OPEN ACCESS
Journal of Financial Management, Markets and Institutions
(2021) 2150006 (21 pages)

DOI: 10.1142/S2282717X21500067

THE DIVERGENCE OF ESG RATINGS: AN ANALYSIS OF ITALIAN LISTED COMPANIES

VINCENZO CAPIZZI

Dipartimento di Studi per l'Economia e l'Impresa
Università degli Studi del Piemonte Orientale
"Amedeo Avogadro"
Via E. Perrone 18, 28100 Novara, Italy
vincenzo.capizzi@uniupo.it

ELEONORA GIOIA*, GIANCARLO GIUDICI† and FRANCESCA TENCA‡

School of Management, Politecnico di Milano
Via Lambruschini 4, 20156 Milano, Italy
**eleonora.gioia@mail.polimi.it*
†*giancarlo.giudici@polimi.it*
‡*francesca.tenca@polimi.it*

Received 27 September 2021
Accepted 30 September 2021
Published 12 November 2021

The increasing attention to sustainability issues in finance has brought a proliferation of environmental, social, and governance (ESG) metrics and rating providers that results in divergences among the ESG ratings. Based on a sample of Italian listed firms, this paper investigates these divergences through a framework that decomposes ESG ratings into a value and a weight component at the pillar (i.e. E, S, and G) and category (i.e. sub-pillar) levels. We find that weights divergence and social and governance indicators are the main drivers of rating divergences. The research contributes to develop a new tool for analyzing ESG divergences and provides a number of recommendations for researchers and practitioners, stressing the need to understand what is really measured by the ESG rating agencies and the need for standardization and transparency of ESG measurement to favor a more homogeneous set of indicators.

Keywords: Environmental social governance; rating; divergence; ESG performance.

JEL Classification: G20, G24, G28

‡Corresponding author.

1. Introduction

Sustainable Responsible Investment (SRI) has grown significantly in recent years, reaching $35.3 trillion in assets under management at the start of 2020, representing an increase of 15% in the last two years and 55% in the last four years and accounting for 36% of total assets under management (GSIA 2020). The United States and, especially, Europe are the leading markets, representing more than 80% of global sustainable investing (GSIA 2020). Environmental, social, and governance (ESG) issues have been more and more included into investors' portfolio selection strategies, through the use of positive or negative screening criteria or explicitly factoring ESG principles into their financial choices (Cellier *et al.* 2016). Thus, a key requirement of SRI is the access to high-quality sustainability-related data, ratings, and methodologies (European Commission 2020). Nevertheless, the ESG rating industry is largely characterized by the lack of transparency (Chatterji *et al.* 2009), clear disclosure rules, and rating standardization (Escrig-Olmedo *et al.* 2014, Bolognesi & Burchi 2021, Conca *et al.* 2021), contrary to financial reporting and credit ratings.

Following this increasing attention to sustainability and social issues, a proliferation of rating providers has characterized the last decades with the development of different SRI products and services, including raw sustainability data, ratings and rankings, indices and benchmarks, consulting services, and reporting practices. Simultaneously, the ESG industry has seen a trend of consolidation with a few big players that, through a series of mergers and acquisitions, currently dominate the ESG rating market (Avetisyan & Hockerts 2017). MSCI, Refinitiv, Bloomberg, Sustainalytics (recently purchased by Morningstar), S&P Global, VigeoEiris, and Inrate are some of the most prominent firms, which are based in Europe and in the United States. Given the increasing relevance of sustainable investment, the wide number of players with proprietary methodologies, and the lack of strict rules to be followed, the inevitable consequence is a variety of ESG ratings and measurement qualities that differ in their dimensionality, reliability, and construct validity (Widyawati 2021). Furthermore, ESG is a concept based on continuously changing indicators and often qualitative information (Paltrinieri *et al.* 2021).

From the previous considerations, it is evident that we need to better understand whether and how the ESG ratings provided by different rating agencies differ and, above all, the underlying drivers of this divergence in order to effectively advise investors and companies about ESG performance. So far, researchers investigating ESG ratings have found mixed results and have mainly looked at aggregated metrics. The convergent validity of the same environmental indicator has been found to have a degree of convergence by some studies (Semenova & Hassel 2015) and a degree of divergence by others (Hedesström *et al.* 2011). Conversely, an assessment of aggregated ESG data reveals a low level of convergence among different ratings (Dorfleitner *et al.* 2015, Chatterji *et al.* 2016). Furthermore, most prior studies on ESG measurement provide an in-depth, but incomplete examination of validity and reliability issues, with research on MSCI (previously MSCI KLD) ratings dominating

the field (e.g. Mattingly & Berman 2006, Chatterji *et al.* 2009, Delmas & Blass 2010, Kang 2015, Mattingly 2017), with only a few studies comparing ESG ratings of different providers (Berg *et al.* 2020, Widyawati 2021).

The aim of this paper is to examine the differences in ESG ratings across a broader sample of rating agencies and understand the sources of divergences. Only a few papers have established a quantitative framework for assessing those disparities. The study most similar to ours is Berg *et al.* (2020), which identifies three causes of ESG rating divergences (i.e. scope, measurement, and weights).

Using the ESG ratings of Italian companies issued by six leading international ESG rating agencies, i.e. MSCI, Refinitiv, S&P Global, Inrate, Arabesque S-Ray (hereafter Arabesque), and Truvalue Labs (FactSet) (hereafter Truvalue), we develop a quantitative framework to study the ESG rating divergences. The framework adopts a top–down approach, starting from the overall ESG scores and examining the contribution of each pillar [Environmental (E), Social (S), and Governance (G)]. Divergences are further investigated by decomposing the score into value and weight components of aggregate ESG ratings and at the pillar and sub-pillar levels (i.e. for different categories).

Our main results can be summarized as follows. First, the weight component, in most of the cases, explains the highest percentages of disagreement across ESG ratings of different agencies. Second, considering the pillar level, the divergences related to the Environmental pillar are the lowest ones. On the other hand, the Social and Governance pillars explain a higher percentage of divergences. Third, going deeper at the category level, the analysis reveals that the level of divergences is mainly due to the weight component and it is more relevant for the Governance pillar categories.

We contribute to the emerging research that has documented the divergence of ESG ratings (Windolph 2011, Chatterji *et al.* 2016, Brandon *et al.* 2019, Berg *et al.* 2020, Widyawati 2021). Our main achievement is to explain why ESG ratings diverge by contrasting the underlying methodologies through a replicable framework and quantifying the main sources of divergence. Our research also provides important empirical foundation for future studies that aim to investigate the relationship between ESG performance, rating disagreement, and stock price performance.

The paper is organized as follows. Section 2 presents the sample used for the analyses and documents some preliminary evidences on the divergence of aggregated ESG rating. Section 3 develops the quantitative framework and illustrates the findings of ESG rating divergences in terms of value and weight components, documenting the discrepancies at the pillar and category levels. Finally, Sec. 4 concludes, highlighting the implications of our research as well as future research directions.

2. Methods

2.1. *Research design and sample*

ESG ratings initially appeared in the 1980s as a means for investors to evaluate firms based on factors other than solely financial performance, such as the social and

environmental performances. Since then, the increasing focus on ESG investing has led to the rise in the number and influence of ESG rating agencies (Lopez *et al.* 2020). According to Li & Polychronopoulos (2020), there are more than 70 different firms around the globe that provide some sort of ESG scoring data. However, each ESG rating agency has developed a proprietary methodology with specific steps followed in the assessment of rated firms. As a result of this variety of approaches, ESG ratings typically are conflicting and are often not comparable due to discrepancies in definitions and evaluations of ESG constructs. In recent years, the ESG industry has also seen a consolidation tendency that, however, had less to do with best practices and more to do with the strategy of increasing market shares through mergers and acquisitions (Dimmelmeier 2020).

We use data from six different ESG rating providers: MSCI, Refinitiv, S&P Global, Arabesque, Truvalue Labs, and Inrate.[a] Together, these rating agencies are major players in the ESG rating space (Eccles & Stroehle 2018) and cover a substantial part of the overall market for ESG ratings. Overall ESG scores, single-pillar (Environmental, Social, and Governmental) scores, and category scores (comprehensive of values and weights for each pillar and category) were retrieved from public sources and proprietary databases for a sample of 210 Italian firms listed on the stock exchange in the years 2019 (188 firms) and 2020 (182 firms).[b] ESG rating score is the general judgment assigned to a company's ESG performance. Pillar scores are assigned to each pillar namely E, S, and G; for instance, the E score refers only to a company's environmental performance and is the summary of the performance of different categories, such as pollution, energy consumption, GHG emissions, etc.

In order to operationalize our model, we had to introduce some adjustments to the raw data. First, we consider the average of the ratings issued during the year when available (i.e. Arabesque, MSCI, and Truvalue) or the rating scores provided at the end of the year when only these figures are available (i.e. Refinitiv, Inrate). Second, different rating scores assigned by different agencies are translated into a homogeneous scale ranging from 0 to 100. Then, specific assumptions are applied to the providers. In detail, Inrate does not disclose single-score weights. Thus, we estimate them through multiple linear regression models applied to all firms covered worldwide by Inrate analysts, controlling for different industrial sectors. MSCI provides two scores: MSCI Industry Adjusted ratings and MSCI Weighted ratings. The latter computes the weighted average of pillars and respective weights; therefore, it is the only one included in the analyses at the pillar and category levels. Finally, Truvalue does not provide the segmentation in pillar scores, but only scores of the different categories, which were therefore grouped into the three main pillars of E, S, and G. More in detail, Truvalue uses Dynamic Materiality percentages; this methodology consists in tracking company data (i.e. number of news) tagged to a specific category over the last 12 months (Truvalue Labs 2020). We use these

[a]We were not able to obtain access to granular data reported by Sustainalytics, which is also a major player in the arena. We thank Arabesque and Inrate for disclosing relevant information for our work.
[b]Some firms are added between 2019 and 2020, while others are no longer evaluated by agencies.

Table 1. Descriptive statistics of the aggregate ESG rating in 2020 for the five rating agencies.

	Arabesque	Refinitiv	Inrate	Truvalue	MSCI Industry Adjusted	MSCI Weighted	S&P Global
Panel A: Full Sample							
No. of firms	107	86	36	75	147	147	101
Mean	52.02	60.11	46.30	57.45	52.17	48.90	35.45
Median	52.12	62.21	54.17	57.57	54	48	25
Standard deviation	7.26	18.51	25.62	9.43	19.41	7.92	24.56
Minimum	33.57	11.82	8.33	24.52	15	21.83	3
Maximum	68.08	91.66	83.33	77.21	100	75.17	90
Panel B: Common Sample							
No. of firms	22	22	22	22	22	22	22
Mean	59.27	49.62	52.74	57.35	74.69	61.35	54.14
Median	59	58.33	54.51	57.93	76.64	59.88	50.96
Standard deviation	25.80	24.86	12.87	10.98	11.76	18.14	8.77
Minimum	19	8.33	0	35.01	52.62	30	42.33
Maximum	90	83.33	62.81	77.21	91.66	100	75.17

Note: Panel A shows the data for the full sample, Panel B for the restricted common sample.

percentages as category weights, which are summed up to obtain the overall weights for different pillars.

Table 1 provides descriptive statistics of the aggregate ratings and their sample characteristics. The baseline year of our analysis is 2020. We tested whether our results are specific to the year of the study by rerunning the analysis for the year 2019 and obtained similar results.[c] Panel A of Table 1 shows the full sample of firms rated by any of the six agencies: this number ranges from 36 to 147. Panel B of Table 1 limits the sample to 22 firms that have been rated by all the agencies. The latter are some of the largest publicly traded Italian companies, for which the transparency and the availability of ESG information are expected to be better. The mean and median ESG ratings are, in fact, higher in the common sample for all rating providers. We may observe that the mean ESG rating issued by different agencies is quite different: Refinitiv has the largest average score, while S&P has the lowest. If the difference can be explained by the change in the sample of the covered companies, this possible explanation cannot be applied to the restricted sample (Panel B of Table 1), where we also see relevant differences: the average scores range from 49.62 (Refinitiv) to 74.69 (MSCI Industry Adjusted).

2.2. *Correlations and disagreement analysis*

In this sub-section, we illustrate the extent of divergence between different rating agencies through a correlation analysis. Table 2 shows the Pearson's correlations between the aggregate ESG rating scores for the full sample (Panel A) and for the common sample (Panel B). It is evident how correlations are low for the majority of

[c]This analysis is not included in the paper, but available upon authors' request.

Table 2. Pearson's correlations between the aggregate ESG rating scores.

Panel A: Full Sample					
	S&P Global	Inrate	Arabesque	Truvalue	Refinitiv
S&P Global	1				
Inrate	0.45	1			
Arabesque	0.29	0.32	1		
Truvalue	0.15	0.24	0.06	1	
Refinitiv	0.64	0.50	0.37	0.05	1
MSCI Industry Adjusted	0.52	0.03	0.28	0.26	0.40
MSCI Weighted	0.57	0.08	0.38	0.36	0.40
Panel B: Common Sample					
	S&P Global	Inrate	Arabesque	Truvalue	Refinitiv
S&P Global	1				
Inrate	0.64	1			
Arabesque	0.32	0.12	1		
Truvalue	0.22	0.40	0.35	1	
Refinitiv	0.68	0.36	0.44	0.24	1
MSCI Industry Adjusted	0.46	0.18	0.60	0.43	0.50
MSCI Weighted	0.50	0.21	0.56	0.56	0.50

Note: Panel A shows the data for the full sample, Panel B for the common sample. The correlation between the two different scores issued by MSCI is not considered.

Table 3. Pearson's correlations between ESG ratings for the three pillars for full sample.

Environmental	S&P Global	Inrate	Arabesque	Refinitiv	MSCI Weighted
S&P Global	1				
Inrate	0.48	1			
Arabesque	0.59	0.56	1		
Refinitiv	0.62	0.42	0.67	1	
MSCI Weighted	0.30	−0.18	0.06	0.02	1
Truvalue	0.25	0.04	0.31	−0.01	0.08
Social	S&P Global	Inrate	Arabesque	Refinitiv	MSCI Weighted
S&P Global	1				
Inrate	0.45	1			
Arabesque	0.46	0.52	1		
Refinitiv	0.54	0.62	0.59	1	
MSCI Weighted	0.21	0.06	0.29	0.11	1
Truvalue	0.16	0.45	0.29	0.36	0.12
Governance	S&P Global	Inrate	Arabesque	Refinitiv	MSCI Weighted
S&P Global	1				
Inrate	0.48	1			
Arabesque	−0.07	0	1		
Refinitiv	0.49	0.31	−0.08	1	
MSCI Weighted	0.17	0.05	−0.08	−0.08	1
Truvalue	0.25	0.17	−0.21	−0.10	−0.01

Note: MSCI computes pillar scores comparable to other ratings only for the MSCI Weighted score.

Table 4. Pearson's correlations between ESG ratings for the three pillars for common sample.

Environmental	S&P Global	Inrate	Arabesque	Refinitiv	MSCI Weighted
S&P Global	1				
Inrate	0.65	1			
Arabesque	0.75	0.55	1		
Refinitiv	0.67	0.38	0.71	1	
MSCI Weighted	0.32	−0.09	0.05	0.08	1
Truvalue	0.32	0.04	0.34	0.14	0.33

Social	S&P Global	Inrate	Arabesque	Refinitiv	MSCI Weighted
S&P Global	1				
Inrate	0.61	1			
Arabesque	0.66	0.71	1		
Refinitiv	0.55	0.48	0.65	1	
MSCI Weighted	0.17	0.20	0.21	0.23	1
Truvalue	0.43	0.70	0.32	0.34	0.20

Governance	S&P Global	Inrate	Arabesque	Refinitiv	MSCI Weighted
S&P Global	1				
Inrate	0.54	1			
Arabesque	−0.15	−0.34	1		
Refinitiv	0.50	0.08	−0.08	1	
MSCI Weighted	0.37	0.01	0.06	0.16	1
Truvalue	0.21	0.19	−0.41	−0.11	−0.06

Note: MSCI computes pillar scores comparable to other ratings only for the MSCI Weighted score.

the pairs of rating providers, ranging from 0.03 to 0.64 for the full sample. The average correlations are 0.32 and 0.41, respectively, for the full and common samples. S&P Global and Refinitiv show the highest level of agreement between them with a correlation of 0.64 in the full sample and of 0.68 in the common sample.

Table 3 reports the correlations for the three different pillars (E, S, and G) for the full sample, while Table 4 shows the same metrics for the common sample. We underline again that MSCI provides this specification only for the MSCI Weighted score. Correlations are confirmed to be quite low, especially for the Governance pillar that shows the lowest correlations, with an average coefficient of 0.09 for the full sample and of 0.06 for the common sample. The Social and Environmental pillars show higher correlation values. The Social pillar correlations are even slightly higher than the overall ones (0.35 and 0.43). The average correlations for the overall ESG rating score and for the three pillars are summarized in Table 5. These results are largely consistent with prior findings (Chatterji *et al.* 2016, Berg *et al.* 2020).

3. Quantitative Divergence Framework

We now turn to the development of our quantitative divergence framework. ESG ratings are scores that combine a variety of parameters (or indicators) into a single number that is used to assess a company's ESG performance. Technically, such a rating can be expressed in terms of a measurement (or value) and a weight component. Measurement refers to the indicators that are used to produce a numerical value for each attribute. Weights refer to the function that linearly combines multiple indicators into one rating. Our goal is to create a model that can be used to compare two distinct ratings from two different agencies, with a characterization of component differences explaining the sources of rating divergence. In particular, we are interested in differences in values and weights between the different E, S, and G pillars (first level) and between different categories within the same pillar (second level). The model, thus, may be used to analyze divergences at the category level, which help to explain where divergences at the pillar level come from.

3.1. *Divergence at the pillar level*

ESG scores are given by a weighted sum of the scores attributed to the three pillars (E, S, and G) by their relative weights (W) for a company i:

$$\mathrm{ESG}(i) = E(i) * W_E + S(i) * W_s + G(i) * W_G. \tag{1}$$

The overall ΔESG difference between the rating attributed by agency a and the one attributed by agency b to the same company i can be decomposed into differences caused by each pillar. The following equations disaggregate ΔESG into the different components:

$$\Delta\mathrm{ESG}_{a-b}(i) = \mathrm{ESG}_a(i) - \mathrm{ESG}_b(i) = \Delta E_{a-b}(i) + \Delta S_{a-b}(i) + \Delta G_{a-b}(i), \tag{2}$$

Table 5. Average correlations between aggregate ESG rating scores and pillar ratings, for the full and common samples.

	Full sample	Common sample
ESG	0.32	0.41
Environmental	0.28	0.35
Social	0.35	0.43
Governance	0.09	0.06

where

$$\Delta E_{a-b}(i) = E_a(i) * W_{Ea} - E_b(i) * W_{Eb} = \Delta E\_\text{Values}(i) + \Delta E\_\text{Weights}(i), \quad (3)$$

$$\Delta S_{a-b}(i) = S_a(i) * W_{Sa} - S_b(i) * W_{Sb} = \Delta S\_\text{Values}(i) + \Delta S\_\text{Weights}(i), \quad (4)$$

$$\Delta G_{a-b}(i) = G_a(i) * W_{Ga} - G_b(i) * W_{Gb} = \Delta G\_\text{Values}(i) + \Delta G\_\text{Weights}(i). \quad (5)$$

To compute the statistics ΔValues and ΔWeights of each pillar, we apply the same logic. ΔValues is computed as the average of the weights multiplied by the difference in the score values, while ΔWeights is calculated as the average of the values multiplied by the difference in the weights. The sum of these two components is equal to the variables introduced in Eqs. (3)–(5). Hence, the overall ΔESG is divided into the following six components:

$$\Delta E\_\text{Values}(i) = \text{Average}(W_{Ea}; W_{Eb}) * (E_a(i) - E_b(i)), \quad (6)$$

$$\Delta E\_\text{Weights}(i) = \text{Average}(E_a; E_b) * (W_{Ea} - W_{Eb}), \quad (7)$$

$$\Delta S\_\text{Values}(i) = \text{Average}(W_{Sa}; W_{Sb}) * (S_a(i) - S_b(i)), \quad (8)$$

$$\Delta S\_\text{Weights}(i) = \text{Average}(S_a; S_b) * (W_{Sa} - W_{Sb}), \quad (9)$$

$$\Delta G\_\text{Values}(i) = \text{Average}(W_{Ga}; W_{Gb}) * (G_a(i) - G_b(i)), \quad (10)$$

$$\Delta G\_\text{Weights}(i) = \text{Average}(G_a; G_b) * (W_{Ga} - W_{Gb}). \quad (11)$$

For each pillar, the percentages of divergence explained by the values and weights, respectively, are computed as follows (referring, for example, to the E pillar):

$$\%\ \text{Divergence for values } (E) = \frac{|\Delta E\_\text{Values}|}{|\Delta E\_\text{Values}| + |\Delta E\_\text{Weights}|}, \quad (12)$$

$$\%\ \text{Divergence for weights } (E) = \frac{|\Delta E\_\text{Weights}|}{|\Delta E\_\text{Values}| + |\Delta E\_\text{Weights}|}. \quad (13)$$

In Table 6, we compute the statistics above for each combination of two different rating agencies, and we find the average divergence between the ESG scores decomposed into the value and the weight components and the average divergence across the three pillars. The numbers are computed for all listed companies that in 2020 are covered by both the agencies.

Table 6. Average divergence decomposition into value and weight components of ESG rating scores and E, S, and G pillar components.

	No. of firms	ΔValues	ΔWeights	ΔE	ΔS	ΔG
Refinitiv–Arabesque	80	44%	56%	24%	45%	32%
Refinitiv–Inrate	28	45%	55%	22%	39%	39%
Refinitiv–S&P Global	69	66%	34%	19%	56%	25%
Arabesque–S&P Global	85	73%	27%	26%	33%	41%
Arabesque–Inrate	30	35%	65%	23%	34%	43%
Inrate–S&P Global	35	36%	64%	22%	29%	49%
Arabesque–MSCI	98	38%	62%	34%	33%	33%
MSCI–Inrate	34	48%	52%	31%	33%	36%
MSCI–Refinitiv	82	58%	42%	26%	45%	29%
MSCI–Truvalue	64	34%	66%	44%	27%	29%
Refinitiv–Truvalue	52	41%	59%	38%	36%	26%
Arabesque–Truvalue	62	26%	74%	42%	23%	35%
Inrate–Truvalue	29	42%	58%	38%	44%	17%
Truvalue–S&P Global	62	46%	54%	43%	30%	27%
Average		45%	55%	31%	36%	33%

Note: MSCI refers to the MSCI Weighted score.

The average ΔWeights is larger than the average ΔValues (55% versus 45%). The Environmental pillar, in most of the cases, is responsible for the lowest percentage of divergence, with the notable exception of Truvalue (when compared to MSCI Weighted, Refinitiv, and Arabesque). Indeed, Truvalue's weights disagree a lot with the others, especially for the Environmental pillar. We should remind that Truvalue weights are based on a different methodology than the weights of other agencies, as they rely on the percentage of public news related to specific ESG factors being evaluated in a given company. Instead, the Social pillar explains the highest percentage of divergence (36%). The Governance pillar explains 33% of divergence on average (lower than the Social pillar), even though it is the major source of divergence for the total variance between two ESG scores in certain cases (mostly involving Arabesque and Inrate paired with other agencies).

Table 7 reports the decomposition of divergence into ΔValues and ΔWeights for the three different pillars of E, S, and G. As expected, ΔWeights is larger than ΔValues for the Environmental (18% versus 13%) and Governance (19% versus 14%) pillars, while differences are comparable for the Social pillar (both values are close to 18%). With the exception of the pairings comprising S&P Global and most of the couples including MSCI, which account for the bulk of the observed variance in value scores, the Social pillar reveals average ΔWeights greater than the average ΔValues.

3.2. *Divergence at the category level*

In this sub-section, we present the lowest level of breakdown of ratings divergence, which is performed at the category level. This represents the single key indicator of sustainability performance tracked by the rating agencies.

Table 7. Average divergence decomposition into value and weight components for the three E, S, and G pillars.

	ΔE_Values	ΔE_Weights	ΔS_Values	ΔS_Weights	ΔG_Values	ΔG_Weights
Refinitiv–Arabesque	10%	14%	18%	26%	16%	16%
Refinitiv–Inrate	11%	10%	23%	15%	10%	30%
Refinitiv–S&P Global	13%	6%	39%	17%	14%	11%
Arabesque–S&P Global	15%	12%	26%	8%	33%	7%
Arabesque–Inrate	9%	13%	12%	22%	14%	29%
Inrate–S&P Global	11%	11%	15%	14%	10%	39%
Arabesque–MSCI	13%	21%	12%	21%	13%	20%
MSCI–Inrate	15%	17%	22%	11%	12%	24%
MSCI–Refinitiv	13%	13%	28%	17%	17%	12%
MSCI–Truvalue	14%	30%	10%	17%	10%	19%
Refinitiv–Truvalue	15%	23%	15%	20%	11%	15%
Arabesque–Truvalue	7%	35%	5%	18%	14%	21%
Inrate–Truvalue	18%	21%	17%	28%	8%	10%
Truvalue–S&P Global	18%	26%	16%	13%	12%	15%
Average	13%	18%	18%	18%	14%	19%

Note: MSCI refers to the MSCI Weighted score.

Table 8. Classification of categories into final categories: MSCI, Refinitiv, and Arabesque.

MSCI	Refinitiv	Arabesque	Final categories
Environmental			
Climate Change	Emissions	Emissions	Emission
		Environmental Management	
Pollution and Waste		Waste	
		Environmental Stewardship	
Natural Capital	Resource Use	Resource Use	Resource Use
		Water	
Environmental Opportunities	Innovation	Environmental Solutions	Innovation
Social			
Human Capital	Workforce	Compensation	Workforce
		Diversity	
		Employment Quality	
		Labor Rights	
		Occupational Health and Safety	
		Training and Development	
	Human Resources	Human Rights	Human Rights
Social Opportunities	Communities	Community Relations	Community
Product Liability	Product Responsibilities	Product Quality and Safety	Product Responsibility
Stakeholder Opposition		Product Access	
Governance			
Corporate Governance	Management Score	Business Ethics	Corporate Governance
	Shareholder Score	Corporate Governance	
Corporate Behavior	CSR Strategy	Transparency	CSR Strategy
		Capital Structure	
		Forensic Accounting	

This analysis was constrained by the availability of data. Only for Arabesque, MSCI, Refinitiv, and Truvalue we have available information about categories and relative weights. As for the aggregate rating, each pillar score is computed as the weighted sum of different categories. Since each rating agency chooses to break down the concept of ESG performance into different indicators and organizes them into different hierarchies, we aggregate the various categories to obtain a final category list common to each pair of agencies.

There a minimum of two to a maximum of 11 single indicators monitored by the different rating agencies when analyzing each of the three pillars. Moreover, Truvalue provides five high-level categories, divided into 26 sub-categories. In order to perform a meaningful comparison of these different rating systems, we develop our categorization of the data using a top–down approach. Refinitiv was chosen as reference. Both Arabesque and MSCI categories and Truvalue sub-categories were assigned to Refinitiv categories on the basis of an accurate screening of rating

Table 9. Truvalue classification of categories and sub-categories into final categories.

Truvalue categories	Truvalue sub-categories	Final categories
Environmental		
Environment	Air Quality	*Emissions*
	Energy Management	*Resource Use*
	Water and Wastewater Management	*Resource Use*
	Ecological Impacts	*Emissions*
	Waste and Hazardous Materials Management	
	GHG Emissions	
Business Model and Innovation	Product Design and Lifecycle Management	*Innovation*
	Business Model Resilience	
	Supply Chain Management	*Resource Use*
	Materials Sourcing and Efficiency	
	Physical Impacts of Climate Change	*Emissions*
Social		
Social Capital	Human Rights and Community Relations	*Human Rights and Community*
	Customer Privacy	*Product Responsibility*
	Data Security	
	Access and Affordability	
	Product Quality and Safety	
	Customer Welfare	
	Selling Practices and Product Labeling	
Human Capital	Labor Practices	*Workforce*
	Employee Health and Safety	
	Employee Engagement, Diversity, and Inclusion	
Governance		
Leadership and Governance	Business Ethics	*Corporate Governance*
	Competitive Behavior	*CSR Strategy*
	Management of the Legal and Regulatory Environment	
	Critical Incident Risk Management	
	Systemic Risk Management	

providers' definition of each category, which was provided in their methodology documentation. We require that each category could be assigned only to one final category. The final classification is shown in Table 8 (for Refinitiv, MSCI, and Arabesque categories) and Table 9 (for Truvalue sub-categories). The Environmental pillar is divided into three final categories, the Social pillar into four final categories (three for Truvalue, since "Human Rights and Community Relations" is one single category for this agency), and the Governance pillar into two final categories.

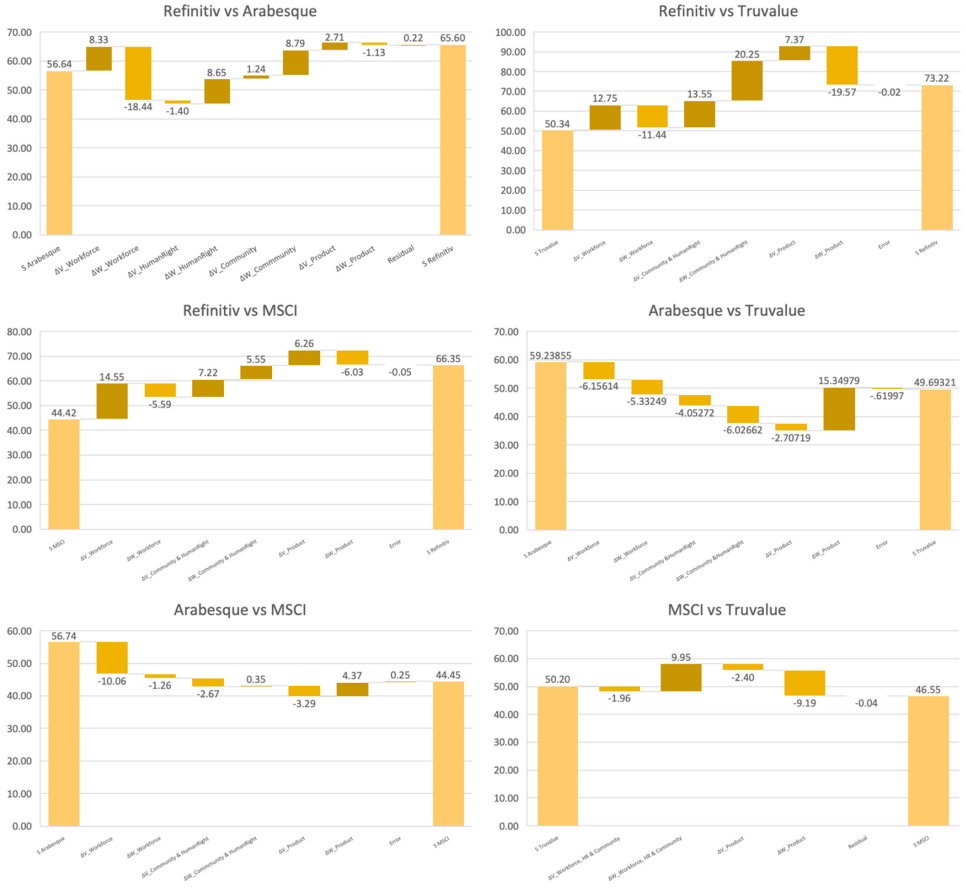
Interestingly, Tables 8 and 9 show that there is some scope divergence between different ESG ratings. For instance, categories can be very broadly defined for one agency (e.g. MSCI "Human Capital") or much more detailed for another, making it difficult to assess the sources of divergence, merely looking at the aggregate pillar scores.

To compute ΔValues and ΔWeights of each final category, the same breakdown logic used for pillars was applied at the category level [see Eqs. (6)–(11)]. In the case

Fig. 1. Average differences in values and weights for the Environmental pillar categories.

of Truvalue, not all the categories and weights are reported by the analysts, since the rating depends on the number of news available related to the specific category. When there was no data, the category is therefore not considered for the computation of the final category (i.e. its weight is equal to zero).

Figures 1– 3 report the average scoring differences computed for each category of the Environmental, Social, and Governance pillars, respectively. The Environmental pillar (Fig. 1) shows the lowest variations across different categories, confirming that moderate agreement can be found about the measurement of the different environmental categories considered. The highest divergences are related to the difference between the weights of the *Emission* and *Innovation* categories. These results are not surprising considering, for instance, that information about the *Emission* indicator are



Note: Truvalue and MSCI when compared with other agencies, we consider only one category for Community and Human Rights. For Truvalue versus MSCI, we consider only one category for Workforce and Community and Human Rights.

Fig. 2. Average differences in values and weights for the Social pillar categories.

quite objective and are generally publicly disclosed by large listed companies, leading to low divergences of category values. However, the differences concerning weights testify that different agencies attribute different importance to environmental issues.

The Social pillar (Fig. 2) shows higher variations across the components of different categories. Even in this case, the weight component shows higher average differences. The disagreement among the values attributed to different categories is also in this case not so relevant, especially for Refinitiv versus Arabesque and MSCI versus Truvalue. This means that, despite the wide number of factors evaluated in this pillar, when categories are aggregated, different agencies tend to assign similar values to one company. Thus, if differences exist at a more granular level, they tend to offset when grouped into the final categories.

Lastly, the Governance pillar (Fig. 3) shows the highest variation across weights and values for both categories. The majority of the variance is again explained by the differences in weights, which are particularly high. However, for the *Corporate*

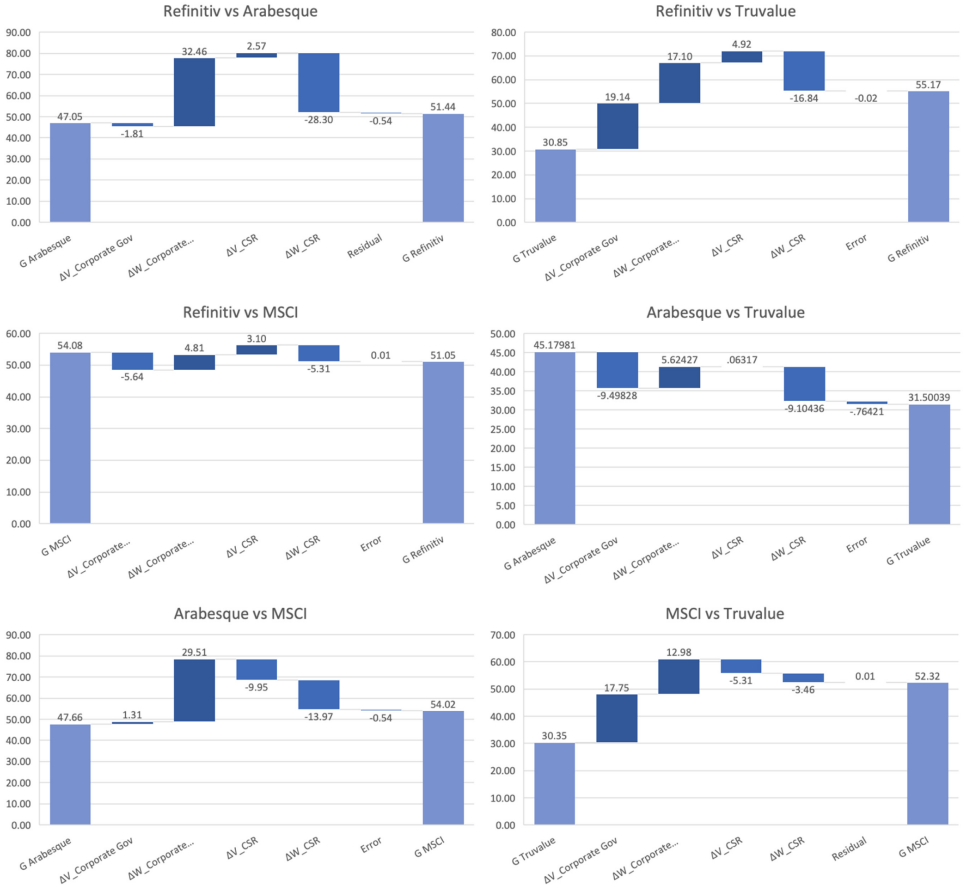


Fig. 3. Average differences in values and weights for the Governance pillar categories.

Governance (CG) category, differences between values are in many cases even larger than those between weights (i.e. Refinitiv versus Truvalue, Arabesque versus Truvalue, and MSCI versus Truvalue). One cause might be the subjectivity involved in determining the relevance of each category in different sectors in terms of governance issues. Indeed, the relevance of a category for a specific industry is more straightforward for the Environmental and Social pillars, and the results demonstrate that a certain consensus among agencies may be reached.

In sum, the category analysis has shown that there is substantial score divergence, especially for the weight component. There is, at least, some level of agreement regarding measurement (i.e. category values) of firms' environmental categories, for which the information are more easily obtained from public records. However, other categories, especially for the Governance pillar, show high levels of disagreement both for values and weights (e.g. *Corporate Governance*). Moreover, disagreement might tend to increase with granularity. This is the case of the Social pillar, where divergences seem to compensate each other to some extent through aggregation.

4. Concluding Discussion and Implications

The aim of this study was to explain why ESG ratings diverge. We developed a framework for comparing different ESG rating methodologies in a systematic way. This framework allows us to split the differences between ESG ratings into two main components: ESG values and ESG weights at the pillar and category (second) levels.

We achieved some important results. First, we confirm a low correlation between ESG ratings of different rating agencies according to prior literature (Chatterji *et al.* 2016, Berg *et al.* 2020). Since our study refers to the scores issued in 2020, it seems that no particular convergence has been experienced in the market compared to previous findings. We also found that the weight component is more relevant than the value component in explaining ESG rating divergences.

Second, the Environmental pillar differences are the lowest ones in the majority of comparisons with the only exception of Truvalue. This result can be attributed to the differences in the indicator weights assigned by Truvalue, which are computed following a different methodology compared to other agencies, depending on the number of company's news related to a specific category (and therefore, pillar).

Third, the Social and Governance pillars explain the majority of differences. The Social categories are the widest in number, although the topics addressed are very similar (e.g. human rights; workforce conditions, health, and safety; product quality; and impact on the society) with the primary differences attributable to the weights. For the Governance pillar, both weights and values account for a significant percentage of divergence for many agency comparisons. This latter finding can be explained by a higher level of subjectivity in the governance category evaluations, which can vary greatly between agencies, with some performance indicators included by one agency, but not rated by another.

Our findings demonstrate that ESG rating divergence is driven not merely by the differences in analysts' evaluations, but also by disagreement about the underlying methodological issues and metrics. Weights divergence is particularly concerning, because it indicates a conflict on the relevance of different ESG performances and how pillars and categories are related to one another. As a result, even if a firm receives the same score value for its ESG performance, the ESG ratings generated by various rating providers might still differ significantly.

4.1. *Implications for scholars and practitioners*

Our results suggest various implications for researchers, investors, companies, and rating agencies. Scholars should take into account that the divergence among different ESG pillar and category scores can affect the results and comparability of their studies. Certain results that have been obtained on the basis of one ESG rating might not be replicable with ESG ratings issued by another rating provider. This is particularly relevant for the stream of researches investigating the relation between firms' ESG and financial performances. Our first recommendation is to rely on data from more than one agency when analyzing those issues to improve the generalizability of findings and detect differences originated by the use of different scores. Another suggestion is to replicate the same study throughout time, because rating methodology and reporting practices are constantly evolving and ESG performance may be influenced. Third, researchers are encouraged to build hypotheses around indicators that are more clearly defined (e.g. at the pillar or category level) than the aggregated ESG rating scores in order to rely on more transparent measurements (such as the Environmental pillar). In this situation, it would be still necessary to evaluate the adoption of a variety of different weighting methodologies, as we saw there is greater disagreement on the weights assigned to indicators.

Considering investors, our framework helps them to understand why a company's ESG performance from different rating agencies may diverge. In this case, the choice of a particular rating provider can affect their investing decisions in unpredictable ways. A first recommendation is not to look at just one agency, but to compare different ESG indicators from different providers. This advice is especially important when it comes to the Governance pillar and related categories, because the level of disagreement is larger and the range of issues covered is broader and more generic than the ones considered in the other pillars. Conversely, investors may also rely on a single rating agency after persuading themselves that the metrics are consistent with their investing goals. Hence, the recommendation is to collect information about what is measured by each rating agency to select the ratings that best fit with their needs.

For companies, our results highlight that there is substantial disagreement about their ESG performance. This divergence not only occurs at the aggregate level, but is actually more pronounced for specific categories of ESG performance. Having that in mind, companies should increase the level of information disclosure in order to facilitate the rating process. The amount of information provided can help them to

enhance their ESG scores because rating agencies often penalize companies that do not supply enough information, and more transparent information would also reduce analysts' subjective judgment. This would require more efforts in terms of resources and time spent for sustainability disclosure, but given the increasing importance of assets invested in SRI (GSIA 2020) and the growing awareness about these concerns, it is critical for businesses to take steps in this direction.

Finally, for rating agencies a higher level of transparency with respect to their ESG scores and methodologies used is necessary to better understand what stands behind the ESG ratings. Recently, there has been an increase in the level of transparency; however, while some agencies have begun to publish comprehensive methodology and even some insights into their ratings, many agencies continue to keep detailed descriptions and data on scores and weights confidential. The standardization process of ESG ratings should be fueled by the introduction of new requirements and standards from policymakers. For instance, the European Commission is increasingly addressing the problem of corporate sustainability disclosure and could promote the standardization of ESG indicators and measurement. This would be especially important for SMEs, for which requirements are currently lacking and are, therefore, frequently excluded from evaluations due to a lack of information. Indeed, the establishment of transparent criteria would reduce the possible greenwashing phenomenon (Mrkajic *et al.* 2019) and limit analysts' subjective evaluation.

4.2. *Limitations and future research directions*

Our study has some limitations which could be the basis for scholars seeking to advance knowledge on the topic. First, as our sample consists of Italian firms and is therefore context-specific, it would be interesting to validate our findings using a wider sample. Indeed, the research may be expanded to include a European or worldwide sample of firms in order to look into regional differences. Moreover, other rating agencies can be included to corroborate the differences in the value and weight components found, as well as their relative importance. Second, another limitation is given by the lack of a complete overview on the indicators used for the definition of category scores, which has determined a classification of the categories based on the descriptions provided by the different agencies. Future studies could apply different taxonomies, using more granular information on the indicators considered by agencies to generate category scores, together with more advanced techniques to classify categories. The development of rating methodologies based on Artificial Intelligence and big data analytics also calls for studies on this topic using more sophisticated approaches (e.g. Lanza *et al.* 2020). Furthermore, research considering the relationship between ESG divergence and stock market performance is surely welcome. Recent evidence has shown that stock returns are positively related to ESG rating disagreement (Brandon *et al.* 2019), however, more research is needed to support these findings and effectively inform investors' financial decisions.

References

E. Avetisyan & K. Hockerts (2017) The consolidation of the ESG rating industry as an enactment of institutional retrogression, *Business Strategy and the Environment* **26**, 316–330.

F. Berg, J. F. Koelbel & R. Rigobon (2020) Aggregate confusion: The divergence of ESG ratings, Working Paper No. 5822–19, MIT Sloan School of Management, Cambridge, MA.

E. Bolognesi & A. Burchi (2021) Non-financial reporting regulation, sell-side financial analysts and the ESG disclosure premium, Working Paper.

R. G. Brandon, P. Krueger & P. S. Schmidt (2019) ESG rating disagreement and stock returns, Research Paper No. 19–67, Swiss Finance Institute — Geneva, Geneva.

A. Cellier, P. Chollet & J.-F. Gajewski (2016) Do investors trade around social rating announcements?, *European Financial Management* **22**, 484–515.

A. Chatterji, D. Levine & M. Toffel (2009) How well do social ratings actually measure Corporate Social Responsibility?, *Journal of Economic & Management Strategy* **18**, 125–169.

A. K. Chatterji, R. Durand, D. I. Levin & S. Touboul (2016) Do ratings of firms converge? Implications for managers, investors and strategy researchers, *Strategic Management Journal* **37**, 1597–1614.

L. Conca, F. Manta, D. Morrone & P. Toma (2021) The impact of direct environmental, social, and governance reporting: Empirical evidence in European-listed companies in the agri-food sector, *Business Strategy and the Environment* **30**, 1080–1093.

M. Delmas & V. D. Blass (2010) Measuring corporate environmental performance: The trade-offs of sustainability ratings, *Business Strategy and Environment* **19**, 245–260.

A. Dimmelmeier (2020) Mergers and Acquisitions of ESG firms: Towards a new financial infrastructure?, Working Paper, SocArxiv, doi:10.31235/osf.io/jt2uk.

G. Dorfleitner, G. Halbritter & M. Nguyen (2015) Measuring the level and risk of corporate responsibility: An empirical comparison of different ESG rating approaches, *Journal of Asset Management* **16**, 450–466.

R. G. Eccles & J. Stroehle (2018) Exploring social origins in the construction of ESG measures, Working Paper, SSRN Electronic Journal, doi: 10.2139/ssrn.3212685.

E. Escrig-Olmedo, M. Munoz-Torres, M. Fernandez-Izquierdo & J. Rivera-Lirio (2014) Lights and shadows on sustainability rating scoring, *Review of Management Science* **8**, 559–574.

European Commission (2020) Study on sustainability-related ratings, data and research, ERM Report, Directorate-General for Financial Stability, Financial Services and Capital Markets Union, Brussels, doi:10.2874/14850.

Global Sustainable Investment Alliance (GSIA) (2020) Global sustainable investment review 2020, Report. Available at: http://www.gsi-alliance.org/wp-content/uploads/2021/07/GSIR-2020.pdf (accessed 20 September 2021).

M. Hedesström, U. Lundqvist & A. Biel (2011) Investigating consistency of judgment across sustainability analyst organizations, *Sustainable Development* **19**, 119–134.

J. Kang (2015) Effectiveness of the KLD social ratings as a measure of workforce diversity and corporate governance, *Business & Society* **54**, 599–631.

A. Lanza, I. Faiella & E. Bernardini (2020) Mind the gap! Machine learning, ESG metrics and sustainable investment, Occasional Paper No. 561, Bank of Italy.

F. Li & A. Polychronopoulos (2020) What a difference an ESG ratings provider makes!, Research Affiliates, January. Available at: https://www.researchaffiliates.com/documents/770-what-a-difference-an-esg-ratings-provider-makes.pdf (accessed 20 September 2021).

C. Lopez, O. Contreras & J. Bendix (2020) ESG ratings: The road ahead, Report, Milken Institute, doi:10.2139/ssrn.3706440.

J. E. Mattingly & S. L. Berman (2006) Measurement of corporate social action: Discovering taxonomy in the Kinder Lydenburg Domini ratings data, *Business & Society* **45**, 20–46.

J. E. Mattingly (2017) Corporate social performance: A review of empirical research examining the corporation–society relationship using Kinder, Lydenberg, Domini social ratings data, *Business & Society* **56**, 796–839.

B. Mrkajic, S. Murtinu & V. G. Scalera (2019) Is green the new gold? Venture capital and green entrepreneurship, *Small Business Economics* **52**, 929–950.

A. Paltrinieri, U. D. Dervi, A. Khan, I. Saba and M. K. Hassan (2021) Green and socially responsible finance: Past, present, and future. In: *Proceedings of the ADEIMF Summer Conference 2021*.

N. Semenova & L. G. Hassel (2015) On the validity of environmental performance metrics, *Journal of Business Ethics* **132**, 249–258.

Truvalue Labs (2020) Truvalue Labs methodology: Scoring.

L. Widyawati (2021) Measurement concerns and agreement of environmental social governance ratings, *Accounting & Finance* **61**, 1589–1623.

S. E. Windolph (2011) Assessing corporate sustainability through ratings: Challenges and their causes, *Journal of Environmental Sustainability* **1**, 5.