with RepRisk. We find that instrumentation does not make a big improvement in RepRisk's estimate. This may be related to the methodology RepRisk uses for computing its ESG ratings which is markedly different from the other rating agencies. RepRisk concentrates on negative news only, and disregards positive news, resulting in a unique signal. As we discuss below both Truvalue Labs and RepRisk are valuable instruments that improve the

estimation for other ratings even though they are not strong predictors by themselves.

The immediate question should be whether the changes in the coefficients are statistically significant. We perform a Hausman specification test following equations (24) and (25). We compare each pair of coefficients given a region and rating agency. Those results are presented in Table 5. When we implement the Hausman specification test, we reject many of the pairwise comparisons: 13 out of 24 are statistically significant at the 5 percent confidence level (i.e. with a t statistic larger than 1.96). In Europe, we see 3 rejections, North America has 4 rejections (one of which being a decrease in the coefficient), and Japan has the most rejections, 6, with one being a decrease in the coefficient. These results show that the 2SLS coefficients are statistically different from the OLS ones in more than half of the cases.

In summary, we find large increases in coefficients that are economically important and statistically different from zero. The average increase in the coefficients is 0.202, from an average OLS estimate of 0.127. This provides a measure of the size of the noise present in the data. After the attenuation bias is corrected with our procedure, the average coefficient more than doubles.

By comparing the OLS and the 2SLS estimates, we can assess how noisy each rating agency's estimates are. From the ratios of the 2SLS and OLS estimates we can extract the variance of the noise to the signal. It is given in equation (26), which we reproduce here for expositional convenience:

$$\frac{\hat{b}_{2SLS}}{\hat{b}_{OLS}} = \frac{c_{impact}}{c_{impact} \cdot \frac{\sigma_Y^2}{\sigma_Y^2 + \sigma_{\epsilon_{Y,i,t}}^2}} = \frac{\sigma_Y^2 + \sigma_{\epsilon_{Y,i,t}}^2}{\sigma_Y^2} = 1 + \kappa.$$

The estimates of the noise-to-signal ratios κ are presented in Figure 2. As mentioned before, the ratings of Truvalue Labs are the noisiest: out of the top-6 noisiest scores, three places are held by Truvalue Labs. In Figure 1, Truvalue Labs' ratings occupies the top-3 positions, i.e., OLS estimates based on Truvalue Labs' ratings have the largest bias. As we have discussed in Section 3.1.1, Truvalue Labs construct their ratings based on media reports, rather than, for example, mandatory disclosure or expert assessment. It seems that this approach results in a relatively noisy rating. As we show below, however, that this does not mean that Truvalue Labs is a bad instrument.

Our analysis reveals that the ESG ratings with the least amount of noise in North America are those of MSCI and Sustainalytics. This finding seems very reasonable to us, since the U.S. is a target market for MSCI and Sustainalytics, and these raters employ ample resources to analyze U.S. firms. In Europe and Japan, ISS, S&P Global, and Vigeo-Eiris have the least proportion of noise; all three started as European raters.

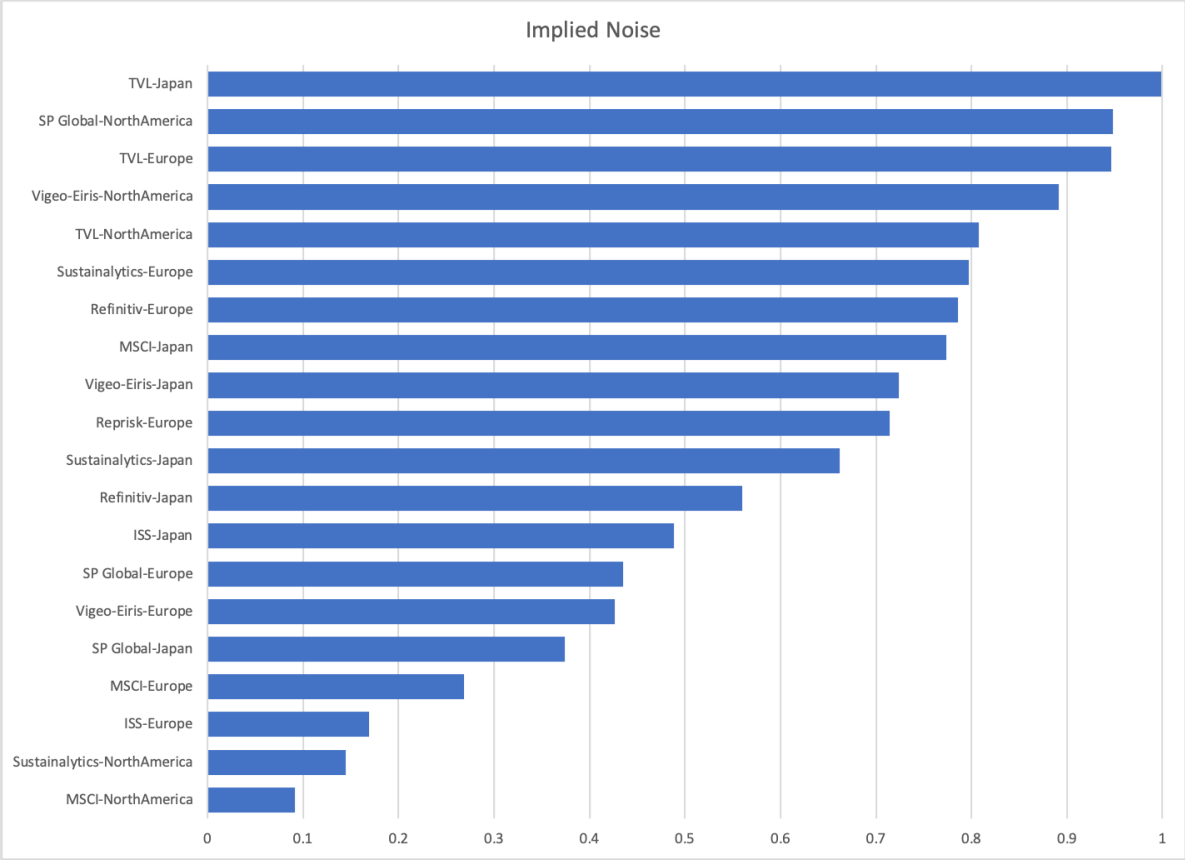
Table 5. OLS vs. 2SLS estimates: Hausman specification test. This table reports differences in estimates of b_{2SLS} from our second-stage regression $r_{k,t+1} = a_{2SLS} + b_{2SLS} \cdot \hat{s}_{k,i,t} + c_{2SLS} \cdot X_{k,t} + \mu_t + \nu_{k,t}$, and estimates of b_{OLS} from the OLS regression $r_{k,t+1} = a_{OLS} + b_{OLS} \cdot s_{k,i,t} + c_{OLS} \cdot X_{k,t} + \mu_t + \nu_{k,t}$, where $r_{k,t+1}$ is denotes excess return of stock k in month $t+1$, $s_{k,i,t}$ is the ESG rating of firm k by rater i in month t, X_k is the vector of stock-level controls (log size, book-to-market, EBIT over total assets, market beta, volatility, momentum, and industry and country fixed effects) and $\hat{s}_{k,i,t}$ denotes the fitted value of the ESG rating of firm k by rater i in month t from the first-stage regression $s_{k,i,t} = a_1 + b_1 \cdot z_{k,i,t} + c_1 \cdot X_{k,t} + \mu_t + \varepsilon_{k,t}$. The reported differences between the coefficients are multiplied by 100, and so are the standard deviations. The regressions are run for each rater, whose names are reported in the left column. TVL stands for Truvalue Labs. The left column also reports regions in which firms are located. The 2SLS estimate is from the two-stage regression with the largest set of instruments that pass the Hansen J test. Stdev refers to the standard deviation in equation (25). Standard errors are computed using the Newey-West procedure. The t-statistics from the Hausman specification test are reported in the last column. T-statistics > 1.96 are highlighted in bold.

		Difference in Coefficients (2SLS − OLS)	Stdev	Hausman Test (t-stat)
North America	MSCI	0.0186	0.0932	0.20
	ISS	-0.1216	0.0606	**2.01**
	RepRisk	-0.0607	0.2287	0.27
	TVL	0.5570	0.2545	**2.19**
	Vigeo-Eiris	0.3705	0.0852	**4.35**
	SP Global	0.1893	0.0682	**2.77**
	Refinitiv	-0.0347	0.0596	0.58
	Sustainalytics	0.0259	0.1628	0.16
Europe	MSCI	0.0761	0.1275	0.60
	ISS	0.0419	0.0862	0.49
	RepRisk	0.0437	0.2379	0.18
	TVL	0.5754	0.2839	**2.03**
	Vigeo-Eiris	0.1115	0.0844	1.32
	SP Global	0.1136	0.0956	1.19
	Refinitiv	0.2771	0.0980	**2.83**
	Sustainalytics	0.4211	0.1820	**2.31**
Japan	MSCI	0.4906	0.1459	**3.36**
	ISS	0.2583	0.1346	1.92
	RepRisk	-0.7033	0.2744	**2.56**
	TVL	1.2007	0.3167	**3.79**
	Vigeo-Eiris	0.2621	0.0660	**3.97**
	SP Global	0.0973	0.0662	1.47
	Refinitiv	0.2433	0.1079	**2.25**
	Sustainalytics	0.3853	0.1379	**2.79**

By examining Figure 2, one might hastily conclude that one should source scores only from ESG raters with the most precise measurement, located at the bottom of the figure. We caution against such conclusion. Disregarding scores of other raters amounts to throwing away valuable information about the unobservable ESG attribute the ratings are trying to measure. Our valid set of instruments (other rating agencies' scores) never includes fewer than four raters and, in many cases, includes all rating agencies. Intuitively, by combining different ratings, and in particular ratings that rely on different information sources and contain different sorts of noise, one can get the most precise signal about the unobservable ESG attribute.

In our sample, both the OLS and 2SLS regression estimates in Table 4 indicate that ESG

Figure 2. Implied noise in ESG ratings. This figure plots our measure of noise in the ESG ratings, computed as $\kappa = 1 - \hat{b}_{2SLS}/\hat{b}_{OLS}$, where $\hat{b}_{2SLS}$ and $\hat{b}_{OLS}$ are 2SLS and OLS regression estimates, resp., reported in Table 4. The vertical axis contains ESG raters names and regions in which rated firms are located. Implied noise can not be computed in four cases where the coefficients of the 2SLS are smaller than the OLS coefficients.



ratings have a *positive* effect on expected stock returns. Our model has the following two key predictions about the sign of the effect. First, if the mass of ESG investors λ is fixed, firms with higher ESG ratings should have higher stock valuations and *lower expected returns* (or the cost of equity). Second, an increase in the mass of ESG investors (an increase in λ) leads to *higher* same-period stock returns. If we were to extend our model to one more period, it is easy to see that an (unanticipated) increase in next-period's mass of ESG investors would have a *positive* effect on next period's stock returns, the outcome variable in our main regression (30). During our sample period, there has been a marked increase in the mass of ESG investors, i.e., strong inflows into ESG funds, and so both channels are relevant for our estimation. We do not have a way to explicitly control for unanticipated investor inflows in our empirical specification, and so the estimated coefficient on the ESG rating captures both (i) the reduction of the cost of equity and (ii) the positive effect of future ESG fund inflows.

Our leading explanation for the positive coefficient on the ESG rating is that the flow effect plays a dominant role in our sample—financial markets did not accurately anticipate the meteoric rise of ESG investing.

In this paper we concentrate on the quality of the estimation and leave the important question of disentangling possible channels through which ESG ratings affect stock returns to future research. Furthermore, ESG ratings change infrequently. If we run the regression contemporaneously, we find the exact same results: the coefficients are all positive and the majority are statistically significant. We conclude that, in our sample period, the ESG fund inflows channel dominates. The attenuation bias that we detect will remain relevant, even if in the future the inflow of ESG investments levels off, so that the expected returns of ESG stocks turn negative (Pastor, Stambaugh and Taylor, 2021*a*). In this case, we would expect the coefficients to become more negative with our 2SLS correction procedure.

4.1 Valid Instruments: Theory

Theoretically, a rejection of an overidentifying restrictions test implies that one of the instruments is not exogenous and hence is invalid and should be excluded from the estimation procedure. In this section, we present possible reasons for why a rating agency's score may end up being an invalid instrument. This happens if our null hypothesis ((19), (20), and (21)) is violated, and, as we argue below, in most cases, the overidentifying restrictions tests would diagnose these violations.

To explain the logic behind the failures of the overidentifying tests in our setting, let us focus on a special case with only three rating agencies. We will use the first ESG rater's score $s_{1,t}$ as the regressor and $s_{2,t}$ and $s_{3,t}$ as two possible instruments. The structural equation determining how stock returns are related to ESG ratings is given by

$$\Delta S_{t+1} = a_0 + c_{impact} \cdot Y_t + c_X \cdot X_t + \eta_t + \nu_t,$$

and $s_{1,t} = Y_t + \epsilon_{Y_{1,t}}$. As before, η_t corresponds to the stock return relevant innovations after controlling for firm characteristics, and ν_t is the expectations error. Substituting $s_{1,t}$ in the structural equation, we derive the reduced form equation one can estimate:

$$\Delta S_{t+1} = a + c_{impact} \cdot s_{1,t} + c_X \cdot X_t \quad \underbrace{-c_{impact} \cdot \epsilon_{Y_{1,t}} + \eta_t + \nu_t}_{\text{error}}. \tag{32}$$

The bias in the OLS estimator comes from the correlation between the regressor $s_{1,t}$ and the error term, through $\epsilon_{Y_{1,t}}$.

Under the null hypothesis (equations (19), (20), and (21)), the scores $s_{2,t} = Y_t + \epsilon_{Y_{2,t}}$ and $s_{3,t} = Y_t + \epsilon_{Y_{3,t}}$ are both valid instruments because the assumptions in (19), (20), and (21) assure that

$$\begin{aligned} E[\epsilon_{Y_{2,t}} \cdot \text{error}] &= 0, \\ E[\epsilon_{Y_{3,t}} \cdot \text{error}] &= 0. \end{aligned} \tag{33}$$

We now analyze the four cases in which our null hypothesis (equations (19), (20), and (21)) is not satisfied. These cases are: (1) correlated errors across rating agencies, (2) errors correlated with stock relevant innovations, (3) the ESG attribute is correlated with cash flows, and (4) non-classical errors-in-variables. In all these cases either the estimated coefficient's interpretation changes, and/or an instrument is invalid.[23] In our analysis below we highlight when candidate instruments are not valid and how we detect this.

- **Case 1. Errors of the rating agencies are correlated with each other:**

$$E[\epsilon_{Y_{1,t}} \cdot \epsilon_{Y_{i,t}}] \neq 0, \quad i \neq 1.$$

  This correlation can occur if either some rating agencies are influenced by scores of other raters or both rating agencies use similar data and similar estimation procedures. We analyze this in detail in Appendix A.3 and show that, in this case, equation (33) is violated. This violation will be detected by the overidentifying restrictions test.

- **Case 2: Measurement error is correlated with the stock return relevant innovations (or the expectation errors):**

$$\begin{aligned} E[\eta_t \cdot \epsilon_{Y_{i,t}}] &\neq 0, \\ E[\nu_t \cdot \epsilon_{Y_{i,t}}] &\neq 0. \end{aligned}$$

  These correlations can happen when the rating agencies decide their scores by observing contemporaneous and past stock return realizations, for example, when ESG ratings are backfilled retroactively. It is immediate that (33) is violated in that case. Again, the Hansen J test of overidentifying restrictions would detect this violation.

- **Case 3. ESG performance Y is correlated with unobservable or omitted firm characteristics:**

$$cov(Y_t, \eta_t) \neq 0.$$

[23] An instrument is invalid if equation (33) fails, i.e., the instrument is correlated with an error term and hence it is not exogenous.

The above correlation could occur through (unobservable) characteristics of the firms, which are not included in our set of controls, e.g., management quality. Suppose, for example, that better managerial quality improves returns and also makes management care about the environment. In this case, a positive innovation η_t is an unexpected improvement in management quality and its correlation with ESG performance Y_t is positive. In other words, in this case ESG performance has pecuniary benefits, in addition to non-pecuniary.

The bias in the standard OLS regression $\Delta S_{t+1} = a + b \cdot s_{1,t} + \eta_t + \nu_t$ would be

$$\hat{b}_{PB} = \frac{cov(\Delta S_{t+1}, s_{1,t})}{var(s_{1,t})} = \left(c_{impact} + \frac{cov(Y_t, \eta_t)}{\sigma_Y^2} \right) \cdot \frac{\sigma_Y^2}{\sigma_Y^2 + \sigma_{\epsilon^2_{Y,i,t}}}, \tag{34}$$

where PB denotes the OLS estimate for the pecuniary benefits case. If the errors in variables are uncorrelated, the instrumental variable estimation produces

$$\hat{b}_{PB,2SLS} = \frac{cov(\Delta S_{t+1}, s_{2,t})}{cov(s_{1,t}, s_{2,t})} = c_{impact} + \frac{cov(Y_t, \eta_t)}{\sigma_Y^2}. \tag{35}$$

There are two important points to highlight in this case. First, the validity of the instrument is not affected by the correlation between the ESG attribute and the stock relevant innovations. This means that the overidentifying restrictions tests will not fail. Second, notice that in this case the coefficient estimated by 2SLS will not be given by c_{impact}, but will include the bias from the correlation between cash flows and ESG performance. This changes the interpretation of the coefficient we estimate but does not invalidate the instrumentation. Importantly, even in this case there will be attenuation bias. By comparing OLS to 2SLS (equations (34) and (35)), one can see that our procedure would still correct the attenuation bias.

- **Case 4. Non-classical Errors-in-Variables:**

$$E[Y_t \cdot \epsilon_{Y_{1,t}}] \neq 0,$$
$$E[Y_t \cdot \epsilon_{Y_{2,t}}] \neq 0.$$

Under these conditions, s_2 may not be a valid instrument. This depends on the form of the correlation of the errors across rating agencies. The exclusion restriction (33) is given by

$$cov(s_{2,t}, -c_{impact} \cdot \epsilon_{Y_{1,t}} + \eta_t + \nu_t) = \underbrace{cov(Y_t, \epsilon_{Y_{1,t}}) + cov(\epsilon_{Y_{2,t}}, \epsilon_{Y_{1,t}})}.$$

The sum of the two covariances on the right-hand side may or may not be zero de-

pending on the correlations. In this case, the 2SLS estimator does not recover the true coefficient. One instructive special case to consider is when the errors are linear in Y_t. Assume the rating agencies' scores are given by

$$\begin{aligned} s_{1,t} &= Y_t + \epsilon_{Y_{1,t}}, \qquad \epsilon_{Y_{1,t}} = \alpha_1 Y_t + \psi_{1,t}, \\ s_{2,t} &= Y_t + \epsilon_{Y_{2,t}}, \qquad \epsilon_{Y_{2,t}}, = \alpha_2 Y_t + \psi_{2,t}, \\ \text{where } E[\psi_{1,t} \cdot \psi_{2,t}] &= 0 \end{aligned} \tag{36}$$

Substituting these into the equation (17), we have

$$\begin{aligned} s_{1,t} &= (1+\alpha_1) Y_t + \psi_{1,t}, \\ s_{2,t} &= (1+\alpha_2) Y_t + \psi_{2,t}, \end{aligned}$$

which implies that when $s_{2,t}$ is used as instrument for $s_{1,t}$ the coefficient on $s_{1,t}$ is scaled by $1/(1+\alpha_1)$. This implies that the instrument variable produces a biased and inconsistent estimate of the true coefficient. The bias is coming from the correlation between the residuals and the fundamental (see equation (36)).

4.2 Valid Instruments: Empirical Findings

Even though it has been widely reported that ESG ratings exhibit low correlations, we find that they perform very well in our first stage estimation. Table 6 reports the F-statistics of the first-stage 2SLS regressions, in which we use all valid instruments for scores of a given ESG rater. The F-statistics range from 95 to 7335, depending on how many and which instruments we include in the first-stage regressions. These are very high values. So all of our instruments have a strong first stage.

Having a strong predictive power for the other agencies' ESG ratings, however, is necessary but not sufficient for an instrument to be valid. If a candidate instrument is correlated with the error term in our main regression (30), it is not a valid instrument. For each regression in Table 4, we use only those instruments that pass the Hansen J test of overidentifying restrictions. Table 6 reports the sets of valid and excluded instruments.

Curiously, we get no rejections of a model with the full set of instruments in Europe. This means that all agencies' ESG scores serve as valid instruments and our assumptions ((19), (20), and (21)) are not violated. This implies that also the versions of these assumptions explained in Appendix A.2 are also not rejected. Therefore, the assumptions about independence of the weights and the individual measurements are not rejected in the case of Europe.

Table 6. Valid Instruments (Pruned IV). This table reports the largest set of valid instruments (other agencies' ratings) to include in the two-stage least squares estimation (29)-(30). RepR stands for RepRisk, TVL for Truvalue Labs, VE for Vigeo-Eiris, SP for SP Global, Ref for Refinitiv, and Sust for Sustainalytics. Our procedure for identifying the set of rating agencies' whose ESG scores serve as valid instruments is as follows. First, for each rating agency reported in the left column, we run the 2SLS regression (29)-(30), using all remaining rating agencies' scores as instruments. We then implement the Hansen J-test of overidentifying restrictions. If the model (29)-(30) passes the Hansen J test test (passed if p-value > 0.05), then all included instruments are valid; otherwise, we exclude instruments, one at a time, until the model passes the J-test. We report the included (✓) and excluded (✗) instruments. The last column reports the F-statistics from the corresponding first-stage regressions.

		Instruments								
	Regressor	MSCI	ISS	RepR	TVL	VE	SP	Ref	Sust	First-stage F-statistic
North America	MSCI		✗	✓	✓	✓	✓	✓	✓	1611
	ISS	✗		✓	✓	✓	✓	✓	✗	6624
	RepR	✗	✗		✓	✓	✓	✓	✗	655
	TVL	✓	✗	✓		✓	✓	✓	✓	397
	VE	✓	✓	✗	✓		✗	✗	✓	6070
	SP	✗	✓	✓	✓	✗		✓	✗	6585
	Ref	✗	✗	✓	✓	✓	✓		✗	7335
	Sust	✗	✗	✓	✓	✓	✓	✓		579
Europe	MSCI		✓	✓	✓	✓	✓	✓	✓	829
	ISS	✓		✓	✓	✓	✓	✓	✓	1940
	RepR	✓	✓		✓	✓	✓	✓	✓	319
	TVL	✓	✓	✓		✓	✓	✓	✓	162
	VE	✓	✓	✓	✓		✓	✓	✓	2610
	SP	✓	✓	✓	✓	✓		✓	✓	1686
	Ref	✓	✓	✓	✓	✓	✓		✓	1393
	Sust	✓	✓	✓	✓	✓	✓	✓		484
Japan	MSCI		✓	✓	✓	✗	✓	✓	✓	554
	ISS	✓		✓	✓	✗	✓	✓	✓	781
	RepR	✗	✗		✓	✗	✓	✓	✗	362
	TVL	✓	✓	✓		✓	✓	✓	✓	95
	VE	✓	✓	✓	✓		✓	✓	✓	2148
	SP	✓	✓	✗	✓	✓		✓	✓	2593
	Ref	✓	✓	✗	✓	✗	✓		✓	1564
	Sust	✓	✓	✓	✓	✓	✓	✓		447

The only assumption that may still be violated is that ESG performance is correlated with stock return relevant information. If so, our 2SLS coefficient is biased due to omitted variable bias. This bias affects the interpretation of the coefficient, but as mentioned before not the validity of the instruments. Notice, however, that the majority of the coefficients go up when we replace OLS by 2SLS. This suggests that the instrumentation is still solving the attenuation bias.

In North America and Japan, however, we see a number of rejections of the model. The instruments that fail the Hansen J test most often in North America are MSCI, ISS, and Sustainalytics and in Japan it is Vigeo-Eiris. There are many possible reasons for the rejections. We do not have enough information to discern between the possible violations. One possibility is that some of the data from these providers is backfilled retroactively, and

so the rejections are occurring in line with Case 2 of Section 4.1. Another interpretation is that the assumption that measurement errors in ESG ratings are independent across raters is violated, for example, due to herding by the ESG rating agencies. As discussed in Case 1 of Section 4.1, this would lead to a failure of the overidentifying restrictions test. Even though there are some rejections, we are still able to perform our 2SLS estimation because we have four or more instruments in each regression that pass the test.

There are three interesting findings about ESG ratings of Truvalue Labs: first, it is the one instrument that never fails the overidentifying restrictions tests; second, as suggested by Figure 1 and discussed below, it is the measurement that has the largest proportion of noise when used as a regressor; and third, it is the regressor that has the weakest first stage, i.e., the lowest F-statistics. It is particularly interesting that, as we show below, Truvalue Labs ESG scores are very noisy, but yet these scores produce a good instrument. The reason Truvalue Labs ratings pass all of our J tests is that, most likely, the noise in the ratings is truly orthogonal to that in other rating agencies' scores, satisfying the assumptions of the classical errors-in-variables problem.[24] One possible explanation for this result is that Truevalue Labs use news outlets and AI to conduct their analysis. This is a very different procedure and data source to most rating agencies.

Refinitiv, RepRisk, and S&P Global are also instruments that are rarely rejected by the overidentifying restrictions tests. In fact, they are rejected in only 5 out of 63 possible specifications. In contrast, MSCI and ISS are rejected in 10 out of the possible 14 specifications in North America.

4.3 An Alternative Procedure for Evaluating Instruments

There is no established procedure (as far as the authors know) for determining the optimal set of instruments in the presence of many instruments, especially when some of them may be invalid. In the previous subsection, we chose the instruments by starting from the largest possible set and pruning instruments one at a time until the model passes the Hansen test of overidentifying restrictions. The main problem with this procedure is that the Hansen test has not being designed for this sequential search. Therefore, the size is likely to be incorrect. In this subsection we implement a different approach. Although it has other weaknesses, the approach does not rely on a sequential search based on the Hansen overidentifying restrictions test. Instead, it uses a very simple Lasso selection. We will show that the results in this section mirror those from our main procedure. It is reassuring that there is a strong

[24] In other words, it is fine for an instrument to be a noisy measure of the true attribute, as long as the noise is orthogonal to that in other agencies' scores. Truvalue Labs is therefore a good candidate instrument for solving the errors-in-variables problem.

correspondence between the two sets of results.

We compare several 2SLS estimates using different subsets of instruments. For any given rating agency, we start with a 2SLS estimator with one instrument, then increase the number of instruments to two, three, etc., all the way to the largest set of available instruments (seven). For each given subset of $M \in \{1, 7\}$ instruments, we select the M rating agencies that maximize the R^2 in the first stage. We perform this selection using Lasso and changing the penalty until M rating agencies are chosen. Once the M rating agencies are decided, we estimate the coefficient using the standard 2SLS procedure. Notice that the set of selected instruments in this subsection is the one that maximizes the explanatory power in the first stage, while that in our main procedure is the largest set which passes the overidentifying restrictions test. The results are presented in Table 7.

Table 7. Lasso-selected subsets of instrumental variables. This table reports the estimates of b_{2SLS} from our second-stage regression $r_{k,t+1} = a_{2SLS} + b_{2SLS} \cdot \hat{s}_{k,i,t} + c_{2SLS} \cdot X_{k,t} + \mu_t + \nu_{k,t}$, where $r_{k,t+1}$ is denotes excess return of stock k in month $t+1$, $s_{k,i,t}$ is the ESG rating of firm k by rater i in month t, X_k is the vector of stock-level controls (log size, book-to-market, EBIT over total assets, market beta, volatility, momentum, and industry and country fixed effects) and $\hat{s}_{k,i,t}$ denotes the fitted value of the ESG rating of firm k by rater i in month t from the first-stage regression $s_{k,i,t} = a_1 + b_1 \cdot z_{k,i,t} + c_1 \cdot X_{k,t} + \mu_t + \varepsilon_{k,t}$. The regressions are run for each rater, whose names are reported in the left column. TVL stands for Truvalue Labs. The left column also reports regions in which firms are located. The 2SLS estimate is from separate two-stage regressions, with one, two, three, etc. instruments. The order in which the added instruments are selected is dictated by Lasso. Coefficients significant at the 5% level are highlighted in bold.

		OLS	IV1	IV2	IV3	IV4	IV5	IV6	IV7
North America	MSCI	**0.186**	0.075	0.114	0.119	0.194	0.201	**0.245**	**0.249**
	ISS	**0.193**	0.06	0.044	0.065	0.092	0.095	0.097	0.097
	RepRisk	0.081	-0.343	-0.023	0.222	0.257	0.313	0.306	0.295
	TVL	0.133	**1.284**	**1.158**	**1.018**	**0.694**	**0.737**	**0.782**	**0.774**
	Vigeo-Eiris	0.045	**0.339**	**0.251**	**0.209**	**0.243**	**0.251**	**0.248**	**0.249**
	SP Global	0.01	**0.282**	**0.211**	**0.17**	**0.2**	**0.188**	**0.194**	**0.194**
	Refinitiv	0.073	0.055	0.041	0.104	0.085	0.114	0.14	0.148
	Sustainalytics	**0.153**	0.249	**0.483**	**0.512**	**0.415**	**0.376**	**0.355**	**0.37**
Europe	MSCI	**0.207**	0.329	**0.318**	0.26	0.253	0.271	0.283	0.283
	ISS	**0.206**	0.236	0.211	0.229	0.237	0.237	0.247	0.247
	RepRisk	0.018	-0.33	0.086	-0.577	0.007	0.065	0.064	0.061
	TVL	0.033	**1.159**	**0.673**	0.577	**0.604**	**0.603**	**0.607**	**0.608**
	Vigeo-Eiris	0.15	**0.307**	**0.29**	**0.24**	**0.262**	**0.261**	**0.261**	**0.261**
	SP Global	0.148	0.235	0.202	0.235	**0.268**	0.237	**0.263**	**0.261**
	Refinitiv	0.076	0.289	**0.31**	**0.303**	**0.352**	**0.353**	**0.353**	**0.353**
	Sustainalytics	0.107	**0.737**	**0.553**	**0.539**	**0.565**	**0.549**	**0.528**	**0.528**
Japan	MSCI	**0.143**	0.197	0.283	**0.353**	**0.393**	**0.376**	**0.39**	**0.393**
	ISS	**0.27**	0.205	0.3	**0.333**	**0.355**	**0.382**	**0.375**	**0.383**
	RepRisk	**0.159**	0.007	-0.473	-0.017	0.095	0.007	-0.043	-0.01
	TVL	0.001	**0.796**	**0.914**	**1.052**	**1.116**	**1.171**	**1.196**	**1.202**
	Vigeo-Eiris	0.1	**0.378**	**0.319**	**0.329**	**0.35**	**0.361**	**0.363**	**0.362**
	SP Global	**0.163**	0.126	**0.193**	**0.223**	**0.188**	**0.21**	**0.215**	**0.214**
	Refinitiv	**0.191**	0.149	0.21	0.159	0.188	0.204	0.2	0.203
	Sustainalytics	**0.197**	**0.445**	**0.537**	**0.565**	**0.578**	**0.579**	**0.583**	**0.582**

Table 7 presents the point estimates from the OLS and 2SLS regressions of stock returns on a rating agency scores and a set of controls. This is exactly the same regression as in

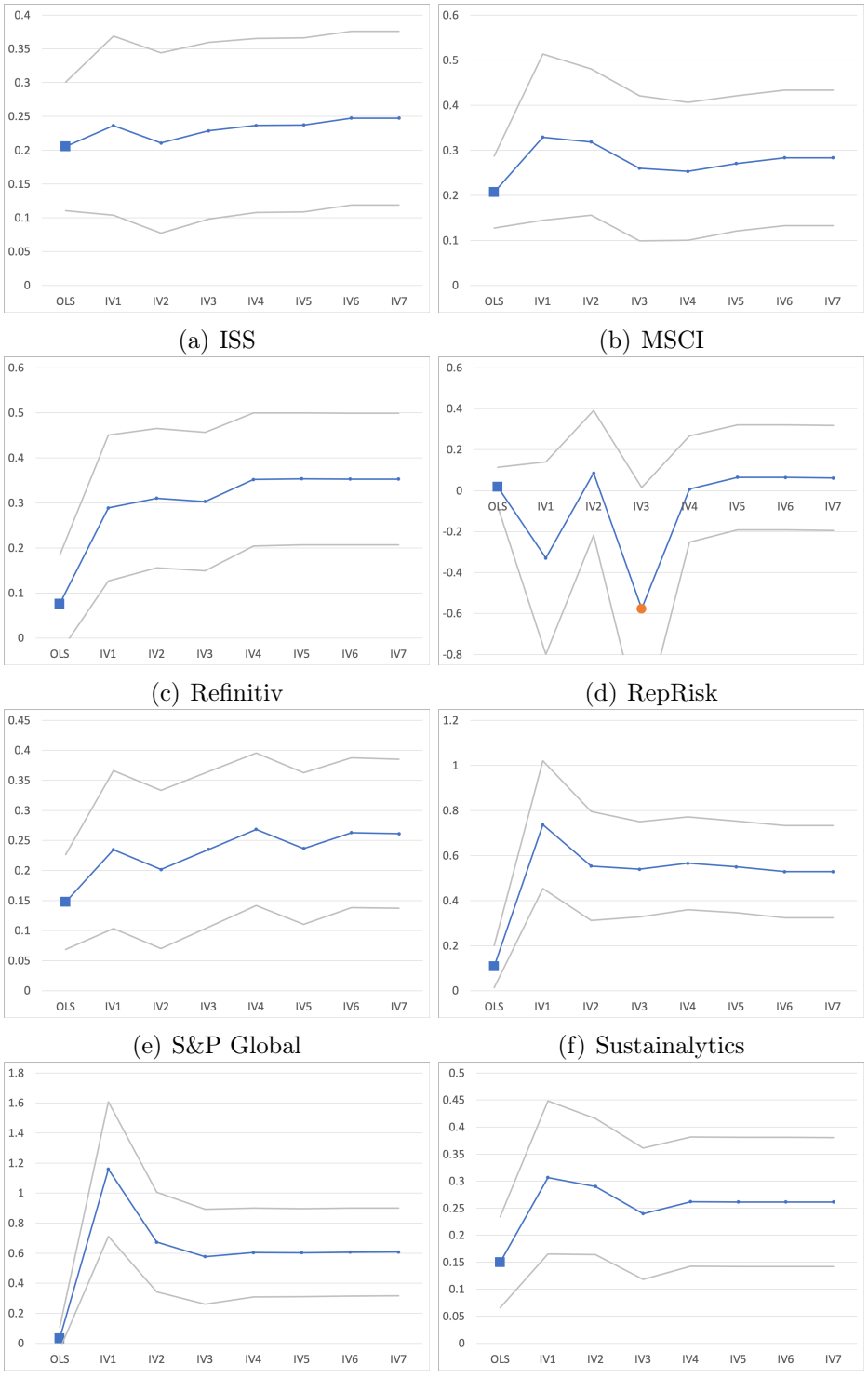
equation (30) except that we have changed the procedure of selecting the instruments. The OLS estimates are shown in column one; the remaining columns report the 2SLS estimates using 1,2,...,7 instruments. To simplify exposition, we only present the point estimates, indicating their significance at the 5% level in bold.[25]

The results are best illustrated using figures. Figures 3, 4, and 5 present the point estimates, one standard error confidence bands, and rejections of the Hansen J test of overidentifying restrictions (OIR) (rejected if p-values are less than 5 percent). The latter are important to report because the procedure in this section does not guarantee that all instruments are valid, and the Hansen J test rejects the model if it encounters an invalid instrument. The quantities on the x-axis are the different subsets of IV (with one, two, three, etc., instruments), and the y-axis measures the estimated coefficient (multiplied by 100, as we have done in prior tables). The blue curve depicts the point estimates; the OLS estimate is indicated with a square, while the different 2SLS point estimates are shown with small blue dots. When the OIR test is rejected (p-value $< 5\%$), the marker becomes an orange circle. The gray curves represent the one standard error confidence bands for each estimate. We present these bands to provide a sense of the precision of the estimates and for comparison across the different estimates. There are eight panels for each region, one for each ESG rating agency.

Figure 3 presents the results for European firms. Similarly to our findings in subsection 4.2, the instruments work very well. Several results are worth highlighting. First, notice that there are very few rejections of the OIR test, reproducing the results from the previous subsection. The only rejection occurs for RepRisk when using a 2SLS estimator with three instruments (marked with an orange circle); the rest of the specifications all pass the Hansen J test. It is possible to interpret this single rejection as a false positive—especially in light of accepting the specification when more instruments are included. Furthermore, instrumenting RepRisk scores is particularly challenging. Notice that for RepRisk the coefficients are very volatile, until at least four instruments are included. As discussed earlier, RepRisk's scoring procedure is markedly different from those of other rating agencies, and it is therefore harder to find good instruments for RepRisk. However, with a large enough set of instruments it is possible to do so and therefore we observe no OIR rejections for the cases of five or more instruments, consistent with the results of our main procedure, applied to European firms.

Second, we replicate our most important result: the attenuation bias. Notice that for almost all the rating agencies, 2SLS estimates exceed the OLS ones, as in our earlier analysis. However, more instruments does not necessarily imply larger coefficients. An instrumental-variables-based approach produces larger coefficients, but the relationship between the num-

[25]The standard errors of the estimates are not reported and are available upon request.



(a) ISS (b) MSCI

(c) Refinitiv (d) RepRisk

(e) S&P Global (f) Sustainalytics

(g) Truvalue (h) Vigeo-Eiris

Figure 3. Estimation for Europe: Coefficients, confidence bands, and OIR test rejections. Point estimates are in blue. Gray bands are one standard error confidence bands for each estimate. Orange circles signify rejections of instrument validity by the OIR test (Hansen J test)

ber of instruments and the magnitudes of 2SLS coefficients is non-monotonic.

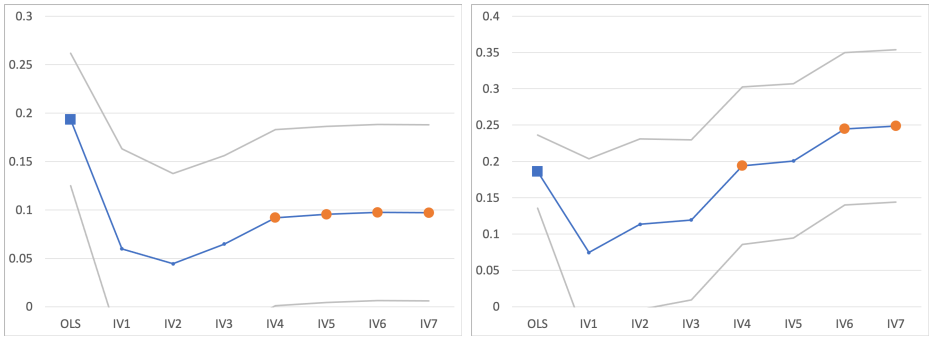
Third, and quite importantly, notice the stability in the coefficients estimated with different subsets of instruments. In general, the point estimates of all 2SLS regressions for a given ESG rating agency tend to be close to each other. More formally, it is easy to see from the figures that, for each rating agency, the one standard error confidence intervals for each 2SLS estimator overlap across all model specifications. Even in the case of estimation for RepRisk with three instruments, when the point estimate becomes negative and the overidentifying restriction is rejected, the confidence band includes every 2SLS point estimate.[26]

The results for North America share many similarities with those for Europe but there are also some noteworthy differences. First, notice that there are many rejections of the overidentifying restrictions tests. For every rating agency, the figures feature several orange circles, especially when many instruments are included. In fact, it is clear that the more instruments are included, the more likely it is that the OIR test rejects the model. The intuition for this is that once an invalid instrument is chosen (because it has high explanatory power in the first stage) by Lasso, it is likely to persist as a regressor as we expand the number of instruments.

Second, even though we also find that in general the OLS estimates are lower than 2SLS ones because of the attenuation bias, that is not the case for all rating agencies. For ISS, all the 2SLS estimates are smaller than the corresponding OLS coefficient. For MSCI, only two estimates are higher, and in both cases the OIR test indicates that the model should be rejected. The case of ISS is an interesting one because, going back to Table 5, the Hausman test for ISS shows a negative difference (meaning that 2SLS is smaller than OLS) and that difference is statistically different from zero. For the other six rating agencies, we do find that the instrumental variables approach produces higher point estimates, consistent with attenuation bias. Additionally, notice that ISS, MSCI, Refinitiv and RepRisk are the four rating agencies in which there are some IV estimates that are lower than the OLS estimate. Similarly, in Table 5 ISS, Refinitiv and RepRisk are the rating agencies for which the difference in the 2SLS and OLS coefficients is negative, although not all are statistically significant. It is reassuring that the results using two very different instrumental variables selection procedures mirror each other.

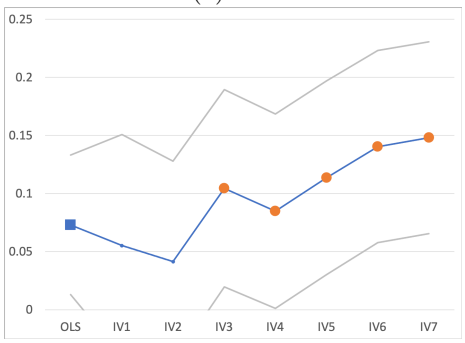
In summary, for the case of North America, most of our main messages delivered for European firms remain true: OLS estimates tend to be biased downward, and 2SLS coefficients estimated with different subsets of instruments are not statistically different from each other.

[26] We did not perform formal specification tests for all these possible comparisons. We believe that showing the one standard error confidence bands should be convincing enough to ameliorate concerns about the stability of the parameters.

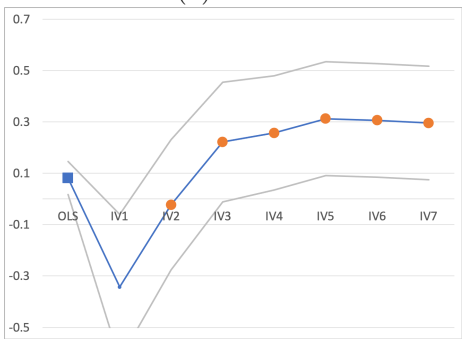


(a) ISS

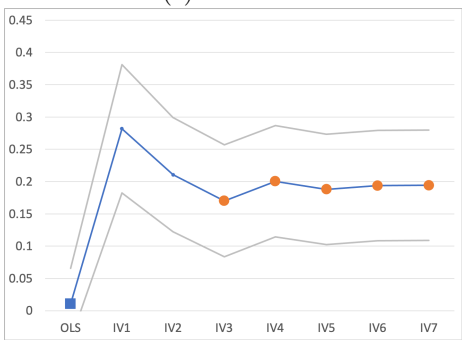
(b) MSCI



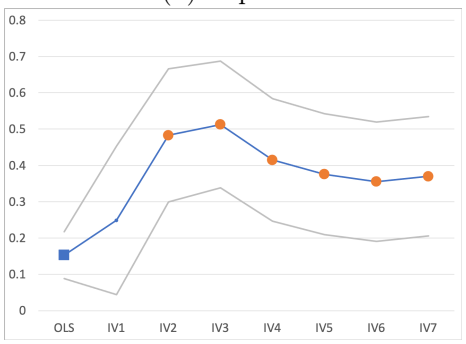
(c) Refinitiv



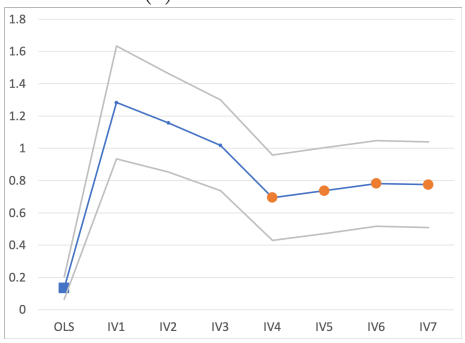
(d) RepRisk



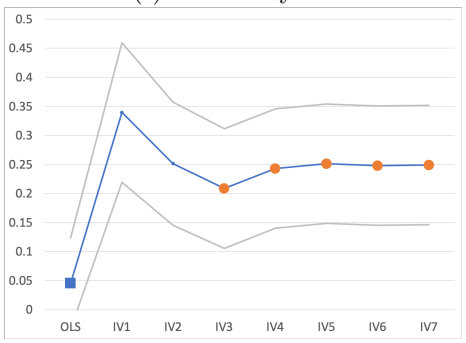
(e) S&P Global



(f) Sustainalytics

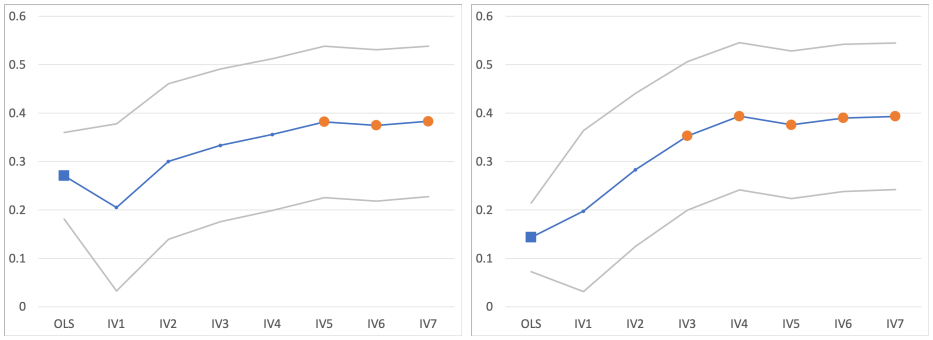


(g) Truvalue



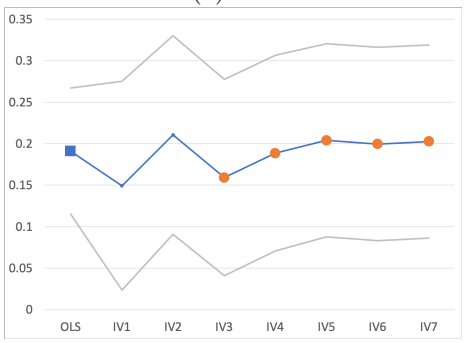
(h) Vigeo-Eiris

Figure 4. Estimation for North America: Coefficients, confidence bands, and OIR test rejections. Point estimates are in blue. Gray bands are one standard error confidence bands for each estimate. Orange circles signify rejections of instrument validity by the OIR test (Hansen J test).

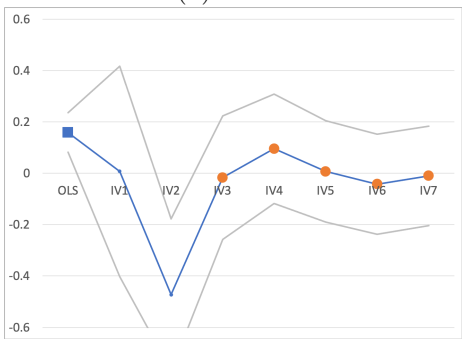


(a) ISS

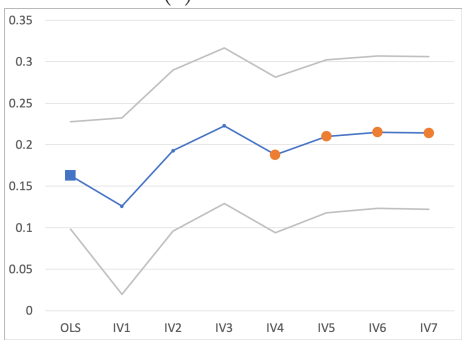
(b) MSCI



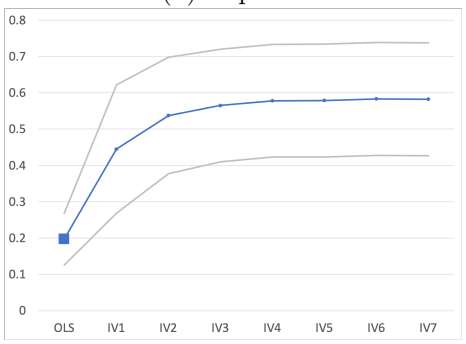
(c) Refinitiv



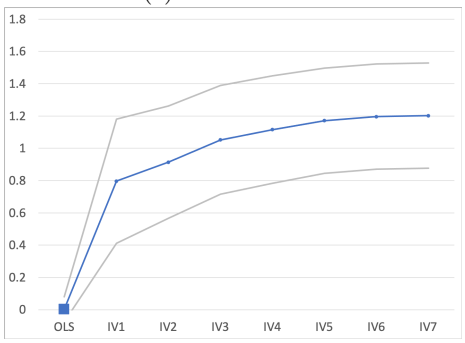
(d) RepRisk



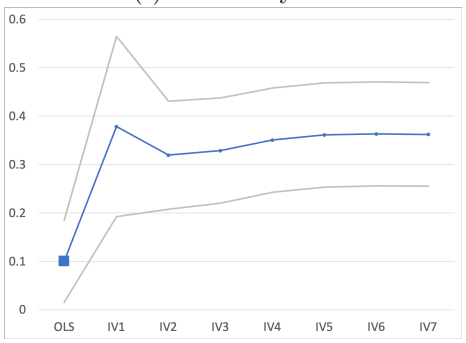
(e) S&P Global



(f) Sustainalytics



(g) Truvalue



(h) Vigeo-Eiris

Figure 5. Estimation for Japan: Coefficients, confidence bands, and OIR test rejections. Point estimates are in blue. Gray bands are one standard error confidence bands for each estimate. Orange circles signify rejections of instrument validity by the OIR test (Hansen J test).

Finally, let us discuss our findings for Japan. Figure 4 reveals that, again, some specifications contain invalid instruments. The model fails the overidentifying restrictions test for ISS, MSCI, Refinitiv, RepRisk, and S&P Global. Those rejections tend to occur when more instruments are included. There is no rejection at all for Sustainalytics, Truvalue and Vigeo-Eiris. Overall, in Japan there are fewer OIR rejections than in North America, but more than in Europe.

Second, with the exception of RepRisk, all the coefficients are positive and there is an attenuation bias in the OLS estimator. For Sustainalytics, Truvalue, and Vigeo-Eiris there is a large increase in the estimates as we move from OLS to 2SLS, whereas for ISS, MSCI, Refinitiv, and S&P Global the increases are smaller, but still there are increases.

Third, similarly to those in our analysis for Europe and North America, the 2SLS confidence intervals tend to overlap with each other, highlighting the stability of the coefficients.

The case of RepRisk shows that with three instruments the point estimate is negative and the model passes the OIR test. This result mirrors the results from our main procedure. Table 5 indicates that RepRisk for Japan is one of the two cases in which the 2SLS is smaller and statistically different from OLS (the difference is negative).

In summary, this subsection has used an alternative instrument selection procedure and the results mirror those from our main procedure. As we said before, our preferred procedure is one in which we start with the maximum number of possible instruments and prune them one by one until the model passes the overidentifying restrictions test. The Hansen test was not designed for this sequential search and therefore in this subsection we perform the estimation using a different procedure. We start by specifying how many instruments are going to be included and choose the rating agencies that form the "best" set. We do this by evaluating the explanatory power of the instruments in the first stage. This procedure can be implemented by comparing the mean squared errors of the different possible combinations or by using Lasso. We have decided on the latter because of computational simplicity. For this set of instruments, the OIR test is properly sized, which is an advantage of the procedure presented in this subsection. The main disadvantage of the procedure is that sometimes a rating agency will be picked when 3 instruments are included, but not picked when 4 are included. As in many Lasso applications, the stability of the included variables is not guaranteed.

Having highlighted the weaknesses of both procedures, we note that it is comforting that the results of both procedures are very close to each other. In other words, the results are quite robust to changes in the instrument selection criteria.

Finally, let us reiterate the important results from the empirical analysis. First, there

is a strong attenuation bias in the OLS coefficients. An instrumental variables estimator delivers higher magnitudes for almost all of them. Second, not all rating agencies are good instruments at all times and for all rating agencies. In other words, we find many combination of instruments in which the overidentifying restrictions are rejected, implying that not all of the instruments are valid. Third, for the instrument selection procedure proposed in this subsection the model fails the overidentifying restrictions tests for the same cases as with our original procedure, presented in the previous subsection.

5 Discussion

One may wonder why we are not using the principle component analysis to extract common factors driving ESG ratings in our sample. The issue with such analysis is that principal components finds the linear combination that maximizes the observed variance. This approach is quite useful when the variables are measured correctly; however, if the variables are measured with noise, it puts higher weights on noisier ratings. Since our goal is to correct for the noise, the principal component analysis is ill-suited for our purposes. Intuitively, all things equal, we would like to put lower weight on noisier ratings, similar to what a Kalman filter would do.

A disadvantage of using filtering to correct measurement error is that a filter could include ESG ratings whose errors are correlated with firms characteristics — i.e. suffer from endogenous bias. In contrast, our pruned IV method explicitly excludes invalid instruments and corrects for some forms of endogeneity.

Additionally, computing an average of the available rating agencies' scores does not fully eliminate the noise, given that invalid instruments will be included in the average.[27]

It is important to correctly interpret the notion of noise. Our procedure disentangles the impact of noise only within the context of the outcome variable that we are measuring. In this study, the noise in ESG ratings is determined with respect to next month's expected stock returns. The reasoning is that ESG investors, collectively, are guided by the true ESG performance, which is a latent variable. As a result, stock returns influenced by ESG investors allow us to separate signal from noise in ESG ratings. However, the latent quality of ESG performance may also enter stock returns over a longer time frame or other outcome variables—such as environmental accidents, regulatory fines, investor sentiment or reputation

[27]The purpose of the averaging of rating agencies' scores is that if the errors are uncorrelated, the average reduces the size of the errors. However, invalid instruments have correlated errors, and therefore the naive averaging of scores will reduce the errors in lesser extent. It is better to exclude the scores that are invalid instruments. Furthermore, the scaling of each rating agency affects the average while it has no impact on the instrumentation.

of a firm, and management quality. Furthermore, investors' collective view may deviate from the way other stakeholders such as governments or civil society organizations view ESG performance. Finding alternative benchmarks to identify noise presents interesting avenues for future research. In fact, it is conceivable that a rating agency that is deemed very noisy for one outcome variable is not noisy for a different one. Within the confines of this study, it is important that noise is determined with respect to the collective view of ESG investors, revealed in monthly stock returns.

Our empirical analysis is not free of limitations. First, our time series are short, and especially so for stock return regressions. We can do very little about this problem because the rating agencies to date have produced data for a relatively short time frame. The longest one we had (KLD) was discontinued, and, even if it were not, having only one rating agency would not have sufficed in our case.

Second, and equally importantly, our estimation relies primarily on cross-sectional variation. The reason is that many rating agencies change their scores at most once a year—with the exception of TruValue and RepRisk who are changing in almost real time. This implies that in practice the time series is even shorter that the time span. Some of the rating agencies change the firm scores not even once a year.

Third, many rating agencies are in the process of consolidation and they often revise their procedures. Some rating agencies back-fill their past scores based on a revised procedure. This is particularly problematic if the back-fill use stock-relevant information. Practitioners and applied work can use our procedure for diagnosing this problem given that it will reject the over-identifying restriction. The back-filling creates a correlation between the score and the stock relevant information.

Fourth, the ESG score could be capturing unobservable firm characteristics, such as management quality. It can be assumed that "better" management increases stock returns, and makes the company more concerned about ESG issues. For example, better managers produce less waste, or better managers are able to motivate their workers better and at the same time are more concerned about labor mistreatment. If that is true, firms with better management quality have both higher returns and higher ESG scores. As we have discussed before, this omitted variable bias has no impact the attenuation bias results we have presented. It changes, however, the structural interpretation of the coefficient. Still there are many practical applications where an accurate estimation of the ESG impact on stock returns is worth performing. For example, a portfolio constructed using the ESG scores will be better constructed when firms are organized by the 2SLS coefficient rather than OLS, which is contaminated by noise. Nevertheless, future research should look into disentangling management quality and ESG performance. An easy implementation would have been to

estimate using the time series variation, adding firm fixed effects. A downside of our sample is that it is short and a very small proportion of the variation comes from the time series.

Finally, we do not observe capital flows toward ESG. In the language of our model, we do not observe λ. This implies that the coefficients we estimate do not have structural interpretation. This is equivalent to having a omitted variable bias—for example, in our model, if the share of ESG investors is constant, the coefficients should be negative, but when capital flows increase unexpectedly the coefficient is positive. One advantage of our procedure is that the error in the rating agencies scores are very unlikely caused by capital flows—although the opposite cannot be assumed. Our specification is using the rating's impact on future expected returns and therefore the attenuation bias is unaffected by the presence of unexpected shocks to capital inflows.

5.1 Simulations

In order to compare our estimation procedure to other noise-reduction methods and to highlight the power of the overidentifying restristion test, we perform several simulations.

We generate 50,000 observations of stock returns and ESG performance from the following structural model:

$$\Delta S_k = \beta \cdot Y_k + \eta_k, \tag{37}$$

where k indexes observations, Y_k is a random normal variable with mean 0 and standard deviation 1; η_k represents the stock return relevant innovations which we assume to be highly volatile, with the standard deviation of 10. The coefficient of interest is β, which we set to be 0.5. We construct 5 ESG rating agencies' scores, modeled as

$$s_{k,i} = Y_k + \epsilon_{Y_i}. \tag{38}$$

All scores have a mean of Y_k and the standard deviations of ϵ_{Y_i}, $i = 1, \ldots, 5$ are $\{5, 3, 1, 2, 1\}$ respectively. In what follows, we suppress the subscript k for expositional reasons.

All the interest is centered around s_1. We explore different correlation structures of the noise of the ESG rating s_1. In the first simulation, it will be correlated with the unobservable firm returns (η) and in the second, correlated with the noise in the ESG rating s_3. The correlations will be fluctuating from -70% to 70%. All other errors are uncorrelated with each other, as in our assumptions (19), (20) and (21).

We estimate several model specifications. First, we run a simple OLS:

$$\Delta S = b_{OLS} \cdot s_1 + \text{error}. \tag{39}$$

Second, one may conjecture that an index constructed as an average of the three raters' scores would be less noisy as each individual rating and recommend using a simple average of the scores instead of an individual score as a regressor. We therefore construct a simple average of the first three scores, $sa = 1/3(s_1 + s_2 + s_3)$, and estimate the following regression:

$$\Delta S = b_{sa} \cdot sa + \text{error}. \tag{40}$$

Third, one may suggest using the principal components analysis as a noise reduction procedure. We therefore estimate the principal component of the three rating agencies (denoted as pc) and estimate the regression

$$\Delta S = b_{pc} \cdot pc + \text{error}. \tag{41}$$

Next, we perform two instrumental variable estimations and compute the corresponding overidentifying restrictions tests. We do know that when s_1 is correlated with the error in s_3, the instrument should be invalid. So, we perform a 2SLS estimation using all valid instruments

$$z = \{s_1, s_2, z_1, z_2\}.$$

Finally, we repeat the above 2SLS estimation with a smaller set of instruments, namely, the two rating agencies that are always excluded from all the analysis (which by construction are valid instruments)

$$z = \{z_1, z_2\}.$$

The first simulation varies the correlation between s_1 and η. The results are presented in Table 8. First, let us concentrate on the row corresponding to the correlation of zero. For this row, the assumptions of the classical errors in variables problem are satisfied. In this case, we clearly see the attenuation bias in the OLS estimation: an estimate of 0.022 instead of 0.5. The bias becomes smaller if the simple average of the three ratings is used as a regressor instead of an individual rating s_1: the coefficient increases to 0.106. But it is still significantly smaller than the true coefficient. Despite fact that the other two scores, s_2 and s_3, contain less noise, the resulting estimate is still far from 0.5. The third column shows the estimates when we use the first principal component as the regressor. It is striking that the estimate is even worse than that from the OLS regression on the individual score s_1. The next two columns show the estimates from the two 2SLS procedures. Notice that both estimates are extremely close to 0.5. The last two columns show the p-values of the overidentifying restrictions tests for the 2SLS estimations and both models pass the test.

We now turn to the remaining rows of Table 8. These rows represent a form of mis-

Correlation of ϵ_{Y_1} and η	(1) OLS	(2) Average	(3) PCA	(4) 2SLS All	(5) 2SLS z's	OIR All (p-value)	OIR z's (p-value)
-0.7	-0.406	-0.653	-0.401	0.497	0.506	0.341	0.164
-0.6	-0.345	-0.545	-0.340	0.496	0.504	0.398	0.197
-0.5	-0.284	-0.436	-0.280	0.495	0.503	0.458	0.234
-0.4	-0.223	-0.327	-0.219	0.495	0.502	0.519	0.276
-0.3	-0.161	-0.219	-0.158	0.494	0.500	0.582	0.321
-0.2	-0.100	-0.111	-0.098	0.493	0.500	0.643	0.369
-0.1	-0.039	-0.003	-0.037	0.493	0.499	0.702	0.421
0	0.022	0.106	0.024	0.493	0.498	0.758	0.476
0.1	0.082	0.214	0.084	0.493	0.497	0.809	0.532
0.2	0.143	0.321	0.144	0.492	0.497	0.855	0.590
0.3	0.204	0.429	0.205	0.492	0.497	0.893	0.647
0.4	0.265	0.537	0.265	0.493	0.496	0.926	0.704
0.5	0.325	0.645	0.325	0.493	0.496	0.951	0.759
0.6	0.386	0.753	0.385	0.493	0.496	0.970	0.811
0.7	0.447	0.860	0.446	0.494	0.496	0.983	0.859

Table 8. Simulation 1: Estimates for the case when the errors in ESG ratings s_1 are correlated with cash-flow relevant innovations η. The true value of the coefficient is 0.5. OIR stands for the Hansen J test of overidentifying restrictions, and corresponding column reports the p-values for the test.

specification of the noise that is particularly interesting—when the noise in rating agencies' scores is correlated with stock returns. First, imagine a rating agency wants to "look good" by backfilling its ratings and retroactively assigning higher scores to those firms that had experienced higher stock returns. This would introduce a positive correlation between stock returns and the error term. The first column reports the OLS coefficient on s_1. Notice that the estimates move with the correlation between the noise in the score s_1 and stock returns. Notice that indeed the OLS coefficient increases, and it can increase quite substantially—we have calibrated the simulation to be able to achieve a coefficient close to the true one. Second, it could be argued that the positive correlation could be also a form of greenwashing, i.e., where the firm with higher returns provides false information to the rater. Symmetrically, a negative correlation biases the coefficient downward.

The PCA (column (3)) produces estimates that are virtually identical to those from the OLS regression with just one rating agency. Of course, we have calibrated the simulation to generate this result. We set the variance of s_1 to be much higher than the variance of s_2 and s_3. Because of this feature, the first principal component loads very highly on s_1. In other words, the only way in which PCA produces a smaller variance of the coefficient is if the most volatile score happens to have the smallest misspecification—something that cannot be tested or assumed. While PCA is useful in many applications, it is not effective

in eliminating the noise in the ESG scores.

Finally, notice that the two 2SLS procedures give the correct estimate regardless of the degree of misspecification. Furthermore, because all the instruments are valid, all the overidentifying restrictions tests are passed and the coefficients in the fourth and fifth columns are virtually identical.

The second simulation considers the case in which some instruments are invalid. Specifically, we change the correlation between the measurement errors in s_1 and s_3 to vary between -0.7 to 0.7. Results are presented in Table 9.

The row corresponding to the correlation of zero coincides with that in Table 8, discussed earlier. In this case, the only problem is the attenuation bias. When the correlations between the errors in s_1 and s_3 are different from zero, the OLS estimate does not change (as one should expect, since s_3 is not used in this regression). The estimate of the coefficient on the average ESG score (column (2)), however, changes. The reason is very simple. When the correlation between the errors is negative, the errors cancel each other, making the average less noisy. Notice that the bias in the estimated coefficient decreases in these cases — or, in other words, that the coefficient increases, moving towards the true value of 0.5.

Correlation of ϵ_{Y_1} and ϵ_{Y_2}	(1) OLS	(2) Average	(3) PCA	(4) 2SLS All	(5) 2SLS z's	OIR All (p-value)	OIR z's (p-value)
-0.7	0.022	0.126	0.021	-0.021	0.506	0.000	0.475
-0.6	0.022	0.123	0.021	-0.014	0.504	0.000	0.476
-0.5	0.022	0.120	0.022	-0.000	0.503	0.000	0.476
-0.4	0.022	0.117	0.022	0.031	0.502	0.000	0.476
-0.3	0.022	0.114	0.023	0.103	0.500	0.000	0.476
-0.2	0.022	0.111	0.023	0.263	0.500	0.000	0.476
-0.1	0.022	0.108	0.023	0.479	0.499	0.000	0.476
0	0.022	0.106	0.024	0.493	0.498	0.758	0.476
0.1	0.021	0.103	0.024	0.357	0.497	0.000	0.476
0.2	0.021	0.100	0.024	0.248	0.497	0.000	0.476
0.3	0.021	0.098	0.024	0.179	0.497	0.000	0.476
0.4	0.021	0.096	0.024	0.137	0.496	0.000	0.476
0.5	0.021	0.093	0.024	0.109	0.496	0.000	0.476
0.6	0.020	0.091	0.024	0.090	0.496	0.000	0.476
0.7	0.020	0.089	0.024	0.076	0.497	0.000	0.476

Table 9. Simulation 2: Estimates for the case when the measurement errors in ESG ratings s_1 and s_3 are correlated. The true value of the coefficient is 0.5. OIR stands for the Hansen J test of overidentifying restrictions, and corresponding column reports the p-values for the test.

The coefficient on the first principal component (column (3)) changes because the change

in the correlation of the scores implies small changes in the eigenvector. However, the improvements are very small. All estimated coefficients remain far from the true value of 0.5.

The estimates of the 2SLS regressions is where this simulation provides some interesting results. First, when all four instruments are used in the first stage regression (column (4)), the coefficient moves away from 0.5, which is in stark contrast to the stability shown in Table 8. The coefficients in column (4) are close to 0.5 only when the correlation is between -0.1 and 0.1. Interestingly, as the correlation moves away from 0, the overidentifying restriction test is rejected very rapidly, detecting a problem with the instruments. This shows that in the simulation the power of the overidentifying restrictions test is quite high. It has 50,000 observations, which is same order of magnitude in our data. Quite robustly, however, when only z_1 and z_2 are used as instruments (column (5)), the overidentifying restrictions are not rejected, and the true coefficient is recovered.

Figure 6 shows how the overidentifying restriction test p-value evolves with the correlation between the measurement errors in the two rating scores, ϵ_{Y_1} and ϵ_{Y_3}. This implies that the invalid instrument is s_3, by construction. To illustrate the power of the OIR test we concentrate on the -10 to +10 percent correlation range, a small deviation around zero. The thick black line is the 2SLS estimate when all four rating agencies' scores are used. For this range, the estimates change slightly with the degree of misspecification. The thick red line is the p-value of the OIR test. The thin red line represents the 0.05 p-value. The power of the overidentifying restrictions test in our setting is striking. Notice that for any correlation larger than 0.03 (in absolute value) has p-values below the threshold. There is a strong rejection of the overidentifying restrictions test, diagnosing the invalidity of s_3 as an instrument. In comparison, there is not a single rejection when z_1 and z_2 are used as instruments.

These result gives us some confidence on our two procedures for selecting instruments. The pruning procedure tends to accept a feasible set too soon. However, given how strong the rejections are in the simulation, even if the size of the test is changed (e.g., from 0.05 to 0.005), we are likely to find a set of instruments that are valid. Of course, this is a conjecture, but it is also supported by our second procedure of selecting instruments, based on Lasso and described in Section 4.3, in which we tend to reject too early.

6 Conclusions

It is notoriously difficult to measure ESG performance of firms. ESG rating agencies often report different estimates for the same attribute. In this paper, we argue that a high level of

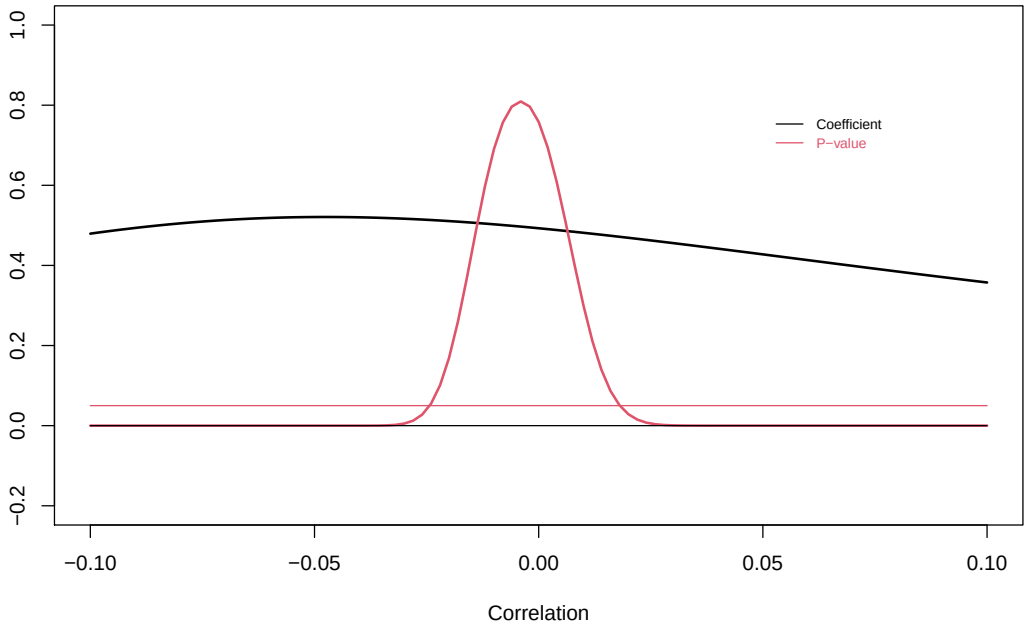


Figure 6. The power of the Hansen J test of overidentifying restrictions. This figure plots the 2SLS coefficient from column (5) of Table 9 (solid black curve), the p-value of the Hansen J test of overidentifying restrictions (thick red curve), and the 0.05 p-value for the Hansen J test (thin red line) as functions of the correlation between measurement errors in the simulated ESG ratings s_1 and s_3.

noise in the estimates leads to a significant bias in the standard regressions that analyze the effects of ESG performance. An important institutional feature of the market for ESG ratings is that there are numerous raters, who use different inputs and methodologies in computing their ratings. We exploit this feature and propose an instrumental variable approach to correcting the bias, which is predicated on using scores of different agencies as noisy measures of a true ESG attribute.

We show that standard regression estimates of the effects of ESG on stock returns are downward biased and, on average, more than double once we apply our noise-correction procedure. We run our estimation separately for all raters in our sample, across three geographical regions, and in the majority of these regressions we observe an increase in the estimates. Our results are therefore supportive of the errors-in-variables interpretation of the discrepancies in scores of ESG rating agencies.

The practical take-away of these results is that it is worthwhile to rely on several complementary ESG ratings. While we find scores of some rating agencies to be very noisy, it does not mean that they are uninformative. Good examples are RepRisk and Truvalue

Labs. These raters construct their ratings based on media reports, rather than, for example, mandatory disclosure or expert assessment. Our estimation procedure shows that while these scores do not perform well as predictors, they are valuable instruments that improve the prediction of other scores.

We provide a ranking of ESG rating agencies' scores, from least noisy to the noisiest. One may be tempted to conclude from our results that one should use ESG scores containing the least amount of noise, rather than scores of other raters. We caution against such interpretation. First, this ranking is specific to our model and regression setup and should not be overgeneralized. Second, our results show that also for the least noisy ratings coefficients increase substantially when instrumented with other ratings. We show that our valid set of instruments never falls below four, and often includes scores of all rating agencies in our sample. In other words, relying on the scores of several complementary ratings yields better results.

We apply our noise-correction procedure to aggregate ESG scores. It would be interesting to perform our estimation for E, S, and G scores separately, as well as for individual indicators that raters use in construction of E, S, and G scores. This would give one a sense which pillar of ESG scores contains the largest amount of noise. It may also further inform the discussion on materiality of certain indicators (see Khan, Serafeim and Yoon, 2016) and the need for harmonized reporting standards.

7 References

Adams, Renée B., and Daniel Ferreira. 2009. "Women in the Boardroom and their Impact on Governance and Performance." *Journal of Financial Economics*, 94(2): 291–309.

Albuquerque, Rui, Yrjö Koskinen, and Chendi Zhang. 2019. "Corporate Social Responsibility and Firm Risk: Theory and Empirical Evidence." *Management Science*, 65(10): 4451–4469.

Avramov, Doron, Si Cheng, Abraham Lioui, and Andrea Tarelli. 2021. "Sustainable investing with ESG rating uncertainty." *Journal of Financial Economics*.

Bauer, Rob, Nadja Guenster, and Rogér Otten. 2004. "Empirical Evidence on Corporate Governance in Europe: The Effect on Stock Returns, Firm Value and Performance." *Journal of Asset Management*, 5(2): 91–104.

Bebchuk, Lucian, Alma Cohen, and Allen Ferrell. 2009. "What Matters in Corporate Governance?" *The Review of Financial Studies*, 22(2): 783–827.

Berg, Florian, Julian F. Kölbel, and Roberto Rigobon. 2020. "Aggregate confusion: The divergence of ESG ratings." *Working Paper, MIT Sloan*. Available at: https://papers.ssrn.com/abstract=3438533.

Berg, Florian, Kornelia Fabisik, and Zacharias Sautner. 2021. "Rewriting History II: The (Un)Predictable Past of ESG Ratings." *Working Paper, ECGI*. Available at: https://ssrn.com/abstract=3722087.

Bolton, Patrick, and Marcin T. Kacperczyk. 2020. "Carbon Premium around the World." *Working Paper, Imperial College*. Available at: https://www.ssrn.com/abstract=3550233.

Broccardo, Eleonora, Oliver Hart, and Luigi Zingales. 2020. "Exit vs. Voice." *Working Paper, NBER*. Available at: https://ssrn.com/abstract=3671918.

Chatterji, Aaron, Rodolphe Durand, David I. Levine, and Samuel Touboul. 2016. "Do Ratings of Firms Converge? Implications for Managers, Investors and Strategy Researchers." *Strategic Management Journal*, 37(8): 1597–1614.

Chava, Sudheer. 2014. "Environmental Externalities and Cost of Capital." *Management Science*, 60(9): 2223–2247.

Christensen, Dane M., George Serafeim, and Anywhere Sikochi. 2021. "Why is Corporate Virtue in the Eye of The Beholder? The Case of ESG Ratings." *The Accounting Review (forthcoming)*.

Christensen, Hans Bonde, Luzi Hail, and Christian Leuz. 2021. "Mandatory CSR and Sustainability Reporting: Economic Analysis and Literature Review." *Working Paper, University of Chicago*. Available at: https://papers.ssrn.com/abstract=3427748.

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): 439–461.

Edmans, Alex. 2011. "Does the Stock Market Fully Value Intangibles? Employee Satisfaction and Equity Prices." *Journal of Financial Economics*, 101(3): 621–640.

El Ghoul, Sadok, Omrane Guedhami, Chuck C.Y. Kwok, and Dev R. Mishra. 2011. "Does Corporate Social Responsibility Affect the Cost of Capital?" *Journal of Banking & Finance*, 35(9): 2388–2406.

Fama, Eugene F, and Kenneth R. French. 2007. "Disagreement, Tastes, and Asset Prices." *Journal of Financial Economics*, 83(3): 667–689.

Friedman, Henry L, and Mirko S. Heinle. 2016. "Taste, Information, and Asset Prices: Implications for the Valuation of CSR." *Review of Accounting Studies*, 22: 740–767.

Gibson, Rajna, Philipp Krueger, and Peter Steffen Schmidt. 2021. "ESG Rating Disagreement and Stock Returns." *Financial Analyst Journal (forthcoming)*.

Gompers, Paul A., Joy L. Ishii, and Andrew Metrick. 2003. "Corporate Governance and Equity Prices." *Quarterly Journal of Economics*, 118(1): 107–155.

Heinkel, Robert, Alan Kraus, and Josef Zechner. 2001. "The Effect of Green Investment on Corporate Behavior." *Journal of Financial and Quantitative Analysis*, 36(4): 431–449.

Kashyap, Anil K, Natalia Kovrijnykh, Jian Li, and Anna Pavlova. 2021. "The Benchmark Inclusion Subsidy." *Journal of Financial Economics*, 142(2): 756–774.

Khan, Mozaffar, George Serafeim, and Aaron Yoon. 2016. "Corporate Sustainability: First Evidence on Materiality." *The Accounting Review*, 91(6): 1697–1724.

Landier, Augustin, and Stefano Lovo. 2020. "ESG Investing: How to Optimize Impact?" *Working Paper, HEC Paris*. Available at: https://www.ssrn.com/abstract=3508938.

Larcker, David F., Peter C. Reiss, and Youfei Xiao. 2015. "Corporate Governance Data and Measures Revisited." *Working Paper, Stanford University*. Available at: https://papers.ssrn.com/abstract=2694802.

Larcker, David F., Scott A. Richardson, and Irem Tuna. 2007. "Corporate Governance, Accounting Outcomes, and Organizational Performance." *The Accounting Review*, 82(4): 963–1008.

Lins, Karl V., Henri Servaes, and Ane M. Tamayo. 2017. "Social Capital, Trust, and Firm Performance: The Value of Corporate Social Responsibility during the Financial Crisis." *Journal of Finance*, 27(4): 1785–1824.

McCahery, Joseph A., Zacharias Sautner, and Laura T. Starks. 2016. "Behind the Scenes: The Corporate Governance Preferences of Institutional Investors." *Journal of Finance*, 71(6): 2905–2932.

Oehmke, Martin, and Marcus M. Opp. 2019. "A Theory of Socially Responsible Investment." *Working Paper, Swedish House of Finance*. Available at: https://www.ssrn.com/abstract=3467644.

Pastor, Lubos, Robert F. Stambaugh, and Lucian A. Taylor. 2021*a*. "Dissecting Green Returns." National Bureau of Economic Research Working Paper 28940.

Pastor, Lubos, Robert F. Stambaugh, and Lucian A. Taylor. 2021*b*. "Sustainable Investing in Equilibrium." *Journal of Financial Economics*, 142(2): 550–571.

A Appendix

A.1 Proofs

PROOF OF LEMMA 1. We start with showing (3)-(4). The conditional distribution of D given s_D is normal. The mean and the variance of that distribution can be computed by a linear regression of D on s_D:

$$D - \overline{D} = \beta_D(s_D - \overline{D}) + \nu_D, \tag{42}$$

where $\nu_D \sim N(0, \sigma^2_{\nu_D})$ and is independent of s_D. We need to determine β_D.

The mean and variance of D conditional on signal s_D are

$$E(D|s_D) = \overline{D} + \beta_D(s_D - \overline{D}) \tag{43}$$

$$Var(D|s_D) = \sigma^2_{\nu_D} \tag{44}$$

The regression coefficient β_D is given by the standard expression:

$$\beta_D = \frac{Cov(D - \overline{D}, s_D - \overline{D})}{Var(s_D - \overline{D})} = \frac{(Cov(D - \overline{D}, D - \overline{D} + \epsilon_D)}{Var(D - \overline{D} + \epsilon_D)} = \frac{\sigma^2_D}{\sigma^2_D + \sigma^2_{\epsilon_D}} \tag{45}$$

The variables D and ϵ_D are independent. Taking variances on each side of (42), we have

$$Var(D - \overline{D}) = Var(\beta_D(s_D - \overline{D}) + \nu_D) = \beta^2_D Var(s_D - \overline{D}) + \sigma^2_{\nu_D} \tag{46}$$

It is easy to see that

$$\sigma^2_{\nu_D} = \sigma^2_D - \left(\frac{\sigma^2_D}{\sigma^2_D + \sigma^2_{\epsilon_D}}\right)^2 (\sigma^2_D + \sigma^2_{\epsilon_D}) = \frac{\sigma^2_D \sigma^2_{\epsilon_D}}{\sigma^2_D + \sigma^2_{\epsilon_D}} \tag{47}$$

Hence, the mean and variance of D, conditional on signal s_D, are

$$E(D|s_D) = \overline{D} + \beta(s_D - \overline{D}) = \overline{D} + \frac{\sigma^2_D}{\sigma^2_D + \sigma^2_{\epsilon_D}}(s_D - \overline{D}) \tag{48}$$

$$Var(D|s_D) = \sigma^2_{\nu_D} = \frac{\sigma^2_D \sigma^2_{\epsilon_D}}{\sigma^2_D + \sigma^2_{\epsilon_D}} \tag{49}$$

The derivation for the mean and variance of Y, conditional on signal s_Y is analogous. In the equations above, we need to replace D by Y and s_D by s_Y. □

PROOF OF LEMMA 2. An ESG-conscious investor chooses their portfolio θ^{ESG} to maximize the expected utility

$$E\left(-\exp(-\gamma(W_1+\theta^{ESG}Y)\,|s_D,s_Y\right).$$

Substituting in their wealth in period 1, we arrive at

$$E\left(-\exp(-\gamma(W_0+\theta^{ESG}(D-S)+\theta^{ESG}Y)\,|s_D,s_Y\right).$$

For a normally distributed random variable x, $E(\exp(x))=\exp\left(E(x)+\frac{1}{2}Var(x)\right)$. Since D and Y are independent, normally distributed random variables, we can show that the above objective is equivalent to the following mean-variance optimization:

$$\max_{\theta^{ESG}}\theta^{ESG}\left[(E(D\,|s_D,s_Y)-S)+E(Y\,|s_D,s_Y)\right]-\frac{1}{2}\gamma(\theta^{ESG})^2\left[Var(D\,|s_D,s_Y)+Var(Y\,|s_D,s_Y)\right].$$

Solving for the portfolio choice θ^{ESG} that maximizes the above objective, we arrive at (8).

To solve for the portfolio of traditional investors, we simply repeat the above derivations, setting Y equal to zero. □

PROOF OF PROPOSITION 1. Substituting in θ^{ESG} and θ^{T} from Lemma 2 into market clearing (9), we derive

$$\begin{aligned}S_0 =&A\lambda Var(D|s_D)(E(D|s_D)+E(Y|s_Y))+A(1-\lambda)(Var(D|s_D)+Var(Y|s_Y))E(D|s_D)\\&-A\gamma\overline{\theta}Var(D|s_D)(Var(D|s_D)+Var(Y|s_Y))\end{aligned}$$

where

$$A=\left[\lambda Var(D|s_D)+(1-\lambda)(Var(D|s_D)+Var(Y|s_Y))\right]^{-1}.$$

Substituting the expressions for the conditional moments from Lemmma 1, we have

$$\begin{aligned}S_0 =&A\lambda\frac{\sigma_D^2\sigma_{\epsilon_D}^2}{\sigma_D^2+\sigma_{\epsilon_D}^2}\left(\overline{Y}+\frac{\sigma_Y^2}{\sigma_Y^2+\sigma_{\epsilon_Y}^2}(s_Y-\overline{Y})+\overline{D}+\frac{\sigma_D^2}{\sigma_D^2+\sigma_{\epsilon_D}^2}(s_D-\overline{D})\right)\\&+A(1-\lambda)\left(\frac{\sigma_D^2\sigma_{\epsilon_D}^2}{\sigma_D^2+\sigma_{\epsilon_D}^2}+\frac{\sigma_Y^2\sigma_{\epsilon_Y}^2}{\sigma_Y^2+\sigma_{\epsilon_Y}^2}\right)\left(\overline{D}+\frac{\sigma_D^2}{\sigma_D^2+\sigma_{\epsilon_D}^2}(s_D-\overline{D})\right)\\&-A\gamma\overline{\theta}\frac{\sigma_D^2\sigma_{\epsilon_D}^2}{\sigma_D^2+\sigma_{\epsilon_D}^2}\left[\frac{\sigma_D^2\sigma_{\epsilon_D}^2}{\sigma_D^2+\sigma_{\epsilon_D}^2}+\frac{\sigma_Y^2\sigma_{\epsilon_Y}^2}{\sigma_Y^2+\sigma_{\epsilon_Y}^2}\right],\end{aligned}$$

where

$$A=\left[\frac{\sigma_D^2\sigma_{\epsilon_D}^2}{\sigma_D^2+\sigma_{\epsilon_D}^2}+(1-\lambda)\frac{\sigma_Y^2\sigma_{\epsilon_Y}^2}{\sigma_Y^2+\sigma_{\epsilon_Y}^2}\right]^{-1}.$$

Simplifying the above expression, we arrive at the statement in the proposition. □

A.2 Possible Sources of Noise in ESG Scores

An ESG score produced by an ESG rating agency is an aggregate of many indicators measuring a variety of attributes, some of which might be unrelated to each other (such as CO2 emissions and labor practices). The ESG attribute Y_t in our model is therefore an aggregate of a multidimensional variable. In this section, we explain how to think about noise and about our noise-correction procedure in this context.

Assume that ESG rating agencies compute the scores as a weighted average of many indicators, corresponding to disaggregated ESG attributes (e.g., CO2 emissions, labor practices):

$$s_{i,t} = \sum_{\alpha \in \{1,n\}} w_{i,\alpha} \cdot I_{i,\alpha,t}, \tag{50}$$

where i indexes ESG rating agencies, a indexes attributes that the agency considers, $I_{i,\alpha,t}$ is rater i's measure of attribute α,[28] and $w_{i,\alpha}$ are the weights.

The true value of Y_t is given by a similar construct,

$$Y_t = \sum_{\alpha \in \{1,n\}} w^\star_\alpha \cdot I^\star_{\alpha,t}, \tag{51}$$

where $I^\star_{\alpha,t}$ are the true values of the indicators and $w^\star_\alpha$ are the true weights—i.e., the weights that the representative ESG investor assigns to individual indicators, which reflect her preferences or social preferences.

At this stage, some discussion about these constructs might be useful. Suppose there are two attributes that are important to investors: labor practices and CO2 emissions. For a given firm, the true values of labor treatment and CO2 emissions are denoted by $I^\star_{\alpha,t}$. As in our model, a rating agency does not observe these true values, it only observes their proxies. Those proxies are the indicators $I_{i,\alpha,t}$ (for rating agency i). For example, an indicator for labor practices could be constructed based on labor turnover as reported by the firm, or the number of complaints in labor courts. Both indicators are correlated with the true value, but they are not identical to it. The difference is the error term. For the case of CO2 emissions, the indicator could be constructed based on the self-reported emissions (which could be noisy due to the self-reporting aspect) or industry estimates (such procedure is typically used to estimate real estate emissions). For the weights, investors have preferences between labor

[28] The indicators $I_{i,\alpha,t}$ are continuous variables. We normalize them so that they are measured on the same scale.

treatment and emissions, represented by $(w^\star_\alpha, 1 - w^\star_\alpha)$. The rating agency does not observe these weights and needs to estimate them or use their own. The weights a rating agency uses are not identical to the true weights and the difference is assumed to be a random variable.

Under our assumptions, it is possible to decompose the measurement error of each rating agency as follows:

$$s_{i,t} = Y_t + \underbrace{\sum_{\alpha\in\{1,n\}} w_{i,\alpha} \cdot \underbrace{(I_{i,\alpha,t} - I^\star_{\alpha,t})}_{\epsilon_{I_{i,\alpha,t}}} + \sum_{\alpha\in\{1,n\}} \underbrace{(w_{i,\alpha} - w^\star_\alpha)}_{\epsilon_{w_{i,\alpha}}} \cdot I^\star_{\alpha,t}}_{\epsilon_{Y_{i,t}}}. \tag{52}$$

There are two sources of noise in this decomposition: the measurement error at the level of the indicator,

$$I_{i,\alpha,t} = I^\star_{\alpha,t} + \epsilon_{I_{i,\alpha,t}}, \tag{53}$$

$$E[\epsilon_{I_{i,\alpha,t}} | I^\star_{\alpha,t}, w^\star_\alpha] = 0, \tag{54}$$

and the discrepancy in the weights

$$w_{i,\alpha} = w^\star_\alpha + \epsilon_{w_{i,\alpha}}, \tag{55}$$

$$E[\epsilon_{w_{i,\alpha}} | I^\star_{\alpha,t}, w^\star_\alpha] = 0. \tag{56}$$

Equation (54) implies that the measurement error in each indicator is mean independent of the true measure and the true weights. In other words, it implies that the difference in the indicators is truly a measurement error. On the other hand, equation (56) implies that the deviations of the weights assigned by the rating agencies relative to the weights that describe the true preferences are orthogonal to the true indicators and the true weights themselves. In this case, the intuition is that the differences in the weights are a mean zero random variable.

The above equations parallel our representation in (17). Additionally, we need to assume that the errors are classical, which is satisfied when the measurement error of the indicators, and the deviations in the weights of the rating agency from the true weights are independent of the true ESG attribute Y. Formally,

$$E\left[\sum_{\alpha\in\{1,n\}} w_{i,\alpha} \cdot (I_{i,\alpha,t} - I^\star_{\alpha,t}) + \sum_{\alpha\in\{1,n\}} (w_{i,\alpha} - w^\star_\alpha) \cdot I^\star_{\alpha,t} \middle| Y_t\right] = 0. \tag{57}$$

Condition (57) is satisfied if conditions (54) and (56) hold.

For one rating agency's score to be a valid instrument for that of another rating agency, one needs to impose two further moment restrictions: (i) the errors-in-variables across two rating agencies are orthogonal (as in equation (21)); and (ii) the errors-in-variables are not correlated with the stock market returns (as in equation (20)).

The discrepancy between ESG ratings of two agencies can be decomposed as follows:

$$s_{i,t} - s_{j,t} = \sum_{\alpha \in \{1,n\}} \underbrace{(w_{i,\alpha} - w_{j,\alpha}) \cdot \bar{I}_{\alpha,t}}_{\text{Scope and Weight}} + \sum_{\alpha \in \{1,n\}} \underbrace{\bar{w}_\alpha \cdot (I_{i,\alpha,t} - I_{j,\alpha,t})}_{\text{Measurement}}, \tag{58}$$

where

$$\bar{w}_\alpha = \frac{w_{i,\alpha} + w_{j,\alpha}}{2}$$
$$\bar{I}_{\alpha,t} = \frac{I_{i,\alpha,t} + I_{j,\alpha,t}}{2}.$$

The first term in (58) captures the weight and scope discrepancies highlighted in Berg, Kölbel and Rigobon (2020). A weight discrepancy occurs when rating agencies assign different weights to the same attribute and the scope discrepancy when one of the agencies disregards a category, assigning it a weight of zero. The second term in (58) is the discrepancy in measurement of the same indicator.

The easiest way to develop an understanding of what the required moment conditions mean in this setting is to study two special cases: pure measurement and pure weights differences.

Assume that the rating agencies only differ in the measurement of the indicators, i.e., their weights are identical to each other and identical to the true weights. In this case, the scores of rating agencies i and j are given by

$$s_{i,t} = Y_t + \sum_{\alpha \in \{1,n\}} w^\star_\alpha \cdot (I_{i,\alpha,t} - I^\star_{\alpha,t}),$$
$$s_{j,t} = Y_t + \sum_{\alpha \in \{1,n\}} w^\star_\alpha \cdot (I_{j,\alpha,t} - I^\star_{\alpha,t}).$$

The two rating agencies' scores are correlated through Y_t and we expect them to strongly predict each other. This is our relevance assumption. Another assumption that we need is the independence assumption, which requires that measurement errors are uncorrelated

across the rating agencies (an analog of (21)), i.e.,

$$E\left[(I_{i,\alpha,t} - I^{\star}_{\alpha,t}) \cdot (I_{j,\alpha,t} - I^{\star}_{\alpha,t})\right] = E\left[\epsilon_{I_{i,\alpha,t}} \cdot \epsilon_{I_{j,\alpha,t}}\right] = 0, \quad \forall i, j. \tag{59}$$

This assumption is natural if the errors in the indicators are purely mistakes that are rating agency specific. There are circumstances in which it can be violated. For example, two rating agencies may use similar data and similar procedures to compute an indicator. Because their models are based on similar principles, it is reasonable to conjecture that the errors in the procedures of some agencies are correlated with each other. The second possible source of failure of independence is that one rating agency's scores are influenced by scores of another, which makes their errors correlated. Both of these violations fall under Case 2 of Section 4.1. As we argue in that section, both of them would be detected by the overidentifying restrictions test.

The second special case is when the measured indicators are all equal to the true indicators and the discrepancy comes exclusively from weight differences. In this case, the scores of ESG rating agencies i and j take a familiar form

$$s_{i,t} = Y_t + \sum_{\alpha \in \{1,n\}} (w_{i,\alpha} - w^{\star}_{\alpha}) \cdot I^{\star}_{\alpha,t},$$
$$s_{j,t} = Y_t + \sum_{\alpha \in \{1,n\}} (w_{j,\alpha} - w^{\star}_{\alpha}) \cdot I^{\star}_{\alpha,t}.$$

As in the previous case, the scores of the two rating agencies are trivially related to each other through Y_t, satisfying the relevance assumption. The main assumption we need to make here is that weight deviations are independent across rating agencies (an analog of (21)), i.e.,

$$E\left[(w_{i,\alpha} - w^{\star}_{\alpha}) \cdot (w_{j,\alpha} - w^{\star}_{\alpha})\right] = E\left[\epsilon_{w_{i,\alpha}} \cdot \epsilon_{w_{j,\alpha}}\right] = 0, \quad \forall i, j. \tag{60}$$

Again, as we show in Case 2 of Section 4.1, this assumption is testable using the overidentifying restrictions test.

Why is it important to distinguish measurement errors in the indicators from the discrepancies in weights? Most commentators would agree that, until standardized ESG disclosure requirements are imposed on firms, there will significant measurement errors at the level of an individual indicator. But the discrepancies—or errors—in weights are equally important. Most rating agencies not only measure individual attributes but by providing a rating they are also reflecting their preferences across those attributes. What matters the most to the rating agencies, is reflected in the weights they use.[29] This service from the rating agencies

[29]See McCahery, Sautner and Starks (2016) for a thorough analysis of different preferences corporate

is important. Investors may not have the detailed understanding of all ESG-related issues, nor the resources needed to achieve such understanding. ESG rating agencies strive to understand these issues deeply; and the weights they assign to individual attributes represent the preferences of many investors and individuals they have interacted with. Their goal is to ascertain the weights of a representative ESG-conscious investor, what we call $w_\alpha^\star$'s. It is quite likely that the assessment of these weights differs across rating agencies. Being able to instrument for these differences is as important as the ability to instrument for the measurement error at the individual attribute level.

Our instrumental variable approach relies on the assumptions presented in this section (or in equations (19), (20), and (21)) and there are instances in which their violations pose a threat to our identification. First, we are using a linear representation of ESG scores (equation (50)). Berg, Kölbel and Rigobon (2020) show that a linear approximation performs very well in and out of sample. However, if the aggregation rules are non-linear, the non-linearity implies a correlation between the measurement error and the true underlying measure and hence conditions (20) and/or (54) will be violated and the errors will not be classical. This violation corresponds to Case 4 of Section 4.1 and it will be detected by the overidentifying restrictions test.

The second possible problem is that two rating agencies impute indicators using similar data and similar models (as argued by Christensen, Serafeim and Sikochi, 2021). In this case, the independence assumption (equation (21) or (59)) is violated. This violation corresponds to Case 1 of Section 4.1 and it will be detected by the overidentifying restrictions test.

Third, firms strategic behavior or manipulation (greenwashing) will produce deviations that are nonclassical errors-in-variables. For example, if the worst firms manipulate the most, there would be a correlation between the measurement error and the true underlying outcome. One needs to assume further that greenwashing is so effective that it fools the rating agencies, and they make the same mistake in assessing an indicator. This again corresponds to Case 1 of Section 4.1. As long as one rating agency sees through the manipulation (or uses a methodology that does not rely on companies' disclosure), the overidentifying restrictions test will detect the correlated instruments. In our dataset, we have two rating agencies, RepRisk and Truvalue Labs, that use public news and machine learning to compute their scores and are therefore less subject to the above manipulation problem.

Finally, our instrumental variable approach may fail because the measurements are correlated with the stock-return relevant cash-flow innovations (η_t). So, when a rating agency looks at the realized returns to determine the value of a particular indicator, or the weights of a particular attribute, the instruments fail the exogeneity assumption. This violation cor-

governance institutional investors have.

responds to Case 2 of Section 4.1 and it will be detected by the overidentifying restrictions test.

A.3 What if Errors are Correlated Across Raters?

In this section we show that if a rater's score is influenced by scores of another rater, leading to a violation of (21), this is diagnosed by the overidentifying restrictions test. For concreteness, suppose that one rater simply follows another, that is,

$$s_{1,t} = Y_t + \epsilon_{Y_{1,t}},$$
$$s_{2,t} = s_{1,t} + u_t,$$

where u_t is an error term. The second rater's score can be expressed as

$$s_{2,t} = Y_t + \underbrace{\epsilon_{Y_{1,t}} + u_t}_{=\epsilon_{Y_{2,t}}}.$$

In this example, errors are correlated because $E[\epsilon_{Y_{1,t}} \cdot \epsilon_{Y_{2,t}}] \neq 0$.

Recall our main structural equation: $\Delta S_{t+1} = a + bY_t + \eta_t + \nu_t$, where η_t corresponds to the stock relevant innovations not included in the controls of the regression and ν_t is the error term. This equation is equivalent to

$$\Delta S_{t+1} = a + b(s_{1,t} - \epsilon_{Y_{1,t}}) + c_X \cdot X_t + \eta_t + \nu_t = a + bs_{1,t} + c_X \cdot X_t + \underbrace{\eta_t + \nu_t - b\,\epsilon_{Y_{1,t}}}_{\text{error}}$$

One can see immediately that s_2 is not a valid instrument for s_1 because it is correlated with the error term. The overidentifying restrictions test, which checks specifically for the correlation of an instrument with the error term, is going to diagnose this.

A similar argument applies to using s_1 as an instrument for s_2. The Hansen J test would fail in this case as well.