Estimating ESG scores separately:

How does noise in each pillar of ESG ratings affect the sensitivity of stock returns to ESG performance?

2053319

Supervisor: Dr Lory Barile

Department of Economics

The University of Warwick

Word Count: 5000

Abstract: ESG (environmental, social, and governance) ratings are noisy signals of ESG performance, causing standard regression estimates to suffer from attenuation bias. Aggregation of the individual pillars exacerbates this issue. To address this, I employ an instrumental variable approach, using the aggregate and pillar-specific scores of other rating agencies to disentangle noise from signal. My results imply that noise in ESG measures is substantial, making up more than 50% of the aggregate score, and over 80% in environmental ratings. Although this noise-correction procedure produces statistically significant coefficients for environmental ratings, social and governance ratings suffer from error correlation issues and a lack of data. This highlights the need for increased transparency in environmental data and greater disclosure on social and governance issues to draw robust conclusions on the impact of ESG on stock returns. [1]

[1] Acknowledgements: I would like to thank Dr Lory Barile for her support and supervision throughout this dissertation.

Contents

1 Introduction **1**

2 Literature Review **3**

3 Data **5**

3.1 ESG Ratings 5

3.2 Financial Data 6

4 Methodology **7**

4.1 Baseline Regression 7

4.2 IV Regression 7

4.3 Noise-to-Signal Ratios 8

5 Results **9**

5.1 Policy Implications and Suggestions 16

6 Concluding Remarks **18**

Appendix **23**

A Correlations Between ESG Raters **23**

A.1 Correlations for Aggregate Scores 23

A.2 Correlations for E Scores 24

A.3 Correlations for S Scores 24

A.4 Correlations for G Scores 24

B Rating Agencies **25**

C Controls **26**

D Descriptive Statistics **27**

E Instrument Validity **28**

1 Introduction

Less than two decades ago, the UN called for "better inclusion of environmental, social, and corporate governance (ESG) factors in investment decisions" (Agnew et al., 2022). Since then, ESG investing has become the fastest-growing segment of the asset management industry: assets in ESG funds reached \$2.5tn in 2022, a 290% increase from 2018 (Lex, 2023). With such exponential growth, ESG information is now considered financially material to investment performance, acting as a key driver of investors' decisions (Amel-Zadeh and Serafeim, 2017). However, investors are unlikely to have a detailed understanding of all ESG-related matters, nor the resources to achieve such understanding. Here, ESG rating agencies come into play: they provide third-party assessments of firms' ESG performance. By taking a weighted average of different ESG indicators, ESG ratings try to aggregate investors' corporate governance preferences to be consistent with the representative ESG-conscious investor (McCahery et al., 2016).

There is often much disagreement regarding firms' ESG ratings across different rating agencies (Berg et al., 2022). This lack of comparability is the primary impediment to the use of ESG information, giving us prima facie indications that ESG ratings are flawed as signals of firms' unobservable ESG performance (Amel-Zadeh and Serafeim, 2017). As such, noise in ESG scores is a first-order problem for regulators, asset managers, and investors alike (Berg et al., 2022). Thus, to uncover the true impact of ESG performance on expected stock returns, it is necessary to disentangle signal from noise in ESG ratings (Berg et al., 2022). This motivates the practical significance of my paper.

Some financial pundits believe that each pillar of the ESG grouping is made up of categories that do not belong together and hence, argue against the aggregation seen in ESG ratings (Populi, 2022). This motivates my original take on investigating the relationship between ESG performance and stock returns: to explore which pillar of ESG ratings contains the largest amount of noise, and how this impacts the sensitivity of stock returns when compared to the aggregated rating. I expect to find that the overall ESG rating has the highest noise-to-signal ratio compared to the individual pillars due to aggregation. Investigating these pillar-specific impacts is markedly significant for informing the discussions on the materiality of specific ESG indicators, and the need for greater, and more harmonised, reporting standards (Khan, Serafiem, and Yoon, 2016).

The results disprove my hypothesis as although the noise implied in ESG ratings is substantial, it makes up 52.51% of the total aggregate score and 85.61%, 98.99%, and 43.16% across the individual E, S, and G pillars. These results suffer from correlated measurement errors across ratings, affecting my ability to determine the noise-to-signal ratios (NSR) in the S and G pillars. Nonetheless, I find that the IV procedure works well to reduce the high level of noise observed in environmental ratings, producing significant results. Overall, this paper demonstrates the need for greater ESG data transparency and disclosure standards for firms and raters.

The rest of the paper is organised as follows: Chapter 2 considers the relevant but limited literature on ESG ratings and stock returns. Chapter 3 defines my sources of ESG and financial data. Chapter 4 describes my methodology, outlining my baseline OLS and 2SLS regression as well as how I calculate the corresponding NSR. Chapter 5 discusses my results and policy suggestions. Chapter 6 concludes the paper.

2 Literature Review

Understanding whether ESG investment affects firms' stock returns and leads to better environmental and social outcomes is still relatively new in economic and financial literature. Yet, existing studies report inconclusive results: studies report both higher stock returns (Edmans, 2011; Khan et al., 2016; Lins et al., 2017; Albuquerque et al., 2019) and lower stock returns for good ESG performers (Chava, 2014; El Ghoul et al., 2011; Bolton and Kacperczyk, 2020).

However, studies find that investors are discerning about ESG ratings. For example, Bernstein, Frydman and Hilt (2023) look at the effect of Moody's ratings on bond yields, finding that attaching ratings to bonds helps marginal investors resolve problems related to asymmetric information. These findings suggest investors' confidence in a firm's sustainability performance increases through ESG ratings. Baker, Egan and Sarkar (2022) examine whether demand for index funds is sensitive to companies' ESG ratings in the fund's underlying portfolio: their results indicate that investors are willing to pay an additional 33 basis points for a one unit increase in a fund's ESG score.

Although interest in ESG seems clear, Berg et al. (2022) argue that because ESG ratings diverge across raters, they obscure researchers' ability to distinguish a relationship between ESG performance and stock returns. Berg, Kölbel and Rigobon (2020) show that inconsistent measurement is the main source of divergence, followed by the choice of ESG attributes, and then aggregation of attributes. The observed divergence across rating agencies raises questions about the accuracy and validity of ESG ratings because they can impact the welfare of investors and firms if used to inform capital decision-making (Chatterji et al. 2016).

An attenuation bias also, in part, explains why the existing literature fails to make consistent conclusions on the relationship between ESG and stock performance (Berg et al., 2022). Aggregate ESG scores are noisy signals of the firm's true unobserved ESG performance, with each ESG attribute containing some noisy ESG signal (Berg et al., 2022). The authors demonstrate that the noisier the ESG signal, the lower its effects on stock prices; this dampening effect is similar to the attenuation bias observed in the OLS regressions used in the literature.

Berg et al. (2022) correct for attenuation bias by employing an instrumental variable (IV) approach: the authors instrument a given ESG rating agency's score with other rating agencies' scores for the same attributes. The authors use a noise-correction procedure to find that the coefficient estimates on the effect of ESG performance on stock returns are stronger than previously estimated: increasing by a factor of 2.6 on average, implying an average NSR of 61.7% across regions, raters, and horizons. Although this procedure effectively addresses issues of measurement error and delivers more robust conclusions, the authors use a weighted sum of indicators (an ESG score of rater i) as an instrument for another weighted sum of indicators (an ESG score of another rater $j{\neq}i$). Thus, these results still suffer from the estimation error introduced in data aggregation. To address this, I extend their model to look at ESG scores at a more granular level – for each E, S, and G pillar – which should see the estimated coefficients much closer to the true parameter. However, a separate instrument for each indicator would be notably more accurate, although such granularity in data is not yet available.

Some literature exists to evaluate the differences between E, S and G on stock returns. Pastor, Stambaugh and Taylor (2021a) find positive returns for high ESG stocks, attributing this to changes in climate concerns over the sample period. Chatterji et al. (2016) and Berg, Koelbel, and Rigobon's (2019) find more consistent results using environmental ratings, over social or governance ratings, across raters. However, the literature does not acknowledge the noise associated with these pillars as explanatory factors for these results. Furthermore, current literature often focuses on the impact of environmental issues, disregarding how social and governance issues may come to determine firms' stock returns. These factors motivate my contribution to the ESG literature.

3 Data

3.1 ESG Ratings

ESG (aggregate and pillar-specific) ratings for S&P500 companies are collected from five rating agencies: Bloomberg; MSCI; Refinitiv; RobecoSAM; and Sustainalytics. These ratings are sourced from the Eikon database (Refinitiv) and the Bloomberg Terminal (Bloomberg; MSCI; RobecoSAM; and Sustainalytics). MSCI's scores are numerically transformed from a letter grade scale to a scale of 0 to 6; Sustainalytics' scores are multiplied by -1 plus 100 to align it with the other ratings. Although more than 600 ESG raters operate in the US market (Deloitte, n.d.), these five agencies are chosen due to data accessibility and prevalence of use amongst corporates and investors (ERM, 2021). Given access to ESG data from other raters, this study could be expanded to include these rating agencies as well; however, doing so would arguably only further highlight the issue of noise in ESG ratings.

These ESG raters provide diverging ratings, as the pair-wise correlations in Appendix A demonstrate. This is likely due to the different methods of choosing, measuring, and aggregating ESG attributes employed by raters; raters may also use different data sources for their assessments. These sources differ in whether the information is: (i) public or private, (ii) self-reported or from a third-party viewer, (iii) mandatorily or voluntarily disclosed, or (iv) if it follows some disclosure standard. The limited amount of standardised and publicly available data about companies' ESG performance is a key challenge for raters (ERM, 2021). Further information regarding each rating agency is summarised in Appendix B.

The S&P 500 is a stock market index made up of the 500 largest American companies by market capitalisation. The companies listed on the S&P 500 are included and evaluated by prominent ESG rating agencies; hence, more ESG rating data are available. S&P 500 companies tend to have a higher degree of ESG disclosure: in 2022, 96% of S&P 500 companies published ESG reports in some form (GA, 2022). Yet, there also tends to be greater disagreement on ESG ratings for companies in the S& P500 (Gibson et al., 2021). Such remarks make S&P 500 companies relevant for this analysis. The data considers the ESG ratings of S&P 500 companies in the most recent fiscal year (2021-2022); this time frame reflects the frequency of changes in ESG ratings, which is mostly annual. Rating agencies often revise their procedures, and in some cases, back-fill their past scores based on these revisions (Berg et al., 2022). As such, analysis of previous years becomes less relevant to understanding the current problem of noise in ESG ratings, further motivating my choice of the most recent fiscal year.

3.2 Financial Data

Financial data for all SP 500 companies is sourced from the Eikon Database. All data is in USD. Year-to-date returns (*YTDReturns*) for the most recent fiscal year (2021-2022) are expressed in percentage points. Stock-level controls follow Lewellen (2015) and are described in Appendix C. Fox Corp and Alphabet Inc appear twice on the S&P as the companies issue two classes of stock: common stock and shareholder stock. As the common stock class is more commonly traded, I use the financial data related to this class of stock to represent both of these companies. Note that there is no difference in ESG data between the two classes of shares. Constellation Energy Corp (CEG.OQ) and GE HealthCare Technologies Inc (GEHC.OQ) do not have ESG ratings and are dropped from the data set. Bunge Ltd, Fair Isaac Corp, and Insulet Corp are also omitted, but because they are all new additions to the SP500 as of April 2023 and the data necessary is not yet available. This totals 499 observations. Descriptive statistics can be found in Appendix D.

4 Methodology

I estimate the OLS regressions of stock returns on separate and aggregate ESG ratings and contrast them to 2SLS regressions, which use the scores of other rating agencies as instruments, to capture the attenuation bias and calculate the NSR. This paper follows a modified version of the methodology used by Berg et al. (2022).

4.1 Baseline Regression

The baseline regression is

$$r_{k,t+1} = \alpha_p + \beta_p * ESGR_{k,t,i,p} + c_X * X_{k,t} + \nu_{k,t,p}, \tag{4.1}$$

where $ESGR_{k,t,i,p}$ denotes the ESG rating for firm k, by rater i, in year t, and for the aggregate or pillar-specific rating $p \in \{AGG, E, S, G\}$. I include stock-level controls $X_{k,t}$ as per Lewellen (2015): this also includes industry fixed effects. This specification does not use firm fixed effects because a firm's ESG rating often changes annually, which also informs why the time horizon of this study is only 12 months. Standard errors are clustered by GICS sub-industry (Bhojraj, 2003).

4.2 IV Regression

I assume that ESG rating agencies produce an imperfect measurement of true ESG performance, Y_p. In other words, I assume that ESG ratings measure ESG performance with noise. Therefore, the OLS estimate of the effect of ESG performance on stock returns β_{OLS_p} suffers from attenuation bias. To address this, I use the scores of other ESG rating agencies as instruments and include the same controls as in (4.1) in the first-stage regression:

$$ESGR_{k,t,i,p} = c_{0,p} + \pi * OESGR_{k,t,i,p} + c_1 * X_{k,t} + \omega_{k,t,p} \tag{4.2}$$

where $\mathrm{OESGR}_{k,t,i,p} := \{ESGR_{j,p} | \forall (j \neq i), ESGR_{j,p}$ is a valid instrument$\}$ are the other rating agencies' scores used as instruments. The Pruning and Lasso procedures are used to determine the valid set of instruments, see Appendix F for further discussion. I take $\hat{ESGR}_{k,t,i,p}$ as the fitted value from the estimation of Equation (4.2). The second stage regression is then as follows:

$$r_{k,t+1} = \alpha_p + \beta_p * \hat{ESGR}_{k,t,i,p} + c_X * X_{k,t} + \eta_{k,t,p} \tag{4.3}$$

To test the validity of my instruments, I conduct overidentifying restrictions (OIR) tests. I test multiple OIR as I have multiple ESG rating agencies in my sample. In my first procedure, which Berg et al. (2022) call Pruning, I 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: failure of this test indicates that some instruments are invalid. In cases where only one instrument is available, I use the Durbin (1954) and Wu–Hausman (Wu 1974; Hausman 1978) statistics to evaluate whether my instruments are exogenous. Lasso, my second IV procedure, starts with a minimal set of instruments and adds instruments until the OIR test is rejected. For both procedures, I find that most of the ESG ratings pass the OIR tests. This can happen if scores of one ESG rater are influenced by another (Berg et al., 2022). Another reason could be that measurement errors are correlated across rating agencies if agencies use similar procedures or rely on imputed data. The OIR tests diagnose these violations and I exclude scores that are invalid instruments from the estimation.

4.3 Noise-to-Signal Ratios

To assess the significance of the attenuation bias, we can take the ratio of the OLS estimates and 2SLS counterparts. Note that an omitted variable would bias both the OLS and IV estimates identically here and therefore cannot be problematic to the robustness of my results. The magnitude of the noise in the individual and aggregate ESG scores can then be estimated as

$$NSR_{i,p} \equiv 1 - \frac{\beta_{OLS_p}}{\beta_{IV_p}} = 1 - \frac{var(Y_p)}{var(Y_p) + var(\omega_{i,p})} = \frac{var(\omega_{i,p})}{var(Y_p) + var(\omega_{i,p})}, \tag{4.4}$$

where $NSR_{i,p}$ is the noise-to-signal ratio in rater i's aggregate or pillar-specific scores. In section 5, I compute the NSRs and compare them across the different pillars and raters.

5 Results

Tables 1-4 report the results. First, I discuss the OLS coefficients. My estimation reveals that 5 of 28 OLS coefficients (17.9%) are significant (at the 10, 5, or 1% level). The overwhelmingly insignificant results observed in the OLS estimation echo the inconclusive relationship often observed between ESG ratings and stock returns.

The most important finding is that the OLS estimator suffers from an attenuation bias. Out of the total 28 specifications considered (by varying rater, pillar, and 2SLS estimation procedure), 22 can be estimated. In other words, 22 possible estimates were found where the OIR was not rejected (78.6%), which guarantees a strong first stage. Of these 22 coefficients, 15 exhibit attenuation bias (68.2%). Of the seven estimates that do not, four (18.2%) result in smaller 2SLS coefficients relative to the corresponding OLS, and thre (13.6%) coefficients switch sign. Note that Equation 4.4 can only be used when attenuation bias exists in the OLS coefficient as it implies negative variances otherwise. Using the Pruning IV and Lasso IV procedures, 16 of 22 2SLS coefficients are larger than their OLS counterparts in absolute terms. Here, the average ratio between the 2SLS and the OLS coefficients is 4.09, that is, almost all coefficients double in size. Thus, for both Pruned and Lasso procedures, there are clear tendencies for 2SLS coefficients to be larger than the corresponding OLS estimates. Furthermore, five of the 2SLS coefficients are significant at the 10% or 5% level. The Lasso procedure confirms the results of the Pruning procedure; overall, both procedures find consistent results of attenuation bias, and suggest that it is consistent across alternative left-hand-side variables.

The results in Table 1-4 illustrate how this 2SLS approach adds value to noisy measures. Refinitiv and MSCI are two examples of raters that do not produce significant OLS coefficients in Table 1 (plus Tables 2-4 for Refinitiv), yet are seldom rejected as instruments. Meanwhile, the 2SLS coefficients of these raters increase substantially after instrumenting with other ratings, although mostly insignificant. The unique methodologies employed by these raters (see Appendix B) may be a related factor: although their differing procedures may add noise, they also add essential information when used in a 2SLS estimation. Thus, despite the low correlations shown in Appendix A, other raters' ESG scores are strong instruments for a given score. F-statistics in Tables 1-4 range from 23.19 to 46.51, demonstrating strong support for the relevance assumption for valid instruments.

However, there are several instances where one or more instrument is rejected by the Sargan-Hansen OIR test. Some instances see Pruned IV and Lasso IV leading to different conclusions as coefficients are sensitive to instrument selection. For example, in the case of RobecoSAM's Governance score, the OLS coefficient is 0.050, Pruned IV is 0.088, and Lasso IV is -0.064. However, none of these 2SLS coefficients are significant. Albeit this indicates violations of the assumptions outlined in Appendix 5, note that the OIR test can detect them. However, the OIR test cannot diagnose the cause of these rejections although I suspect that invalid instruments may be detected due to correlations in raters' measurement errors. This is specifically notable in the case of RobecoSAM, which is a valid instrument in only three of the possible 20 specifications across the aggregate and individual pillars. There are also cases where no instrument passes the Sargan-Hansen OIR test, preventing estimation of the relevant coefficients. This is especially problematic in the social pillar ratings, which either yield no relevant instrument or a 2SLS coefficient that switches sign. The governance pillar also suffers from the problem of sign-switching. Measurement errors could also be correlated across ratings here, as the data are limited given the lack of understanding in quantifying social and governance indicators.

The difference between the OLS and 2SLS coefficients for each rating agency calculates the implied noise and is derived by $NSR_{i,p}$, the noise-to-signal ratio, in equation (4.4), and reported in Table 5. The overall average noise-to-signal ratio for the two 2SLS estimation procedures is 52.51% across the five raters for ESG scores, and 85.61%, 98.99%, and 43.16% across the individual E, S, and G pillars respectively. Note that I can only calculate one noise-to-signal ratio in the S and G pillars. For the aggregated rating, there are marked differences between raters. RobecoSAM has the lowest noise-to-signal ratio (28.96%), whilst Refinitiv and MSCI have the highest noise-to-signal ratios (60.86% and 92.31%). This may be because the noise in these two ratings is truly orthogonal to that in other ratings, explaining their frequent acceptance as instruments. In the same vein, RobecoSAM's low SNR may indicate that its ratings are not orthogonal to that of other ratings as it is most frequently rejected as an instrument.

An interesting observation is that Refinitiv's NSR is similar for their ESG and E ratings, but distinctly different from their S ratings. This may imply that the attenuation bias is consistent for ESG and E due to similar estimation techniques, suggesting that Refinitiv rely heavily on data from the E component to make up their aggregate score. This can be verified by the correlation of 0.782 observed between the two scores. More generally, I find that not only do the coefficients for environmental score become larger in magnitude, but also become statistically significant after I correct for noise. My results suggest that the noise implied in the E measure is substantial with more than 80% of the total score; however, this also means there is a signal in environmental ratings.

These results disprove my hypothesis, which expected a higher noise-to-signal ratio in the overall ESG score as a result of aggregation. The higher noise-to-signal ratio in environmental ratings may be a result of the general push for greater corporate sustainability reporting, which has created a plethora of environmental data. Therefore, perhaps the vast number of indicators used to assess this dimension leads to redundancies and insignificance of certain attributes in explaining environmental performance. This is due to considerable differences in disclosure practices and reporting standards. These results align with the findings of Dunn, Fitzgibbons and Pomorski (2018), who hypothesise that because the data on environmental exposures is noisier than data on the other two pillars, it prevents more precise, statistically significant estimates. However, this goes against the belief of surveyed asset managers that E scores are more trustworthy than S or G, although the evidence seemingly suggests otherwise (Avramov et al., 2022).

Nonetheless, I caution against solely focusing on the E pillar of ESG given it contains the most noise: this recommendation simply reflects that the environmental dimension is currently more widely measured than the social or governance dimensions. The inability to instrument these ratings against one another does not suggest they are irrelevant; rather, it points out that these indicators are difficult to measure.

Table 1: Stock returns and agggregate ESG ratings. This table reports estimates of β_{AGG} from the OLS regression (4.1) and the 2SLS regression (4.3). The first set of columns shows OLS estimates, while the second and third set of columns show 2SLS coefficients from the Pruning and Lasso procedures, respectively. Reported coefficients and standard errors are multiplied by 100.The regressions are run for each rater; their names are reported in the left column. In the column titled "Valid IV", Y indicates the instruments that passes the Sargan-Hansen OIR test, N indicates those that do not. Standard errors are clustered by GICS sub-industry. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

	OLS			Pruned IV					Lasso IV				
Rater	Coeffs	StdErr		Coeffs	StdErr		Valid IV	Ftest	Coeffs	StdErr		Valid IV	Ftest
Refinitiv	0.062	0.054		0.158	0.198		- N Y Y N	42.09	0.158	0.198		- N Y Y N	42.09
RobeccoSAM	0.038	0.034		0.054	0.049		Y - N N Y	43.36	0.054	0.049		Y - N N Y	43.36
Bloomberg	1.687	0.630	***	2.480	1.528	*	Y N - Y Y	43.57	0.828	1.689		Y N - N Y	42.69
Sustainalytics	-0.222	0.102	**	-0.359	0.415		Y N Y - Y	31.68	-0.359	0.415		Y N Y - Y	31.68
MSCI	0.194	0.482		2.519	2.608		N N Y Y -	34.46	2.519	2.608		N N Y Y -	34.46

(5.1)

Table 2: Stock returns and E ratings. This table reports estimates of β_E from the OLS regression (4.1) and the 2SLS regression (4.3). The table structure follows that of Table 1. Standard errors are clustered by GICS sub-industry. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

	OLS			Pruned IV					Lasso IV				
Rater	Coeffs	StdErr		Coeffs	StdErr		Valid IV	Ftest	Coeffs	StdErr		Valid IV	Ftest
Refinitiv	0.052	0.035		0.283	0.116	**	- Y N	23.19	0.283	0.116	**	- N Y	23.19
RobeccoSAM	0.024	0.032		0.227	0.095	**	N - Y	36.87	0.227	0.095	**	N - Y	36.87
Bloomberg	1.005	0.409	**	0.963	0.086		Y Y -	36.72	0.963	0.086		Y Y -	36.72

(5.2)

Table 3: Stock returns and S ratings. This table reports estimates of β_S from the OLS regression (4.1) and the 2SLS regression (4.3). The table structure follows that of Table 1. Standard errors are clustered by GICS sub-industry. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

	OLS			Pruned IV				Lasso IV			
Rater	Coeffs	StdErr		Coeffs	StdErr	Valid IV	Ftest	Coeffs	StdErr	Valid IV	Ftest
Refinitiv	0.001	0.045		-	-	- N N	-	0.113	0.091	- Y N	43.27
RobeccoSAM	0.034	0.029		-0.004	0.047	Y - N	41.38	-	-	N - N	-
Bloomberg	0.534	0.309	*	-	-	N N -	-	-	-	N N -	-

(5.3)

Table 4: Stock returns and G ratings. Stock returns and S ratings. This table reports estimates of β_G from the OLS regression (4.1) and the 2SLS regression (4.3). The table structure follows that of Table 1. Standard errors are clustered by GICS sub-industry. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

	OLS			Pruned IV				Lasso IV			
Rater	Coeffs	StdErr		Coeffs	StdErr	Valid IV	Ftest	Coeffs	StdErr	Valid IV	Ftest
Refinitiv	0.073	0.032	**	-0.062	0.177	- N Y	36.42	-0.062	0.177	- N Y	36.42
RobeccoSAM	0.050	0.036		0.088	0.144	Y - Y	46.51	-0.064	0.157	N - Y	36.87
Bloomberg	-0.266	0.758		-	-	N N -	-	-	-	N N -	-

(5.4)

Table 5: Noise-to-Signal Ratios

	Rater	Pruned	Lasso
Aggregate (ESG)	Refinitiv	60.86	60.86
	RobeccoSAM	28.96	28.96
	Bloomberg	32.00	-
	Sustainalytics	38.19	38.19
	MSCI	92.31	92.31
E pillar	Refinitiv	81.58	81.58
	RobeccoSAM	89.65	89.65
	Bloomberg	-	-
S pillar	Refinitiv	-	98.99
	RobeccoSAM	-	-
	Bloomberg	-	-
G pillar	Refinitiv	-	-
	RobeccoSAM	43.16	-
	Bloomberg	-	-

(5.5)

Noise-to-signal ratios are described in per cent (%)

5.1 Policy Implications and Suggestions

Noise in ESG ratings has normative implications on policymaking. For instance, a prima facie suggestion would be to standardise how and what ESG ratings measure. Although this would lead to an improvement in the correlations between raters' scores, it would also likely result in less reliable information as ratings would not necessarily be accurate, despite being precise. Instead, regulators should aim to standardise disclosure. For example, if firms measure biodiversity impact differently, it becomes hard to interpret that data credibly: this issue exists across all three pillars. It is here that disclosure frameworks, like those proposed by the Global Reporting Initiative and the Sustainability Accounting Standards Board, become useful. If firms publish their ESG data against these frameworks, rating agencies no longer have to compete to collect this information privately; instead, divergence in ratings becomes solely a result of diverging opinions and interpretations, and not a question of data validity. Industry-specific frameworks are also likely to improve data validity as the financial materiality of different ESG issues varies across sectors (Amel-Zadeh and Serafeim, 2017; Pless, 2023).

Furthermore, it is important for regulators to improve transparency about their measurement and aggregation systems. This would encourage competition between raters to find best measurement practices and in turn, improve the signal strength of ESG ratings. Such competition may also result in lower costs to information gathering, further bolstering the incentive to be best-in-class. With regards to the individual pillars, a taxonomy about how the issues within E, S, and G are broken down should be developed so that the sub-scores, and the categories within, are comparable across rating agencies.

In application, noise in ESG ratings implies that sustainable investing is riskier, hence potentially reducing investor participation and economic welfare (Avramov et al., 2022). Thus, integrating the aforementioned practices is likely to mitigate current ESG uncertainty. Nonetheless, it is clear that producing ESG ratings is conceptually challenging as agencies need to assess a multitude of issues for each of the three pillars. Furthermore, although this investigation makes no claim to the causal effect of ESG ratings on stock performance, improving the noise-to-signal ratio through a clear taxonomy of ESG performance, and unified disclosure standards will improve robustness of such claims in future research.

6 Concluding Remarks

The construction of ESG ratings is not regulated, and the methodologies used can be opaque and proprietary: this hinders the ability to compare ESG data between agencies and companies (Mackintosh, 2018; Berg, Kölbel, Rigobon, 2020). My results demonstrate how this leads to the substantial issue of noise in ESG ratings as signals of ESG performance. However, I caution against only using scores from ESG raters with the most precise measurement. This is because disregarding other raters' scores would mean disregarding valuable information about the unobservable ESG performance these ratings aim to measure. This informs the relevance of an IV procedure to disentangle signal from noise used in this paper. My paper enriches academic and policy discussions to the extent that noise decreases with a better understanding of a firm's true ESG performance. Additionally, my contribution to assessing the individual pillars of ESG further highlights the need for increased transparency of environmental data, and greater disclosure on social and governance issues.

Whilst existing literature often focuses on ESG in a single-period environment, it would be worthwhile to extend ESG rating analysis to a dynamic setting; in a recent study, Berg, Heeb, and Kölbel (2023) find that stock returns response to ESG ratings is slow, unfolding gradually of up to two years. However, to investigate noise, such an extension would be uninformative until raters adopt uniform disclosure and reporting standards. Investigating further stocks and stock indices is an obvious next step for future research. Furthermore, greater access to ESG ratings from other rating providers would likely increase the significance, robustness and precision of my results as more instruments could be employed for noise reduction. All withstanding, the ultimate aim is to see future research on this issue disappear with signal strength for ESG performance overwhelming potential issues of attenuation bias.

References

Agnew, H., Klasa, A., Mundy, S. (2022). How ESG investing came to a reckoning. Financial Times. Available at: https://www.ft.com/content/5ec1dfcf-eea3-42af-aea2-19d739ef8a55

Albuquerque, R., Koskinen, Y., Zhang, C. (2019). Corporate Social Responsibility and Firm Risk: Theory and Empirical Evidence. Management Science, 65(10), 4451-4469.

Amel-Zadeh, A. Serafeim, G. (2017). Why and How Investors Use ESG Information: Evidence from a Global Survey (Working Paper No. 17-079) Harvard Business School.

Avramov, D., Cheng, C., Lioui, A. Tarelli, A. (2022) 'Sustainable investing with ESG rating uncertainty', Journal of Financial Economics, 145(2), pp. 642-664.

Baker, M., Egan, M. L., Sarkar, S. K. (2022). 'How Do Investors Value ESG?' (Working Paper No. 30708). National Bureau of Economic Research).

Bernstein, A., Frydman, C., and Hilt, E. (2023) 'The Value of Ratings: Evidence from their Introduction in Securities Markets'. Working Paper No. 31064). National Bureau of Economic Research.

Berg, F., Koelbel, J. F., Rigobon, R. (2020). Aggregate confusion: The divergence of ESG ratings. Working Paper. MIT Sloan.

Berg, F., Koelbel, J. F., Pavlova, A., Rigobon, R. (2022). ESG confusion and stock returns: Tackling the problem of noise. Working Paper. National Bureau of Economic Research.

Berg, F., Heeb, F., Kölbel, J. (2022). The Economic Impact of ESG Ratings.

Bhojraj, S., Lee, C.M.C. and Oler, D.K. (2003), What's My Line?

A Comparison of Industry Classification Schemes for Capital Market Research. Journal of Accounting Research, 41: 745-774.

Bloomberg. (n.d.). Bloomberg Launches Proprietary ESG Scores. Available at: https://www.bloomberg.com/company/press/bloomberg-launches-proprietary-esg-scores/

Bolton, P., Kacperczyk, M. T. (2020). Carbon premium around the world. Working Paper. Imperial College.

Chatterji, A.K., Durand, R., Levine, D.I. and Touboul, S. (2016), Do ratings of firms converge? Implications for managers, investors and strategy researchers. Strat. Mgmt. J., 37: 1597-1614.

Chava, S. (2014). Environmental externalities and cost of capital. Management Science, 60, 2223-2247.

Deloitte. (n.d.). ESG Ratings: Do They Add Value? Available at: https://www2.deloitte.com/ce/en/pages/about-deloitte/articles/esg-ratings-do-they-add-value.html

Dunn, J., Fitzgibbons, S., Pomorski, L. (2018). Assessing risk through environmental, social, and governance exposures. Journal of Investment Management, 16(1), 4-17.

Durbin, J. (1954). Errors in Variables. Revue de l'Institut International de Statistique / Review of the International Statistical Institute, 22(1/3), 23–32.

Edmans, A. (2011). Does the stock market fully value intangibles? Employee satisfaction and equity prices. Journal of Financial Economics, 101, 621-640.

El Ghoul, S., Guedhami, O. , Kwok C., Mishra, D. (2011). Does corporate social responsibility affect the cost of capital? Journal of Banking and Finance, 35, 2388-2406.

ERM. (2021). New ERM report ranks ESG ratings agencies and urges action to maintain business and investor trust. ERM. Available at: https://www.erm.com/news/new-erm-report-ranks-esg-ratings-agencies-and-urges-action-to-maintain-business-and-investor-trust/

GA Institute. (2022). Sustainability Reporting In Focus. Available at: https://www.ga-institute.com/research/ga-research-directory/sustainability-reporting-trends/2022-sustainability-reporting-in-focus.html

Gibson, R., and Krueger, P. and Schmidt, P. (2019). ESG Rating Disagreement and Stock Returns. Swiss Finance Institute Research Paper No. 19-67, European Corporate Governance Institute – Finance Working Paper No. 651/2020, Financial Analyst Journal.

Hausman, J. A. (1978). Specification Tests in Econometrics. Econometrica, 46(6), 1251–1271.

Khan, M., Serafeim, G., Yoon, A. (2016). Corporate sustainability: First evidence on materiality. The Accounting Review, 91, 1697-1724.

Lewellen, J. (2015). The cross-section of expected stock returns. Critical Finance Review, 4, 1-44.

Lins, K. V., Servaes, H., Tamayo, A. M. (2017). Social capital, trust, and firm performance: The value of corporate social responsibility during the financial crisis. Journal of Finance, 27, 1785-1824.

Mackintosh, J., 2018. Is Tesla or Exxon more sustainable? It depends whom you ask. Wall Street Journal, 17. Available at https://on.wsj.com/2MQCC4m

McCahery, J. A., Sautner, Z., Starks, L. T. (2016). Behind the scenes: The corporate governance preferences of institutional investors. Journal of Finance, 71, 2905-2932.

MSCI ESG Research LLC. (2020). MSCI ESG Fundamentals Calculation Methodology. Available at: https://www.msci.com/documents/-

10199/1283513/MSCIESGMetricsCalcMethodologyDec2020.pdf/92a299cb-0dbc-63ba-debb-e821bd2e2b08

Pastor, L., Stambaugh, R. F., and Taylor, L. A. (2021). Dissecting green returns. Working Paper, National Bureau of Economic Research.

Pástor, Ľ., Stambaugh, R. F., Taylor, L. A. (2021). Sustainable investing in equilibrium. Journal of Financial Economics, 142(2), 550-571.

Pless, J. (2023). To Starve or to Stoke? Understanding Whether Divestment vs. Investment Can Steer (Green) Innovation (Working Paper No. 30942). National Bureau of Economic Research.

Populi, L. (2022). Lex Populi: Pets at Home unleashes diversification. Financial Times. Available at: https://www.ft.com/content/d2f4475a-e4fc-4834-8cb6-29efdec6fc36.

Refinitiv. (n.d.). ESG Scores. Available at: https://www.refinitiv.com/en/sustainable-finance/esg-scores

Robeco. (n.d.). ESG Integration. Available at: https://www.robeco.com/en-int/sustainable-investing/journey/esg-integration

Sustainalytics. (n.d.). ESG Risk Ratings Methodology. Available at: https://connect.sustainalytics.com/esg-risk-ratings-methodology

Wu, D.-M. (1974). Alternative Tests of Independence between Stochastic Regressors and Disturbances: Finite Sample Results. Econometrica, 42(3), 529–546. https://doi.org/10.2307/1911789

Appendix

A Correlations Between ESG Raters

The average pairwise correlations are 0.47, 0.27, and 0.24 for the corresponding E, S, and G pillars; the environmental pillar is the only pillar with a greater average correlation than the aggregate rating (0.29). The variation in correlations between the three pillars and against the composite score is likely due to the discrepancies in aggregation and weighting procedures across raters. The higher correlations observed in the environmental pillar can be explained by the increasingly ability for environmental issues to be measured and quantified. There are more systematic, regulation-driven, firm-level attempts to quantify environmental indicators. For instance, all SP500 companies report on their greenhouse gas (GHG) emissions, water, and electricity use as these dimensions are recognised as important to a firm's environmental performance. There is also at least basic guidance about how to measure these environmental indicators (see, for instance, the Greenhouse Gas Protocol). In contrast, the criteria applied to governance and social ratings may likely differ among rating providers as they require more value judgments. For example, there is little agreement on what the most important social and governance issues are, as well as a little understanding of how to quantify these issues. Thus, S and G scores are inherently more subjective than E scores, suggesting more disagreement (i.e., lower correlations) among rating agencies.

A.1 Correlations for Aggregate Scores

	Refinitiv	RobecoSAM.	Bloomberg	Sustainalytics	MSCI
Refinitiv	1.00				
RobecoSAM	0.57	1.00			
Bloomberg	0.37	0.39	1.00		
Sustainalytics	0.12	0.22	-0.06	1.00	
MSCI	0.36	0.39	0.29	0.22	1.00

Table 6: Pair-wise correlations for aggregate ESG scores

A.2 Correlations for E Scores

	Refinitiv	RobecoSAM	Bloomberg
Refinitiv	1.00		
RobecoSAM	0.59	1.00	
Bloomberg	0.38	0.43	1.0000

Table 7: Pair-wise correlations for environmental (E) scores

A.3 Correlations for S Scores

	Refinitiv	RobecoSAM	Bloomberg
Refinitiv	1.00		
RobecoSAM	0.51	1.00	
Bloomberg	0.11	0.18	1.00

Table 8: Pair-wise correlations for social (S) scores

A.4 Correlations for G Scores

	Refinitiv	RobecoSAM	Bloomberg
Refinitiv	1.00		
RobecoSAM	0.17	1.00	
Bloomberg	0.29	0.26	1.00

Table 9: Pair-wise correlations for governance (G) scores

B Rating Agencies

I use the ESG scores from raters Bloomberg, MSCI, Refinitiv, RobecoSAM, and Sustainalytics in this paper. These rating agencies mainly use data from five sources: (i) companies' own ESG reports, (ii) regulatory filings, (iii) the media, (iv) questionnaires rating agencies send to companies, and (v) modelled data.

Bloomberg Covers 3,700+ ESG data fields including from third-party providers. Rating takes a numerical score out of 100. ESG ratings are updated daily.

MSCI Examines 35 key issues across the three pillars and ten overarching themes. Rating scale is from AAA to CCC. Only considers publicly available data. Does not publish individual pillar ratings. Governance issues make up a third of the aggregate rating; environmental and social issues have varying weights in the overall ESG score. ESG ratings are updated weekly and a company-level review is performed annually.

Refinitiv Subsidiary of the London Stock Exchange. Employs 630+ measures to evaluate over 10 categories, with a separate score assessing 23 controversy topics. Publishes both aggregate and individual scores, as well as a ESG controversies score. Rating takes a numerical score out of 100. ESG ratings are updated weekly.

RobecoSAM Bought and used by S&P. Relies mostly on private information gathered from questionnaires: this information is scored and weighed into E, S, and G standalone scores as well as overall ESG score. Rating takes a numerical score out of 100. Performs a daily assessment of ESG controversies and a company-level review is performed annually.

Sustainalytics Owned by Morningstar. The overall ESG rating is a sum of three main areas: corporate governance, material ESG issues, and idiosyncratic ESG issues. Does not publish individual pillar ratings. Ratings take a numerical score out of 100 and are industry-based. ESG ratings are updated annually.

C Controls

Control	Definition
Beta	Market beta estimated by monthly returns from t_{-60} to t_{-1}
Dividends	Dividends per share over t_{-12} divided by price at the end of t_{-1}
Market Value	Logarithm of market value of equity at the end of t_{-1}
Book-to-market	Logarithm of book equity minus the logarithm of market value of equity at t_{-1}
Asset Growth	Logarithm of growth in total assets in the last fiscal year
ROA	Income before extraordinary items divided by average total assets in the last fiscal year
Momentum	Return from t_{-12} to t_{-2}
Volatility	Monthly standard deviation, estimated from daily returns from t_{-12} to t_{-1}

Table 10: Definitions for stock-level controls, where t represents months

D Descriptive Statistics

	Mean	StDev
Refinitiv (ESG)	68.66	12.64
Refinitiv (E)	62.23	21.46
Refinitiv (S)	71.20	15.27
Refinitiv (G)	69.07	16.04
RobecoSAM (ESG)	66.14	21.69
RobecoSAM (E)	64.62	22.88
RobecoSAM (S)	59.77	25.92
RobecoSAM (G)	69.78	19.07
Bloomberg (ESG)	4.48	1.27
Bloomberg (E)	3.71	2.10
Bloomberg (S)	3.66	2.04
Bloomberg (G)	7.24	0.75
Sustainalytics (ESG)	78.52	7.27
MSCI (ESG)	4.00	1.14
Year to Date Returns	0.04	0.15
Beta	1.06	0.43
Dividends	0.26	0.17
Market Value	24.26	1.05
Book-to-market	1.34	1.07
Asset Growth	0.24	1.85
ROA	0.09	0.08
Volatility	0.10	0.04

E Instrument Validity

Assume that (i) ESG ratings are related to each other only through $Y_{k,t,p}$, and (ii) that their measurement error is white noise.

To determine the valid set of instruments, we need to fulfil the relevance and independence conditions. The relevance condition outlines that instruments are correlated with the regressor, which is easy to defend given that ESG rating scores of different agencies are positively correlated as seen in Appendix A. In Tables 1-4, I show that I do not have a problem of weak instruments through F-statistics. The independence condition assumes that all errors $_{k,t,i,p}$ are independent across rating agencies. This assumption is found to be the most probable threat to my IV estimation as noise in ESG scores is likely to be correlated across raters.

See Berg at al. (2022) for a theoretical and empirical discussion that measurement errors are truly additive and orthogonal to true ESG performance Y_k, t, p, as in the classical errors-in-variables problem.