Centro UC
CLAPES UC
Centro Latinoamericano de Políticas Económicas y Sociales

ESG Ratings: An Industry in Need of a Major Overhaul

Arturo Cifuentes

Ventura Charlin

Jorge Alfaro

www.clapesuc.cl

Documento de Trabajo Nº 111 (marzo 2022)

ESG Ratings: An Industry in Need of a Major Overhaul

Arturo Cifuentes

Senior Research Associate, CLAPES–UC, Pontificia Universidad Católica de Chile, Alameda 440, Piso 13, Santiago, Chile; email: ac@arturocifuentes.com

Ventura Charlin

Principal, V.C. Consultants, 230 E. 73rd Street, Suite 7D, New York, NY, 10021, United States; email: ventcusa@gmail.com

Jorge Alfaro

Research Associate, CLAPES–UC, Pontificia Universidad Católica de Chile, Alameda 440, Piso 13, Santiago, Chile; email: jealfaro@uc.cl

Abstract

The impetus for adopting ESG-sensitive investment policies has increased steadily since 2006, when the United Nations outlined its Principles of Responsible Investment (PRI). More recently, a new industry aimed at helping investors to make sound ESG-driven decisions has flourished: ESG rating agencies. We investigated the ratings provided by four leading rating agencies (ISS, MSCI, S&P, and Sustainalytics) to the companies included in the S&P500 index. Using a number of measurement theory techniques, we concluded that ESG ratings currently exhibit an abysmally low level of reliability (18.3%) and agreement (5.4%). This situation differs sharply not only with the levels of reliability and agreement found in credit ratings, but also with the reliability and agreement found in areas in which subjectivity plays an important role, for example, wine ratings and clinical psychology diagnosis. These findings challenge the claim that ESG ratings can be helpful to make investment decisions. Moreover, it strongly suggests that the ESG ratings industry as a whole is in much disarray.

Keywords: ESG; ESG ratings; ESG investments, ratings reliability, ratings agreement, socially responsible investments, rating agencies

JEL classification codes: C10; G24; M14; Q56

1. Introduction

In 2006, under the leadership of Kofi Annan, the United Nations (UN) introduced the Principles of Responsible Investment (PRI), a set of six general guidelines aimed at incorporating E(nvironmental), S(ocial), and G(overnance) factors into investment-decision processes (United Nations, 2006). Strictly speaking, the idea of introducing ethical considerations into investment decisions is not a new concept. In fact, the Bible, the Jewish law, and the Islamic tradition, all make references to the importance of making ethical investment decisions. And the concept of Socially Responsible Investments (SRI), which is closely related to ESG factors, appears to have been already expressed at least as back as the 18th century, albeit in a less formal fashion, by the quakers. Whatever its origins, the fact of the matter is that following the UN initiative, a growing number of investors have chosen to adhere to the PRI. As of this writing, more than 7,000 corporate entities in 135 countries have become signatories to these principles; and currently, ESG assets are estimated at $35 trillion, roughly one-third of global total (Henze and Boyd, 2021).

Notwithstanding its conceptual clarity (e.g., "*we will incorporate ESG issues into investment analysis and decision-making process*"), the reality is that the PRI are vague in relation to particulars. Not surprisingly, different investors have adopted different –and sometimes conflicting– interpretations of what ESG compliance means. For example, Vanguard, a U.S. investment advisor, under the S-umbrella, considers whether a company is involved in the tobacco and opioid sectors, two industries that invite almost universal scorn. But Vanguard also includes adult entertainment and casinos, two sectors that in general people find less distasteful (Grim and Berkowitz, 2018). The CFA Institute, on the other hand, makes no reference to any of these industries but considers customer satisfaction and protection of data privacy as key S-considerations; two elements not mentioned by Vanguard (Hayat and Orsagh, 2015). The

Governance and Accountability Institute (G & A) includes many factors under the G-umbrella, although nothing is related to lobbying activities, an issue which appears to be import for the CFA Institute (G & A, 2021). In short, although there is much common ground in reference to what ESG entails –at least in theory– a precise definition is still lacking. Additionally, many of the factors involved in ESG assessments are qualitative in nature (e.g., commitment to community relations) and therefore involve some degree of subjectivity; it is not then surprising that investors feel often confused when it comes to implementing ESG initiatives (Mackintosh, 2022; Cheng, 2022). And to these considerations we need to add the fact that carrying out ESG-due diligence is not free: it involves spending time and resources that are certainly beyond the scope of retail investors and even for large institutions can be taxing.

Let us consider now the ESG rating agencies (RA). Clearly, these RAs came into being with the goal of providing much needed guidance to investors. There are currently 125 organizations providing ESG ratings and research (Paul Weiss, 2021); and although many ESG RAs are niche players and/or operate domestically, many offer global coverage. To name a few: Asset4, Bloomberg, CDP, Corporate Knights, FTSE Russell, ISS, MSCI, S&P, Refinitiv, RepRisk, Sustainalytics, Thomson Reuters, Vigeo Eiris.

Considering the important role played by the credit RAs in the fixed income market, it is reasonable to ask ourselves if the ESG RAs could eventually achieve a similar level of relevance. Recall that the three leading credit RAs (Fitch, Moody's, and S&P) dominate the global bond market as their ratings are entrenched in both, the regulatory framework and the investment policies of most institutional investors. Notice that their ratings seldom disagree, and they all use a similar (equivalent) rating scale. And more relevant, their language and symbols (e.g., investment grade, non-investment grade, AAA, BBB) have become the lingua franca of the fixed

income market. All this begs the question: are the ESG RAs in a position to achieve a similar role in the ESG investment arena?

When it comes to ESG-conscious investing, most investors seem to prefer a policy based on integration rather than exclusion; i.e., instead of outright eliminating certain companies from the set of investment options, they favor rules based on meeting on average some meaningful portfolio-level ESG criteria. For instance: don't invest in a company whose ESG score is below certain threshold. Or, maintain an overall ESG score above a given limit. In principle, ESG ratings would be a natural match to facilitate the implementation of such policies. This, of course, as long as ESG ratings themselves meet certain minimum requirements. What are those requirements?

Measurement theory, a body of knowledge developed to assess the validity and reliability of different instruments, can be helpful to evaluate the potential usefulness of ESG ratings. In fact, measurement theory has been successfully used in the health sciences, psychology, educational testing, and other disciplines to address questions similar to the questions faced by ESG investors, namely, can ESG ratings be trusted? Strangely enough, the concepts and tools employed in this field have been notoriously absent in discussions regarding the suitability or usefulness of ESG ratings. Our paper is an attempt to remedy this situation. And it represents an effort to show that bringing measurement theory into the ESG ratings conversation can add much clarity to the discussion. More specifically, we seek to evaluate the reliability and agreement (defined more formally later) among the different ESG RAs.

2. Literature Review

The academic literature has paid lots of attention to the performance of portfolios based on ESG-criteria. Friede et al., (2015) summarized the findings of more than 2,000 studies and offered a good overview of this topic. Interest on this topic has not abated as it is evident by the many recent

articles addressing different aspects of this issue (e.g., Steen et al., 2019; Dolvin et al., 2019; Engelhardt et al., 2021).

However, ESG ratings have received much less attention. In fact, initially many authors did not even use the term "rating" and simply referred to "ESG measures", "environmental performance metrics" and the like. One of the first papers to look into this topic recognized the fluidity of the term "corporate social responsibility" and the difficulties associated with its assessment (Hill et al., 2007). Sandberg et al. (2009), who preferred the term SRI, addressed the heterogeneity implicit in this definition and the potential benefits of introducing some standardization. They remained skeptical, nevertheless, regarding the likelihood of achieving this goal due to, among other things, different cultural values among the many stakeholders. Delmas et al. (2013) focused on what they termed "corporate environmental performance" and attempted to identify the factors that explained most of the variance in relation to this metric. Semenova and Hassel (2015) looked at the "environmental performance metrics" informed by different companies and concluded that in general they do not converge even though they are supposed to be based on similar factors. Doyle (2018), who explicitly used the term ESG ratings, has argued that in general these ratings tend to favor large companies and firms domiciled in Europe instead of the U.S. Walter (2020) argued vigorously that much needs to improve in the ESG ratings industry. Among his suggestions: the necessity to create metrics tied to a normative improvement (what are the outcomes that one seeks to achieve?), and the need to establish a certification procedure akin to what exists in the credit ratings arena.

In relation to the disagreement observed in ESG ratings, it seems that the first authors to deal explicitly (and more quantitively) with this problem were Chatterji et al. (2015) and Dorfleitner et al. (2015). Chatterji et al. did not use the term "ESG", they preferred the term "corporate social

responsibility" and "socially responsible investments". They looked at the ratings provided by six RAs (Asset4, Calvert, DJSJ, FTSE4Good, Innovest, and KLD) and concluded that they exhibited little correlation among them (values between -0.10 and 0.40). They attributed the lack of "ratings convergence" (their term) to the fact that different RAs were using different methods to evaluate the same construct. Dorfleitner et al. (2015) arrived at similar conclusions by looking at the correlation between ratings and their respective distributions which they considered very different.

An MIT Sloan school study, which took as a given the lack of agreement among the RAs, attempted to identify the causes behind this phenomenon (Berg et al., 2022). The authors concluded that the main reason for the divergence was also that different RAs were using different approaches to assess the performance of a firm under the same category. Slightly less important was what the authors described as "scope divergence", namely, the fact that the RAs considered different factors when assessing the E, S, or G merits of a firm. The study was based on ratings from six RAs (Asset4, KLD, MSCI, RobecoSAM, Sustainalytics, and Vigeo Eiris).

Dimson et al., (2020), based only on ratings from three RAs (FTSE Russell, MSCI, and Sustainalytics) reported several cases in which a company was rated very highly by one RA and at the bottom of the scale by another RA. Billio et al., (2020), based on ratings by MSCI, Refinitiv, RobecoSAM and Sustainalytics, also confirmed the heterogeneity of ratings reported by previous studies. These authors, to their credit, employed rank correlations (a more appropriate correlation metric than Pearson's correlation since ratings are based on ordinal scales) and also reported what they called "percentage of observed agreement," which at 24% they judged it to be very low. Regrettably, they did not correct this metric for chance agreement. These authors also suggested that the disagreement in ratings was due, in part, due to the lack of a common definition of ESG (i.e., different RAs measure different things).

Another study, curiously concluded that larger ratings discrepancies were positively correlated with higher returns (Brandon et al., 2021); this finding is rather tragic for it could encourage investors to avoid firms that are unanimously rated highly on ESG metrics. Also, intriguing is the conclusion of another recent article on ESG ratings disagreement: greater level of ESG disclosure, results in greater ESG rating disagreement (Christensen et al., 2022); this somehow counterintuitive conclusion is unfortunate for it seems to indicate that transparency –something most regulators push for– appears to contribute to create confusion.

A more recent paper by Gyönyörová et al. (2021) however, stands out, in our view, due to two distinctive features. First, it provides the most complete review of the literature on ESG ratings that we have seen. And second, and more important, it addresses the ESG ratings disagreement using an approach which differs substantially from all previous studies. These authors based their analysis on the companies in the S&P1200 index and five RAs (Bloomberg, CDP, ISS, RobecoSAM and Sustainalytics). They relied first on an exploratory analysis using principal axis factoring and oblique rotation, which they then complemented with a confirmatory factor analysis using out-of-sample data. They concluded that the ratings did not exhibit convergence validity (i.e., different RAs were not measuring the same construct). And suggested that investors would be better off using several ratings simultaneously.

The financial press and some business outlets have also reported concerns regarding the lack of agreement among ESG ratings. Typically, these articles show examples in which the ratings diverge a lot and/or estimate some correlation between the ratings. See, for instance, Doyle (2018), Wigglesworth (2018), Temple-West (2019), The Economist (2019), Matos (2020), Moore (2020), Nauman (2020), Paul Weiss (2021), Prall (2021), Tarnavsky (2021), Mackintosh (2022) and Schwartzkopff (2022).

In brief, previous studies have found that ESG ratings show a great deal of discrepancy depending on the RA used, and have attributed this phenomenon to the fact that different RAs measure different things with different methods, and then combine them using different weights. Yet, notwithstanding their merits and the commonality of their findings, many of these studies suffer from a few shortcomings.

First, most studies used correlation as a proxy for agreement. However, we should note that if a RA gives consistently lower ratings than a competitor (e.g., always two notches below), the correlation will be one. This, of course, despite the constant disagreements in their ratings. Hence, correlation is not a good metric to assess ratings agreement. And second, ratings, despite the fact that they are normally expressed with numbers (MSCI is one of the few exceptions), represent ordinal categories and not quantitative scales. (A quantitative scale is such that the distances between all neighboring categories are the same.) Under these circumstances a Pearson's correlation is not the right correlation to employ. Nevertheless, many authors have failed to consider this issue. Moreover, normalizing rating scales to distributions with mean zero and standard deviation equal to one, is a dubious practice when dealing with ordinal scales. Finally, and more relevant, none of the previous studies –Gyönyörová et al. (2021) is the exception– relied on any of the techniques employed by measurement theory to address the problem at hand. Most just reported correlations and, in some cases, a few ad hoc metrics without any reference to statistical significance. On the other hand, it is reasonable to assume that the magnitude of the ratings disagreement reported is a valid cause of concern.

To summarize, the idea of using measurement theory concepts to assess the degree of agreement and reliability of ESG ratings is warranted. We think this study would be a nice complement to the paper by Gyönyörová et al. (2021) which focused on validity.

3. Data and Methods

For this study we considered all the companies included in the S&P500 index as of January 2022 (501 in total). We reasoned that in terms of detecting anomalies in ESG ratings, these companies were a better choice (tougher test) than a much bigger group that included less liquid, and possibly less scrutinized, names.

In terms of ratings, we decided, in principle, to consider all the ratings provided under the ESG banner in Bloomberg terminals. The rationale was straightforward: most (if not all) market participants obtain their information via Bloomberg. A rating not accessible via Bloomberg, we think, is likely not to have much impact on investment decisions. This left us with five ESG ratings (although their providers often use a slightly different terminology, e.g., scores): Bloomberg, ISS, MSCI, S&P, and Sustainalytics. Notice that S&P acquired RobecoSAM in 2019; after that date the name RobecoSAM was dropped and the ESG ratings became part of the S&P platform. After a first inspection we detected that Bloomberg ESG score unfortunately covered only 66% of the names in the S&P500 (332 out of 501) while ISS, MSCI, S&P, and Sustainalytics covered, respectively, 492, 476, 501, and 489 of the 501 names in the S&P500. Hence, we decided to drop Bloomberg and carry out the analyses with the four remaining RAs.

This choice might seem questionable given the huge number of RAs operating in this space. However, these four names appear most consistently in all studies and, more important, MSCI and Sustainalytics have been identified as the sources most frequently used by institutional investors (see Wong and Petroy, 2020). Paul Weiss (2021) also cites these four RAs among the most commonly used. Another reputable report, this from the Harvard Law School Forum on Corporate Governance, includes MSCI, S&P and ISS among the most well-known providers of ESG ratings

(Moy Huber and Comstock, 2017). Thus, our choice of RAs to study ESG ratings seems reasonable.

We should also note that these RAs all use different ordinal scales. Thus, before doing any analyses it was necessary to transform all the ratings from the original (raw) scale to a common seven-ordered-category (transformed) scale. (Note: MSCI employs only seven categories.)

ISS's scale, which goes from 1 (best) to 10 (worst), presented some challenges due to the seven-ten mismatch. We opted for the following transformation: (1, 2) → (1); (3) → (2); (4) → (3); (5, 6) → (4); (7) → (5); (8) → (6); (9, 10) → (7).

MSCI employs a seven-category scale (AAA, AA, A, BBB, BB, B, and CCC), in which AAA is best and CCC is worst. For convenience, we coded these categories from 1 (AAA) to 7 (CCC).

S&P ratings go from 0 (worst) to 100 (best). Thus, we needed to reverse the order of the scale (subtracting from 100 the current rating) to make it compatible with the direction of the other scales; then, we divided them into seven buckets, namely: [0-13], [14-27], …, up to [84-100] in which [0-13] is best.

The scale used by Sustainalytics (S-A in what follows) goes from 0 (best) to 45 (worst). Hence, the seven-category scale became: [0-11], [12-16], [17-21], …, [32-36], [37-45].

MSCI, S&P, and S-A ratings, when plotted in histogram-form, display a unimodal distribution. Curiously enough, ISS ratings follow a uniform distribution. This suggests that these ratings were allocated by benchmarking the companies in relation to each other, which resulted in an equal number of firms in each bucket.

Table 1 summarizes the data. The industry sectors correspond to the Standard Industry Classification (SIC) sector codes. The first line in the table (percentage in the top four categories) attempts to capture what in the fixed income market would be described as "investment grade." In this context, it amounts to a pass, in a pass/fail distinction based on ESG criteria.

Table 1 Basic Data: Descriptive Information

Variable	Companies Rated by:			
	ISS	MSCI	S&P[a]	S–A
% Companies Rated in the Top 4 Categories	60.4	83.6	81.6	75.1
Average Raw Rating	5.5	3.4	33.8	21.9
Average Scale-Transformed Rating	4.0	3.4	3.0	3.5

SIC Sector	Number of Companies Rated by:				Market Capitalization[b] (in billions)
	ISS	MSCI	S&P	S–A	
Construction	7	6	7	7	$156
Finance, Insurance, Real Estate	97	93	98	96	$6,345
Manufacturing	196	195	201	197	$17,975
Mining	14	13	15	13	$468
Retail Trade	32	31	32	32	$4,401
Services	75	71	76	73	$10,514
Transportation & Public Utilities	61	57	62	61	$3,072
Wholesale Trade	10	10	10	10	$237
Total	492	476	501	489	$43,168

[a] The rating scale was reversed for this RA

[b] This is the total market capitalization of all the S&P500 companies in the sector

Table 2 reports the correlations between the ratings of any two RAs, based on Kendall's Tau-b, an appropriate metric to assess the strength and direction of association when dealing with ordinal scales. The lower triangle shows the correlation based on the transformed scales (that is, after collapsing all the scales into a seven-category scale); the upper triangle shows the correlation based on the raw data (original scales). Two observations are in order. First, the similarity between both correlation indicates that the transformation applied to have one common scale (with seven

categories) did not introduce undesirable distortions. And second, the low correlation figures (average ~ 16%), hints, although not conclusively, to potential disagreements among the four RAs.

Table 2 Kendall Tau–b Correlation Coefficients between Raw Ratings (Upper Triangle) and Scale–Transformed Ratings (Lower Triangle)

	ISS_R	$MSCI_R$	$S\&P^a_R$	$S–A_R$
ISS	1	0.209	0.103	0.056
MSCI	0.208	1	0.251	0.151
S&P	0.112	0.270	1	0.146
S–A	0.060	0.165	0.166	1

[a] The rating scale was reversed for this RA

Note: The average Kendall Tau–b Correlations are 0.156 and 0.160 for the Upper Triangle and the Lower Triangle respectively

The goal of this project was to examine the extent to which the ratings given by the four RAs are equivalent. "Equivalent," of course, is not a well-defined concept. This situation, nevertheless, is analogous to a situation frequently encountered in measurement theory and biostatistics, namely, a case in which several judges are asked to offer a diagnosis based on the same information. There are several tests that, taken together, can be used to answer this question. Each one, based on a different criterion, illuminates a different aspect of the problem. These criteria are: (i) inter-rater reliability; (ii) inter-rater agreement, and (iii) differences of rankings in paired observations. The next section describes the tests, all carried out based on the seven-category scale.

4. Analyses

4.1. Inter-rater reliability

This concept (often confused with inter-rater agreement) refers to the degree to which ratings by different judges (RAs) are similar when expressed as deviations from their means. Alternatively, we can say it refers to the degree to which the order relationship implied by the ratings of one RA

is analogous to the order relationship implied by the ratings of another RA (regardless of the numerical value assigned to the ratings).

The inter-rater reliability can be examined using the R_{SF} coefficient (Shrout and Fleiss, 1979) defined as

$$R_{SF} = \frac{MS_{cr} - MS_{error}}{MS_{cr} + MS_{error}\ (K-1)} \quad [1]$$

where the analysis involves using the standard two-way analysis of variance (ANOVA) to compute the mean square for company ratings (MS_{cr}) and the mean square for error (MS_{error}). These two components are then inserted into the standard equation for reliability where *K* denotes the number of RAs compared. Values closer to 1 indicate a high degree of reliability, whereas values closer to 0 show the opposite. Table 3 offers a more precise semantic interpretation based on the suggestions made by Cicchetti and Sparrow (1981).

Table 3 Semantical Interpretation of Different Agreement/Reliability Levels, based on Cicchetti and Sparrow (1981)

Magnitude of agreement or reliability coefficient	Strength of Agreement (interpretation)
< 0.40	Poor
0.40 – 0.59	Fair
0.60 – 0.74	Good
0.74 – 1.00	Excellent

Table 4 displays in the upper triangular section the R_{SF} values for all six possible two-pair comparisons. The single value in the lower triangle is the overall inter-rater reliability *(K*= 4 in this case).

Clearly, we have a very low degree of reliability. This result indicates that the ordinal relationships implied by the ratings of the four RAs are very dissimilar. In other words, if we want to know

how three companies, say *X, Y* and *Z*, are ranked, based on their ESG merits, we will arrive at very different conclusions depending on which RA we use.

Table 4 Inter–Rater Reliability: Comparisons of Ratings

	ISS	MSCI	S&P	S–A	Average Reliability
ISS	—	0.234	0.151	0.074	0.153
MSCI		—	0.328	0.208	0.257
S&P	**Overall Inter–Rater Reliability among all 4 RAs = 0.183 (N=468)**		—	0.207	0.228
S–A				—	0.163

4.2. Inter-rater agreement

This concept refers to the degree to which two RAs assign the same ratings to the companies considered. Thus, inter-rater agreement also captures the differences between the ratings assigned. Inter-rater agreement can be tested by means of the T_{TW} index coefficient (Tinsley and Weiss, 1975), which is defined as

$$T_{TW} = \frac{N_A - N \times P_C}{N - N \times P_C} \quad [2]$$

where N_A is the number of agreements, N is the number of companies being rated, and P_C is the probability of having a chance agreement on the rating of a given company. Positive (negative) values are associated with levels of agreement that are higher (lower) than the agreement which could have been obtained simply by chance. Values closer to 1 indicate higher agreement (Table 3 also applies). Table 5 shows the T_{TW} values for several comparisons.

The *POA* column is the percentage of observed agreement without correcting for chance agreement. This metric (imperfect because it overestimates the actual value) already suggests a

gloomy degree of agreement. Note that these values are very much in line with the values reported by Billio et al. in the bottom section of Table 3 of their paper (Billio et al., 2020). For example, in our case the *POA* between MSCI and S&P is 22.3%; Billio et al. reported 19.5%.

Table 5 Inter–Rater Agreement: Comparisons of Ratings

[1] RA 1	[2] RA 2	[3] *POA*	[4] T_{TW} (0-discrepancy)	[5] T_{TW} (1-discrepancy)	[6] T_{TW} (Pass/Fail)	[7] κ_W
MSCI	**ISS**	14.3%	–0.001	0.096	0.251	0.121
MSCI	**S&P**	22.3%	0.093	0.372	0.529	0.181
MSCI	**S–A**	20.7%	0.075	0.378	0.383	0.099
S&P	**ISS**	18.1%	0.044	0.087	0.236	0.097
S&P	**S–A**	21.9%	0.088	0.272	0.362	0.123
S–A	**ISS**	16.3%	0.024	0.034	0.157	0.047
	Average		0.054	0.206	0.320	0.111

.

The fourth column is a comparison in which the T_{TW} coefficient was calculated based on a definition of agreement that called for having exactly the same rating (0-discrepancy). As expected, these estimates are much lower than their *POA* counterparties and offer a dismal picture of the agreement level.

The fifth column is based on a more relaxed version of agreement (total coincidence or a one-notch discrepancy). Still, this forgiving version of agreement results is very low T_{TW}'s.

The sixth column attempts to perform the equivalent of an investment grade versus non-investment grade comparison, commonly used in the fixed income arena. Accordingly, we grouped the ratings into two sets: the first set consists of categories 1, 2, 3, and 4, similar to AAA, AA, A, and BBB; the second set gathers the remaining categories, 5, 6, and 7. Again, this comparison, which is only based on the ability to separate the companies in two groups according to ESG merits, e.g., "good"

or “bad,” or “pass” or “fail”, results in very low agreement. Even the S&P-MSCI comparison yields an unimpressive 53%.

The final column shows Cohen’s weighted Kappa (κ_w), another metric to assess agreement. The κ_w is defined as follows

$$\kappa_W = \frac{POA - P_C}{1 - P_C} \quad [3]$$

where *POA* denotes the percentage of observed agreement and P_C denotes the percentage agreement expected by chance alone (Cohen, 1960). The κ_w involves a more relaxed definition of agreement which does not require to have exactly the same rating. In this case we employed a linear weighting scheme (Cohen 1968; Cicchetti and Allison, 1971). Again, values closer to 1 denote higher levels of agreement, values below 0.4 suggest an unacceptable level of agreement (see Table 3). The figures reported are self-explanatory.

Overall, these results have a troubling practical implication: investors wishing to structure an ESG-compliant portfolio based on guidelines built around ESG ratings are likely to end up with very different portfolios depending on which RA they choose to believe.

Table 6 Wilcoxon Pairwise Comparisons

Pairwise Comparisons	# of Companies	Average Rating[a]			Std. Dev. of Rating[a]			Wilcoxon Signed–Rank Test	
		RA 1	RA 2	Gap[b]	RA 1	RA 2	Gap[b]	Signed–Rank Score	p value
MSCI (RA 1) vs. ISS (RA 2)	470	3.37	3.94	–0.57	1.21	2.15	2.16	–12705	<.0001
MSCI (RA 1) vs. S&P (RA 2)	476	3.37	2.93	0.44	1.21	1.54	1.61	11456	<.0001
MSCI (RA 1) vs. S-A (RA 2)	473	3.37	3.48	–0.11	1.22	1.45	1.68	–3112	0.1301
S&P (RA 1) vs. ISS (RA 2)	492	2.96	3.98	–1.02	1.55	2.14	2.44	–19668	<.0001
S&P (RA 1) vs. S-A (RA 2)	489	2.95	3.49	–0.54	1.55	1.46	1.89	–12943	<.0001
S-A (RA 1) vs. ISS (RA 2)	484	3.50	3.98	–0.48	1.45	2.15	2.50	–9429	<.0001

[a] While averages and standard deviations are not valid for ordinal scales, they are presented here for informative purposes only.
[b] A positive gap indicates that RA 1 has been more severe in its ratings; a negative gap indicates that RA 2 has been more severe.

4.3 Ranking differences in paired observations

The Wilcoxon signed-rank test is the nonparametric alternative method to the paired-sample t -test (Wilcoxon, 1945). Paired observations X_1 and X_2 (ratings) are presumed to be drawn at random from a single population; this test makes the additional assumption that the distribution of the differences between X_1 and X_2 is symmetric about zero (null hypothesis). Table 6 describes the six possible RA comparisons.

Let d_j denote the difference in any matched pair of observations, that is, $d_j = X_{1j} - X_{2j}$ for $j = 1, \ldots, N$; where N is the total number of pairs and M is the reduced sample size after excluding the pairs such that $\left| X_{1j} - X_{2j} \right| = 0$.

The Wilcoxon matched-pairs signed-ranks test statistic is defined as

$$W = \sum_{j=1}^{M} sgn(d_j)\ rank(\left| d_j \right|) \qquad [4]$$

Note that since the rating scales are ordinal scales (ranked data), not interval (quantitative) scales, that is, scales with equally-spaced ordered categories, a comparison using a parametric paired-differences t -test would be inappropriate.

With the exception of the MSCI vs S-A comparison, all other paired-comparisons shown in Table 6 yielded significant differences. At first glance, it might appear that these differences –about a half-notch discrepancy in most cases– might not amount to much. But in practice, that is, in building portfolios that need to meet, on average, a certain ESG rating-specified constraint, a half-notch difference is substantial. It can result, again, in very different asset selection decisions. Evidence from the fixed income markets, where portfolio construction is normally constrained by credit ratings-based rules, supports this observation. Notice also that the standard deviation of differences falls in the 1.61−2.50 range. This is consistent with the many instances of huge

discrepancies in ratings reported elsewhere and also observed in our database. For example, Abiomed (a medical devices company) and Molson Coors (a drinks and brewing company) were rated by MSCI and ISS almost at the extreme opposite of their rating scales, AA and 10, and AAA and 9, respectively.

Finally, we also conducted the non-parametric Friedman test of differences (Friedman, 1937) among the rankings given by the four RAs (N=468). The analysis resulted in a Friedman test statistic of 57.61 ($p<.0001$) indicating that there were significant differences among the different RA's ESG ratings and therefore confirming the findings from the pairwise comparisons.

4.4 Industry sector comparisons

To gain additional insight into these ratings differences, we conducted the same analyses just described at an industry level. Table 1 indicates that only four industry sectors have enough observations to carry out the analyses: (a) Finance, insurance, and real estate (finance in what follows); (b) Manufacturing; (c) Services; and (d) Transportation and public utilities (transportation in what follows). We also decided to limit these analyses to MSCI, S&P and S-A and dropped from these comparisons ISS due to its low inter-rater reliability and agreement.

Table 7 Inter–Rater Reliability: Comparisons of Ratings by SIC Sector

SIC Sector	MSCI vs. S&P	MSCI vs. S-A	S&P vs. S-A	Average Reliability by Sector
Finance	0.409	0.156	0.322	0.296
Services	0.424	0.583	0.395	0.467
Manufacturing	0.280	0.217	0.137	0.211
Transportation	0.171	0.055	0.239	0.155

The right-most column in Table 7 suggests an interesting outcome: reliability is much higher (although by no means high) in services and finance than manufacturing and transportation. Table

8 (analogous to Table 5) reinforces this tendency: manufacturing and transportation show in general lower levels of agreement. This is especially clear in the fifth and sixth columns (0- and 1-discrepanycy, respectively).

Table 8 Inter–Rater Agreement: Comparisons of Ratings by SIC Sector

[1] SIC Sector	[2] RA 1	[3] RA 2	[4] *POA*	[5] T_{TW} (0 notch discrepancy)	[6] T_{TW} (1 notch discrepancy)	[7] T_{TW} (Pass/Fail)	[8] κ_W
	MSCI	**S&P**	23.7%	0.109	0.420	0.527	0.231
Finance	**MSCI**	**S–A**	18.3%	0.046	0.508	0.484	0.052
	S&P	**S–A**	32.3%	0.210	0.557	0.500	0.241
Average for Finance sector				0.122	0.495	0.504	0.174
	MSCI	**S&P**	22.5%	0.096	0.471	0.606	0.253
Services	**MSCI**	**S–A**	36.6%	0.260	0.724	0.746	0.335
	S&P	**S–A**	24.7%	0.121	0.485	0.589	0.229
Average for Services sector				0.159	0.560	0.647	0.272
	MSCI	**S&P**	24.1%	0.114	0.372	0.549	0.167
Manufacturing	**MSCI**	**S–A**	18.1%	0.045	0.306	0.223	0.097
	S&P	**S–A**	17.3%	0.035	0.204	0.208	0.063
Average for Manufacturing sector				0.109	0.420	0.527	0.231
	MSCI	**S&P**	12.3%	−0.024	0.197	0.368	0.005
Transportation	**MSCI**	**S–A**	12.3%	−0.024	0.111	0.123	−0.071
	S&P	**S–A**	21.9%	0.055	0.145	0.246	0.109
Average for Transportation sector				0.003	0.151	0.246	0.014

Table 9, analogous to Table 6, shows the results of the Wilcoxon test (paired observations). The previously mentioned tendency is less apparent. Yet we should note that in the finance-services sectors only three paired comparisons are significant whereas in the manufacturing-transportation sectors five are significant. Again, this seems to validate the finding that in the manufacturing-transportation sectors ratings discrepancies are more pronounced than in the finance-services sectors.

Table 9 Wilcoxon Pairwise Comparisons by SIC Sector

SIC Sector	Pairwise Comparisons	# of Companies	Average Rating[a]			Std. Dev. of Rating[a]			Wilcoxon Signed–Rank Test	
			RA 1	RA 2	Gap[b]	RA 1	RA 2	Gap[b]	Signed–Rank Score	p value
Finance	MSCI (RA 1) vs. S&P (RA 2)	93	3.55	3.03	0.52	1.12	1.58	1.49	526	0.0015
	MSCI (RA 1) vs. S-A (RA 2)	93	3.55	2.90	0.65	1.12	1.18	1.49	706	<.0001
	S&P (RA 1) vs. S-A (RA 2)	96	3.04	2.93	0.11	1.57	1.19	1.62	78	0.6091
Services	MSCI (RA 1) vs. S&P (RA 2)	71	3.28	2.99	0.30	1.20	1.64	1.54	210	0.0707
	MSCI (RA 1) vs. S-A (RA 2)	71	3.28	2.79	0.49	1.20	1.01	1.01	314	<.0001
	S&P (RA 1) vs. S-A (RA 2)	73	3.03	2.81	0.22	1.66	1.04	1.52	143	0.2244
Manufacturing	MSCI (RA 1) vs. S&P (RA 2)	195	3.24	2.73	0.51	1.26	1.49	1.66	2068	<.0001
	MSCI (RA 1) vs. S-A (RA 2)	193	3.24	3.88	–0.64	1.26	1.54	1.77	–2773	<.0001
	S&P (RA 1) vs. S-A (RA 2)	197	2.75	3.85	–1.10	1.50	1.54	2.00	–4063	<.0001
Transportation	MSCI (RA 1) vs. S&P (RA 2)	57	3.46	3.14	0.32	1.27	1.49	1.78	129	0.2084
	MSCI (RA 1) vs. S-A (RA 2)	57	3.46	4.05	–0.60	1.27	1.17	1.68	–265	0.0078
	S&P (RA 1) vs. S-A (RA 2)	61	3.16	4.10	–0.93	1.47	1.15	1.63	–381	<.0001

[a] While averages and standard deviations are not valid for ordinal scales, they are presented here for informative purposes only.
[b] A positive gap indicates that RA 1 has been more severe in its ratings; a negative gap indicates that RA 2 has been more severe.

What is so special about the manufacturing-transportation sectors that we observe less reliability and agreement than in the finance-services sectors? Previous authors have argued that differences in ratings can be attributed to the fact that different RAs evaluate the E, S and G components of their ratings by focusing on different elements, which, in turn, they assess using different approaches (Chatterji et al., 2015; Berg et al., 2022). However, we are not aware of any study that has attempted to look at ratings differences based on a sector-by-sector approach. We speculate that in the manufacturing-transportation sectors there are many environmental-related factors (e.g., water management, deforestation, pollution, land contamination, climate change vulnerability) that are largely absent in the finance-services sectors, and whose assessment brings an additional element of subjectivity and/or uncertainty. This is perhaps the reason we see a more acute manifestation of poor reliability and agreement in the manufacturing-transportation sectors. This, however, is only a conjecture.

In summary, taken together, all these analyses paint a dismal level of reliability and agreement among the four RAs studied.

5. Discussion

The previous analyses validate the concerns that have been expressed regarding ESG ratings. In short, at least in the case of the four RAs considered (ISS, MSCI, S&P, and S-A) the level of reliability is very low, the degree of agreement is also very low, and the magnitude of the ratings discrepancies is in general significant. This situation presents a real challenge for investors. In essence, investors who might wish to decide between any two companies based on their ESG ratings, or, more broadly, investors willing to rely on portfolio-level rules and investment policies based on ESG ratings, will arrive at very different conclusions depending on whose ratings they decide to use. These findings, all based on generally accepted and widely used techniques from

measurement theory, paint a troubling picture of an industry that has largely presented itself as instrumental to helping investors to make sound ESG-related decisions.

Although the first ESG RA (Vigeo Eiris) was founded in France in 1983, the ESG ratings industry is relatively young and only started to gather momentum in the last fifteen years. In contrast, by way of comparison, Moody's and S&P, the leading RAs in the credit field, were founded in 1909 and 1860 respectively. In fact, many of the recent acquisitions that have taken place in the ESG ratings industry reflect the typical consolidation that marks a young industry (e.g., S&P bought the ESG ratings business from RobecoSAM in 2019; the same year Moody's bought Vigeo Eiris; and in 2020 Morningstar acquired Sustainalytics). Unfortunately, consolidation in other areas, namely, the adoption of compatible (equivalent) rating scales, an agreement on what factors should be captured by the ESG rating, or even the precise meaning of an ESG rating, has not yet arrived. This has also contributed to the confusion investors have expressed in relation to these ratings. This situation is in clear contrast with the credit-ratings industry in which all RAs communicate their ratings using comparable scales (e.g., AAA/Aaa; AA/Aa, A/A, BBB/Baa) even if their approach to determine the ratings are different.

There is, however, something more problematic about the ESG ratings industry that cannot be attributed to its youth: the element of subjectivity implicit in any ESG assessment. A look at other industries in which ratings (that is, opinions given by different judges) play a role is illuminating. Consider an article which summarizes several findings from different disciplines (Ashton, 2012). The study reported that on average reliability in wine ratings studies was around 0.50, and agreement was a mere 0.34. In other disciplines such as meteorology, business, auditing, personnel management, medicine, and clinical psychology reliability was in the 0.70-0.91 range while

agreement fluctuated between 0.49 and 0.75. Somehow expected, reliability and agreement were highest in meteorology and lowest in clinical psychology.

The Mexican fixed income market also provides an interesting comparison. Credit ratings in this market are dominated by the three U.S. leading RAs (Fitch, Moody's, and S&P). A study reported that reliability of ratings was around 0.90 while agreement, using a seven-category scale similar to the one employed for the comparisons described in Table 5 (column 4), resulted in values in the 0.62-0.78 range (Charlin and Cifuentes, 2017). (We are not aware of any similar study in the U.S. bond market. Anecdotal information, however, suggests that in the U.S. market reliability and agreement are comparable to that of the Mexican bond market.)

It seems clear that ESG ratings –with an average reliability of 18.3% (Table 4) and agreement level of 5.4% (Table 5, fourth column)– are dismal even by wine industry standards. This should not surprise: in the ESG ratings arena, it has not yet been possible to establish a clear link or a causal effect between any specific ESG factor and a measurable financial performance metric (Damoradan, 2020). In the credit ratings field, on the contrary, RAs devoted themselves to assess only one element (creditworthiness) which, ultimately, can result in a very clear-cut outcome: default. Not only that, but there is also a much deeper consensus and understanding on what are the factors that affect a company creditworthiness, namely, debt-to-equity ratios, debt-to-free cashflow ratios, etc., all quantities that involve a much lesser degree of subjectivity than ESG-related assessments. Given this background, we should not be surprised that credit ratings exhibit much higher levels of reliability and agreement than ESG ratings.

Then, the critical question is: what level of reliability and agreement is realistic to expect in ESG ratings.? The wine industry provides some guidance.

Rating a wine reduces to making an assessment based on many factors, most of them quite subjective, namely, aroma, typicity, intensity, sweetness, acidity, color, bouquet, flavor, and finish. Not unlike the situation of an ESG rater who must consider factors such as biodiversity, employee engagement, political contributions, oversight of strategy, executive compensation, office perks and human rights. The wine industry, mindful of these challenges, has made an effort to at least agree on which factors a rater must consider. This, obviously, has not changed the fact that there is an element of subjectivity inherent in wine tasting that can only be mitigated but not eliminated. Hence, wine ratings are probably destined to be less reliable, and show less agreement than ratings in other spheres. Wine ratings, nevertheless, provide a reasonable benchmark we can use to judge ESG ratings. In other words, ESG ratings should at least aspire to meet the levels of reliability and agreement observed in the context of wine ratings. Granted, this is not a very high standard. But it would certainly represent an improvement given the current situation.

In short, this study shows that ESG ratings display very low reliability and agreement not only in absolute terms, but also in comparison with other industries. The unfavorable comparison with wine ratings, given the inherent subjectivity involved in tasting wines (somehow analogous to the subjectivity involved in ESG-related evaluations) is especially damming for the credibility of the ESG ratings industry. It also hints that ESG ratings are unsuitable to play a role similar to the important role played by credit ratings in the fixed income market.

6. Conclusions and Recommendations

ESG ratings were designed with one purpose in mind: to help ESG-conscientious investors to make sound decisions. At present, considering the low reliability and agreement we found, at least in the case of the four RA investigated, it is unrealistic to think that ESG ratings could be useful for this purpose. A recent investor survey supports this view (Wong and Petroy, 2020). It reported

that almost all the companies that participated in the survey believe that their research teams know better than the RA analysts, and, overall, most participants expressed skepticism regarding the ESG ratings industry. This leaves retail investors, who do not enjoy the luxury of having their own teams of analysts, in the dark.

It is naïve to believe that ESG ratings, given the inherent subjectivity they entail, could ever achieve the high level of reliability and agreement that has been observed in the context of credit ratings in the bond market. Or, for that matter, in fields such as meteorology, accounting, or even clinical psychology. However, a reasonable aspiration should be to achieve at least the level of reliability and agreement observed in the wine market. To this end, we suggest two initiatives: (i) the ESG RAs should come to an agreement on what are the factors (criteria) they will focus on when making their E, S and G evaluations. This does not mean that they will have to agree on the methods employed to assess these criteria; and (ii) they should inform their ratings using comparable scales (with clear equivalence rules). A modest goal should be to achieve at least a 75% agreement (see Table 3) when using only a two-category scale (like pass/fail, something akin to the investment grade/non-investment grade characterization commonly used in the bond market). Recall that in this study we reported an abysmal 32% for this estimate (Table 5, sixth column). Anything short of meeting these goals will do little to improve the credibility of the ESG ratings industry.

Finally, a necessary clarification. Our conclusions should not be regarded as an indictment of the idea of incorporating an ethical dimension into investment decisions. They simply indicate that under the current state of affairs, the ESG ratings industry is not well positioned to offer useful advice regarding this matter.

REFERENCES

Ashton, R., 2012. Reliability and Consensus of Experienced Wine Judges: Expertise Within and Between, *Journal of Wine Economics*, 7(1): 70–87. DOI: 10.1017/jwe.2012.6

Berg, F., Kölbel, J. and Rigobon, R. 2022. Aggregate Confusion: The Divergence of ESG Ratings, Available at SSRN, DOI: 10.2139/ssrn.3438533

Billio, M., Costola, M., Hristiva, I. and Latino, C., 2021. Inside the ESG Ratings: Disagreement and Performance, *Corporate Social Responsibility and Environmental Management*, 28(5): 1426–1445. DOI: 10.1002/csr.2177

Brandon, R. G., Krueger, P., Schmidt, P. S., 2021. ESG Rating Disagreement and Stock Returns, *Financial Analysts Journal*, 77(4): 104–127. DOI: 10.1080/0015198X.2021.1963186

Charlin, V. and Cifuentes, A., 2017. Reliability and Agreement of Credit Ratings in the Mexican Fixed-Income Market, *Journal of Credit Risk*, 13(3): 21–45. DOI: 10.21314/JCR.2017.227

Chatterji, A., Durand, R., Levine, D. and Touboul, S., 2015. Do Ratings of Firms Converge? Implications for Managers, Investors and Strategy Researchers, *Strategic Management Journal*, 38(8): 1597–1614. DOI: 10.1002/smj.2407

Cheng, S., 2022. Peering Through the Kaleidoscope of ESG Rating Confusion, South China Morning Post, January 14, https://www.scmp.com/presented/news/hong-kong/education/topics/rising-tide-environmental-awareness/article/3162460

Christensen, D. M., Serafeim, G. and Sikochi, A. 2022. Why Is Corporate Virtue in the Eye of the Beholder? The Case of ESG Ratings, *The Accounting Review*, 97(1): 147–175. DOI: 10.2308/TAR-2019-0506

Cicchetti, D. V. and Allison, T., 1971. A New Procedure for Assessing Reliability of Scoring EEG Sleep Recordings, *American Journal of EEG Technology*, 11(3): 101–110. DOI: 10.1080/00029238.1971.11080840

Cicchetti, D. V. and Sparrow, S. A., 1981. Developing Criteria for Establishing Interrater Reliability of Specific Items: Applications to Assessment of Adaptive Behavior. *American Journal of Mental Deficiency*, 86(2): 127–137. PMID: 7315877.

Cohen J., 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1): 37–46. DOI:10.1177/001316446002000104

Cohen, J., 1968. Weighted Kappa: Nominal Scale Agreement Provision for Scaled Disagreement or Partial Credit. *Psychological Bulletin*, 70(4): 213–220. DOI: 10.1037/h0026256

Damoradan, A., 2020. Sounding Good or Doing good? A Skeptical Look at ESG, Professor A. Damoradan/s Blog, https://aswathdamodaran.blogspot.com/2020/09/sounding-good-or-doing-good-skeptical.html

Delmas, M. A., Etzion, D. and Nairn-Birch, N., 2013. Triangulating Environmental Performance: What Do Corporate Social Responsibility Ratings Really Capture?. *Academy of Management Perspectives*, 27(3): 255–267.DOI: 10.5465/amp.2012.0123

Dimson, E., Marsh, P. and Staunton, M., 2020. Divergent ESG Ratings, *Journal of Portfolio Management*, 47(1):.75–87. DOI: 10.3905/jpm.2020.1.175

Dolvin, S., Fulkerson, J. and Krukover, A., 2019. Do "Good Guys" Finish Last? The Relationship between Morningstar Sustainability Ratings and Mutual Fund Performance. *The Journal of Investing*, 28(2): 77-91. DOI: 10.3905/joi.2019.28.2.077

Dorfleitner, G., Halbritter, G. and 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. DOI: 10.1057/jam.2015.31

Doyle, T., 2018. Ratings that Don't Rate: The Subjective World of ESG Ratings Agencies, Harvard Law School Forum on Corporate Governance, https://corpgov.law.harvard.edu/2018/08/07/ratings-that-dont-rate-the-subjective-world-of-esg-ratings-agencies/

Engelhardt, N. Ekkenga, J. and Posch, P. 2021. ESG Ratings and Stock Performance during the COVID-19 Crisis. *Sustainability*, 13(13): 7133. DOI: 10.3390/su13137133

Friede, G., Busch, T., and Bassen, A. 2015. ESG and Financial Performance: Aggregated Evidence from More than 2000 Empirical Studies, *Journal of Sustainable Finance & Investment*, 5(4): 210–233, DOI: 10.1080/20430795.2015.1118917

Friedman, M., 1937. The Use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of Variance. *Journal of the American Statistical Association*, 32(200): 675–701. DOI: 10.1080/01621459.1937.10503522

Governance and Accountability Institute (G & A), 2021. SustainabilityHQ, ESG Factors, Master List, https://www.sustainabilityhq.com/esg-matters/esg-factors-master-lists-categories/

Grim, M. and Berkowitz, D., 2018. ESG, SRI and Impact Investing: a Primer for Decision-Making, Vanguard Research, https://personal.vanguard.com/pdf/ISGESG.pdf

Gyönyörová, L., Stachoň, M. and Stašek, D. 2021. ESG Ratings: Relevant Information or Misleading Clue? Evidence from the S&P Global 1200. *Journal of Sustainable Finance & Investment*. DOI: 10.1080/20430795.2021.1922062

Hayat, U. and Orsagh, M., 2015. Environmental, Social and Governance Issues in Investing, CFA Institute, https://www.cfainstitute.org/-/media/documents/article/position-paper/esg-issues-in-investing-a-guide-for-investment-professionals.ashx

Henze V. and Boyd, S., 2021. ESG Assets Rising to $50 Trillion Will Reshape $140.5 Trillion of Global AUM by 2025, Bloomberg Intelligence

Hill, R.P., Ainscough, T., Shank, T. and Manullang, D., 2007. Corporate Social Responsibility and Socially Responsible Investing: A Global Perspective. *Journal of Business Ethics*, 70: 165–174. DOI: 10.1007/s10551-006-9103-8

Mackintosh, J., 2022. Why the Sustainable Investment Craze Is Flawed. The Wall Street Journal, January 23

Matos, P., 2020. ESG and Responsible Institutional Investing Around the World: A Critical Review, CFA Institute Research Foundation

Moore, S., 2020. ESG Investing Takes Off, But Classification Confusion Remains, Forbes, November 16, https://www.forbes.com/sites/simonmoore/2020/11/16/esg-investing-takes-off-but-classification-confusion-remains/?sh=4b78a53e263f

Moy Huber, B. and Comstock, M., 2017. ESG Reports and Ratings: What They Are, Why They Matter, Harvard Law School Forum on Corporate Governance,

https://corpgov.law.harvard.edu/2017/07/27/esg-reports-and-ratings-what-they-are-why-they-matter/

Nauman, B., 2020. Heavy Flows into ESG Funds Raise Questions over Ratings, Financial Times, March 4

Paul Weiss, 2021. ESG Rating and Data: How to Make Sense of Disagreement, Paul Weiss Client Memorandum, https://www.paulweiss.com/insights/esg-thought-leadership/publications/esg-ratings-and-data-how-to-make-sense-of-disagreement?id=39303

Prall, K., 2021. ESG Ratings: Navigating Through the Haze, CFA Institute Blog, https://blogs.cfainstitute.org/investor/2021/08/10/esg-ratings-navigating-through-the-haze/

Sandberg, J., Juravle, C., Hedesström, T.M. and Hamilton, I., 2009. The Heterogeneity of Socially Responsible Investment. *Journal of Business Ethics*, 87: 519–533. DOI: 10.1007/s10551-008-9956-0

Schwartzkopff, F., 2022. Investors Seek Clearer Definition of What ESG Means in EU, Bloomberg News, https://www.bloombergquint.com/global-economics/investors-plead-for-clearer-definition-of-what-esg-means-in-eu

Semenova, N. and Hassel, L.G., 2015. On the Validity of Environmental Performance Metrics. *Journal of Business Ethics*, 132, 249–258. DOI:10.1007/S10551-014-2323-4

Shrout, P. and Fleiss, J., 1979. Intraclass Correlations: Uses in Assessing Rater Reliability. *Psychological Bulletin*, 86(2): 420–428. DOI: 10.1037//0033-2909.86.2.420

Steen, M., Taghawi, J., Moussawi, T. and Gjolberg, O. 2019. Is There a Relationship Between Morningstar's ESGRratings and Mutual Fund Performance? *Journal of Sustainable Finance & Investment*, 10(4): 349–370. DOI: 10.1080/20430795.2019.1700065

Tarnavsky, M., 2021. Too Many Cooks in the Kitchen: Why Do ESG Scores Differ So Much? Cyan Reef B.V., https://cyanreef.com/too-many-cooks-in-the-kitchen-why-the-esg-scores-differ-so-much/

Temple-West, P., 2019. Companies Struggle to Digest 'Alphabet Soup' of ESG Arbiters, Financial Times, October 6

The Economist, 2019. Poor Scores: Climate Change Has Made ESG a Force in Investing, The Economist, December 7

Tinsley, H. and Weiss, D., 1975. Interrater Reliability and Agreement of Subjective Judgments. *Journal of Counseling Psychology*, 22(4): 358–376. DOI: 10.1037/h0076640

United Nations, 2006. What Are the Principles for Responsible Investment? https://www.unpri.org/about-us/what-are-the-principles-for-responsible-investment

Walter, I., 2020. Sense and Nonsense in ESG Ratings. *Journal of Law, Finance, and Accounting*, 5(2): 307–336. DOI: 10.1561/108.00000049

Wigglesworth, R., 2018. Rating Agencies Using Green Criteria Suffer from Inherent Biases, Financial Times, July 20

Wilcoxon, F., 1945. Individual Comparisons by Ranking Methods. *Biometrics Bulletin*, 1(6): 80–83. DOI: 10.2307/3001968

Wong, C. and Petroy, E. 2020. Rate the Raters: Investor Survey and Interview Results, SustainAbility, https://www.sustainability.com/globalassets/sustainability.com/thinking/pdfs/sustainability-ratetheraters2020-report.pdf

Centro UC
CLAPES UC
Centro Latinoamericano de
Políticas Económicas y Sociales

clapesuc

@clapesuc

clapes_uc

clapesuc