

ESG: Credibility behind the scores – The reliability and transparency of ESG ratings

Rita Mayer and Anita Reizingerné Ducsay *

Corvinus University of Budapest. Budapest, Hungary

* Correspondence: anita.ducsay@uni-corvinus.hu

Abstract: The emergence of sustainability in business has led to a growing number of market players becoming concerned about ESG rating, which includes environmental, social and governance aspects. Investors are also pricing in the scores generated by ESG indicators. The growing interest in ESG data raises the question of the reliability of the scores provided by different ESG rating agencies. This research explores the differences in the methodologies used by the most reputable ESG rating agencies through a content analysis. The inconsistency of ESG scores will be investigated by taking a random sample of companies in eight pre-selected industries and subjecting them to a correlation test. The purpose of this is to highlight the uncertainties associated with ESG ratings and to draw attention to the fact that ensuring the reliability, consistency, and transparency of ESG information remains an unresolved issue and challenge for both decision makers and users.

Keywords: corporate sustainability; ESG ratings; ESG rating agencies; ESG reliability

1. Introduction

The emergence of sustainability in business has led to an increasing number of market players having concerns about ESG rating, which encompasses environmental, social and governance aspects. As accountants, our main focus is to ensure that reports, statements and ratings about a company present a true and fair picture. At the level of sustainability disclosures, this is also important because the phenomenon of greenwashing is becoming more and more common as companies try to appear sustainable. As such, ESG ratings, which are produced by external rating companies, should aim to paint an objective picture of a company's sustainability performance. Trends in ESG also show the need for reliable and realistic information on companies' sustainability performance. Recently, there has been a growing demand from investors for transparent and clear non-financial data on companies in addition to financial data. This growing demand has led to the emergence of a growing number of players in the ESG rating market, which has led to increasingly fierce competition. In this competitive environment, one of the competitive advantages of rating firms may come from the fact that they may be able to rate a company more favourably than other rating agencies, following their own methodology. The flexibility on the regulatory side also allows this, as there is no generally accepted and established practice with regard to ESG disclosures, including the methodology behind ratings, which has regulatory power. Thus, it can be problematic for a potential investor how to navigate between different scales and ratings, which one to consider, and whether the same decision is even reached when considering different ESG ratings.

Citation:

Mayer, R., & Reizingerné Ducsay, A. (in press). ESG: Credibility behind the scores – The reliability and transparency of ESG ratings. *Prosperitas*. Budapest Business School. https://doi.org/10.31570/prosp_2022_0041

History:

Received: 10 Dec 2022
Last revised: 7 Feb 2023
Accepted: 9 Feb 2023
Published: 7 Mar 2023



Copyright:

© 2023 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY-NC) license.

2. Literature review

According to Wong (2018), ESG rating can be understood as an assessment based on a comparative evaluation of performance on environmental, social and governance issues. Accordingly, ESG scores serve as a measure of corporate sustainability performance. There

is a growing literature on what exactly is the appropriate composition of these metrics, which are the financial or non-financial indicators that best explain a company's ESG score.

According to Kocmanová et al. (2017), sustainability is a phenomenon that cannot be measured simply by one indicator but requires the use of composite indicators. The disadvantage of this, however, is that a properly constructed indicator system needs to be developed and weights associated with each indicator need to be found. Therefore, as part of their research, they have attempted to identify factors that can be used to quantify the ESG performance of companies using factor analysis. In their analysis, they used a total of 39 financial and non-financial indicators, resulting in the SEESG model.

This research shows that currently there is no universally accepted, mandatory set of indicators to be used when calculating ESG scores. One of the reasons for this is that there are also gaps in the regulations on sustainability disclosure in this area. Although non-financial reporting requirements have changed a lot in recent years, this reporting format is much less systematic than financial reporting, as there is no standardised scheme, only guidelines, recommendations, and frameworks (GRI, IR, SASB, CDP, ISO 26000) regulate the information to be disclosed and the reporting practices (Lippai-Makra, 2020).

In addition to the usage of indicators, their appropriate measurement, and the lack of regulation, four main issues affecting the reliability of ESG data can be identified in the literature.

Company size: companies with larger resources have the opportunity to measure their sustainability performance in more detail, using more indicators, and thus regularly achieve higher ESG scores (Doyle, 2018; Drempeć et al., 2020; Liang & Renneboog, 2020).

Geography: the regulatory specificities of companies' home countries create a unique environment that influences disclosure and reporting levels. This means that, as the degree of regulation on sustainability reporting varies from country to country, ESG performance may be measured in more depth and with more indicators in some areas, while in others ESG performance is disclosed much more superficially. This is why one study found that companies in Europe have a much higher ESG rating than in the US or elsewhere (Doyle, 2018; Liang & Renneboog, 2020).

Industry sector: several rating companies try to standardise ratings by industry, but in doing so, they simply weight the scores for the three dimensions within the industry and do not take into account company-specific risks (Doyle, 2018).

Rating agencies: in their research, Berg et al. (2022) found that the rating companies themselves can have an impact on ESG scores. They observed the existence of the so-called "rater effect", which means that raters can be influenced in their ratings by the reputation and marketing strategy of the firm. On the other hand, the influence of rating firms on ESG scores may also depend on what the firm is willing to disclose and what information it provides to the raters. What is more, it is possible that if a question is not answered within the questionnaire, the company will automatically receive a poor rating for that aspect.

Another aspect that strongly questions the reliability of ESG scores is the consistency of the rating companies available on the market and the services they offer. Li and Polychronopoulos (2020), for example, identified 70 ESG rating firms, not counting investment banks, governmental organisations and research organisations that conduct ESG-related research. This number is significant in itself and it is reasonable to think that the ratings of the different rating agencies are not fully aligned.

In the literature, several researchers have attempted to compare different rating companies, both in terms of the used methodology and the indicators are taken into account. One, perhaps most important finding, which also affects the question of reliability, is that the metrics used by rating companies come from different sources: company registers, non-financial reports, media, third-party data, and questionnaires, which fundamentally calls into question the extent to which the different ratings measure the same phenomenon (Gyönyörövá et al., 2021).

In addition, another factor that may undermine the comparability of ESG scores issued by different evaluators is that these evaluators use different methodologies and approaches. This involves not only considering exactly which indicators they use to try to measure sustainability performance, but also the weighting they give to each indicator (Gyönyörövá et al., 2021). For example, Chatterji et al. (2016) compared 6 rating companies and found that there was a low correlation between the ESG scores they calculated. Similar results were obtained by Berg et al. (2022) and Capizzi et al. (2021). According to them, the main

differences are due to the scope (what is measured), the measurement itself (what indicators are used) and the weighting of each indicator. Other studies have also shown that there is inconsistency in the classifications (Delmas et al., 2013; Dorfleitner et al., 2015; Semenova & Hassel, 2015). According to Dimson et al. (2020), another reason for the discrepancies is that the three ESG dimensions are considered with different weights in the cumulative score.

Contrary to the above, Utz (2019) took a different approach to the reliability of ESG ratings. He analysed how unexpected scandals are reflected in ESG scores. As a result, he found that the retrospective scores decrease significantly in the year of the scandal, which led him to conclude that ESG ratings appear to be reliable in this respect, as they track changes.

3. Methodology and data

It is clear from the research in the literature that there is a lot of uncertainty in the information on ESG performance, so the scores calculated by some rating agencies cannot be trusted blindly. This research attempts to explore the main methodological differences that may lead to differences in ESG scores, by reviewing the assessment methodologies of eight well-known ESG rating companies in a content analysis. The main objective of this is to identify similarities and differences between the selected rating firms in terms of the methodology used and the information used. This may predict the reliability of the overall ESG scores, as possible differences may lead to certain rating companies considering different aspects with different weights. This could make it impossible or uncertain to compare companies on an ESG basis.

Q1: What significant differences can be identified between the methodologies used by each rating company?

Content analysis of the methodology of the raters

To answer this question (Q1), the methodology used by ESG rating agencies was first examined to highlight the main differences that may lead to a lack of reliable comparison of scores between rating agencies. To this end, eight ESG rating agencies with the largest coverage have been selected. These rating agencies are MSCI, Sustainalytics, Refinitiv, S&P, FTSE Russell, ISS, CDP and Bloomberg. The information for the content analysis has been gathered from the brochures published on the official websites of each rating agency.

Following the content analysis, quantitative research is conducted to find out to what extent a significant relationship can be observed between the ESG scores calculated by the selected ESG rating companies. The analysis is conducted not only for the cumulative ESG scores but also for the scores of the three dimensions (environmental, social, governance) separately, as Dimson et al. (2020) found that the scores of the rating agencies differ because they use different weightings when aggregating the three scores. Thus, this dimension-by-dimension analysis would eliminate this bias. The research is also extended to include an additional question that ESG scores may also be affected by the industry in which the rated company operates, as for example a pharmaceutical company or the energy sector may have a much higher importance on environmental impacts. Thus, it is clear that industry specificities have an impact on ESG scores. The question is to what extent the ratings of each rating agency are industry specific and to what extent companies within a given industry can be compared based on ESG scores.

Q2: Can a significant relationship be observed between the scores published by each ESG rating company?

Q3: Are there any significant differences observed along the ESG dimensions compared to the cumulative scores?

Q4: Does the relationship between ESG ratings differ at industry level?

Several studies have found that ESG scores can be related to the size and geographical location of a company. Consequently, it has been preliminarily assumed that larger, more capital-intensive companies tend to have higher ESG scores, as their resources and greater capacity allow them to monitor their ESG performance on a regular basis and thus provide more ESG information to the rating agencies. A further preliminary hypothesis is that European

companies perform better from an ESG perspective than their counterparts in the US or other continents.

- **H1: Larger companies regularly score better on ESG because their greater resources allow them to better measure their sustainability performance.**
- **H2: European companies have better ESG scores than US or other companies.**

Numerical analysis of reliability

In the second phase of the research, the reliability of the scores provided by ESG rating companies will be quantified. As a first step, the data available for free on the internet was examined. This research found that of all the ESG rating companies presented earlier, S&P is the only one whose scores are available in aggregated form and by dimension on the rating company's public website. However, it is important to point out that S&P's database includes a number of companies for which access to data is restricted. MSCI and Sustainalytics also provide access to the ESG scores of rated companies, with the difference that for these rating agencies only the aggregated ESG scores are published on the website. Refinitiv's database is much larger than the previous ones. Eikon provides not only aggregated and dimensional scores, but also the possibility to display key background indicators. This service is available through the website by subscription.

This research, relying primarily on the Refinitiv database, used Eikon to collect the companies assessed and their ESG scores for 8 industries (commodities, consumer cyclicals, non-cyclicals, energy, finance, healthcare, industrial products, technology). From the resulting database, only those companies for which aggregated ESG scores were available were filtered. As the research investigates not only the relationship between aggregate scores, but also the correlation between the scores of each ESG dimension. The database generated after data cleaning was suitable for further analysis, so using Excel, a simple random sampling was carried out, with all elements of the population having the same probability of being included in the sample. This was the appropriate sampling procedure in this case, as it was a finite, homogeneous population where no prior filtering was needed to narrow down the elements to be selected. Thus, the eight industries were sampled separately so that the results could be interpreted in a comparable way across industries. This method is also useful because no industry will be under- or over-represented in the sample. This is advantageous because each industry represents a completely different aspect of sustainability, so this method can also be used to filter out any bias due to this factor. Prior to selection, the number of sample items per sector was determined in 5 batches and then sampling was performed using Excel's random function. Subsequently, I collected the ESG scores of the 40 companies selected. This was done by simply extracting the scores published by Refinitiv from the original database. The Bloomberg ESG scores were then collected using Bloomberg Terminal, downloading the ESG data of the selected sample companies as of 31.12.2021. The analysis was then extended to compare the scores of three other ESG rating companies, S&P, Sustainalytics and MSCI. In this case, the ESG information required could only be found one by one using the search engine on the rating companies' official websites. However, in many cases this search did not yield any results or simply required a subscription to obtain the relevant ESG scores. Overall, 20 companies' ESG scores were collected from the original sample of 40 items. Thus, at the aggregate score level, it was possible to compare the scores of all five evaluators across the 20-item sample.

Sample

In examining the composition of the sample, the main focus was on the ESG score, geographical location and size of the selected companies.

Regarding the ESG scores of the companies, the ratings published by Refinitiv show a predominantly medium (B, C) ESG performance for the companies in the sample. This is advantageous for the analysis because it means that there are no companies in the sample with extremely good or poor performance, which could influence the ratings of the raters because of their size or market power. The following figure (Figure 1) shows the distribution of the number of companies in the sample by ESG rating.

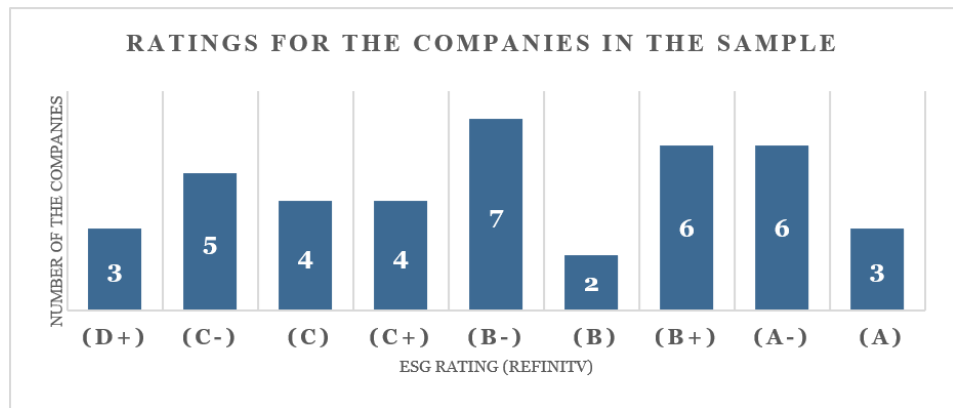


Figure 1. Ratings for the companies in the sample. *Source: Authors' own*

However, the geographical location of the companies in the sample is less favourable. In fact, when looking at the location of the selected companies, it appears that Asian and Oceanian companies are over-represented in the sample, as they represent a larger share of the sample than the majority. In contrast, European and US companies are under-represented in the sample, with both continents accounting for less than 25% of the sample (Figure 2). This type of bias could then pose a problem when drawing final conclusions, as a number of previous studies have shown that companies' ESG scores are significantly affected by their geographical location.

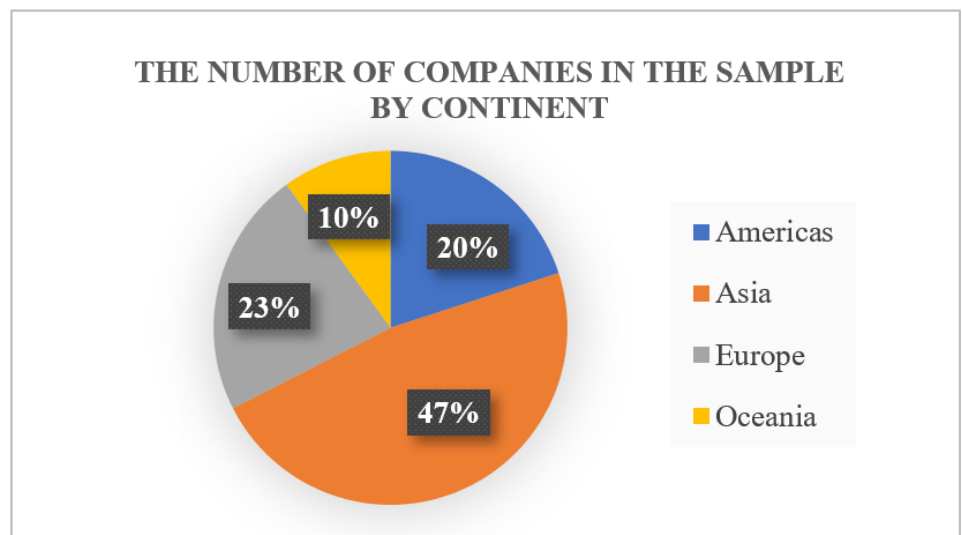


Figure 2. The number of companies in the sample by continent. *Source: Authors' own*

Table 1. The size of the companies in the sample. *Source: Own work*

Total assets (\$ billion)	Number of companies
Less than 10	24
10 – 20	5
20 – 30	5
30 – 40	1
40 – 50	4
More than 50	1

For the selected companies, analyses were also conducted on the size of the sample based on total asset value. As shown in Table 1, the sample was predominantly composed of companies with assets less than \$10 billion on 31 December 2021. However, large companies such as LG Electronics Inc, L'Oreal SA, Tata Motors Ltd, UMB Financial Corp and Koninklijke Ahold Delhaize NV, with assets of more than \$40 billion at the end of 2021, make up 1/6 of the sample. As research also aims to understand the relationship between the size of

companies and the ESG score calculated, it is particularly beneficial to include some large companies in the sample in addition to mid-sized companies.

Methodology

To explore the reliability of the scores published by each ESG rating company, a correlation analysis was performed on the selected sample using IBM SPSS Statistics 27. For this research, correlation analysis is an appropriate methodology because if a sufficiently strong relationship is observed between the ESG scores of each rater, it can in fact be said that the score from two different sources measures the same phenomenon and is therefore suitable for comparing companies based on their sustainability performance.

4. Results

4.1. Content analysis of the methodology of the raters

The results of the qualitative analysis show that the methodology behind ESG ratings differs in several aspects (Table 2), suggesting that the measurement methodology of the rating agencies does not necessarily provide a reliable basis for quantifying the sustainability performance of companies.

Table 2. Comparison of the methodology of the rating companies. *Source: Own work*

Rating company	Number of indicators	Weighting	Scale	Scoring system	Rating
MSCI	more than 350	industry-adjusted weighted average	0–10	AAA – CCC	Laggard, Average, Leader
Sustainalytics	more than 300	weighted average	0–100	0 – 40+	Negligible, Low, Medium, High, Severe
Refinitiv	more than 630	weighted average by materiality	0–1	A+ – D-	Laggard - Leader
S&P	more than 130	predetermined weights	0–100	0 – 100	–
FTSE Russell	more than 300	weighted average by materiality	0–5	0 – 5	–
ISS	more than 700	weighted average by materiality	1–4	A+ – D-	Poor, Average, Good, Excellent
CDP	–	weighted average by materiality	0 or 1	A – D-	–
Bloomberg	more than 600	weighted average by materiality	0–1	0 – 1	–

Number of indicators

The examination of the methodology used by the ESG rating companies showed that there are differences on how many indicators each ESG rating company uses and what topics they cover within sustainability. Understanding this is important because it can cause a discrepancy in the scores calculated by providers if different topics are assessed at different depths during the assessment. A comparison of the evaluation methodologies revealed that there are three groups of raters based on the number of indicators used. S&P forms the group of ESG providers that use relatively few indicators to assess the sustainability performance of companies. Some rating providers use slightly more, between 300 and 400 indicators to determine their ESG score. Such ESG rating companies include MSCI, Sustainalytics and FTSE Russell. In the third category, we are talking about well over 700 indicators, which is typical of the methodology used by Refinitiv, ISS and Bloomberg. This high number is also due to the fact that ISS, for example, specifically highlights that in addition to the indicators for the three dimensions of ESG, a number of industry-specific indicators are used in the rating.

Weighting

In addition to the indicators used, the methodologies for ESG ratings also differ in the way in which the indicators are aggregated and thus the final pillar and aggregate scores are calculated. A review of the measurement methodology of each ESG rating company shows that, in general, all the ESG rating companies surveyed use the weighted averaging method to determine the final ESG score. One of the main differences, however, is what a particular rater considers to be a material, relevant sustainability issue. According to MSCI's methodology, a topic is considered relevant if it is likely to incur significant costs or contribute to revenue growth. In contrast, Sustainalytics considers a topic to be material to a company if it is associated with a management initiative or oversight. S&P determines the materiality of ESG-related issues by assessing the potential impact and likelihood of an issue occurring on an industry-by-industry basis. In contrast, the FTSE Russell uses an Industry Classification Index (ICB), Refinitiv uses the Materiality Matrix, while ISS uses the Industry Classification Matrix to assess the relevance of a given topic. It is therefore clear that the relevance of each ESG factor may vary from industry to industry so the weighting tends to give more weight to more pronounced sustainability issues in the final score. Refinitiv, FTSE Russell, ISS, CDP and Bloomberg follow this methodology. MSCI's weighting methodology is somewhat more specific, following similar principles. In these ratings, the impact of the company on the environment and society is assessed in relation to the rest of the industry, and whether this impact is realised in the short or long term. As a result, the more significant the short-term impact, the higher the weights. Thus, in this case, we can speak of a kind of industry-adjusted weighting.

Scale

Besides the methodological uncertainties and differences identified above, the main reason for the inconsistency of ESG ratings is the significant differences in the measurement scales used by the rating companies. Based on the comparative analysis carried out in my research, the rating companies under investigation use completely different scales to determine the final ESG score of companies, which can be very misleading for the users of the information. However, some similarities can be found between the scores of some ESG rating agencies. Sustainalytics and S&P, for example, measure companies' sustainability performance on a scale of 0 to 100, while Refinitiv and Bloomberg have ESG scores between 0 and 1. The ratings of the other ESG providers surveyed are completely different. MSCI measures the ESG performance of companies on a scale of 0 to 10, while FTSE Russell scores between 0 and 5. In contrast, ISS scores 1 for worst performance and 4 for best performance. The CDP assessment is specific in that answers to each question are rewarded with a score of 0 or 1.

Scoring system

In addition to the different scales used by each rating company to determine a company's ESG score, there is also uncertainty about the conversion of companies' scores according to their own scoring system. Accordingly, in MSCI's ratings, CCC represents the worst ESG performance and AAA the best. But among the ESG rating companies under review, there are several that use letter scale to determine companies' ultimate sustainability performance. The methodologies of Refinitiv and ISS are quite similar in this respect, as both ESG providers' ratings give the final score on a scoring scale between D- and A+. CDP's scoring system is somewhat similar, where companies' ESG performance is ranked on a scale from D- to A. For the other rating companies, it is observed that the scores are not converted but the score obtained in the scoring is reflected in the rating. Accordingly, the S&P ratings show a value between 0 and 100, the FTSE Russell ratings a value between 0 and 5 and the Bloomberg ratings a value between 0 and 1. An exception to the above is the Sustainalytics scoring system, where a specific methodology for converting scores is also observed. Thus, in the final assessment, 0 represents the best ESG performance, while scores above 40 indicate an extremely high ESG risk. The scoring discrepancies identified suggest that the scores produced by the specific methodologies of ESG rating companies cannot be easily

compared, it is obviously difficult to draw a parallel between a C rating and a 12-point ESG performance. Thus, it is not an easy task to base an investment or financing decision on this. Another difficulty is that some of the rating agencies under review define ESG performance on a relatively narrow scale (0-5), while others rank companies on a much wider scale (0-100) based on their sustainability performance. This raises the question of whether what might be a score of 30-40 in one rating system might be equivalent to a 2 or 3 in another.

It is also worth considering what constitutes the best and worst performance on the scales used by each rating company. In the case of the rating agencies studied, it can be observed that the scale used is associated with progressively worse ESG performance as the scale increases. One of the reasons for this is that Sustainalytics, for example, assesses sustainability performance in terms of the extent to which a company's operations expose it to ESG risk. Thus, the higher the score, the higher the risk. For other ESG rating companies using a numerical scale, such as S&P, FTSE Russell and Bloomberg, higher scores have traditionally been associated with better ESG performance. In contrast, MSCI, Refinitiv, ISS and CDP show an inverse relationship between score and final rating. Accordingly, a higher score is indicated by the letters higher up in the alphabet.

Rating

Besides different scoring scales, some ESG rating companies classify the companies they rate into categories. Among the rating agencies examined, MSCI, Sustainalytics, Refinitiv and ISS have methodologies for making such classifications. For example, MSCI classifies the companies it rates into three categories. Accordingly, it distinguishes between laggards, averages and outperformers based on their ESG performance. Refinitiv's methodology is very similar. In this case, there are no specific categories, as the companies are ranked on a scale from laggards to pioneers. In contrast, Sustainalytics classifies companies into five groups. It is important to stress, however, that this categorisation is about how risky a company is from an ESG perspective. Accordingly, a company can be rated as severe, high, medium, low, or negligible. And according to the ISS rating, a company's ESG performance can be poor, medium, good, or excellent. Presumably, this further categorisation is intended to allow external stakeholders using ESG ratings to better understand the results of the assessment and thus to better rank companies based on their ESG performance. However, in our view, this kind of reassessment of results only makes the ratings produced by individual rating agencies more complex and less transparent.

Topics covered

In addition, another important issue is the main sustainability topics covered by the rating companies with these indicators. Table 3 illustrates that for each of the ESG dimensions, variations can be found in the topics covered.

For the environmental factor, the majority of the rating companies surveyed consider climate change, biodiversity, water security, resources used and waste to be the most relevant. While energy management, air pollution and environmental opportunities are only included in the methodology of a minority of the qualifiers examined.

In terms of the social factor, health, safety, human rights, product stewardship and suppliers are addressed. In other words, the methodologies examined tend to focus on the social problems of the company's internal stakeholders, with less emphasis on external stakeholders.

In the case of the corporate governance aspect, there are much sharper differences than in the previous two dimensions. The table also shows that among the issues examined, the only indicator most widely used by companies that have carried out an ESG rating is that related to management roles. In addition, the topic of business ethics is worth mentioning, which is considered by more than half of the surveyed rating agencies. While taxation is only part of the methodology of the two rating agencies.

Table 3. Topics covered by the rating agencies. Source: Own work

Topic	MSCI	Sustainalytics	Refinitiv	S&P	FTSE Russell	ISS	Bloomberg	S
Climate change	x	x			x	x	x	5
Biodiversity	x	x	x	x	x		x	6
Water	x		x	x	x	x	x	6
Resources used	x	x	x	x	x			5
Energy			x	x		x	x	4
Air pollution (CO2)		x		x			x	3
Waste	x	x	x		x		x	5
Environmental opportunities	x		x	x			x	4
Health and safety	x	x	x	x	x	x	x	7
Workforce development	x		x	x			x	4
Human rights		x	x	x	x	x	x	6
Community		x	x		x		x	4
Product responsibility	x	x	x	x			x	5
Customer protection	x	x		x	x			4
Supply chain	x	x			x	x	x	5
Social opportunities	x			x				2
Vezetői szerepek	x	x	x	x	x		x	6
Independence						x	x	2
Compensation	x		x				x	3
Business ethics	x	x		x		x		4
Corruption		x		x	x			3
Tax	x				x			2
Shareholder rights			x			x	x	3
CSR strategy		x	x	x				3
S	16	15	15	16	13	9	17	

The table also highlights the variation in the number of topics covered among the rating agencies. A rating agency will cover roughly 15 to 16 sustainability-related topics. In contrast, ISS only covers 9 topics. However, it is important to underline that these are mostly the topics that most rating agencies use.

4.2. Numerical analysis of reliability

From the correlation tests carried out, it can be basically seen that for the 40-item sample, there is a relatively strong positive relationship between the Refinitiv and Bloomberg scores, with a positive correlation above $r=0.7$ (Table 4). Extending the analysis to the scores of the other ESG rating companies, such an exact statement can no longer be made. Indeed, the relationship analysis on the 20-item sample paints a mixed picture of ESG raters' ratings. A strong correlation between Refinitiv's and Bloomberg's scores is also found in this case, with a correlation coefficient of $r=0.728$, significant at 1%. A similar result is found for the correlation between Refinitiv and S&P scores with a strength of $r=0.702$. In contrast, the strength of the correlation between Bloomberg and S&P scores is lower than the previous one. Here the correlation is somewhat weaker, with a value of $r=0.644$. A significant relationship of 1% is also observed between the MSCI and Sustainalytics ESG scores. In this case, the relationship is very strong ($r=-0.815$) with a negative direction. This comes from the fact that Sustainalytics measures the ESG riskiness of companies, so for these ratings, the lower the score, the better the ESG performance. In addition to the above, there are two other cases where we can speak of a significant correlation between the ESG scores of the rating companies included in the analysis. In these cases, in contrast to the previous ones, significantly weaker relationships can only be detected at 5%. Accordingly, a negative relationship of strength $r=-0.503$ is observed between Refinitiv and Sustainalytics scores, while a negative relationship of strength $r=-0.524$ is observed between S&P and Sustainalytics scores. Thus, Refinitiv's score is not significantly related to MSCI's score ($r=0.339$) among the ESG rating companies studied, while Sustainalytics is not significantly related to Bloomberg's score ($r=-0.426$). It can also be observed that MSCI scores indicate a significantly strong relationship only for Sustainalytics scores, while this is absolutely not the case for the other ESG rating companies.

Table 4. Results of the correlation analysis at the aggregate score level. Source: Own work

Correlations					
	Ref_ESG	BBG_ESG	SP_ESG	MSCI_ESG	Sustainalytics_ESG
Ref_ESG	1	.728**	.702**	0.339	-.503*
BBG_ESG	.728**	1	.644**	0.304	-.426*
SP_ESG	.702**	.644**	1	0.383	-.524**
MSCI_ESG	0.339	0.304	0.383	1	-.815**
Sustainalytics_ESG	-.503*	-.426*	-.524**	-.815**	1

** Correlation is significant at the 0.01 level (1-tailed).
 * Correlation is significant at the 0.05 level (1-tailed).

Overall, therefore, aggregate scores do not always show a strong positive relationship. This may be due, among other things, to the fact that the ESG rating companies surveyed measure their sustainability performance on different scales. The correlation analysis suggests that rating agencies that use the same scale of measurement for their ESG scores are more strongly correlated. Accordingly, there is a relatively strong significant relationship between Refinitiv, Bloomberg and S&P, which is because each of these rating agencies issues a score between 0 and 100. Similar arguments can be used to explain the strong significant relationship between MSCI and Sustainalytics scores.

A closer look at the scores of the ESG rating companies reveals that, compared to the aggregate scores, further differences can be observed between the scores published by the different rating agencies along the different dimensions of ESG. However, it is also important to highlight that in this case, in contrast to the above, data were available for only three ESG rating agencies (Refinitiv, Bloomberg, S&P) to allow the relationship between scores by pillar to be examined.

For the environmental dimension, there is still a strong ($r=0.732$) significant relationship between Refinitiv and Bloomberg scores at the 1% confidence level (Table 5). In contrast, the relationship between the Refinitiv and the S&P scores is only significant at a medium level ($r=0.460$). While there is no significant relationship between Bloomberg and S&P scores.

Table 5. Results of the correlation analysis on the environmental dimension. Source: Own work

Correlations			
	Ref_Env	BBG_Env	SP_Env
Ref_Env	1	.732**	.460*
BBG_Env	.732**	1	0.441
SP_Env	.460*	0.441	1

** Correlation is significant at the 0.01 level (2-tailed).
 * Correlation is significant at the 0.05 level (2-tailed).

The social dimension shows somewhat weaker links than the previous ones (Table 6). The correlation between the Refinitiv and Bloomberg scores is only medium ($r=0.565$), while the relationship between the Refinitiv and S&P scores is strong ($r=0.695$). In contrast, the relationship between the Bloomberg and S&P ESG scores is medium ($r=0.622$). In all three cases, the relationship is significant at the 1% level.

Table 6. Results of the correlation analysis on the social dimension. *Source: Own work*

Correlations			
	Ref_Soc	BBG_Soc	SP_Soc
Ref_Soc	1	,565**	,695**
BBG_Soc	,565**	1	,622**
SP_Soc	,695**	,622**	1

** . Correlation is significant at the 0.01 level (2-tailed).

When looking at the corporate governance dimension, it appears that this aspect has the weakest correlations between the ratings of companies with ESG ratings among the three aspects of ESG (Table 7). It is important to note, however, that even in this case, there is a significant medium-strength relationship between Bloomberg and S&P scores, with a confidence level of 1%. In contrast, although there is a significant relationship between Refinitiv and S&P scores, this relationship is characterised by a much weaker correlation of 0.458. There is no significant relationship at all between Refinitiv and Bloomberg ESG scores.

Table 7. Results of the correlation analysis on the corporate governance dimension. *Source: Own work*

Correlations			
	Ref_Gov	BBG_Gov	SP_Gov
Ref_Gov	1	0,270	,458*
BBG_Gov	0,270	1	,617**
SP_Gov	,458*	,617**	1

*. Correlation is significant at the 0.05 level (2-tailed).
 **. Correlation is significant at the 0.01 level (2-tailed).

Overall, the correlation analysis along each ESG dimension shows that of the three segments, it is the environmental factor that shows a relatively strong correlation between the scores of the rating companies. This may be due to the fact that environmental responsibility and measures taken to protect the environment have played a major role in business for many years. For example, if we look at carbon emissions and the regulations that apply to them, the various EU standards, companies are under some regulatory pressure to monitor their environmental impact. And this kind of development could also have a positive impact on ESG scores through a more accurate measurement of the environmental dimension with more indicators, i.e., rating companies may now have a wider range of information at their disposal. This is not at all the case for the social dimension, where the very quantification of a social factor can be problematic, so it is not surprising that the relationship between the rating agencies surveyed is somewhat weaker. In the case of the corporate governance segment, this phenomenon is particularly true. As we have seen in the methodological analysis, this is the dimension where the range of sustainability-related issues that ESG rating companies attempt to cover at all is the widest. Thus, there is no question that if different topics are associated with the corporate governance aspect, the relationship between scores will be much weaker.

Perhaps even more interesting than the previous questions is how the relationship between the scores reported by each ESG rating company evolves in certain industries. This is also important because, as we have seen in the chapter on rating methodology, in many cases rating companies include specific, industry-specific elements in their ratings, and it may be worth considering the extent to which they arrive at the same result as a result. The linkage analysis by industry was carried out on the original sample of 40 items. This has the advantage that 5 to 5 companies from each industry are equally represented in the sample, so that no industry is over-represented. However, one of the main disadvantages of using the 40-item sample is that it only allows for a comparison of Refinitiv and Bloomberg scores. However, it is also worth pointing out that I found a stable and strong relationship between the scores of

these two rating agencies in most of the cases presented above, so it may be interesting to see how this result evolves at the level of individual industries.

The link analysis shows that only three of the eight pre-selected industries (energy, finance, health) show a significant link. And this relationship is typically strong, with correlation coefficients of $r=0.9$ or above. This result can be explained by the fact that the energy, financial and healthcare industries are more important from a sustainability perspective and that the nature of these activities may provide a number of industry-specific indicators that can help ESG rating companies to better assess the sustainability performance of companies in these industries.

In addition to the aggregate scores, it is also worth looking at the relationship between the ratings given by the rating agencies in the industries under review along each ESG dimension.

The results of the correlation calculation carried out show that within the environmental dimension, only one industry, the energy sector, shows a significant and rather close ($r=0.948$) relationship between the Refinitiv and Bloomberg scores. These results may be explained by the fact that the environmental factor is of major importance in the energy sector, and it is therefore assumed that companies have built up a set of indicators to measure the environmental impact of their operations as accurately as possible. This may in turn contribute to a similar assessment of the sustainability performance of companies in the energy sector by the individual rating companies. For the whole energy industry, however, the correlation study has produced some surprising results. It was previously thought that since the healthcare sector, particularly pharmaceuticals, has a significant environmental burden, the ESG raters' scores would be heavily weighted towards this area. The results show, however, that the ESG rating companies surveyed have a very mixed view of the environmental performance of companies in the healthcare sector, which may be due to several methodological differences.

Also, for the social dimension, the relationship between Refinitiv and Bloomberg scores was found to be significant for a single industry, the financial sector, with a correlation coefficient of $r=0.966$. Thus, it can be stated that the relationship is very strong and positive. The significant relationship for the financial sector can be explained by the fact that financial institutions in this industry are, by the nature of their activities, in constant contact with society, as they need the trust of their customers to operate successfully. It can therefore be assumed that rating companies attach greater importance to the social factor along these lines and arrive at similar results when assessing the sustainability performance of companies. However, it is also important to look at the results for the health sector. Although in this case there is a strong positive correlation between the scores of companies with ESG ratings, the correlation is not significant. This is interesting because the social importance of the health sector is undoubtedly unquestionable. This raises the question of how it is possible that the scores of individual rating agencies do not show a significant relationship. Presumably, this is also a case of methodological differences in the assessment.

As regards the corporate governance dimension, it can be concluded that only the energy sector shows a significant correlation between the Refinitiv and Bloomberg scores, with a strength of $r=0.920$. The results, therefore, show that, as with the aggregate scores, it can be observed that, across the three ESG segments, the corporate governance dimension is the one where the surveyed rating agencies show a rating inconsistency.

As several studies have shown that the measurement of corporate sustainability performance, and hence ESG scores, can be influenced by company size and geography in addition to general financial and non-financial indicators, I have also examined the validity of these assumptions. These relationship tests were also carried out on the 40-item sample.

For the size of the companies, both Refinitiv ($r=0.578$) and Bloomberg ($r=0.499$) found a significant relationship of medium strength between the value of the company's assets and the aggregate ESG scores. This may be because larger, better-capitalised companies tend to have better information systems and thus rating companies can paint a more accurate picture of a company's sustainability performance. In addition, linked to company size, larger companies may be able to exert pressure on rating agencies due to their market power or bargaining power, to ultimately obtain a more favourable ESG score. So, hypothesis H1 can be accepted.

In terms of geographic location, no significant relationship between the continent of a company's headquarters and its aggregate ESG score is found for either Refinitiv or

Bloomberg. In addition, the results obtained for the two raters studied show a contradiction. Indeed, the Refinitiv scores indicate a positive relationship of less than $r=0.1$, while the correlation coefficient for Bloomberg is $r=-0.041$, which essentially indicates a very weak negative relationship. So, hypothesis H2 should be rejected.

5. Limitations and future directions

The results of my research provide sufficient incentive to carry out further studies. On the one hand, it may be worthwhile to expand the number of rating agencies included in the study, as the present research typically examines the reliability of ESG scores published by Refinitiv and Bloomberg, and only at the aggregate score level was it possible to examine the scores of three additional rating agencies, S&P, Sustainalytics and MSCI. This is also a major limitation of the research, as the data available for each of the rating agencies and the accessibility of the scores was very limited, which may have an indirect impact on the results of the research, and it may be worthwhile to extend the study to additional rating agencies in the future. A further opportunity lies in the industries included in the study, where it may be worthwhile to investigate the reliability of sustainability scores of companies operating in the construction sector.

Most companies today use the GRI standards to report on their sustainability performance. Thus, a future direction could also be to examine the extent to which the methodology of individual rating agencies is in line with GRI recommendations.

6. Conclusions

Overall, it can therefore be concluded that the scores published by each ESG rating company do not allow for a reliable comparison of companies based on their sustainability performance. The results of the relationship analysis show that for the aggregate scores, there are some raters whose scores are closely and significantly related to each other. However, this can be misleading, as correlations along each aspect of ESG show that only the environmental segment shows a strong significant relationship between scores, while the social segment shows a medium correlation and the corporate governance segment a weak correlation, indicating differences in the rating agencies' scoring methodologies. The comparability of the scores is further undermined by the fact that the correlation analysis by the industry sector also only yielded significant results in one case. These findings therefore suggest that the scores published by the ESG rating companies under review do not necessarily provide a reliable basis for investment or financing decisions. The research findings highlight that a reliable comparison of ESG scores also requires an understanding of the assessment methodology used by individual rating agencies. Accordingly, it would be useful for investors to give preference to the ratings of those rating agencies whose methodologies are best suited to the expectations and preferences of the investor.

Funding: This research received no external funding.

Conflicts of Interest: The author declares no conflict of interest.

References

- Berg, F., Kölbel, J. F., & Rigobon, R. (2022). Aggregate Confusion: The Divergence of ESG Ratings. *Review of Finance*, 26(6), 1315–1344. <https://doi.org/10.1093/rof/rfac033>
- Capizzi, V., Gioia, E., Giudici, G., & Tenca, F. (2021). The divergence of ESG ratings: An analysis of Italian listed companies. *Journal of Financial Management, Markets and Institutions*, 9(2), 2150006. <https://doi.org/10.1142/S2282717X21500067>
- Chatterji, A., Durand, R., Levine, D. I., & Touboul, S. (2016). Do ratings of firms converge? Implications for managers, investors and strategy researchers. *Strategic Management Journal*, 37(8), 1597–1614. <https://doi.org/10.1002/smj.2407>
- Delmas, M. A., Etzion, D., & Nairn-Birch, N. (2013). Triangulating Environmental Performance: What Do Corporate Social Responsibility Ratings Really Capture?. *Academy of Management Perspectives*, 27(3), 255–267. <https://doi.org/10.5465/amp.2012.0123>

- Dimson, E., Marsh, P., & Staunton, M. (2020). Divergent ESG Ratings. *The Journal of Portfolio Management*, 47(1), 75–87. <https://doi.org/10.3905/jpm.2020.1.175>
- Dorfleitner, G., Halbritter, G., & Nguyen, M. (2015). Measuring the level and risk of corporate responsibility – An empirical comparison of different ESG rating approaches. *Journal of Asset Management*, 16(7), 450–466. <https://doi.org/10.1057/jam.2015.31>
- Doyle, T. M. (2018). *Ratings that don't rate: The subjective world of ESG ratings agencies*. American Council for Capital Formation. <https://corpgov.law.harvard.edu/2018/08/07/ratings-that-dont-rate-the-subjective-world-of-esg-ratings-agencies/>
- Drempetic, S., Klein, C., & Zwergel, B. (2020). The Influence of Firm Size on the ESG Score: Corporate Sustainability Ratings Under Review. *Journal of Business Ethics*, 167, 333–360. <https://doi.org/10.1007/s10551-019-04164-1>
- Gyönyörová, L., Stachoň, M., & Stašek, D. (2021). ESG ratings: relevant information or misleading clue? Evidence from the S&P Global 1200. *Journal of Sustainable Finance and Investment*. <https://doi.org/10.1080/20430795.2021.1922062>
- Kocmanová, A., Dočekalová, M. P., & Šimanavičienė, Ž. (2017). Corporate Sustainability Measurement and Assessment of Czech Manufacturing Companies Using a Composite Indicator. *Engineering Economics*, 28(1), 88–100. <https://doi.org/10.5755/ie.28.1.15323>
- Li, F., & Polychronopoulos, A. (2020). *What a difference an ESG ratings provider makes*. Research Affiliates Publication. <https://www.researchaffiliates.com/content/dam/ra/publications/pdf/770-what-a-difference-an-esg-ratings-provider-makes.pdf>
- Liang, H., & Renneboog, L. (2020). Corporate social responsibility and sustainable finance: A review of the literature. *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.3698631>
- Lippai-Makra, E. (2020). A nem pénzügyi jelentéstétel szabályozása Magyarországon. In: Kovács, T., Szóka, K. (Eds.) XIV. Soproni Pénzügyi Napok: „Gazdaságvédelem és pénzügyi kiutak” pénzügyi, adózási és számviteli szakmai és tudományos konferencia. Sopron: Soproni Felsőoktatásért Alapítvány (2020). pp. 113-125., 13 p.
- Semenova, N., & Hassel, L. G. (2015). On the Validity of Environmental Performance Metrics. *Journal of Business Ethics*, 132(2), 249–258. <https://doi.org/10.1007/s10551-014-2323-4>
- Utz, S. (2019). Corporate scandals and the reliability of ESG assessments: Evidence from an international sample. *Review of Managerial Science*, 13(2), 483–511. <https://doi.org/10.1007/s11846-017-0256-x>
- Wong, C. (2018). *Rate the Raters 2018: Ratings Revisited*. The Sustainability Institute. https://www.sustainability.com/globalassets/sustainability.com/thinking/pdfs/sa-ratetheraters_ratings-revisited_march18.pdf