NBER WORKING PAPER SERIES

ESG CONFUSION AND STOCK RETURNS: TACKLING THE PROBLEM OF NOISE

Florian Berg
Julian F. Koelbel
Anna Pavlova
Roberto Rigobon

Working Paper 30562
http://www.nber.org/papers/w30562

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
October 2022, revised July 2024

Florian Berg, Julian Kölbel, and Roberto Rigobon are grateful to the 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 Ing-Haw Cheng, Aaron Pancost, and Jonathan Parker and seminar participants at the 2022 American Finance Association meetings, ARCS 2022, the 2022 ASU Sonoran Winter Finance Conference, George Washington University, London Business School, MIT GCFP, MIT Sloan, the 2021 PCOB colloquium on ESG Disclosure, the University of Amsterdam, the University of Mannheim, and the University of Michigan for excellent feedback. All remaining errors are ours. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

ESG Confusion and Stock Returns: Tackling the Problem of Noise
Florian Berg, Julian F. Koelbel, Anna Pavlova, and Roberto Rigobon
NBER Working Paper No. 30562
October 2022, revised July 2024
JEL No. C26,G12,Q56

ABSTRACT

Existing measures of ESG (environmental, social, and governance) performance ESG ratings are noisy and, therefore, standard regression estimates of the effect of ESG performance on stock returns are biased. Addressing this as a classical errors-in-variables problem, we develop a noise-correction procedure in which we instrument ESG ratings with ratings of other ESG rating agencies. With this procedure, the median increase in the regression coefficients is a factor of 2.1. The results are similar when we use accounting profitability measures as outcome variables. In simulations, our noise-correction procedure outperforms alternative approaches such as simple averages or principal component analysis.

Florian Berg
MIT Sloan School of Management
E62-520100 Main Street
fberg@mit.edu

Julian F. Koelbel
University of St. Gallen
PLD-G 02
Plattenstr. 32
St. Gallen 9000
Switzerland
julian.koelbel@unisg.ch

Anna Pavlova
London Business School
Regents Park
London NW1 4SA
UK
apavlova@london.edu

Roberto Rigobon
MIT Sloan School of Management
100 Main Street, E62-516
Cambridge, MA 02142
and NBER
rigobon@mit.edu

1 Introduction

Ever since the onset of ESG (environmental, social, and governance) investing, academics, practitioners, and policymakers have been trying to understand the relationship between ESG performance and financial variables, particularly stock returns.[1] An important problem obscuring this relationship that the literature has not remedied so far is the problem of accurately measuring ESG performance. Available measures of ESG performance – ESG ratings – are noisy. The average pairwise correlation of the ESG ratings in our sample is only 0.36.[2] Berg, Kölbel, and Rigobon (2022) show that the main source of this disagreement is differences in measurement, which suggests that ESG attributes are measured imperfectly. Nonetheless, ESG ratings guide ESG fund investments and link investor preferences to portfolio choices.[3] Pastor, Stambaugh, and Taylor (2023) show that ESG investors do indeed tilt their holdings towards firms with high ESG ratings.

In this paper, we posit that ESG ratings suffer from measurement error and find strong support for this hypothesis. We depart from the existing literature because we do not assume that ESG ratings provide accurate measurements. Instead, we assume that they measure ESG performance with noise. This creates a problem for standard regressions seeking to establish a relationship between financial variables and ESG performance. As is well established in the literature, the estimated coefficients on regressors measured with noise are biased. We consistently find that the estimates suffer from attenuation bias, i.e., the estimated coefficient is biased towards zero. To address this bias, we propose a noise-correction procedure to 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 stock returns and accounting profitability measures, relying on the seven largest ESG rating agencies. The corrected estimates demonstrate that the effect of firms' ESG performance on stock returns is positive, in line with prior results, but substantially larger than

[1] A recent meta-study by Atz, Bruno, Liu, and Van Holt (2022) identifies 1141 peer-reviewed academic papers written between 2015 and 2020 that investigate the link between ESG and financial performance.

[2] The ESG rating agencies in our dataset include ISS ESG (majority stake owned by Deutsche Boerse), MSCI IVA (owned by MSCI), RepRisk (independent), Refinitiv (formerly known as Asset4), SP Global CSA (formerly known as RobecoSAM), Sustainalytics (owned by Morningstar), Truvalue Labs (owned by FactSet), and Moody's (formerly known as Vigeo-Eiris). See Section 2 for more details.

[3] For example, MSCI, the largest ESG rating provider by revenue (see https://www.opimas.com/research/742/detail/) has more than 1700 clients worldwide (see https://www.msci.com/our-clients/corporates, accessed on April 11, 2024).

previously estimated: after correcting for attenuation bias, the median increase in the regression coefficients is a factor of 2.1. The results are similar when we use accounting profitability measures as outcome variables. The attenuation bias is stable over different time horizons over which we measure returns. We estimate the noise-to-signal ratio for each ESG rating agency (some of which are very large) and show in simulations that our noise-correction procedure outperforms alternative noise-reducing approaches such as simple averages or principal component analysis.

We start with standard cross-sectional regressions of stock returns on ESG ratings, which have been widely used in the literature. To tackle attenuation bias, we replace standard ordinary least squares (OLS) regressions with two-stage least squares (2SLS) regressions. In the first stage of the 2SLS estimation, we instrument the rating of one agency by the ratings of other agencies in our sample. We document that the first stage is strong – each rating agency's ESG score is well predicted by a combination of other rating agencies' scores. Our 2SLS regressions reveal that the effect of ESG performance on stock returns is much stronger relative to those from the standard regressions — nearly all coefficients increase substantially in magnitude. This result is consistent with the hypothesis that ESG ratings are measured with error and that the bias we see in the standard regressions is indeed an attenuation bias.[4] Furthermore, the statistical significance of the estimates from the 2SLS regressions is typically higher than from their OLS counterparts, so the link between ESG performance and stock returns is stronger than previously documented.

Measurement error in ESG ratings can significantly affect ESG portfolio construction. In a long-short portfolio based on the first and fifth quintiles of the MSCI ratings, 46.2% of the stocks in the first quintile are replaced by stocks in other quintiles after we instrument the rating using our procedure.

Furthermore, the ratios of 2SLS and OLS coefficients can be estimated for each rating agency, and they provide a measure of the *implied noise* in the rating agency's score. We find that the noise-to-signal ratio ranges from 23.7 to 98.5% across ESG raters.

[4] In Appendix A.1, we develop a theoretical model that establishes a linear relationship between ESG performance and stock returns and shows that the noisier the measurement of ESG performance, the lower the sensitivity of stock returns to ESG performance. Moreover, 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 toward zero. The noisier the measurement, the larger the bias.

There is agreement in the literature that over our (short) sample period, green stocks outperformed brown, and our estimates are consistent with this finding. We also find a positive relationship between ESG performance and stock returns, albeit higher in magnitude. This is puzzling in light of the theory that green stocks attract higher investor demand and, therefore, should have higher prices and lower expected returns (Heinkel, Kraus, and Zechner, 2001; Pastor, Stambaugh, and Taylor, 2021). The literature has attributed the outperformance to a strengthening of climate concerns over the sample period and to inflows into ESG funds, as in, e.g., van der Beck (2021); Pastor, Stambaugh, and Taylor (2022); Karolyi, Wu, and Xiong (2023). We do not duplicate their performance decomposition in this paper. Our point is orthogonal. Regardless of whether the sign of the relationship is positive or negative, an OLS regression mismeasures the coefficient because of the noise in ESG ratings. Our 2SLS estimation approach recovers the true coefficient under certain assumptions. In future work, as the time dimension of our panel grows and realized returns become more representative of expected returns, we expect the coefficient to become negative.

While the leading application of our procedure is to the relationship between ESG performance and stock returns, we also consider other outcome variables. The accounting literature has focused on the effects of ESG performance on accounting measures of performance. We apply our procedure to study how ESG performance affects the accounting profitability measures and demonstrate that standard regressions in the literature suffer from attenuation bias, which our procedure corrects, leading to a median increase in the coefficients by a factor of 1.9. The estimated coefficients are all positive[5] and their statistical significance is higher than in the OLS regressions. Finally, we apply our procedure to the E, S, and G components of ESG ratings, and our results are consistent with what we find for aggregate ESG ratings, with a median expansion factor of 1.7.

How do we ensure that the ESG ratings are *valid* instrumental variables (IVs)? Although each rating agency's ESG rating is well predicted by a combination of other ratings, this is insufficient to guarantee that ESG scores are valid instruments. While it is not possible to directly test for instrument validity, we conduct overidentifying restrictions (OIR) tests to establish whether our

[5] In line with Lins, Servaes, and Tamayo (2017); Liang and Renneboog (2017); Cornett, Erhemjamts, and Tehranian (2016)

instruments are coherent with each other.[6] The OIR test compares IV estimates from two models with different sets of instruments. If the instruments are valid, the two estimates must be the same.

What makes an ESG rating an invalid instrument? The exclusion restriction is likely to be violated if ESG scores used as instruments are, for example, backfilled retroactively by a provider or if the scores of one ESG rater are influenced by another. Another issue could be that measurement errors are correlated across rating agencies because agencies use similar procedures or rely on imputed data to arrive at the scores. The OIR tests diagnose these violations.

As we have seven ESG rating agencies in our sample, we can test multiple OIR. In our noise-correction procedure, which we term *2SLS pruning*, we choose the instruments by starting from the largest possible set and pruning instruments one at a time until the model passes the Sargan-Hansen test of OIR. The failure of the test indicates that the set of chosen instruments is not coherent, which can happen when some instruments violate the exclusion restriction.

To provide further validation for our procedure, we follow Pancost and Schaller (2021) in arguing theoretically that the attenuation bias, captured by the ratio between the 2SLS and OLS coefficients, should be invariant to the horizon over which stock returns are measured. We, therefore, estimate the model for stock returns measured over months $t+1$, $t+2$, and $t+3$ and find that the ratios between the 2SLS and OLS coefficients in these three tests are indeed statistically indistinguishable from each other.

It is important to highlight that all raters' scores are valuable. Disregarding the scores of some raters amounts to discarding valuable information about the imperfectly measured ESG attributes. However, researchers may not have access to 7 different ratings like we do. The minimum number of alternative ratings required for our procedure is two. We demonstrate that our procedure performs well even with only two alternative ratings, both empirically and in simulations.

We run simulations to compare our noise-correction procedure to common alternative approaches such as a simple average or principal component analysis. The simulations show that our procedure

[6] Valid instruments are always coherent, but the reverse is not always true. The OIR test would classify invalid instruments as coherent in a knife-edge case when the entire set of instruments suffers from the exact same form of misspecification (Parente and Santos Silva, 2012).

performs significantly better than the alternatives. Suppose, for example, that one ESG rating is noisier than another. A simple average ignores this information and puts equal weight on the two ratings, which is not optimal. The principal component analysis is designed to explain observed variance. It would, therefore, put the largest weight on the signal with the highest variance and most likely the highest noise. An ideal approach should instead put the lowest weight on the noisiest variable, which is what our procedure does. In additional simulations, we focus on potential model misspecifications and find that the OIR test has significant power in our setting.

Finally, we recognize that, in practice, ESG ratings are aggregates of multiple indicators (e.g., carbon emissions, labor practices, etc.), and raters choose different sets of indicators in constructing their scores. This means that we have fewer instruments than possible sources of noise. In simulations, however, we show that our procedure still goes a long way in recovering the effect of ESG performance on stock returns, i.e., the 2SLS estimates are much closer to the true coefficient than their OLS counterparts.

Our paper is related to several strands of literature. First, numerous studies have explored the link between ESG performance and stock returns. However, the evidence is not conclusive; studies report both higher stock returns for ESG performers (Edmans, 2011; Lins, Servaes, and Tamayo, 2017; Albuquerque, Koskinen, and Zhang, 2019; Karolyi, Wu, and Xiong, 2023) as well as lower stock returns (Chava, 2014; El Ghoul, Guedhami, Kwok, and Mishra, 2011; Bolton and Kacperczyk, 2022). Pastor, Stambaugh, and Taylor (2022) stress the importance of distinguishing between expected and realized stock returns, and argue that the expected stock returns of high ESG performers are lower (see also van der Beck, 2021). While our model is fully consistent with these studies, we add an important point that regardless of whether one focuses on expected or realized returns, the noisy measurement will tend to attenuate the effect and develop a procedure for correcting the attenuation bias.

Second, our analysis is related to the literature on measurement error, which is too vast to summarize here. We rely on the classical errors-in-variables approach, and, following Pancost and Schaller (2021), evaluate the robustness of this approach in our setting by changing the dependent variable. Our approach is related to Gillen, Snowberg, and Yariv (2019) but adds a strategy to weed

out incoherent instruments, which is valuable when more than one instrument is available. Future research could examine measurement error consistent high-order moment estimators (Erickson and Whited, 2010).

Finally, our paper is related to the literature on ESG rating divergence (Berg, Kölbel, and Rigobon, 2022; Christensen, Hail, and Leuz, 2021; Christensen, Serafeim, and Sikochi, 2022). Two recent papers that study the consequences of ESG rating divergence at the firm level, Avramov, Cheng, Lioui, and Tarelli (2022) and Gibson Brandon, Krueger, and Schmidt (2021), suggest that uncertainty about ESG performance leads to a higher risk premium. Our perspective is different. We interpret ESG rating divergence as measurement error, which attenuates the true effect of ESG performance on stock returns in standard regressions.

2 Data

2.1 ESG Ratings

We rely on seven different ESG ratings: ISS's Numeric ESG Overall Rating, Moody's Global score, MSCI's IVA Industry Weighted score, Refinitiv's TRESG score, Sustainalytics' ESG Risk Rating, S&P Global's ESG score (SPGlobal), and the Insight Score from Truvalue Labs (TVL). A detailed overview of names and ownership is shown in Table A1.

In line with prior literature, different ESG raters provide diverging ESG ratings. Table 1 shows the pairwise correlations between the ESG scores. The correlations are all positive, ranging from 0.04 for the TVL-Sustainalytics pair to 0.75 for the Moody's-ISS pair. The divergence arises because each ESG rating is generated by a unique methodology. Methodologies differ due to different ways of measuring ESG attributes and different ways of choosing and aggregating ESG attributes. Berg, Kölbel, and Rigobon (2022) show that measurement is the main source of divergence, followed by the choice (scope) and the aggregation (weights) of attributes.

ESG rating agencies resort to a variety of data sources for their assessment. A key challenge is that there is only a limited amount of standardized and publicly available data about companies'

ESG performance. Mainly, data comes from five distinct sources: (1) companies' own ESG reports, (2) regulatory filings, (3) the media, (4) questionnaires that rating agencies send to companies, and (5) modeled data. 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), whether the disclosure is mandatory or voluntary, and whether it follows disclosure standards such as the Global Reporting Initiative (GRI) or the Sustainability Accounting Standards Board (SASB). As a result, the indicators that ESG ratings are built upon are noisy.

ESG rating agencies determine which attributes should be evaluated as part of their scoring procedure and how important they are relative to each other. The list of relevant attributes typically includes greenhouse gas 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 determined 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.[7] 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.

For example, at an individual indicator level (e.g., CO_2 emissions), "true" means precisely the actual CO_2 emissions that occurred. The noise is the difference between what the rating agency uses for CO_2 emissions and the actual one. For ESG ratings, which are themselves weighted averages of indicators, "true" means that the indicators are free of measurement error and that the weights assigned to the indicators coincide with the weights that the representative ESG investor assigns to individual ESG attributes.

[7] See McCahery, Sautner, and Starks (2016) for an analysis of the corporate governance preferences of different institutional investors.

2.2 Financial Data

Financial data comes from Compustat's Capital IQ. *Return* is the stock return expressed in percentage points. *Beta* is the market beta estimated from monthly returns from month -60 to month -1. *Dividends* are the dividends per share over the prior 12 months divided by price at the end of the prior month. *Market Value* is the logarithm of the market value of equity at the end of the prior month. *Book-to-market* is the logarithm of book equity minus the logarithm of market value of equity at the end of the prior month. *Asset Growth* is the logarithm of growth in total assets in the prior fiscal year. *ROA* is the income before extraordinary items divided by average total assets in the prior fiscal year. *Momentum* is the return from month -12 to month -2. *Volatility* is the monthly standard deviation, estimated from daily returns from month -12 to month -1. All financial variables are winsorized at the 1% level. *Flows* is the ratio of fund flows into ESG funds in the Eurozone, the U.S., the U.K., and Japan according to Morningstar. The unit of measurement is U.S. dollars.

2.3 Descriptives

Descriptive statistics are provided in Appendix Table A2. Refinitiv, SPGlobal, Sustainalytics, Truvalue Labs, as well as Moody's have a rating on a scale from 0 to 100, ISS from 1 to 4, and MSCI from 0 to 10. We multiply Sustainalytics' scores by -1 and add 100 to align them with the other ratings. A high rating signifies a good performance, and a low rating signifies a bad performance.

The analysis is performed from January 2015 to December 2020, the starting point being determined by Sustainalytics' data, which starts in December 2014. We excluded the social score from ISS (ISS-S) because they started offering their S sub-component at the end of 2017. The sample consists of 1297 firms and 70246 firm-month observations.

The descriptive statistics of the financial variables in the U.S. are in line with Lewellen (2015)'s large stocks sample. The average market value is 11.3 billion U.S. dollars. This skews the sample

slightly towards larger firms. A reason for this might be that ESG rating agencies have better coverage for larger firms.

Table 1. Pairwise correlations of ESG raters' scores. This table presents the pairwise correlations of ESG ratings. We multiplied Sustainalytics' scores by -1 and added 100 so that a higher value corresponds to a better ESG performance for all ratings. Pairwise correlations for E, S, and G subscores are shown in Appendix Table A3.

ESG	ISS	Moody's	MSCI	Refinitiv	Sustainalytics	SPGlobal	TVL
ISS	1						
Moody's	0.75	1					
MSCI	0.49	0.50	1				
Refinitiv	0.64	0.67	0.40	1			
Sustainalytics	0.22	0.23	0.29	0.22	1		
SPGlobal	0.61	0.65	0.40	0.65	0.21	1	
TVL	0.14	0.12	0.20	0.08	0.04	0.07	1

3 Errors-in-Variables and Stock Returns

In this section, we empirically address the problem of noise in ESG ratings. To do this, we examine the impact of ESG measurement errors on the relationship between ESG performance and stock returns. We depart from the rapidly growing literature on the impact of ESG on financial outcomes in that we do not assume that ESG ratings provide accurate measurements. Instead, we assume that they measure ESG performance with noise.

A simplified asset pricing representation of stock returns and their relation to ESG performance can be written as follows:

$$r_{k,t+1} = \alpha + \beta \cdot Y_{k,t} + M_{k,t} + \epsilon_{k,t}, \tag{1}$$

where $r_{k,t+1}$ is the stock return between time t and $t+1$ for firm k. $Y_{k,t}$ is the true ESG performance at time t, $M_{k,t}$ is an omitted variable that affects stock returns and may be correlated with ESG performance, and $\epsilon_{k,t}$ are the mean zero innovations assumed to be orthogonal to all the regressors. There are T time periods and K firms. In vector notation, $\mathbf{r} := \{\tilde{r}_{k,t+1}, \quad \forall k \wedge t+1\}$, $\mathbf{Y} := \{\tilde{Y}_{k,t}, \quad \forall k \wedge t\}$, $\mathbf{M} := \{\tilde{M}_{k,t}, \quad \forall k \wedge t\}$, and $\boldsymbol{\epsilon} := \{\epsilon_{k,t}, \quad \forall k \wedge t\}$, where all vectors are $KT \times 1$, tilde indicates that a variable has been demeaned, and the symbol $\wedge$ denotes "and." In Section 4,

in which we estimate the effects of ESG performance on stock returns, we will additionally control for standard asset pricing characteristics.

The ESG performance $\mathbf{Y}$ can be correlated with the omitted variable $\mathbf{M}$. The omitted variable can be interpreted in many different ways: (i) as unexpected capital flows into ESG stocks, (ii) as an investors' sentiment shift toward ESG stocks, or (iii) as shifts in management quality. The presence of the omitted variable is important for interpreting the coefficients. We limit ourselves to addressing attenuation bias and design our empirical method to be robust to the presence of omitted variable bias.

In the previous section, we argued that ESG ratings are noisy. Therefore, in our specification below we assume that ESG rating agencies produce an imperfect measurement of the true ESG performance. Suppose that there are N ESG rating agencies indexed by i. The score of rating agency i for firm k is given by $s_{k,t,i}$ and this score contains measurement error (or noise), denoted as $\eta_{k,t,i}$. Formally,

$$s_{k,t,i} = Y_{k,t} + \eta_{k,t,i}, \qquad i \in \{1, \ldots, N\}. \tag{2}$$

We assume the measurement error ($\eta_{k,t,i}$) is as in the classical errors-in-variables problem, that is, orthogonal to $(\mathbf{Y}, \mathbf{M}, \boldsymbol{\epsilon})$. (A complete list of assumptions follows in Section 3.1). As before, we demean and stack these observations into vectors $\mathbf{s_i} := \{\tilde{s}_{k,t,i}, \quad \forall k \wedge t\}$ and $\boldsymbol{\eta_i} := \{\eta_{k,t}, \quad \forall k \wedge t\}$.

While the true ESG performance is not observable in the data, ESG scores are. The reduced-form of the structural model (1) that one can take to the data is as follows:

$$r_{k,t+1} = \alpha + \beta \cdot s_{k,t,i} + \nu_{k,t}, \tag{3}$$

where $\nu_{k,t} = M_{k,t} + \epsilon_{k,t} - \eta_{k,t,i} \cdot \beta$. In other words, we have replaced the true ESG performance $Y_{k,t}$ with an ESG score, which is a noisy version of $Y_{k,t}$.

The coefficient of interest is β, whose OLS estimate, derived in Appendix A.2.1, is given by

$$\beta_{OLS} = \frac{\mathbf{Y'Y}}{\mathbf{Y'Y} + \boldsymbol{\eta_i}'\boldsymbol{\eta_i}} \cdot \left[\beta + \frac{\mathbf{Y'M}}{\mathbf{Y'Y}}\right]. \tag{4}$$

It is easy to see that the OLS estimate of β is biased for two reasons. The first bias is the attenuation bias, $\mathbf{Y}'\mathbf{Y}/(\mathbf{Y}'\mathbf{Y} + \boldsymbol{\eta_i}'\boldsymbol{\eta_i})$, which occurs because of the measurement error in the regressor, while the second bias is the omitted variable bias, $\mathbf{Y}'\mathbf{M}/\mathbf{Y}'\mathbf{Y}$.

We concentrate on the attenuation bias instead of the omitted variable bias for two reasons. First, given the substantial disagreement of ESG scores across rating agencies, noise in ESG scores is a first-order problem for regulators, asset managers, and investors. Second, the fact that there are multiple ESG ratings that disagree, but are nonetheless correlated with each other, provides a unique opportunity to address the attenuation bias that results from noisy measurement. If an omitted variable bias is present, the interpretation of the coefficient changes, but our procedure will still tackle the attenuation bias.

One of the known approaches for tackling attenuation bias is to use alternative noisy measures of the regressor as instruments. Consider again the reduced-form regression (3), in which the score $\mathbf{s_i}$ of ESG rating agency i is a noisy measure of $\mathbf{Y}$. We will use the scores of other ESG ratings as instruments. In the empirical section, we will use several instruments, but here, for expositional purposes, let us focus on one instrument $\mathbf{s_j}$, which is the score of firm k at time t from rater j (for $j \neq i$).

For the moment, assume that $\mathbf{s_j}$ is a valid instrument for $\mathbf{s_i}$. We will discuss this assumption in detail in Section 3.1 below. The IV estimate of β, derived in Appendix A.2.1, is then given by

$$\beta_{IV} = \left[\beta + \frac{\mathbf{Y}'\mathbf{M}}{\mathbf{Y}'\mathbf{Y}}\right]. \tag{5}$$

It is instructive to compare the OLS and IV estimates (Equations (4) and (5)). Notice that the omitted variable biases both the OLS estimate and the IV estimate in exactly the same manner. Hence, to isolate the attenuation bias, we simply need to compute the ratio of the two estimates. An implicit assumption behind the IV estimation is that the omitted variable bias does not affect the measurement error (we formalize this in the next section in assumption (8)). The magnitude of

the noise in the ESG score, which is used as a regressor, can then be estimated as

$$\kappa_i := 1 - \frac{\beta_{OLS}}{\beta_{IV}} = 1 - \frac{\mathbf{Y}'\mathbf{Y}}{\mathbf{Y}'\mathbf{Y} + \boldsymbol{\eta_i}'\boldsymbol{\eta_i}} = \frac{\boldsymbol{\eta_i}'\boldsymbol{\eta_i}}{\mathbf{Y}'\mathbf{Y} + \boldsymbol{\eta_i}'\boldsymbol{\eta_i}}, \tag{6}$$

where κ_i is the noise-to-signal ratio in the rater i's scores. In Section 4.2.1, we will compute the noise-to-signal ratios and compare them across different raters.

3.1 Identifying assumptions for the IV procedure

Under which conditions are ESG ratings valid instruments for each other? In a nutshell, we need to assume that ESG ratings are related to each other *only* through $\mathbf{Y}$ and that their measurement error is white noise. We now formally spell out the assumptions under which ESG ratings are valid instruments for each other. We do so for the case in which there are multiple instruments.

As a baseline, we need the *relevance* assumption, i.e., that instruments are correlated with the regressor. This assumption is easy to defend. In Section 2, we demonstrate that ESG rating scores of different agencies are positively correlated and that some of these correlations are relatively high. In Section 4, we will show formally that we do not have a problem of weak instruments (see the first-stage F-statistics in Tables 2–4).

Given relevance, we require three further assumptions regarding the measurement errors $\eta_{k,t,i}$. First, the errors are *classical*, i.e., additive and orthogonal to $\mathbf{Y}$, as in the classical errors-in-variables problem: [8]

$$E[\eta_{k,t,i}|Y_{k,t}] = 0, \quad \forall i. \tag{7}$$

Second, the error terms $\eta_{k,t,i}$ are independent of both the stock cash-flow innovations $\epsilon_{k,t}$ (not captured by the firm-level controls that we introduce later) and the omitted variable $\mathbf{M}$, i.e.,

$$E[\eta_{k,t,i} \cdot \epsilon_{k,t}] = 0 \quad \text{and} \quad E[\eta_{k,t,i} \cdot M_{k,t}] = 0, \quad \forall i. \tag{8}$$

[8] In Appendix Figure A1, we plot the distributions of the ESG rating variables. We note that they are all close to a binomial distribution. Even though the distributions are bounded, there is very little mass in the tails. This lends support to the assumption that the errors are classical.

Furthermore, when we introduce controls $X_{k,t}$ in our baseline regression, we will assume that $E[\eta_{k,t,i} \cdot X_{k,t}] = 0$. These three assumptions are the *exclusion restriction.*

Third, we assume that all errors ($\eta_{k,t,i}$) are independent across rating agencies:

$$E[\eta_{k,t,i} \cdot \eta_{k,t,j}] = 0 \quad \forall j \neq i. \tag{9}$$

This is the *independence* assumption.

Assumptions (7), (8), and (9) imply that the measurement error of each ESG rating is effectively white noise. This is a strong assumption, and we review below some possibilities regarding how it could be violated. While we cannot rule out those violations in principle, we can test for them empirically using the Sargan-Hansen OIR test, as we explain in Section 3.2.

The most probable threat to our IV estimation is a violation of the independence assumption (9), i.e., noise in ESG scores may be correlated across raters. This can occur if several rating agencies use similar data and similar estimation procedures to arrive at their scores. Also, rating agencies may rely on the same imputation method for missing data. As imputation always approximates the missing true value with some error, this error would then be correlated across those ratings that use the same procedure. It is also possible that the exclusion restriction (8) is violated. This can occur when ESG rating agencies retroactively adapt their scores after observing past stock return realizations (Berg, Fabisik, and Sautner, 2021). Another possibility is that errors are driven by other firm characteristics, such as size. However, Gibson Brandon, Krueger, and Schmidt (2021) show that ESG rating divergence is very difficult to explain with other firm characteristics. Finally, errors could be non-classical, causing a violation of assumption (7). This could occur when errors are related to the true ESG performance in an asymmetric or non-linear fashion. For instance, it could be that bad performance is easier to detect than good performance.

3.2 Testing the Coherence of Instruments

In our setting, we have several ESG rating agencies producing ratings that intend to capture a firm's true ESG performance. This implies that there are several rating agencies that could be used as

instruments. When there are two or more instruments, it becomes possible to run overidentifying restriction tests to check the coherence of the instruments.

The first step is to argue that different ESG ratings can be used as both regressors and instruments. This is a counter-intuitive implication of the errors-in-variables setting that is sometimes missed in applied work. In many applications, the regressors and the instruments are not interchangeable. In the case of errors-in-variables, they are, as long as the instruments are valid. This is because model misspecification is not in the structural form but in the measurement of the variables.

Let us now show what happens to the OLS and IV estimates if assumptions (7), (8), and (9) are violated. Suppose that the measurement error in the regressor s_i is correlated with the true ESG performance, with the omitted variable, with the stock market innovations, and with the measurement errors of another rating agency s_j, which we use as an instrument. The OLS and the IV estimates, derived in Appendix A.2.2, are given by:

$$\beta_{OLS} = \frac{\beta \cdot \mathbf{Y'Y} + \mathbf{Y'M} + \beta \cdot \mathbf{Y'}\boldsymbol{\eta_i} + \mathbf{M'}\boldsymbol{\eta_i} + \boldsymbol{\epsilon'}\boldsymbol{\eta_i}}{\mathbf{Y'Y} + \boldsymbol{\eta_i}'\boldsymbol{\eta_i} + \mathbf{Y'}\boldsymbol{\eta_i}}, \tag{10}$$

$$\beta_{IV} = \frac{\beta \cdot \mathbf{Y'Y} + \mathbf{Y'M} + \beta \cdot \mathbf{Y'}\boldsymbol{\eta_j} + \mathbf{M'}\boldsymbol{\eta_j} + \boldsymbol{\epsilon'}\boldsymbol{\eta_j}}{\mathbf{Y'Y} + \mathbf{Y'}\boldsymbol{\eta_i} + \mathbf{Y'}\boldsymbol{\eta_j} + \boldsymbol{\eta_i}'\boldsymbol{\eta_j}}. \tag{11}$$

where the term in purple is what causes attenuation bias in the OLS estimate, the terms in red are the ones due to violation of the classical errors-in-variables (Equation (7)), the term in green is due to the violation of the exclusion restriction (Equation (8)), and the term in blue corresponds to the violation of the independence of the instruments (Equation (9)).[9]

The OIR test compares IV estimates from two models, with different sets of instruments. If the identifying assumptions (7), (8), and (9) hold, the two IV estimates should be the same. However, when the identifying assumptions fail, the two IV estimates are different from each other.[10] We use the Sargan-Hansen OIR test to probe the equality of the two different IV estimates. The Sargan-

[9] See Appendix A.2.1 for the derivations.

[10] There is one knife-edge possibility that all the covariances are different from zero but identical across the two models.

Hansen test uses the two instruments (or all the available instruments) simultaneously in the first-stage regression and then compares the correlations between the instruments and the residuals from the regression.

Unfortunately, the OIR test cannot diagnose which assumption is violated, i.e., it cannot determine which of the covariances is different from zero. Instead, we can speculate about this based on economic arguments. However, the OIR test indicates which instruments cause violations, thus allowing us to estimate the model with a subset of instruments for which the OIR test does not reject the model.

3.3 Estimation Procedure

In our empirical implementation, we find that of the many instruments we have available, some are coherent, and some are not. How do we choose, then, which instruments to include in the estimation? Unfortunately, the Sargan-Hansen OIR test does not discern which instrument is valid. In essence, it tests whether the coefficients are different from each other for each subset of instruments, but it does not identify which coefficient is correct. However, when the OIR test is not rejected, all the subsets of instruments are coherent. Coherent here means that either all instruments in the set are valid or they all have the same form of misspecification, which is a knife-edge case.

Our procedure, which we term *2SLS pruning*, starts with the full set of instruments and reduces it if the OIR test is rejected. First, we select a rating agency whose scores we want to instrument, i.e., the regressor. We then use all remaining rating agencies' scores as instruments, so the IV estimator we have been discussing technically becomes a 2SLS estimator. Second, we estimate specification (3) using 2SLS and run the Sargan-Hansen OIR test. If the model passes the Sargan-Hansen test, then all included instruments are coherent; otherwise, we exclude instruments, one at a time, until the model passes the test. We report the included and excluded instruments. The 2SLS pruning procedure identifies the maximum number of coherent instruments for each ESG rating i, $i = 1, \ldots, 7$, and provides the 2SLS estimate of the effect of ESG performance i on stock returns.

There are two concerns, however, regarding the 2SLS pruning procedure. First, it is unclear whether the OIR tests have sufficient power in our application. The short answer is that they have. We indeed find many rejections in our empirical implementation (Section 4). Additionally, we evaluate the power of the OIR test in simulations (see Section 5) and confirm that the test has sufficient power to detect coherent instruments. Second, the 2SLS pruning procedure is a sequence of OIR tests and the size should be adjusted to reflect the fact that the tests are not independent. We indeed estimate many OIR tests in this procedure. For example, when 6 ESG ratings are included as instruments, we run only one OIR test. If the OIR test rejects that model, our next step is to perform the OIR tests on 6 combinations of 5 instruments. If all of these 6 tests reject the model, we consider 15 possible combinations of 4 instruments and therefore run 15 OIR tests and so forth until a set of instruments passes the test. If our model passes the OIR tests, this is evidence that instruments are coherent, implying that the OIR tests do not diagnose violations of our identifying assumptions (7), (8), and (9).

To address the issue of serial testing, we select an increasing sequence of rejection thresholds, namely 2.5%, 3.33%, 4.17%, 5%, and 5.83%.[11] We have estimated even larger bands, and the results are virtually identical. The reason why we increase the thresholds for the p-values is as follows. For example, when we are testing all the combinations of 5 raters as instruments, we only proceed to the case of 4 raters-combinations when we find a rejection in all the combinations of 5 raters. The usual procedure that corrects for serial testing lowers the p-value thresholds when more testing is performed because the testing purpose is to find at least one rejection. In contrast, our purpose is to find no rejections, and hence we use an increasing sequence of rejection thresholds.

4 Empirical Results

In this section, we estimate the OLS regressions of stock returns on ESG ratings and contrast them to 2SLS regressions, which use scores of other rating agencies as instruments.

[11] Note that we only have 5 thresholds as we need at least two instruments for the OIR test.

4.1 Model Specifications

The OLS regression is

$$r_{k,t+h} = \alpha + \beta \cdot s_{k,t,i} + c_X \cdot \mathbf{X}_{k,t} + \nu_{h,k,t}, \tag{12}$$

where $s_{k,t,i}$ denotes the ESG rating of firm k, by rater i, in month t, h denotes the horizon, i.e., the number of months over which the returns are measured and all returns are observed at monthly frequency. Following Lewellen (2015), the vector $\mathbf{X}_{k,t}$ includes stock-level controls consisting of *Beta*, *Dividends*, *Market Value*, *Book-to-market*, *Asset Growth*, *ROA*, *Momentum*, and *Volatility*. In addition, we control for *ESG Flows*. $\mathbf{X}_{k,t}$ also includes industry and month fixed effects.[12] We cluster standard errors by month and firm.

As argued in Section 3, the OLS estimate of the effect of ESG performance on stock returns, β_{OLS}, suffers from attenuation bias. To assess the significance of the bias, we compare the OLS estimates with their 2SLS counterparts. The first-stage regression uses ESG scores of other rating agencies as instruments for a given rater's score and includes the same controls as in (12):

$$s_{k,t,i} = c_0 + \pi \cdot \mathbf{Z}_{k,t,i} + c_1 \cdot \mathbf{X}_{k,t} + \eta_{k,t}, \tag{13}$$

where $\mathbf{Z}_{k,t,i} := \{s_j | \forall (j \neq i),\ s_j \text{ is a valid instrument}\}$ are other rating agencies' scores that are being used as instruments. Denote by $\hat{s}_{k,t,i}$ the fitted value from the estimation of Equation (13). Then the second stage regression is

$$r_{k,t+h} = \alpha + \beta \cdot \hat{s}_{k,t,i} + c_X \cdot \mathbf{X}_{k,t} + \nu_{h,k,t}. \tag{14}$$

Provided that our assumptions are satisfied, we expect that $|\beta_{2SLS}| > |\beta_{OLS}|$. In the empirical implementation, we partial out all the controls from returns and the ESG rating scores, allowing us to use equations from Section 3 directly.

[12] We do not use firm fixed effects because the frequency of most ESG ratings changes is annual, and we have a very limited time series.

4.2 Main Results

Table 2. Stock returns and ESG ratings.
This table reports estimates of β from the OLS regression (12) and the 2SLS regression (14). The first set of columns shows OLS estimates. The second set of columns shows 2SLS estimates produced with the 2SLS pruning procedure as described in Section 3.3. The check marks in the columns titled "Coherent IVs" indicate the selection of instruments that pass the Sargan-Hansen OIR test. We cluster standard errors by month and firm. All reported coefficients and standard errors are multiplied by 100. The regressions are run for each rater, whose names are reported in the column "Rater". The first panel presents results for a return horizon of one month, i.e., computed between t and $t+1$, while ESG ratings are observed in month t. The second panel presents results for a horizon of two months and the third for a horizon of 3 months. *p<0.1; **p<0.05; ***p<0.01.

		OLS			2SLS Pruning											
Horizon	Rater	β_{OLS}	StdErr		β_{2SLS}	StdErr		$\frac{\beta_{2SLS}}{\beta_{OLS}}$	Coherent IVs							F–test
									IS	MS	Re	SP	Su	TV	Mo	
1	ISS	0.009	0.005	*	0.026	0.008	***	3.0		✓	✓	✓	✓	✓	✗	10429
	MSCI	0.012	0.004	***	0.024	0.008	***	2.0	✓		✓	✓	✓	✓	✗	4792
	Refinitiv	0.007	0.004	*	0.021	0.007	***	3.0	✓	✓		✓	✓	✓	✗	8598
	SPGlobal	0.011	0.004	***	0.019	0.007	***	1.8	✓	✓	✓		✓	✓	✗	8822
	Sustainalytics	0.023	0.006	***	0.038	0.012	***	1.6	✓	✓	✓	✓		✓	✗	4301
	TVL	0.001	0.004		0.065	0.019	***	95.1	✓	✓	✓	✓	✓		✗	497
	Moody's	-0.003	0.005		0.021	0.007	***	-7.0	✓	✓	✓	✓	✓	✓		11841
2	ISS	0.009	0.005	*	0.024	0.008	***	2.7		✓	✓	✓	✓	✓	✗	10253
	MSCI	0.011	0.004	***	0.024	0.008	***	2.2	✓		✓	✓	✓	✓	✗	4715
	Refinitiv	0.007	0.004	*	0.020	0.007	***	2.7	✓	✓		✓	✓	✓	✗	8448
	SPGlobal	0.009	0.004	**	0.019	0.007	***	2.1	✓	✓	✓		✓	✓	✗	8696
	Sustainalytics	0.024	0.006	***	0.027	0.012	**	1.1	✓	✓	✓	✓		✓	✓	3707
	TVL	0.001	0.004		0.062	0.019	***	111.1	✓	✓	✓	✓	✓		✗	489
	Moody's	-0.001	0.005		0.020	0.007	***	-13.5	✓	✓	✓	✓	✓	✓		11671
3	ISS	0.006	0.005		0.021	0.008	***	3.5		✓	✓	✓	✓	✓	✗	10076
	MSCI	0.012	0.004	***	0.020	0.008	***	1.8	✓		✓	✓	✓	✓	✗	4637
	Refinitiv	0.005	0.004		0.015	0.007	**	2.9	✓	✓		✓	✗	✓	✗	10197
	SPGlobal	0.007	0.004	*	0.012	0.007	*	1.7	✓	✓	✓		✗	✓	✗	10142
	Sustainalytics	0.023	0.006	***	0.023	0.012	*	1.0	✓	✓	✓	✓		✓	✓	3652
	TVL	0.002	0.004		0.064	0.020	***	40.4	✓	✓	✓	✓	✓		✗	481
	Moody's	-0.003	0.005		0.017	0.007	**	-5.5	✓	✓	✓	✓	✓	✓		11493
Median								2.1								8598

Our main finding is that the 2SLS estimates are substantially larger than the OLS estimates. Table 2 compares the OLS and 2SLS estimates for the effect of ESG ratings on stock returns for various return horizons. Column $\frac{\beta_{2SLS}}{\beta_{OLS}}$ reports the expansion factor, which has a median of 2.1 across all specifications. Coefficient magnitudes tend to double when implementing our estimation procedure.

The expansion is also very consistent. In nearly all specifications, the 2SLS coefficient is larger than the OLS coefficient. The exception is Moody's, where the sign of the coefficient flips. However, the OLS estimate for Moody's is not statistically different from zero, while the 2SLS estimate is significant and in line with the coefficients from other raters. Thus, considering the uncertainty in the OLS point estimates, these cases may also be expansions. Finally, there is Sustainalytics over a three-month horizon, where the two coefficients are exactly the same.

The substantial and consistent expansion of the coefficients strongly suggests the presence of an attenuation bias in OLS regressions with ESG ratings. The 2SLS estimates also tend to have higher statistical significance. While the standard errors are, of course, larger for the 2SLS estimates, also the coefficients are substantially larger. The gain in coefficient magnitude from correcting the attenuation bias outweighs the efficiency loss of the 2SLS estimator.

The second important observation is that there are instances in which an instrument is rejected by the Sargan-Hansen OIR test. Table 2 shows accepted instruments as green check marks and excluded instruments as red crosses. Most ESG ratings are accepted as instruments, suggesting that the procedure is empirically feasible in many cases. Moody's ESG is rejected throughout as an instrument with only two exceptions. Sustainalytics is rejected as an instrument in two cases. While the OIR test does not provide a diagnosis of the reasons for rejection, based on our discussion in Section 3.1, we believe that the most probable reason is that measurement errors are correlated across rating agencies. In Section 5, we will provide evidence from simulations that show that, in our application, the OIR test is quite sensitive to violations of our identifying assumptions (7), (8), and (9).

The third observation is that, despite their low correlations with each other, ESG ratings are strong instruments for each other. F-statistics in Table 2 have a median of 8598 and range from 497 to 11841. This demonstrates strong support for our relevance assumption. While some individual scores could be very noisy, taken together, they work extremely well in predicting any given rater's scores.

Finally, we note that all of the significant coefficients, both from OLS and 2SLS, are positive. Positive coefficients are not what one would expect based on an equilibrium asset pricing model, but they are consistent with empirical findings covering a similar period. For instance, Pastor, Stambaugh, and Taylor (2022) and Karolyi, Wu, and Xiong (2023) also find positive returns for high ESG stocks, attributing this to changes in climate concerns over the sample period, as well as unexpected flows into high ESG stocks. Over a longer time series, one would expect a negative effect of ESG performance on stock returns to appear. We provide a model making this prediction in Appendix A.1

4.2.1 Consistent expansion for different return horizons

Running the analysis for stock returns at different horizons provides a robustness check of our approach (Pancost and Schaller, 2021).[13] If the stock returns are computed at different horizons, one can argue that the variances of the innovations change, that the importance of the omitted variable shifts, etc., but the noise in the ESG score remains the same. Therefore, the attenuation bias should be the same regardless of the horizon over which the returns are computed. For a detailed discussion, see Appendix A.2.3.

The ratios between 2SLS and OLS estimates in Table 2 are stable across horizons. Figure 1 provides a further illustration of that, relying on the noise-to-signal ratio κ_i, as defined in Equation (6). Ratios from different horizons are shown in different colors, and the dots tend to be very close to each other in most cases. These results provide evidence that attenuation bias is stable across different horizons over which the left-hand-side variable is measured.

The values plotted in Figure 1 can interpreted as the implied noise in ESG scores. When the 2SLS estimate is much larger than the OLS estimate, the noise-to-signal ratio of rater i, κ_i, tends towards 1, indicating that the score is very noisy. This is the case for TVL, followed by ISS and Refinitiv. On the other end of the spectrum, we see MSCI, SPGlobal and Sustainalytics with relatively low levels of implied noise. Note that ratios below zero and above 1 cannot be interpreted as implied noise, and for this reason Moody's ESG is not shown in the plot.

[13] We thank Aaron Pancost for suggesting this test to us.

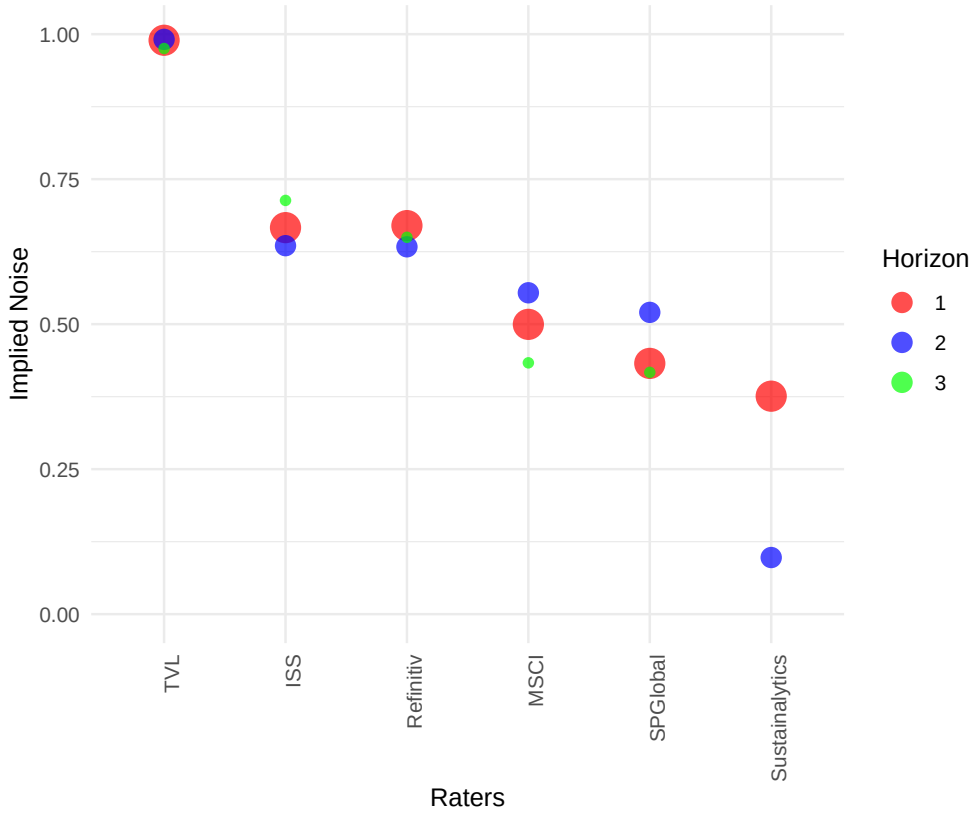


Figure 1. Implied noise in ESG ratings. This figure illustrates the implied noise for stock returns measured over different horizons (from time t to $t+1$ in red, from time $t+1$ to $t+2$ blue, and from time $t+2$ to $t+3$ in green). The vertical axis measures the noise-to-signal ratio κ_i, defined in Equation (6). Raters are sorted by the median observation along the horizontal axis. Observations above one and below zero are not shown as there is no economic interpretation. This results in an exclusion of one observation for Sustainalytics and all three observations for Moody's ESG ratings.

One might hastily conclude that one should use only the scores with the lowest implied noise. We caution against this conclusion. Disregarding scores of other raters amounts to discarding valuable information about the unobservable ESG performance the ratings are trying to measure. 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 performance. For example, the ESG rating by Truvalue Labs does not produce any significant OLS coefficient in Table 2, yet it is never rejected as an instrument. At the same time, after instrumenting with other ratings, the 2SLS coefficients of TVL increase substantially and are consistently

significant. This may indicate that this rating contains substantial noise and yet conveys essential information.[14]

4.3 Stock returns and E, S, and G components

The attenuation bias is also apparent for the E, S, and G components, as shown in Table 3. These estimations are performed only for those raters that offer E,S, and G components explicitly.[15] The median expansion factor is 1.7. For the E component, four out of five coefficients expand as expected. For the S component, two out of four coefficients expand. For the G dimension, four out of five coefficients expand. We do not count Moody's-G as an expansion because the sign switches, and both coefficients are significant. The results for the individual components provide reassurance that the noise is not mainly a result of different aggregation procedures in the sense that one rater puts more weight on E and another on G. Instead, it suggests that the results are driven by measurement noise also within the E, S, and G categories.

4.4 Accounting Variables as Outcome Variables

The attenuation bias is a general problem in the literature on ESG factors, affecting not only regressions on stock returns but also many other outcome variables, such as accounting performance and other measures of cash flows.

The true coefficients may be very different for different outcome variables, and there may also be different omitted variables. But as argued in section 3.1, the potential presence of an omitted variable does not interfere with our correction for the attenuation bias as long as our assumptions hold. In the following, we test the performance of our method by regressing ESG ratings on accounting variables.

[14] This pattern could well be related to this ratings' unique methodology. Truvalue Labs employs a methodology that relies strongly on computer algorithms to process large amounts of online information while strongly emphasizing new information. It is the only rater in our study that has daily variations in their score. One could thus say that the high variation comes at the expense of noise.

[15] We excluded ISS-S because they started offering their S sub-component at the end of 2017.

Table 3. Stock returns and E, S, and G components.
This table reports estimates of β from the OLS regression (12) and the 2SLS regression (14). ESG ratings are observed in month t, stock returns are computed between t and $t+1$. The first set of columns shows OLS estimates. The second set of columns show 2SLS estimates produced with the 2SLS pruning procedure as described in Section 3.3. The check marks in the columns titled "Coherent IVs" indicate the selection of instruments that passes the Sargan-Hansen OIR test. We cluster standard errors by month and firm. All reported coefficients and standard errors are multiplied by 100. The regressions are run for each rater, whose names are reported in the column "Rater". The first set of rows presents results for the E (environmental) component, the second panel for the S (social) component, and the third panel for the G (governance) component. *p<0.1; **p<0.05; ***p<0.01.

Indicator	Rater	OLS β_{OLS}	OLS StdErr		2SLS Pruning β_{2SLS}	2SLS Pruning StdErr		$\frac{\beta_{2SLS}}{\beta_{OLS}}$	Coherent IVs	Ftest
									IS MS Re SP Mo	
E	ISS-E	0.009	0.005	*	0.014	0.007	**	1.6	✓ ✓ ✓ ✓	13342
	MSCI-E	0.010	0.005	*	0.023	0.013	*	2.3	✓ ✓ ✓ ✓	3270
	Refinitiv-E	0.005	0.004		0.015	0.007	**	2.7	✓ ✓ ✓ ✓	11976
	SPGlobal-E	0.013	0.004	***	0.010	0.006		0.7	✓ ✓ ✓ ✓	15061
	Moody's-E	0.002	0.005		0.017	0.007	***	8.9	✓ ✓ ✓ ✓	17004
									MS Re SP Mo	
S	MSCI-S	0.006	0.004		0.008	0.016		1.5	✓ ✓ ✓	1703
	Refinitiv-S	0.006	0.004		0.006	0.008		1.0	✓ ✓ ✓	12046
	SPGlobal-S	0.009	0.004	**	0.003	0.008		0.4	✓ ✓ ✓	11229
	Moody's-S	-0.003	0.005		0.018	0.008	**	-6.0	✓ ✓ ✓	13958
									IS MS Re SP Mo	
G	ISS-G	0.001	0.005		0.036	0.013	***	30.1	✓ ✓ ✓ ✗	3068
	MSCI-G	0.009	0.004	**	0.028	0.017		3.1	✓ ✓ ✓ ✗	1584
	Refinitiv-G	0.005	0.004		0.029	0.012	**	6.0	✓ ✓ ✓ ✗	2317
	SPGlobal-G	0.009	0.004	**	0.015	0.012		1.7	✓ ✓ ✓ ✗	3523
	Moody's-G	-0.011	0.005	*	0.024	0.011	**	-2.3	✓ ✓ ✓ ✓	5687
Median								1.7		8458

We again estimate the OLS specification (12) and contrast it to the 2SLS specification (13)–(14). We consider return on equity (RoE) as an outcome variable. Unlike stock returns, this variable is highly persistent. We, therefore, consider three alternative specifications of our model. The first one mimics our earlier specification using RoE as the outcome variable. The second one additionally controls for the lag of the outcome variable. The third is an error-correction model: the outcome variables are *changes* in RoE, and the regressors include lagged RoE as controls. The remaining control variables are capital expenditures scaled by total assets, the natural logarithm of total assets, and the long-term debt scaled by total assets. The limitation of the first one is a persistent outcome variable, which may introduce a bias in the estimate of the coefficient of interest. The second and third specifications are two different ways to address this. The reason we are presenting all three specifications is to see how the estimates of the coefficient of interest vary and, hence, infer how

much the potential biases affect the different estimates. Importantly, those biases should not affect the ability of our method to correct the attenuation bias.

Table 4 shows the results for the outcome variable Return on Equity. We find that all coefficients expand except for two cases involving TVL. In the first case, the 2SLS estimate is not feasible, and in the second case, the OLS estimate is non-significantly negative and thus could also represent an expansion. The median expansion factor is 1.9 across the estimations. We also consider alternative outcome variables that proxy for cash flows, such as Return on Assets (RoA) and Return on Sales (RoS). The corresponding tables are shown in Appendix A.5, giving qualitatively the same result. Figure A2 compares the noise-to-signal ratio across the accounting variables RoE, RoA, and RoS. It is evident that across these very different outcome variables, the attenuation bias is relatively stable.

Table 4. Return on Equity and ESG ratings.
This table reports estimates of β from the OLS regression (12) and the 2SLS regression (14) with Return on Equity (RoE) as the outcome variable. The observation frequency is annual; ESG ratings are observed in year $t-1$. The first panel shows results for RoE observed in year t. In the second panel, the lagged variable RoE_{-1} is added as a regressor. In the third panel, the dependent variable is switched to ΔRoE, which is $RoE_t - RoE_{t-1}$. The first set of columns shows OLS estimates. The second set of columns shows 2SLS estimates produced with the 2SLS pruning procedure as described in Section 3.3. The check marks in the columns titled "Coherent IVs" indicate the selection of instruments that pass the Sargan-Hansen OIR test. We cluster standard errors by month and firm. All reported coefficients and standard errors are multiplied by 100. *p<0.1; **p<0.05; ***p<0.01.

Model	Rater	OLS Coeffs	OLS StdErr		2SLS Pruning Coeffs	2SLS Pruning StdErr		$\frac{\beta_{2SLS}}{\beta_{OLS}}$	Coherent IVs IS	MS	Re	SP	Su	TV	Mo	Ftest
RoE	ISS	0.018	0.004	***	0.032	0.006	***	1.8		✓	✓	✓	✗	✓	✓	1999
	MSCI	0.012	0.004	***	0.043	0.008	***	3.7	✓		✓	✓	✗	✓	✓	620
	Refinitiv	0.022	0.004	***	0.027	0.005	***	1.3	✓	✓		✓	✗	✓	✓	2139
	SPGlobal	0.015	0.004	***	0.033	0.006	***	2.2	✓	✓	✓		✗	✓	✓	1949
	Sustainalytics	0.030	0.004	***	0.053	0.010	***	1.8	✓	✓	✓	✓		✓	✓	296
	TVL	0.003	0.003		-	-										
	Moody's	0.020	0.004	***	0.033	0.006	***	1.6	✓	✓	✓	✓	✗	✓		2751
RoE inc. RoE_{-1}	ISS	0.014	0.004	***	0.019	0.005	***	1.4		✓	✓	✓	✗	✓	✓	1582
	MSCI	0.008	0.003	**	0.032	0.006	***	4.2	✓		✓	✓	✓	✓	✓	445
	Refinitiv	0.013	0.003	***	0.017	0.004	***	1.3	✓	✓		✓	✗	✓	✓	1670
	SPGlobal	0.008	0.003	***	0.023	0.005	***	2.8	✓	✓	✓		✓	✓	✓	1296
	Sustainalytics	0.014	0.003	***	0.033	0.008	***	2.2	✓	✓	✓	✓		✓	✓	235
	TVL	0.001	0.002		0.056	0.015	***	56.1	✓	✓	✓	✓	✗		✓	62
	Moody's	0.012	0.003	***	0.022	0.005	***	1.8	✓	✓	✓	✓	✗	✓		2171
ΔRoE inc. RoE_{-1}	ISS	0.009	0.004	**	0.014	0.006	**	1.6		✓	✓	✓	✓	✓	✓	1320
	MSCI	0.007	0.003	**	0.022	0.008	***	3.3	✓		✓	✓	✓	✓	✓	445
	Refinitiv	0.010	0.003	***	0.012	0.005	**	1.2	✓	✓		✓	✓	✓	✓	1399
	SPGlobal	0.006	0.003	*	0.016	0.005	***	3.0	✓	✓	✓		✓	✓	✓	1296
	Sustainalytics	0.013	0.003	***	0.026	0.009	***	2.00	✓	✓	✓	✓		✓	✓	235
	TVL	-0.002	0.003		0.043	0.017	**	-20.5	✓	✓	✓	✓	✗		✓	62
	Moody's	0.008	0.004	**	0.017	0.006	***	2.0	✓	✓	✓	✓	✓	✓		1829
Median								1.9								1302

4.5 Practical implications

4.5.1 2SLS Estimates with two Instruments

Our method relies on using information from several ESG ratings. Since obtaining access to ESG ratings is costly[16], many users of our method may not have access to seven different ESG ratings. Thus, in this section, we show how the method performs when fewer ratings are available.

[16]In contrast to credit ratings, which follow an issuer-pays model, ESG ratings commonly follow an investor-pays model, meaning that the users of the ratings need to pay.

The minimum is three ESG ratings so that one can serve as the focal regressor and the other two as instruments. At least two instruments are needed to perform an OIR test. We attempt to estimate 2SLS coefficients for all possible sets of 2 instruments. With six ESG ratings to choose from, there are 15 potential sets available.

Table 5 shows the coherent pairs of instruments. Many pairs fail the OIR test. However, at least one pair of instruments passes for each rating. Sustainalytics, MSCI, and SPGlobal are the most frequently accepted instruments in our application to stock returns. The sets that pass the OIR test all yield β estimates larger than the OLS estimates, in line with the main results. A detailed comparison is shown in Appendix Figure A3.

Table 5. Coherent pairs of instruments. This table presents an overview of which pairs of instruments are identified as coherent by the OIR test when the number of instruments is limited to two. It summarizes regressions over horizons of one, two, and three months. The first column provides the regressor that is instrumented, the second column provides the number of pairs that are coherent, i.e., pass the OIR test, and the following columns present the number of times each rating is included in a pair as an instrument. The last line reports the share of pairs in which instruments are included. Since each pair contains two instruments, the shares sum to 200%.

Regressor	**Coherent Pairs**	**Instruments**						
		ISS	MSCI	Refinitiv	SPGlobal	Sustainalytics	TVL	Moody's
ISS	4	0	2	0	0	4	2	0
MSCI	7	0	0	2	3	7	2	0
Refinitiv	4	0	2	0	0	4	2	0
SPGlobal	5	0	3	0	0	5	2	0
Sustainalytics	15	4	9	5	8	0	4	0
TVL	2	0	0	0	2	2	0	0
Moody's	34	9	15	9	12	15	8	0
Total	71	13	31	16	25	37	20	0
Share		18%	44%	23%	35%	52%	28%	0%

4.5.2 Impact on portfolio sorts

Is the difference between the original ratings and instrumented ratings economically meaningful? Investors often rank firms by ESG score to construct ESG portfolios, and many tests in the asset pricing literature rely on portfolio sorts. We, therefore, sort stocks in quintiles based on the original MSCI scores as well as on the instrumented scores and compare the resultant quintiles. In Figure 2, we look at the transitions between quintiles. The horizontal axis shows the quintiles for the original ratings, and the vertical axis shows the quintiles for the instrumented ratings. The instru-

mented ratings are markedly different from the original ratings: Many firms transition into different quintiles. For instance, the fifth quintile of the original ratings has an overlap of only 45.8% with the instrumented ratings. The picture is similar for the first quintile, where the overlap is 46.2%. Put differently, in a long-short portfolio based on the first and fifth quintiles, nearly half of the stocks move to a different quintile after instrumentation. Based on this, we recommend running our procedure to obtain instrumented ratings before conducting standard portfolio sorts common in empirical asset pricing.

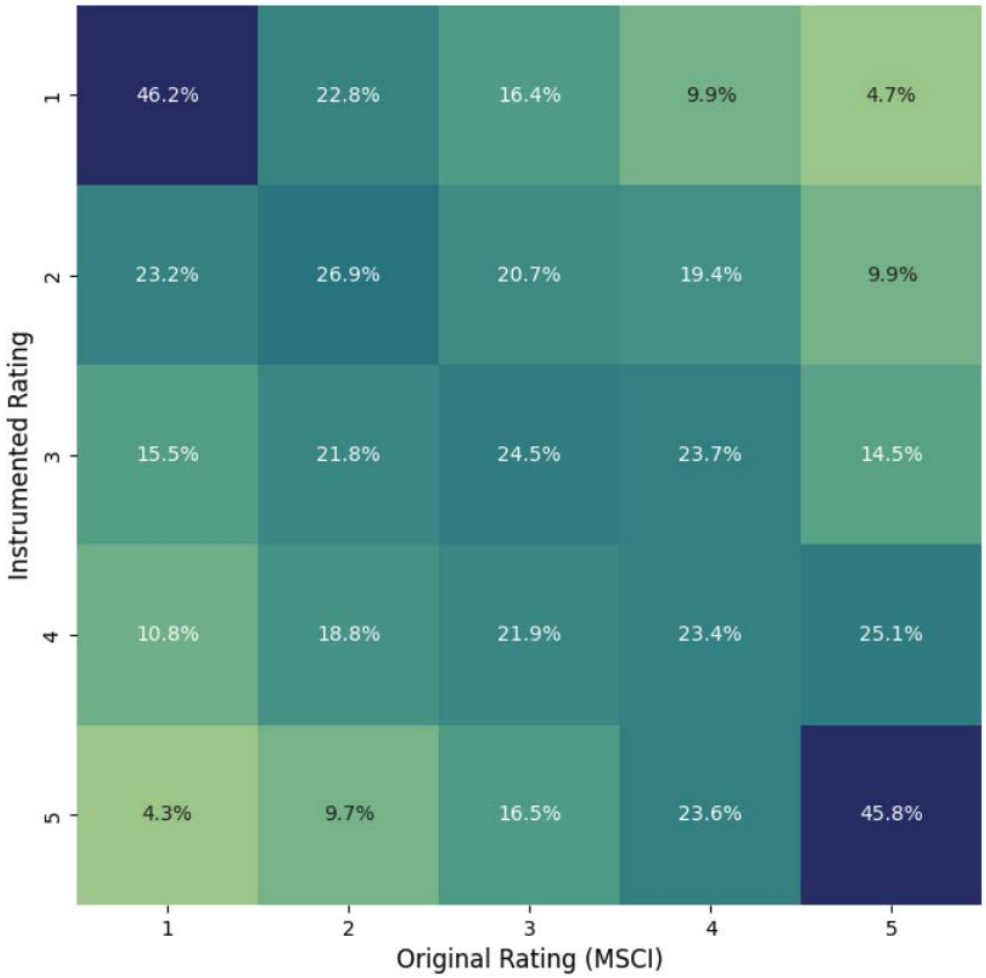


Figure 2. Transition matrix for original MSCI vs. instrumented ratings. The figure depicts how firms change ESG rating quintiles when ESG ratings are instrumented using the 2SLS pruning procedure, combining data for all raters for the ESG dimension. The horizontal axis shows the quintiles for the original ratings, and the vertical axis shows the quintiles for the instrumented ratings. Each row or column sums up to 100%, and if both ratings were strictly the same, each diagonal element would equally sum up to 100%. The first cell at the top-left, for example, indicates that 46.2% of the original ratings in the first quintile are still in the first quintile after instrumentation, and the cell below indicates that 23.2% moved to the second quintile.

5 Simulations

In this section, we present a series of simulations to test the robustness of our IV estimation strategy. First, we compare the 2SLS procedure to alternative noise-reduction procedures that are frequently used by practitioners: averaging the rating scores and using the principal component analysis. We demonstrate that, in our application, the 2SLS procedure is superior to the other two. Second, we show that the OIR test is highly sensitive to simulated violations of our identifying assumptions for 2SLS estimation, providing reassurance that the selection for valid instruments is robust. Third, we explore the case of multiple noisy indicators to the point where there are more sources of noise than instruments, as well as noise coming from measurement in indicators versus noise coming from the aggregation of indicators. We can show that an IV estimation based on multiple instruments tends to come close to the true coefficients in all those settings.

5.1 2SLS versus Alternative Noise-Reduction Techniques

For the simulations, we generate 15,000 observations of stock returns and ESG performance from the following structural model:

$$r_k = \beta \cdot Y_k + \epsilon_k, \tag{15}$$

where k indexes observations. Y_k is a normal random variable with a mean of 0 and a standard deviation of 1. The coefficient of interest is β, which we set equal to 0.5. We assume that there are N rating agencies indexed by i, with scores modeled as in our preceding analysis:

$$s_{k,i} = Y_k + \eta_{k,i}. \tag{16}$$

The focus is on rating $s_{k,1}$, which is our "problematic" regressor, i.e., it is measured with noise. For the remainder of this section, we suppress the subscript k for expositional clarity.

We perform two simulations. In the first, we generate $N = 3$ ratings and benchmark the 2SLS procedure to two alternative noise-reduction approaches, simple averaging and principal component analysis. Here, all instruments are valid, i.e., satisfy assumptions (7), (8), and (9). In the second

simulation, we add an invalid instrument s_4, whose errors η_4 correlate with η_1, thereby violating assumption (9).

In each simulation, we compare several procedures. First, we run a simple OLS:

$$r = \alpha + \beta \cdot s_1 + \nu. \tag{17}$$

Second, one may conjecture that an index constructed as an average of the raters' scores would be less noisy than each rating individually. We therefore construct a simple average of the rating scores, $s^{avg} = 1/N \sum_{i=1}^{N} s_i$, and estimate the following regression:

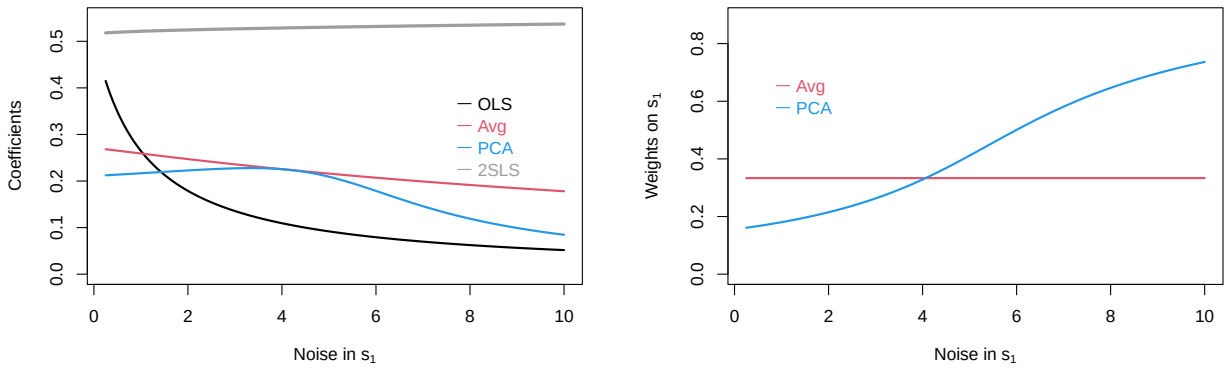
$$r = \alpha_{avg} + \beta_{avg} \cdot s^{avg} + \varepsilon. \tag{18}$$

Third, one may suggest using principal components analysis as a noise reduction procedure. We therefore estimate the first principal component of the N rating agencies' scores, denoted as s^{pc}, and estimate the following regression:

$$r = \alpha_{pc} + \beta_{pc} \cdot s^{pc} + \varepsilon. \tag{19}$$

Finally, we perform our instrumental variable estimation and the corresponding OIR test, using the ratings $i = \{2, \ldots, N\}$ as instruments for s_1 in the first stage.

The first simulation varies σ_{η_1}, the variance of the noise in s_1, from 0.25 to 10. The variance of the noise for s_2 and s_3 is kept constant at a value of 4. For each value of σ_{η_1}, we perform the above noise-reduction procedures and summarize the estimation results in Figure 3. Panel (a) plots the estimate of β for the OLS, simple average, PCA, and 2SLS estimation methods. The 2SLS estimate, shown in gray, is very close to the true β of 0.5 for any level of noise in s_1. It varies from 0.518 to 0.537 over the range of simulated noise in s_1. The OLS estimate displays attenuation bias, which, as expected, increases with the level of noise in s_1. The estimate based on the simple average is inferior to OLS for low levels of noise. As the noise in s_1 increases, it becomes advantageous to include



(a) Attenuation bias

(b) Weight on s_1

Figure 3. Simulation 1: Comparison of OLS, 2SLS, simple average, and PCA. This figure compares the estimation approaches OLS, (17), simple average (18), principal component analysis (PCA) (19), and our 2SLS procedure. Panel (a) shows how the coefficient *beta* varies over different levels of noise in rating s_1. Panel (b) shows the weights that different estimation procedures put on the rating s_1. All instruments are valid by construction.

scores of other raters. This is true even when the other scores are noisier than the regressor because the error terms are independent. The estimate based on the first principal component is never better than the simple average which is because the PCA finds the linear combination of raters' scores that maximizes the observed variance. This approach is quite useful when the variables are measured correctly; however, if the variables are measured with noise, the noise becomes part of the observed variance and a PCA puts greater weight on noisier ratings. This pattern is apparent from Panel (b) of Figure 3. Simple average puts the same weight on all ratings, regardless of their noisiness. Intuitively, one should put a lower weight on a noisier rating. This is, in fact, what our 2SLS procedure does, as noisier indicators have smaller coefficient estimates in the first stage.

Our second simulation adds one instrument, s_4, which is invalid by construction. Specifically, we vary the correlation between η_1 and η_4 from -0.5 to 0.5. As a result, s_1 and s_4 are related not only through Y but also through their errors. The standard deviations of the noise, σ_{η_i}, $i = 1, 2, 3, 4$, are set to be $\{3, 1, 1, 1\}$, respectively. This reflects the case where the potential instruments are less noisy than the regressor.[17] Table 6 presents the results of this simulation.

[17] The qualitative results do not depend on the choice of the standard deviations of the noise, except in the case where the noise on the regressor is exactly zero. The level of variance in the instruments changes the region where the OIR is rejected. The general message, though, remains the same. First, when the noise in instruments is sufficiently correlated, the OIR test is rejected; and second, when there is no rejection, even in the cases where the instruments are invalid, the point estimates of the 2SLS are very close to the true coefficient.

Let us first concentrate on the row corresponding to the correlation of zero. For this row, the assumptions of the classical errors-in-variables problem are satisfied. As before, we see the attenuation bias in the OLS estimation (17): an estimate of 0.060 instead of 0.5. The bias becomes smaller if we use the simple average s^{avg} (18) as a regressor: the coefficient increases to 0.307. Even though the other three scores, s_2, s_3, and s_4, contain less noise than s_1, the resulting estimate is still far from 0.5. The third column shows the estimates when we use the first principal component s^{pc} (19) as the regressor. With 0.072, the estimate is nearly as biased as the OLS estimate. The next two columns show the estimates from the 2SLS procedures, the first one including all available instruments, $\mathbf{Z} = \{s_2, s_3, s_4\}$, the second one just two instruments, $\mathbf{Z} = \{s_2, s_3\}$. Notice that both estimates are very close to 0.5, as the correlation between errors is zero, and thus, all instruments are valid. The last two columns show the p-values of the OIR tests for the 2SLS estimations, and both models pass the test.

Let us now turn to the rows in Table 6 where the correlations between the errors η_1 and η_4 are different from zero. The OLS estimate does not change, since s_4 is not used in this regression. The estimate of the coefficient on the average s^{avg} in column (2), however, increases for negative correlations and decreases for positive correlations. This is because negatively correlated errors cancel each other out. As a result, the average is becoming less noisy, and the coefficient is moving slightly towards the true value of 0.5. The coefficient on the first principal component in column (3) varies because the change in the correlation of the scores implies small changes in the eigenvector. However, all estimated coefficients remain far from the true value of 0.5.

The estimates of the 2SLS regressions in column (4) highlight the outcome of our estimation procedure for the case in which an invalid (and hence also incoherent) instrument is present. The estimated coefficient moves away quickly from the true value of 0.5. The coefficients in column (4) are close to 0.5 only when the correlation is between -0.1 and 0.1. The 2SLS estimates that exclude s_4 in column (5), however, recover the true coefficient for any level of correlation. Most importantly, the OIR test in column (6) shows that models with invalid instruments are rejected with high reliability. The OIR tests in column (7) for the set of invalid instruments are never rejected. Furthermore, cases where the OIR test is not rejected (even though the instruments are

invalid) are instances in which the point estimates of the 2SLS procedure are very close to the true coefficient.

Table 6. Simulation 2: Estimates for the case where the measurement errors η_1 and η_4 are correlated. The true value of the coefficient is 0.5. Columns (1)-(3) report the estimates from regressions (17)–(19), respectively. 2SLS All reports the estimates with $\mathbf{Z} = \{s_2, s_3, s_4\}$ and 2SLS Z reports the estimates with only $\mathbf{Z} = \{s_2, s_3\}$, i.e., the subset of valid instruments. OIR stands for the Sargan-Hansen OIR test, and the corresponding column reports the p-values for the test.

Correlation of η_1 and η_4	(1) OLS	(2) Average	(3) PCA	(4) 2SLS All	(5) 2SLS Z	OIR All (p-value)	OIR Z (p-value)
-0.5	0.060	0.341	0.069	0.107	0.524	0.000	1.000
-0.4	0.060	0.334	0.071	0.177	0.524	0.000	1.000
-0.3	0.060	0.327	0.073	0.285	0.524	0.000	1.000
-0.2	0.060	0.320	0.075	0.422	0.524	0.000	1.000
-0.1	0.060	0.313	0.077	0.528	0.524	0.000	1.000
0	0.060	0.307	0.078	0.530	0.524	0.951	1.000
0.1	0.060	0.301	0.078	0.455	0.524	0.003	1.000
0.2	0.060	0.296	0.079	0.365	0.524	0.000	1.000
0.3	0.060	0.290	0.079	0.290	0.524	0.000	1.000
0.4	0.060	0.285	0.079	0.234	0.524	0.000	1.000
0.5	0.060	0.280	0.079	0.193	0.524	0.000	1.000

Figure 4 shows how the p-value from the OIR test evolves with the correlation between η_1 and η_4. Notice that any correlation larger than 0.1, in absolute value, produces p-values below the threshold. This simulation suggests that the OIR test has sufficient power to reject when an invalid instrument is present.

This result gives us confidence in the ability of our two procedures to select instruments. The 2SLS 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 changes (e.g., from 0.05 to 0.005), we are likely to find a set of instruments that are valid.

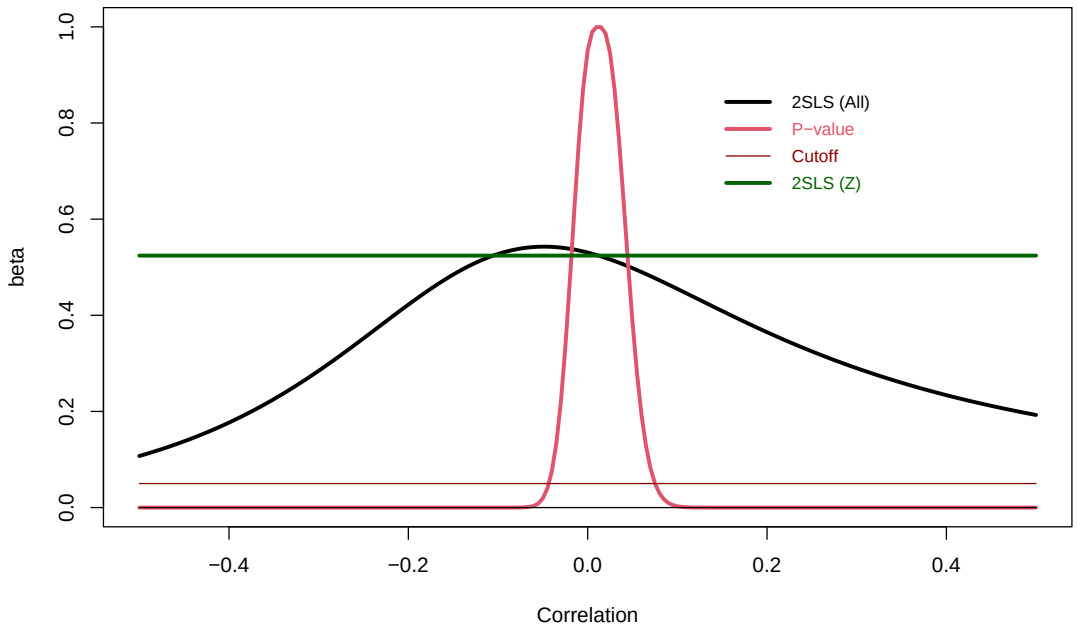


Figure 4. The power of the Sargan-Hansen OIR test. This figure plots the 2SLS coefficients from columns (4) and (5) of Table 6 as functions of the correlation between measurement errors in the simulated ESG ratings s_1 and s_4. The green line is the 2SLS estimate with valid instruments. The thick black line is the 2SLS estimate when including invalid instruments. The thick red line is the p-value of the Sargan-Hansen OIR test when invalid instruments are present, analogous to column (6) in Table 6. The thin red line represents the 0.05 p-value.

5.2 Aggregation of Many ESG Indicators

One potential criticism of our procedure is that ESG performance is a complicated aggregate, which different agencies define in different ways. For example, one agency may include water pollution among the attributes it measures, and another agency may not. Would our procedure still recover the true effect of ESG performance? This section presents a simulation that addresses this question.

As explained in detail in Appendix A.6, the rating agencies' scores are computed as a weighted average of many indicators, corresponding to disaggregated ESG attributes (e.g., carbon emissions, labor practices):

$$s_i = \sum_{a \in \{1,n\}} w_{a,i} \cdot I_{a,i}, \tag{20}$$

where i indexes ESG rating agencies, a indexes attributes that the agency considers, $I_{a,i}$ is rater i's measure of attribute a, and $w_{a,i}$ are the weights.

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

$$Y = \sum_{a \in \{1,n\}} w_a^\star \cdot I_a^\star,$$

where $I_a^\star$ are the true values of the indicators and $w_a^\star$ are the true weights—i.e., the weights that the representative ESG investor assigns to individual indicators, which reflect her preferences.

The measurement error of each rating agency can be decomposed as follows:

$$s_i = Y + \underbrace{\sum_{a \in \{1,n\}} w_{a,i} \cdot \underbrace{(I_{a,i} - I_a^\star)}_{\eta_{I_{a,i}}} + \sum_{a \in \{1,n\}} \underbrace{(w_{a,i} - w_a^\star)}_{\eta_{w_{a,i}}} \cdot I_a^\star}_{\eta_{Y_i}}.$$

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

$$I_{a,i} = I_a^\star + \eta_{I_{a,i}}, \tag{21}$$

and the discrepancy in the weights,

$$w_{a,i} = w_a^\star + \eta_{w_{a,i}}. \tag{22}$$

The validity of instruments requires orthogonality of $\eta_{I_{a,i}}$ and $\eta_{w_{a,i}}$ across rating agencies. The measurement error in the aggregated rating, η_{Y_i}, parallels our η_i in Section 3 (see Equation (2)). Appendix A.6 provides more detail.

Even under the above assumptions, it is unclear how our 2SLS procedure performs when there are many sources of noise and a limited number of instruments. If there are 8 possible attributes, then there are 8 sources of measurement error and 7 different weights. Would our procedure that relies on only 6 other raters' scores to use as instruments recover the true coefficient?

There are two sources of noise in rating agencies' scores that serve as our instruments. First, every individual attribute in (20) is measured with noise. For simplicity, in this simulation, we

assume that measurement errors in each indicator have the same variance equal to 1. Second, each rating agency measures only a subset of the attributes. Therefore, the scores from the rating agencies that will be used as instruments present an incomplete picture of the true ESG performance. Such incomplete coverage, which Berg, Kölbel, and Rigobon (2022) term differences in scope, is a source of potential bias. Third, we assume that all agencies use equal weighting in their ESG scores. Hence, our simulation explores the effects of two sources of noise present in the ratings: (i) the noise with which individual indicators are measured and (ii) the noise resulting from the coverage problem.

We continue using the following structural model to simulate observations of stock returns:

$$r_k = \beta \cdot Y_k + \epsilon_k,$$

where k indexes observations, $Y_k = \frac{1}{n} \sum_{a \in \{1,n\}} I^{\star}_{a,k}$, and $\beta = 0.5$. The corresponding reduced-form model is

$$r_k = \alpha + \beta \cdot s_{1,k} + \nu_k. \tag{23}$$

We again estimate Equation (23), first by OLS and then by 2SLS, using other rating agencies' scores as instruments. We assume that there are 8 attributes and that these attributes are orthogonal to each other. We further assume that the rating of interest, s_1, includes all 8 attributes in its ESG rating and that it measures each one with noise. We assume that only 5 other rating agencies' ESG scores are available for use as instruments. As in our previous simulation, we generate 15,000 draws. Errors with which each attribute is measured are classical and satisfy the identifying assumptions spelled out in Appendix A.6. These assumptions imply that all instruments are valid.

The results of the simulation are presented in Table 7, which has the following structure. In the first column, we report the coefficient from our baseline OLS regression (17). In the remaining columns, we present 2SLS estimates, in which we vary the number of instruments used in the first stage from 1 to 5, labeled as IV1 to IV5. In rows, we vary the number of attributes. In the first set of rows, the instruments cover only 1 of the possible 8 attributes, in the second set 2 of the possible 8, and so on, until the last set of rows in which each instrument covers all the attributes.

We randomly choose the subset of attributes that is covered by each rater, with replacement. The rating is then computed as a simple average of the simulated indicators.

Table 7. Simulation 3: Fewer instruments than sources of noise. The true value of the coefficient is 0.5. In the first column, we report the coefficient from our baseline OLS regression (17). In the remaining columns, we present 2SLS estimates, in which we vary the number of instruments used in the first stage from 1 to 5, labeled as IV1 to IV5. In rows, we vary the number of attributes covered by each rating. The rating s_1 that we are instrumenting covers all 8 attributes.

Number of Attributes per Rating		OLS	IV1	IV2	IV3	IV4	IV5
1	Coefficient	0.07	0.681	0.839	0.826	0.686	0.702
	Std Error	0.031	0.28	0.21	0.186	0.171	0.169
	1st stage F-stat		188	169	144	129	105
2	Coefficient	0.07	0.797	0.804	0.786	0.757	0.687
	Std Error	0.031	0.219	0.194	0.175	0.168	0.162
	1st stage F-stat		310	198	164	133	114
3	Coefficient	0.07	0.466	0.503	0.427	0.354	0.378
	Std Error	0.031	0.209	0.179	0.168	0.157	0.145
	1st stage F-stat		341	234	179	154	146
4	Coefficient	0.07	0.814	0.641	0.639	0.621	0.63
	Std Error	0.031	0.209	0.169	0.157	0.152	0.147
	1st stage F-stat		342	264	205	165	140
5	Coefficient	0.07	0.729	0.727	0.707	0.707	0.727
	Std Error	0.031	0.179	0.166	0.15	0.144	0.139
	1st stage F-stat		466	274	226	184	158
6	Coefficient	0.07	0.405	0.452	0.437	0.492	0.46
	Std Error	0.031	0.167	0.148	0.142	0.137	0.135
	1st stage F-stat		540	347	254	206	170
7	Coefficient	0.07	0.636	0.487	0.494	0.498	0.492
	Std Error	0.031	0.173	0.151	0.144	0.137	0.134
	1st stage F-stat		500	333	246	205	172
8	Coefficient	0.07	0.645	0.448	0.477	0.451	0.488
	Std Error	0.031	0.163	0.148	0.139	0.135	0.133
	1st stage F-stat		569	348	265	210	175

Let us now discuss the first set of (three) rows presented in Table 7, in which each instrument measures only one attribute. Notice that the OLS coefficient is 0.07. The 2SLS coefficient with one instrument is 0.681 and 0.702 when 5 instruments are used. The second row in the set presents the standard errors of the estimates. Notice that the precision of the estimates improves with the number of instruments. The OLS coefficient is statistically different from zero and is statistically smaller than 0.5 at the 1 percent significance level. The coefficients in IV1 to IV5 estimations are consistently higher than their OLS counterpart and fairly close to the true parameter $\beta = 0.5$. The

third row reports the F-statistic of the first stage, which indicates that the instruments still have relevance, which is expected given that all the scores are correlated through the fundamental ESG attributes. This indicates that in the extreme case where all instruments only cover each 1 of the 8 indicators that make up Y, a 2SLS approach alleviates attenuation bias, and comes fairly close to the true parameter.

The simulations get even closer to the true parameter when the instruments contain more information about Y. We have also run OIR tests for all IV specifications that use two or more instruments, and observed no rejections. This is not surprising. While our instruments have incomplete coverage, they are still valid by construction, irrespective of whether the noise comes from the measurement of individual indicators or from an omitted fundamental.

In this section, we have studied a simulation in which ESG ratings include 8 possible attributes. Estimation performed on more granular ratings, E, S, or G, or, better yet, at an individual indicator level (measuring just one attribute) would reduce estimation error introduced by data aggregation. Ideally, we would like to have a separate instrument for each individual indicator included in a rating. Instead, we are using a weighted sum of indicators (e.g., an ESG score) as an instrument for another weighted sum (e.g., an ESG score of another rater).

6 Limitations

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. Our estimation relies primarily on cross-sectional variation because many rating agencies change their scores once a year at most, which implies that in practice the time series is even shorter than the time span of our sample.

Second, many rating agencies are in the process of consolidation, which often involves revising their procedures. Thus, some rating agencies back-fill their past scores based on a revised procedure. This is particularly problematic if the back-fill is based on stock-relevant information. If point-in-

time scores are not available, practitioners and applied work can use our procedure to diagnose this problem, which will show up as a rejection of the OIR tests.

Finally, the ESG score could be capturing unobservable firm characteristics. In our specification this corresponds to the omitted variable. As we have discussed before, the omitted variable has no impact on our results regarding the attenuation bias. However, it does change the structural interpretation of the coefficient. A potential omitted variable could be management quality. Better managers might be able to foster more collaborative work environments that result in less labor mistreatment at the same time. This could impact both stock returns and ESG performance. Thus, future research should look into disentangling management quality and ESG performance by adding firm fixed effects, for instance. Of course, though, this would require much longer time series than are currently commercially available.

7 Conclusions

It is notoriously difficult to measure the ESG performance of firms. ESG rating agencies often report different estimates for the same attribute. We argue that noise in the estimates leads to a significant attenuation bias in the standard regressions that analyze the effects of ESG performance.

In this paper, we turn the problem of divergence to our advantage by exploiting the fact there is information in observing multiple ESG ratings at once. We assume that each ESG rating measures ESG performance with noise. We propose an instrumental variable approach that uses the information from several ESG ratings to reduce the bias from noisy measurement.

We show that coefficients more than double when we apply our noise-correction procedure. We run our estimation separately for the seven raters in our sample and across different return horizons. In most of these regressions, we observe an increase in the estimates. The median increase is a factor of 2.1.

The practical takeaway of these results is that it is worthwhile to rely on several complementary ESG ratings. While we find the scores of some rating agencies to be very noisy, it does not mean

that they are uninformative. One illustrative example is Truvalue Labs. Our estimation procedure shows that while these scores do not perform well as predictors of stock returns, they are nevertheless valuable instruments that enhance the prediction of other scores.

One may be tempted to conclude from our results that one should use ESG scores containing the least noise rather than scores of other raters. We caution against such an interpretation for two reasons. First, the implied noise we compute is specific to our model and regression setup and should not be overgeneralized. Second, our results show that coefficients also increase substantially for the least noisy ratings when instrumented with other ratings. In other words, relying on the scores of several complementary ratings yields better results.

Our paper offers a practical solution to deal with the divergence of ESG ratings. Whenever an ESG rating is used as a regressor, attenuation bias is likely to become a problem. If a second ESG rating is available, it can be used as an instrument, which reduces the attenuation bias. In this case, one must defend the assumption that the measurement error of the other ESG rating is orthogonal. If more than one additional ESG rating is available, one can rely on the OIR test to check the coherence of instruments. This 2SLS approach to noise reduction is superior to using the averages of ESG scores or principal component analysis. And if noise is indeed a problem, it will make the empirical results stronger.

References

Albuquerque, Rui, Yrjö Koskinen, and Chendi Zhang, 2019, Corporate social responsibility and firm risk: Theory and empirical evidence, *Management Science* 65, 4451–4469.

Atz, Ulrich, Christopher Bruno, Zongyuan Zoe Liu, and Tracy Van Holt, 2022, Does sustainability generate better financial performance? Review, meta-analysis, and propositions, *Journal of Sustainable Finance and Investment (Forthcoming)*.

Avramov, Doron, Si Cheng, Abraham Lioui, and Andrea Tarelli, 2022, Sustainable investing with esg rating uncertainty, *Journal of Financial Economics* 145, 642–664.

van der Beck, Philippe, 2021, Flow-driven ESG returns, *Working Paper,* Available at: https://papers.ssrn.com/abstract=3929359.

Berg, Florian, Kornelia Fabisik, and Zacharias Sautner, 2021, Is history repeating itself? The (un)predictable past of ESG ratings, *Working Paper, ECGI* Available at: https://ssrn.com/abstract=3722087.

Berg, Florian, Julian F Kölbel, and Roberto Rigobon, 2022, Aggregate confusion: The divergence of ESG ratings, *Review of Finance* 26, 1315–1344.

Bolton, Patrick, and Marcin T. Kacperczyk, 2022, Global pricing of carbon-transition risk, *Journal of Finance,* forthcoming.

Chava, Sudheer, 2014, Environmental externalities and cost of capital, *Management Science* 60, 2223–2247.

Christensen, Dane M., George Serafeim, and Anywhere Sikochi, 2022, Why is corporate virtue in the eye of the beholder? The case of ESG ratings, *The Accounting Review* 97, 147–175.

Christensen, Hans Bonde, Luzi Hail, and Christian Leuz, 2021, Mandatory CSR and sustainability reporting: Economic analysis and literature review, *Review of Accounting Studies* 26, 1176–1248.

Cornett, Marcia Millon, Otgontsetseg Erhemjamts, and Hassan Tehranian, 2016, Greed or good deeds: An examination of the relation between corporate social responsibility and the financial

performance of U.S. commercial banks around the financial crisis, *Journal of Banking & Finance* 70, 137–159.

Edmans, Alex, 2011, Does the stock market fully value intangibles? Employee satisfaction and equity prices, *Journal of Financial Economics* 101, 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, 2388–2406.

Erickson, Timothy, and Toni M. Whited, 2010, Measurement error and the relationship between investment and q, *Journal of Political Economy* 118, 1252–1257.

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 Brandon, Rajna, Philipp Krueger, and Peter Steffen Schmidt, 2021, ESG rating disagreement and stock returns, *Financial Analysts Journal* 77, 104–127.

Gillen, Ben, Erik Snowberg, and Leeat Yariv, 2019, Experimenting with Measurement Error: Techniques with Applications to the Caltech Cohort Study, *Journal of Political Economy* 127, 1826–1863 Publisher: The University of Chicago Press.

Heinkel, Robert, Alan Kraus, and Josef Zechner, 2001, The effect of green investment on corporate behavior, *Journal of Financial and Quantitative Analysis* 36, 431–449.

Karolyi, George Andrew, Ying Wu, and Wei (William) Xiong, 2023, Understanding the global equity greenium, *Working Paper,* Available at: https://papers.ssrn.com/abstract=4391189.

Lewellen, Jonathan, 2015, The cross-section of expected stock returns, *Critical Finance Review* 4, 1–44.

Liang, Hao, and Luc Renneboog, 2017, Corporate donations and shareholder value, *Oxford Review of Economic Policy* 33, 278–316.

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, 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, 2905–2932.

Pancost, N. Aaron, and Garrett Schaller, 2021, Measuring measurement error, *Working Paper* Available at: https://www.ssrn.com/abstract=4045772.

Parente, Paulo M.D.C., and J.M.C. Santos Silva, 2012, A cautionary note on tests of overidentifying restrictions, *Economics Letters* 115, 314–317.

Pastor, Lubos, Robert F. Stambaugh, and Lucian A. Taylor, 2021, Sustainable investing in equilibrium, *Journal of Financial Economics* 142, 550–571.

——— , 2022, Dissecting green returns, *Journal of Financial Economics* 146, 403–424.

——— , 2023, Green tilts, *NBER Working Paper 31320.*

A Internet Appendix

A.1 Model

In this section, we present a simple, stylized model with traditional and ESG investors that highlights the attenuation effect when ESG signals are noisy. Our focus is on a single ESG signal, measured with noise. We will show that such measurement error leads to bias in a standard regression analysis of the relationship between stock returns and ESG performance. The noisier the ESG signal, the larger the bias.

We consider a two-period model, with $t = 0, 1$. Investment opportunities are represented by a risky stock of a single firm and a riskless bond, with the risk-free rate normalized to zero.[18] 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 p_t, where $p_1 = D$.

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 $W_0^i + \theta^i(D - p_0)$, 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, $U(W_1, Y) = -exp(-\gamma(W_1 + \theta^{ESG}Y))$.[19] 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.

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

[19] Our approach to modeling ESG investors is similar to that of Pastor, Stambaugh, and Taylor (2021) and Friedman and Heinle (2016).

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

$$s_D = D + \eta_D, \tag{24}$$

$$s_Y = Y + \eta_Y, \tag{25}$$

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

A.1.1 Portfolio Choice and Asset Prices

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 can be found in the Appendix A.1.2).

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

$$E(D|s_D) = \overline{D} + \beta(s_D - \overline{D}) = \overline{D} + \frac{\sigma^2_D}{\sigma^2_D + \sigma^2_{\eta_D}}(s_D - \overline{D}), \tag{26}$$

$$Var(D|s_D) = \sigma^2_{\nu_D} = \frac{\sigma^2_D \sigma^2_{\eta_D}}{\sigma^2_D + \sigma^2_{\eta_D}}. \tag{27}$$

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

$$E(Y|s_Y) = \overline{Y} + \beta(s_Y - \overline{Y}) = \overline{Y} + \frac{\sigma^2_Y}{\sigma^2_Y + \sigma^2_{\eta_Y}}(s_Y - \overline{Y}), \tag{28}$$

$$Var(Y|s_Y) = \sigma^2_{\eta_Y} = \frac{\sigma^2_Y \sigma^2_{\eta_Y}}{\sigma^2_Y + \sigma^2_{\eta_Y}}. \tag{29}$$

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*

$$\theta^T = \frac{1}{\gamma}\frac{E(D|s_D) - p_0}{Var(D|s_D)}, \tag{30}$$

$$\theta^{ESG} = \frac{1}{\gamma}\frac{E(D|s_D) + E(Y|s_Y) - p_0}{Var(D|s_D) + Var(Y|s_Y)}. \tag{31}$$

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 (31)).

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

$$\lambda\theta^{ESG} + (1-\lambda)\theta^T = \overline{\theta}. \tag{32}$$

To solve for the equilibrium stock price, we substitute the optimal portfolios from Lemma 2 into the market clearing condition (32). We report the resulting period-0 stock price in the following proposition.

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

$$p_0 = \overline{D} + \frac{\sigma_D^2}{\sigma_D^2 + \sigma_{\eta_D}^2}(s_D - \overline{D}) + A\lambda\frac{\sigma_D^2\sigma_{\eta_D}^2}{\sigma_D^2 + \sigma_{\eta_D}^2}\left[\overline{Y} + \frac{\sigma_Y^2}{\sigma_Y^2 + \sigma_{\eta_Y}^2}(s_Y - \overline{Y})\right] \tag{33}$$

$$- A\gamma\overline{\theta}\frac{\sigma_D^2\sigma_{\eta_D}^2}{\sigma_D^2 + \sigma_{\eta_D}^2}\left[\frac{\sigma_D^2\sigma_{\eta_D}^2}{\sigma_D^2 + \sigma_{\eta_D}^2} + \frac{\sigma_Y^2\sigma_{\eta_Y}^2}{\sigma_Y^2 + \sigma_{\eta_Y}^2}\right], \tag{34}$$

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

The ESG performance Y does not affect fundamentals (i.e., the firm's cash flow D). However, it does affect asset prices because there is a group of investors that care about it. A positive signal

s_Y about the ESG performance Y boosts the stock price. Y can be interpreted as the true ESG performance, 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.

Proposition 1 establishes a linear relationship between the ESG signal and stock returns. Let us now examine a *realized* per-share return on the stock in period 0:

$$\begin{aligned} p_0 - p_{-1} =& \overline{D} - S_{-1} + \frac{\sigma_D^2}{\sigma_D^2 + \sigma_{\eta_D}^2}(s_D - \overline{D}) + A\lambda \frac{\sigma_D^2 \sigma_{\eta_D}^2}{\sigma_D^2 + \sigma_{\eta_D}^2} \left[\overline{Y} + \underbrace{\frac{\sigma_Y^2}{\sigma_Y^2 + \sigma_{\eta_Y}^2}}_{\text{attenuation effect}} (s_Y - \overline{Y}) \right] \\ & - A\gamma\overline{\theta} \frac{\sigma_D^2 \sigma_{\eta_D}^2}{\sigma_D^2 + \sigma_{\eta_D}^2} \left[\frac{\sigma_D^2 \sigma_{\eta_D}^2}{\sigma_D^2 + \sigma_{\eta_D}^2} + \frac{\sigma_Y^2 \sigma_{\eta_Y}^2}{\sigma_Y^2 + \sigma_{\eta_Y}^2} \right]. \end{aligned} \tag{35}$$

We think about the constant p_{-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.

$$\begin{aligned} E(D) - p_0 =& - \frac{\sigma_D^2}{\sigma_D^2 + \sigma_{\eta_D}^2}(s_D - \overline{D}) - A\lambda \frac{\sigma_D^2 \sigma_{\eta_D}^2}{\sigma_D^2 + \sigma_{\eta_D}^2} \left[\overline{Y} + \underbrace{\frac{\sigma_Y^2}{\sigma_Y^2 + \sigma_{\eta_Y}^2}}_{\text{attenuation effect}} (s_Y - \overline{Y}) \right] \\ & + A\gamma\overline{\theta} \frac{\sigma_D^2 \sigma_{\eta_D}^2}{\sigma_D^2 + \sigma_{\eta_D}^2} \left[\frac{\sigma_D^2 \sigma_{\eta_D}^2}{\sigma_D^2 + \sigma_{\eta_D}^2} + \frac{\sigma_Y^2 \sigma_{\eta_Y}^2}{\sigma_Y^2 + \sigma_{\eta_Y}^2} \right]. \end{aligned} \tag{36}$$

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.[20]

Both the expected and realized returns depend on the noise in the ESG signal, σ_{η_Y}. The noisier the signal s_Y, the lower its passthrough to stock returns. Put differently, noise in the signal creates an attenuation effect (see the highlighted term in Equations (35) and (36)). In the limit of $\sigma_{\eta_Y} \to \infty$, the effect of s_Y on stock returns is fully attenuated. These observations will become important in our empirical analysis, which uses data on ESG scores, which we interpret as noisy ESG signals. The noise in the reported ESG scores is apparent from the discrepancies in measuring the same ESG performance by different ratings providers. In the next section, we treat this as a classical errors-in-variables problem and propose a procedure that tackles the attenuation bias in standard regressions of stock returns on noisy measures of ESG performance.

A.1.2 Proofs

Proof of Lemma 1. We start by showing (26)-(27). 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:

$$D - \overline{D} = \beta_D(s_D - \overline{D}) + \varepsilon_D, \tag{37}$$

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

The mean and variance of D conditional on signal s_D are

$$E(D|s_D) = \overline{D} + \beta_D(s_D - \overline{D}) \tag{38}$$

$$Var(D|s_D) = \sigma^2_{\varepsilon_D} \tag{39}$$

[20] 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.

The regression coefficient β_D is given by the following standard expression:

$$\beta_D = \frac{Cov(D - \overline{D}, s_D - \overline{D})}{Var(s_D - \overline{D})} = \frac{(Cov(D - \overline{D}, D - \overline{D} + \eta_D)}{Var(D - \overline{D} + \eta_D)} = \frac{\sigma_D^2}{\sigma_D^2 + \sigma_{\eta_D}^2} \tag{40}$$

The variables D and η_D are independent. Taking variances on each side of (37), we have

$$Var(D - \overline{D}) = Var(\beta_D(s_D - \overline{D}) + \varepsilon_D) = \beta_D^2 Var(s_D - \overline{D}) + \sigma_{\varepsilon_D}^2 \tag{41}$$

It is easy to see that

$$\sigma_{\varepsilon_D}^2 = \sigma_D^2 - \left(\frac{\sigma_D^2}{\sigma_D^2 + \sigma_{\eta_D}^2}\right)^2 (\sigma_D^2 + \sigma_{\eta_D}^2) = \frac{\sigma_D^2 \sigma_{\eta_D}^2}{\sigma_D^2 + \sigma_{\eta_D}^2} \tag{42}$$

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

$$E(D|s_D) = \overline{D} + \beta(s_D - \overline{D}) = \overline{D} + \frac{\sigma_D^2}{\sigma_D^2 + \sigma_{\eta_D}^2}(s_D - \overline{D}) \tag{43}$$

$$Var(D|s_D) = \sigma_{\varepsilon_D}^2 = \frac{\sigma_D^2 \sigma_{\eta_D}^2}{\sigma_D^2 + \sigma_{\eta_D}^2} \tag{44}$$

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 with Y and s_D with 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

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

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:

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

Solving for the portfolio choice θ^{ESG} that maximizes the above objective, we arrive at (31).

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 (32), we derive

$$\begin{aligned} p_0 =& A\lambda Var(D|s_D)(E(D|s_D) + E(Y|s_Y)) + A(1-\lambda)(Var(D|s_D) + Var(Y|s_Y))E(D|s_D) \\ & - A\gamma\bar{\theta} Var(D|s_D)(Var(D|s_D) + Var(Y|s_Y)) \end{aligned}$$

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

$$\begin{aligned} p_0 =& A\lambda \frac{\sigma_D^2 \sigma_{\eta_D}^2}{\sigma_D^2 + \sigma_{\eta_D}^2} \left(\overline{Y} + \frac{\sigma_Y^2}{\sigma_Y^2 + \sigma_{\eta_Y}^2}(s_Y - \overline{Y}) + \overline{D} + \frac{\sigma_D^2}{\sigma_D^2 + \sigma_{\eta_D}^2}(s_D - \overline{D}) \right) \\ & + A(1-\lambda) \left(\frac{\sigma_D^2 \sigma_{\eta_D}^2}{\sigma_D^2 + \sigma_{\eta_D}^2} + \frac{\sigma_Y^2 \sigma_{\eta_Y}^2}{\sigma_Y^2 + \sigma_{\eta_Y}^2} \right) \left(\overline{D} + \frac{\sigma_D^2}{\sigma_D^2 + \sigma_{\eta_D}^2}(s_D - \overline{D}) \right) \\ & - A\gamma\bar{\theta} \frac{\sigma_D^2 \sigma_{\eta_D}^2}{\sigma_D^2 + \sigma_{\eta_D}^2} \left[\frac{\sigma_D^2 \sigma_{\eta_D}^2}{\sigma_D^2 + \sigma_{\eta_D}^2} + \frac{\sigma_Y^2 \sigma_{\eta_Y}^2}{\sigma_Y^2 + \sigma_{\eta_Y}^2} \right], \end{aligned}$$

where

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

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

A.2 Derivation of Estimates

In this section, we derive the IV estimator and show how it resolves the attenuation bias of the OLS estimator. For expositional simplicity, we use the asymptotic notation — i.e., we use *var* and *cov* as opposed to the matrix notation $\mathbf{X}'\mathbf{X}$.

A.2.1 OLS and IV

Under the assumption that the controls are orthogonal to all RHS variables, the OLS estimate of β in Equation (3) is given by

$$\beta_{OLS} = \frac{cov(r_{k,t+1}, s_{k,t,i})}{var(s_{k,t,i})}.$$

Using the structural Equations (1) and (2), the covariance between the stock returns and the ESG score is

$$\begin{aligned} cov(r_{k,t+1}, s_{k,t,i}) &= cov(\beta Y_{k,t} + M_{k,t}, Y_{k,t}) \\ &= \beta var(Y_{k,t}) + cov(M_{k,t}, Y_{k,t}) \\ &= \beta var(Y_{k,t}) + \gamma var(M_{k,t}). \end{aligned}$$

The variance of the score is given by

$$\begin{aligned} var(s_{k,t,i}) &= var(Y_{k,t} + \eta_{k,t,i}) \\ &= var(Y_{k,t}) + var(\eta_{k,t,i}). \end{aligned}$$

The ratio of these two quantities is the OLS estimate:

$$\begin{aligned} \beta_{OLS} &= \frac{\beta \cdot var(Y_{k,t}) + \gamma \cdot var(M_{k,t})}{var(Y_{k,t}) + var(\eta_{k,t,i})} \\ &= \underbrace{\frac{var(Y_{k,t})}{var(Y_{k,t}) + var(\eta_{k,t,i})}}_{\text{Attenuation}} \cdot \left[\beta + \underbrace{\gamma \frac{var(M_{k,t})}{var(Y_{k,t})}}_{\text{Omitted Variable}} \right]. \end{aligned}$$

The sample equivalent of this estimate is in equation (4) in the main text.

The derivation of IV estimate is as follows: The structural equations of the instrumental variable estimate are

$$\begin{aligned}
\beta_{IV} &= \frac{cov(r_{k,t+1}, s_{k,t,j})}{cov(s_{k,t,i},, s_{k,t,j})}, \\
r_{k,t+1} &= \alpha + \beta \cdot Y_{k,t} + M_{k,t} + \epsilon_{k,t}, \\
s_{k,t,i} &= Y_{k,t} + \eta_{k,t,i}, \\
s_{k,t,j} &= Y_{k,t} + \eta_{k,t,j},
\end{aligned}$$

where the score from the rating agency i is the regressor, while the score from the rating agency $j \neq i$ is the instrument. The covariance between the stock returns and the ESG score used for the instrumentation ($s_{k,t,j}$) when the measurement errors are orthogonal to all controls and innovations is as follows:

$$\begin{aligned}
cov(r_{k,t+1}, s_{k,t,j}) =& cov(\beta Y_{k,t} + M_{k,t} + \epsilon_{k,t}, Y_{k,t} + \eta_{k,t,j}) \\
=& \beta var(Y_{k,t}) + cov(M_{k,t}, Y_{k,t}) \\
=& \beta var(Y_{k,t}) + \gamma var(M_{k,t}).
\end{aligned}$$

The covariance between the ESG score from rating i and the ESG score used for the instrumentation ($s_{k,t,j}$) is as follows:

$$\begin{aligned}
cov(s_{k,t,i}, s_{k,t,j}) &= cov(Y_{k,t} + \eta_{k,t,i}, Y_{k,t} + \eta_{k,t,j}) \\
&= var(Y_{k,t}).
\end{aligned}$$

The IV estimate is then

$$\begin{aligned}\beta_{IV} =& \frac{\beta var(Y_{k,t}) + \gamma var(M_{k,t})}{var(Y_{k,t})} \\ =& \beta + \gamma \frac{var(M_{k,t})}{var(Y_{k,t})}.\end{aligned}$$

The sample equivalent of this estimate is in equation (5) in the main text.

A.2.2 Biased and Inconsistent Estimates of OLS and IV

In this section, we derive the IV estimator, allowing for omitted variable bias. All variables are assumed to have mean zero, we drop the constants and the controls to simplify the notation. Including them in the derivation does not change the results. Under the assumption that the controls are orthogonal to all RHS variables, the OLS estimate of β in Equation (3) is given by

$$\begin{aligned}\beta_{OLS} &= \frac{cov(r_{k,t+1}, s_{k,t,i})}{var(s_{k,t,i})}, \\ r_{k,t+1} &= \alpha + \beta \cdot Y_{k,t} + M_{k,t} + \epsilon_{k,t}, \\ s_{k,t,i} &= Y_{k,t} + \eta_{k,t,i}.\end{aligned}$$

The covariance between the stock returns and the ESG score is

$$\begin{aligned}cov(r_{k,t+1}, s_{k,t,i}) &= cov(\beta Y_{k,t} + M_{k,t} + \epsilon_{k,t}, Y_{k,t} + \eta_{k,t,i}) \\ &= \beta var(Y_{k,t}) + cov(M_{k,t}, Y_{k,t}) + \beta cov(Y_{k,t}, \eta_{k,t,i}) + cov(M_{k,t}, \eta_{k,t,i}) + cov(\epsilon_{k,t}, \eta_{k,t,i}) \\ &= \beta var(Y_{k,t}) + \gamma var(M_{k,t}) + \beta cov(Y_{k,t}, \eta_{k,t,i}) + cov(M_{k,t}, \eta_{k,t,i}) + cov(\epsilon_{k,t}, \eta_{k,t,i}).\end{aligned}$$

The variance of the score is given by

$$\begin{aligned}var(s_{k,t,i}) &= var(Y_{k,t} + \eta_{k,t,i}) \\ &= var(Y_{k,t}) + var(\eta_{k,t,i}) + cov(Y_{k,t}, \eta_{k,t,i}).\end{aligned}$$

The ratio of these two quantities is the OLS estimate:

$$\beta_{OLS} = \frac{\beta var(Y_{k,t}) + \gamma var(M_{k,t}) + \beta cov(Y_{k,t}, \eta_{k,t,i}) + cov(M_{k,t}, \eta_{k,t,i}) + cov(\epsilon_{k,t}, \eta_{k,t,i})}{var(Y_{k,t}) + var(\eta_{k,t,i}) + cov(Y_{k,t}, \eta_{k,t,i})}. \tag{45}$$

The structural equations of the instrumental variable estimate are

$$\begin{aligned}
\beta_{IV} &= \frac{cov(r_{k,t+1}, s_{k,t,j})}{cov(s_{k,t,i}, , s_{k,t,j})}, \\
r_{k,t+1} &= \alpha + \beta \cdot Y_{k,t} + M_{k,t} + \epsilon_{k,t}, \\
s_{k,t,i} &= Y_{k,t} + \eta_{k,t,i}, \\
s_{k,t,j} &= Y_{k,t} + \eta_{k,t,j},
\end{aligned}$$

where the score from the rating agency i is the regressor, while the score from the rating agency $j \neq i$ is the instrument. The covariance between the stock returns and the ESG score used for the instrumentation ($s_{k,t,j}$) is as follows:

$$\begin{aligned}
cov(r_{k,t+1}, s_{k,t,j}) =& cov(\beta Y_{k,t} + M_{k,t} + \epsilon_{k,t}, Y_{k,t} + \eta_{k,t,j}) \\
=& \beta var(Y_{k,t}) + cov(M_{k,t}, Y_{k,t}) + \beta cov(Y_{k,t}, \eta_{k,t,j}) + \\
& cov(M_{k,t}, \eta_{k,t,j}) + cov(\epsilon_{k,t}, \eta_{k,t,j}) \\
=& \beta var(Y_{k,t}) + \gamma var(M_{k,t}) + \beta cov(Y_{k,t}, \eta_{k,t,j}) + cov(M_{k,t}, \eta_{k,t,j}) + cov(\epsilon_{k,t}, \eta_{k,t,j}).
\end{aligned}$$

The covariance between the ESG score from rating i and the ESG score used for the instrumentation ($s_{k,t,j}$) is the following

$$\begin{aligned}
cov(s_{k,t,i}, s_{k,t,j}) &= cov(Y_{k,t} + \eta_{k,t,i}, (Y_{k,t} + \eta_{k,t,j}) \\
&= var(Y_{k,t}) + cov(Y_{k,t}, \eta_{k,t,j}) + cov(Y_{k,t}, \eta_{k,t,i}) + cov(\eta k, t, i, \eta_{k,t,j}).
\end{aligned}$$

The IV estimate is then

$$\beta_{IV} = \frac{\beta var(Y_{k,t}) + \gamma var(M_{k,t}) + \beta cov(Y_{k,t}, \eta_{k,t,j}) + cov(M_{k,t}, \eta_{k,t,j}) + cov(\epsilon_{k,t}, \eta_{k,t,j})}{var(Y_{k,t}) + cov(Y_{k,t}, \eta_{k,t,j}) + cov(Y_{k,t}, \eta_{k,t,i}) + cov(\eta_{k,t,i}, \eta_{k,t,j})}. \tag{46}$$

The sample equivalents of the OLS and IV estimates in (45)–(46) are in equations (10)–(11) in the main text.

A.2.3 Attenuation Across Horizons

In this section, we derive the ratio between the OLS and the IV coefficients and show that it is the same for different horizons of stock returns. Denote the (monthly) stock return from time $t+h-1$ to $t+h$ by $r_{k,t+h}$, where h is the horizon over which the return is measured. The structural forms of the relationship between stock returns $r_{k,t+h}$, for each horizon h, and ESG performance are

$$\begin{aligned}
r_{k,t+1} &= \alpha_1 + \beta_1 \cdot Y_{k,t} + M_{1,k,t} + \epsilon_{1,k,t}, \\
r_{k,t+2} &= \alpha_2 + \beta_2 \cdot Y_{k,t} + M_{2,k,t} + \epsilon_{2,k,t}, \\
&\vdots \\
r_{k,t+h} &= \alpha_h + \beta_h \cdot Y_{k,t} + M_{h,k,t} + \epsilon_{h,k,t}.
\end{aligned}$$

Even though we use the same ESG performance for each of the different return horizons, we allow for the omitted variable to change with the horizon. The reduced forms of the above structural equations are

$$\begin{aligned}
r_{k,t+1} &= \alpha_1 + \beta_1 \cdot s_{k,t,i} + \nu_{1,k,t}, \\
r_{k,t+2} &= \alpha_2 + \beta_2 \cdot s_{k,t,i} + \nu_{2,k,t}, \\
&\vdots \\
r_{k,t+h} &= \alpha_h + \beta_h \cdot s_{k,t,i} + \nu_{h,k,t}.
\end{aligned}$$

The regressors are the same across specifications, and we allow the coefficients, the errors, and the omitted variable to have different variances and covariances. This implies that the covariance between the ESG score and the omitted variable is also likely to change.

The OLS and the IV estimates for any horizon h are given by

$$\beta_{OLS,h} = \left[\frac{var(Y_{k,t})}{var(Y_{k,t}) + var(\eta_{k,t,i})}\right]\left[\beta_h + \frac{cov(Y_{k,t}, M_{h,k,t})}{var(Y_{k,t})}\right], \tag{47}$$

$$\beta_{IV,h} = \left[\beta_h + \frac{cov(Y_{k,t}, M_{h,k,t})}{var(Y_{k,t})}\right]. \tag{48}$$

It is entirely possible that $\beta_{OLS,1} \neq \beta_{OLS,h}$ and that $\beta_{IV,1} \neq \beta_{IV,h}$. However, the ratios of the IV and the OLS coefficients (which are our estimates of the attenuation bias) are identical across all horizons.

$$\frac{\beta_{OLS,1}}{\beta_{IV,1}} = \frac{\beta_{OLS,2}}{\beta_{IV,2}} = \frac{\beta_{OLS,h}}{\beta_{IV,h}} = \left[\frac{var(Y_{k,t})}{var(Y_{k,t}) + var(\eta_{k,t,i})}\right]. \tag{49}$$

Of course, if there is a misspecification in the instruments, then the ratio will not remain the same across the entire range of our left-hand-side variables. In the empirical implementation (Section 4), we change the horizon over which the stock returns are measured and evaluate how stable the ratio of the IV and the OLS coefficients is.

A.3 Additional Figures

In this section, we present additional figures that show the density plots of the ESG ratings in our sample as well as figures pertaining to the implied noise ratios based on the regressions with accounting variables in Section 4.4 and to the estimations with only two instruments in Section 4.5.1.

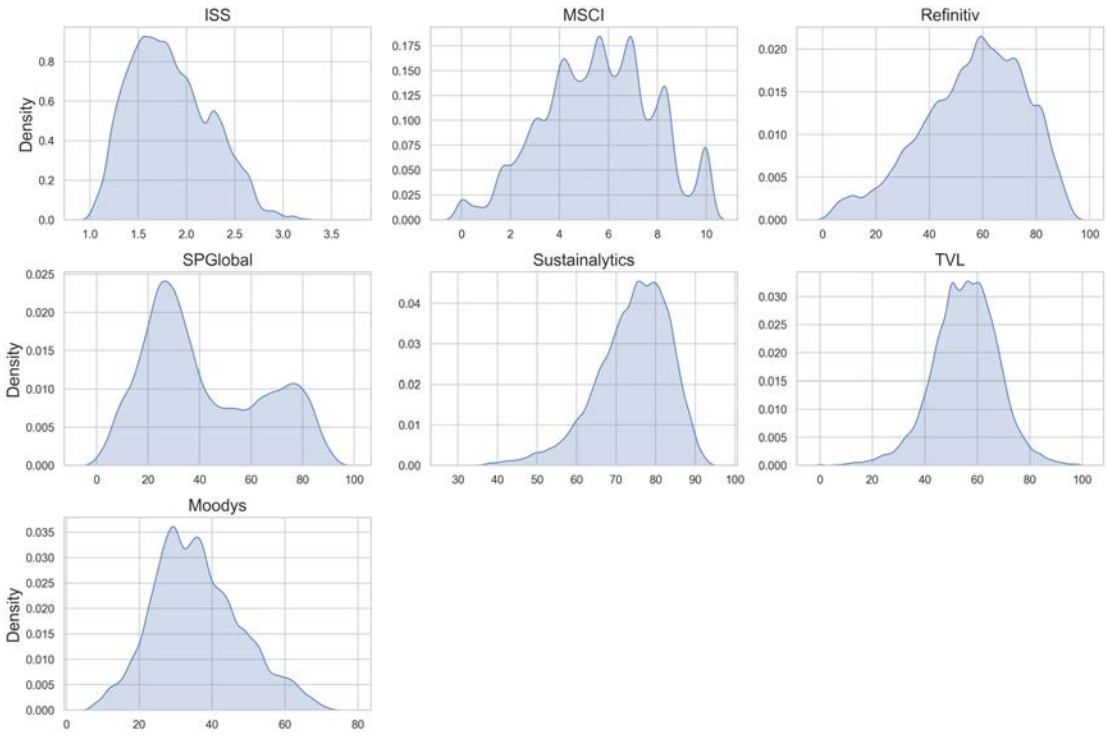


Figure A1. Distribution of ESG Scores. This figure shows the distribution of firms' ESG scores for each of the seven ESG rating agencies in our sample.

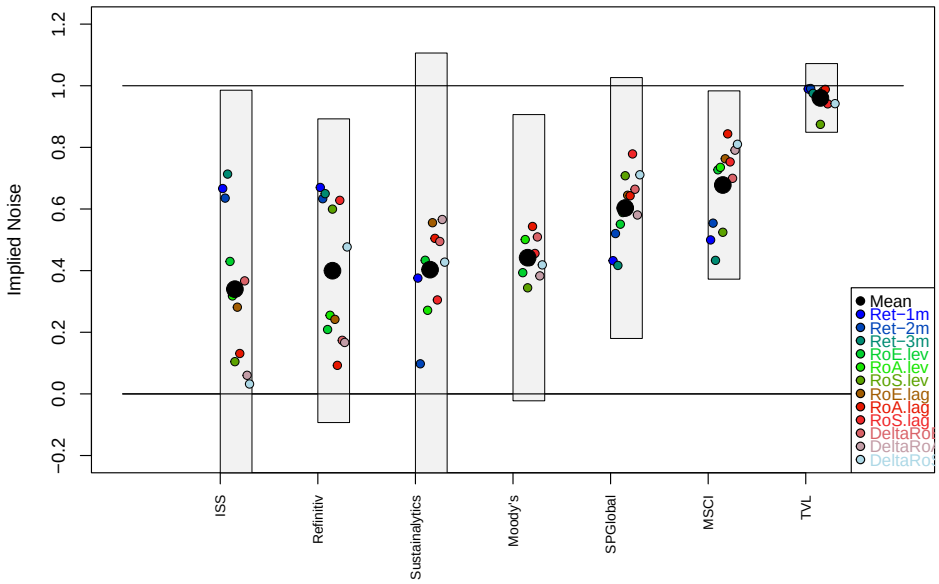


Figure A2. Implied noise in ESG ratings. This figure illustrates the implied noise for the regressions in which the outcome variables are accounting variables. The vertical axis measures the noise-to-signal ratio κ_i, defined in Equation (6). Raters are sorted by the median observation along the horizontal axis. The smaller dots represent the estimates of the noise-to-signal ratio κ_i for different raters, with the estimates computed from regressions with the outcome variables presented in the legend. As we use the ratios of the corresponding 2SLS and OLS coefficients in the estimation of each κ_i, we estimate the error bands for each ratio using the delta method. The large black dots are the means of these estimates for each rating agency and the rectangle is computed using the standard deviation of the mean. We use the simple assumption that the estimates are independent across the different specifications. The variance of the mean, therefore, is the mean of the individual estimated variances.

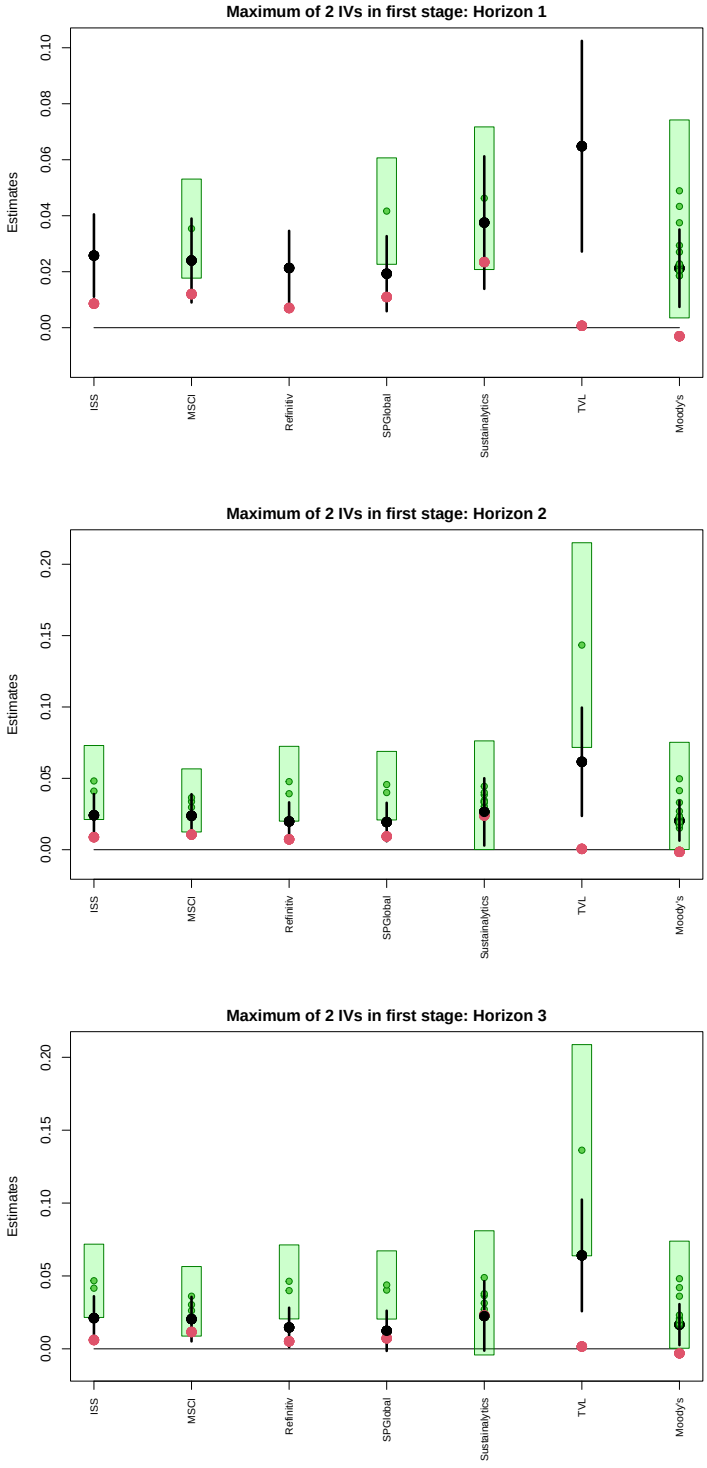


Figure A3. Coefficient estimates with two instruments. This figure compares the baseline stock return regression results from OLS (red dot) and 2SLS pruning (black dot) to the 2SLS estimates obtained with 2 instruments (green dots). The black bar is the 90% confidence interval of the 2SLS estimate, the green bar envelops the confidence intervals of the 2SLS estimates with two instruments. For estimates with two instruments, all combinations of instruments are attempted, but 2SLS estimates are only plotted when they pass the OIR test. The three graphs show results separately for stock return horizons of one, two, and three months.

A.4 Additional Tables

In this section, we provide additional tables clarifying names and ownership of included ESG raters, descriptive statistics, and an overall summary of estimates.

Table A1. ESG Scores Overview. This table shows the names of the ESG raters, previous names, ownership, and the exact name of the scores used in the analysis.

Rater Name	Previous Name	Owner	Score Name
ISS	Oekom Research	ISS Inc	Numeric ESG Overall Rating
Moody's	Vigeo-Eiris	Moody's	Global Score
MSCI	Innovest	MSCI Inc.	IVA Industry Weighted Score
Refinitiv	Asset4	London Stock Exchange Group	TRESG Score
Sustainalytics	–	Morningstar	ESG Risk Rating
SPGlobal	RobecoSAM	S&P Global	ESG Score
Truvalue Labs	–	FactSet	Insight Score

Table A2. Descriptive Statistics. This table shows the descriptive statistics. We multiplied Sustainalytics's ESG scores by -1 and added 100 so that a higher value corresponds to a better ESG performance for all ratings. *Return* is the monthly returns in percentage, *Beta* is the market beta estimated from monthly returns from month -60 to month -1, *Dividends* are the dividends per share over the prior 12 months divided by price at the end of the prior month, *Market Value* is the logarithm of the market value of equity at the end of the prior month, *Book-to-market* is the logarithm of book equity minus the logarithm of market value of equity at the end of the prior month, *Asset Growth* is the logarithm of growth in total assets in the prior fiscal year, *ROA* is the income before extraordinary items divided by average total assets in the prior fiscal year, *Momentum* is the return from month -12 to month -2, and *Volatility* is the monthly standard deviation, estimated from daily returns from month -12 to month -1. *Flows* indicates the ratio of inflows into ESG funds for a given region. The sample consists of 1297 firms and 70246 firm-month observations. All financial variables are expressed in U.S. dollars and are winsorized at the 1% level. *Mean* corresponds to the mean, *StDev* to the standard deviation, *Min* to the minimum, and *Max* to the maximum.

	Mean	StDev	Min	Max
ESG				
ISS	1.84	0.41	1.01	3.28
Moody's	35.50	11.73	7.00	76.00
MSCI	5.58	2.23	0.00	10.00
Refinitiv	56.49	18.86	1.58	92.97
Sustainalytics	73.98	9.46	29.47	94.00
SPGlobal	41.19	22.68	0.00	93.92
TVL	55.73	12.12	0.32	99.00
Environmental				
ISS	1.81	0.49	1.00	3.46
Moody's	33.96	16.78	0.00	85.00
MSCI	5.74	2.17	0.00	10.00
Refinitiv	53.59	26.95	0.00	98.53
SPGlobal	41.80	27.92	0.00	99.83
Social				
Moody's	32.42	12.78	7.00	84.00
MSCI	4.69	1.68	0.00	10.00
Refinitiv	57.00	22.70	0.68	98.64
SPGlobal	36.28	24.27	0.00	96.58
Governance				
ISS	2.28	0.53	1.00	3.66
Moody's	41.83	12.88	5.00	82.00
MSCI	5.30	1.66	0.00	10.00
Refinitiv	56.94	21.43	0.25	98.40
SPGlobal	45.18	20.96	0.00	96.35
Financial Variables				
Return	0.89	8.63	-25.13	28.06
Beta	1.00	0.44	0.09	2.31
Dividends	0.02	0.02	0.00	0.12
Market Value	9.33	1.17	6.57	12.34
Book-to-market	2.20	3.58	-2.48	8.82
Asset Growth	0.06	0.15	-0.23	0.94
ROA	0.05	0.05	-0.12	0.24
Momentum	0.04	0.25	-0.59	0.77
Volatility	0.07	0.03	0.03	0.19
Flows Euro	0.20	0.03	0.17	0.26
Flows U.S.	0.01	0.00	0.01	0.01
Flows U.K.	0.08	0.03	0.07	0.12
Flows Japan	0.01	0.03	0.00	0.05

Table A3. Pairwise correlations of ESG raters' scores. This table presents the pairwise correlations of all four subsamples: ESG, E, S, and G. We multiplied Sustainalytics' scores by -1 and added 100 so that a higher value corresponds to a better ESG performance for all ratings.

Environmental	ISS	Moody's	MSCI	Refinitiv	SPGlobal
ISS	1				
Moody's	0.72	1			
MSCI	0.26	0.34	1		
Refinitiv	0.63	0.68	0.22	1	
SPGlobal	0.67	0.72	0.30	0.70	1
Social	Moody's	MSCI	Refinitiv	SPGlobal	
Moody's	1				
MSCI	0.25	1			
Refinitiv	0.63	0.16	1		
SPGlobal	0.63	0.19	0.58	1	
Governance	ISS	Moody's	MSCI	Refinitiv	SPGlobal
ISS	1				
Moody's	0.62	1			
MSCI	0.36	0.48	1		
Refinitiv	0.28	0.34	0.19	1	
SPGlobal	0.31	0.43	0.15	0.32	1

A.5 Alternative accounting variables

In this section, we present additional results from the regressions on accounting variables described in Section 4.4.

Table A4. Return on Assets and ESG Ratings. This table reports estimates of β from the OLS regression (12) and the 2SLS regression (14). ESG ratings are observed in year $t-1$; RoA is observed in year t. In the second panel, the lagged variable RoA_{-1} is included as a regressor. In the third panel, the only change relative to the second panel is that the dependent variable is ΔRoA, which is $RoA_t - RoA_{t-1}$. The first set of columns shows OLS estimates. The second set of columns shows 2SLS estimates produced with the 2SLS pruning procedure as described in Section 3.3. The check marks in the columns titled "Coherent IVs" indicate the selection of instruments that pass the Sargan-Hansen OIR test. We cluster standard errors by month and firm. All reported coefficients and standard errors are multiplied by 100. *p<0.1; **p<0.05; ***p<0.01.

Model	Rater	OLS Coeffs	OLS StdErr		2SLS Pruning Coeffs	2SLS Pruning StdErr		Coherent IVs IS	MS	Re	SP	Su	TV	Mo	Ftest
RoA	ISS	0.007	0.001	***	0.010	0.001	***	1.5		✓	✗	✓	✗	✓	✓	2317
	MSCI	0.005	0.001	***	0.018	0.002	***	3.7	✓		✓	✗	✗	✗	✗	1247
	Refinitiv	0.007	0.001	***	0.010	0.001	***	1.3	✓	✓		✓	✗	✓	✗	2121
	SPGlobal	0.005	0.001	***	0.012	0.001	***	2.2	✓	✓	✓		✗	✓	✗	2067
	Sustainalytics	0.012	0.001	***	0.017	0.002	***	1.8	✗	✓	✗	✓		✓	✓	425
	TVL	0.001	0.001		-	-										
	Moody's	0.006	0.001	***	0.013	0.001	***	1.6	✓	✓	✓	✗	✗	✓		2998
RoA inc. RoA_{-1}	ISS	0.004	0.001	***	0.004	0.001	***	1.2		✓	✓	✓	✗	✓	✓	1568
	MSCI	0.001	0.001	***	0.009	0.001	***	6.4	✓		✓	✗	✓	✓	✗	595
	Refinitiv	0.003	0.001	***	0.004	0.001	***	1.1	✗	✓		✓	✗	✓	✓	1926
	SPGlobal	0.002	0.000	***	0.005	0.001	***	2.8	✓	✓	✓		✗	✓	✓	1513
	Sustainalytics	0.003	0.001	***	0.007	0.001	***	2.0	✗	✓	✗	✓		✓	✓	330
	TVL	0.000	0.000		0.010	0.003	***	84.7	✗	✓	✗	✓	✗		✓	93
	Moody's	0.003	0.001	***	0.006	0.001	***	2.2	✓	✓	✓	✗	✗	✓		2352
ΔRoA inc. RoA_{-1}	ISS	0.004	0.001	***	0.004	0.001	***	1.1		✓	✓	✓	✗	✓	✓	1568
	MSCI	0.001	0.001	**	0.007	0.001	***	4.8	✗		✓	✓	✗	✗	✓	682
	Refinitiv	0.003	0.001	***	0.004	0.001	***	1.2	✗	✓		✓	✗	✓	✓	1926
	SPGlobal	0.002	0.000	***	0.005	0.001	***	2.4	✗	✓	✓		✗	✓	✓	1720
	Sustainalytics	0.003	0.001	***	0.007	0.001	***	2.3	✗	✓	✓	✓		✓	✓	270
	TVL	0.000	0.000		0.010	0.003	***	-115.1	✗	✓	✗	✓	✗		✓	93
	Moody's	0.003	0.001	***	0.005	0.001	***	1.6	✗	✓	✓	✓	✗	✓		2236
Median								1.9								1568

Table A5. Return on Sales and ESG ratings. This table reports estimates of β from the OLS regression (12) and the 2SLS regression (14). ESG ratings are observed in year $t-1$; RoS is observed in year t. In the second panel, the lagged variable RoS_{-1} is included as a regressor. In the third panel, the only change relative to the second panel is that the dependent variable is ΔRoS, which is $RoS_t - RoS_{t-1}$. The first set of columns shows OLS estimates. The second set of columns shows 2SLS estimates produced with the 2SLS pruning procedure as described in Section 3.3. The check marks in the columns titled "Coherent IVs" indicate the selection of instruments that pass the Sargan-Hansen OIR test. We cluster standard errors by month and firm. All reported coefficients and standard errors are multiplied by 100. *p<0.1; **p<0.05; ***p<0.01.

Model	Rater	OLS Coeffs	OLS StdErr		2SLS Pruning Coeffs	2SLS Pruning StdErr		Coherent IVs IS	MS	Re	SP	Su	TV	Mo	Ftest
RoS	ISS	0.004	0.002	**	0.005	0.003	*	1.1		✓	✓	✓	✗	✓	✓	1999
	MSCI	0.004	0.002	**	0.008	0.004	**	2.1	✓		✓	✓	✗	✓	✓	620
	Refinitiv	0.002	0.002		0.005	0.002	**	2.5	✓	✓		✓	✗	✓	✓	2139
	SPGlobal	0.002	0.002		0.005	0.003	**	3.4	✓	✓	✓		✗	✓	✓	1949
	Sustainalytics	0.015	0.002	***	0.008	0.004	*	0.6	✓	✓	✓	✓		✓	✓	296
	TVL	0.002	0.001	*	0.019	0.012		8.0	✗	✗	✓	✓	✗		✓	62
	Moody's	0.003	0.002	*	0.005	0.003	*	1.5	✓	✓	✓	✓	✗	✓		2751
RoS inc. RoS_{-1}	ISS	0.004	0.001	***	0.003	0.002	*	0.8		✓	✓	✓	✓	✓	✓	1325
	MSCI	0.002	0.001		0.007	0.002	***	4.0	✓		✓	✓	✓	✓	✓	448
	Refinitiv	0.001	0.001		0.004	0.002	**	2.7	✓	✓		✓	✓	✓	✓	1409
	SPGlobal	0.001	0.001		0.004	0.002	**	4.5	✓	✓	✓		✓	✓	✓	1303
	Sustainalytics	0.004	0.001	***	0.005	0.003	*	1.4	✓	✓	✓	✓		✓	✓	241
	TVL	0.001	0.001		0.014	0.006	**	17.0	✓	✓	✓	✓	✗		✓	62
	Moody's	0.002	0.001	*	0.004	0.002	**	1.8	✓	✓	✓	✓	✓	✓		1839
ΔRoS inc. RoS_{-1}	ISS	0.005	0.001	***	0.005	0.002	***	1.0		✓	✓	✓	✓	✓	✓	1325
	MSCI	0.002	0.001		0.009	0.002	***	5.3	✓		✓	✓	✓	✓	✓	448
	Refinitiv	0.003	0.001	**	0.005	0.002	***	1.9	✓	✓		✓	✓	✓	✓	1409
	SPGlobal	0.002	0.001		0.006	0.002	***	3.5	✓	✓	✓		✓	✓	✓	1303
	Sustainalytics	0.004	0.001	***	0.007	0.003	**	1.7	✓	✓	✓	✓		✓	✓	241
	TVL	0.001	0.001		0.015	0.006	***	17.2	✓	✓	✓	✓	✗		✓	62
	Moody's	0.003	0.001	***	0.006	0.002	***	1.7	✓	✓	✓	✓	✓	✓		1839
Median								2.1								1303

A.6 Possible Sources of Noise in ESG Scores

In this section, we explore the implications of differences in measurement versus differences in aggregation for our procedure. 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 the following, we explain how to think about noise and 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):

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

where i indexes ESG rating agencies, a indexes attributes that the agency considers, $I_{a,t,i}$ is a measure of attribute a by rater i,[21] and $w_{a,i}$ are the weights.

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

$$Y_t = \sum_{a \in \{1,n\}} w_a^\star \cdot I_{a,t}^\star, \tag{51}$$

where $I_{a,t}^\star$ are the true values of the indicators and $w_a^\star$ 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 $I_{a,t}^\star$. As in our model, a rating agency does not observe these true values; it only observes their proxies. For each agency i, those proxies are the indicators $I_{a,t,i}$. 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

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

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 nature of this data), or industry estimates (such a procedure is typically used to estimate real estate emissions), or imputed data (e.g., a large fraction of emissions reported by Refinitiv is imputed data). For the weights, investors have preferences between labor treatment and emissions, represented by $(w_a^\star, 1 - w_a^\star)$. 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 i as follows:

$$s_{t,i} = Y_t + \underbrace{\sum_{a\in\{1,n\}} w_{a,i} \cdot \underbrace{(I_{a,t,i} - I_{a,t}^\star)}_{\eta_{I_{a,t,i}}} + \sum_{a\in\{1,n\}} \underbrace{(w_{a,i} - w_a^\star)}_{\eta_{w_{a,i}}} \cdot I_{a,t}^\star}_{\eta_{Y_{t,i}}}. \tag{52}$$

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

$$I_{a,t,i} = I_{a,t}^\star + \eta_{I_{a,t,i}}, \tag{53}$$

$$E[\eta_{I_{a,t,i}} | I_{a,t}^\star, w_a^\star] = 0, \tag{54}$$

and the discrepancy in the weights:

$$w_{a,i} = w_a^\star + \eta_{w_{a,i}}, \tag{55}$$

$$E[\eta_{w_{a,i}} | I_{a,t}^\star, w_a^\star] = 0. \tag{56}$$

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 (2). 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,

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

Condition (57) is satisfied if conditions (54) and (56) hold. We therefore now have two sufficient conditions, (54) and (56), that play the same role as our classical errors-in-variables assumption (7) in the main text.

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 each indicator (Equation (53)) and each weight (Equation (55)) are independent across any two rating agencies (as in the independence assumption (9) in the main text); and (ii) these errors are not correlated with the stock market returns (as in the exclusion restriction (8) in the main text).

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

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

where

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

The first term in (58) captures the weight and scope discrepancies highlighted in Berg, Kölbel, and Rigobon (2022). A weight discrepancy occurs when rating agencies assign different weights to the

same attribute and the a scope discrepancy occurs 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

$$s_{t,i} = Y_t + \sum_{a \in \{1,n\}} w_a^\star \cdot (I_{a,t,i} - I_{a,t}^\star),$$
$$s_{t,j} = Y_t + \sum_{a \in \{1,n\}} w_a^\star \cdot (I_{a,t,j} - I_{a,t}^\star).$$

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 (9)), i.e.,

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

This assumption is natural if the errors in the indicators are purely mistakes that are specific to individual rating agencies. There are circumstances in which it can be violated. For example, two rating agencies may use similar (possibly imputed) 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 the scores of another, which makes their errors correlated. Both of these violations would be detected by the

Sargan-Hansen OIR test (see Appendix A.7 for a detailed explanation of why correlated errors lead to a rejection of the OIR 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:

$$s_{t,i} = Y_t + \sum_{a \in \{1,n\}} (w_{a,i} - w_a^\star) \cdot I_{a,t}^\star,$$
$$s_{t,j} = Y_t + \sum_{a \in \{1,n\}} (w_{a,j} - w_a^\star) \cdot I_{a,t}^\star.$$

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 (9)), i.e.,

$$E\left[(w_{a,i} - w_a^\star) \cdot (w_{a,j} - w_a^\star)\right] = E\left[\eta_{w_{a,i}} \cdot \eta_{w_{a,j}}\right] = 0, \quad \forall i, j. \tag{60}$$

Again, this assumption is testable using the OIR test.

Most rating agencies not only measure individual attributes; they also, by providing a rating, reflect their preferences across those attributes. What matters the most to the rating agencies is reflected in the weights they use.This service from the rating agencies is important. Investors may not have a 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 $w_a^\star$'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 identifying assumptions presented in this section (or in Equations (7), (8), and (9)) 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 (2022) 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 (9) and/or (54) will be violated and the errors will not be classical.

The second potential problem is that two rating agencies have correlated errors. This can occur in a range of scenarios, such as rating agencies using similar data sources and procedures, or rating agencies relying on self-reported data that could have been manipulated by the firms. If the rating agencies impute indicators using similar data and similar models (as argued by Christensen, Serafeim, and Sikochi, 2022), this is likely to result in the violation of the independence assumption (equation (9) or (59)). This violation will be detected by the OIR test. Another possibility is when firms strategic behavior (greenwashing) will produce deviations that are common to the rating agencies. For example, if rating agencies rely on self-reported data, and a firm manipulates its announcements, then the rating agencies share the error. However, as long as one rating agency sees through the manipulation (or uses a methodology that does not rely on companies' disclosure), the OIR test will detect the correlated instruments.

Finally, our instrumental variable approach may fail because the measurements are correlated with the stock-return relevant cash-flow innovations (ϵ_t). Therefore, 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 will be detected by the OIR test.

A.7 What if Errors are Correlated Across Raters?

In this section, we show that if a rater's score is influenced by the scores of another rater, leading to a violation of (9), this is diagnosed by the OIR test. For concreteness, suppose that one rater

simply follows another, that is,

$$s_{k,t,1} = Y_{k,t} + \eta_{k,t,1},$$

$$s_{k,t,2} = s_{k,t,1} + u_{k,t},$$

where $u_{k,t}$ is an error term. The second rater's score can be expressed as

$$s_{k,t,2} = Y_{k,t} + \underbrace{\eta_{k,t,1} + u_{k,t}}_{=\eta_{k,t,2}}.$$

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

Recall our main structural equation: $r_{k,t+1} = a + \beta Y_{k,t} + M_{k,t} + \epsilon_{k,t}$, where $\epsilon_{k,t}$ is the error term. This equation is equivalent to

$$r_{k,t+1} = a + \beta(s_{k,t,1} - \eta_{k,t,1}) + \nu_{k,t} = a + \beta s_{k,t,1} + \underbrace{\nu_{k,t} - \beta\left(\eta_{k,t,2} - u_{k,t}\right)}_{\text{error}}$$

One can see immediately that s_2 is not a valid instrument for s_1 because it is correlated with the error term. The OIR 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 Sargan-Hansen test would fail in this case as well.