# On Accountable Sustainability Measurement[a]

*Anders Westlund*[b], Department of Entrepreneurship, Innovation and Technology, and Center for data Analytics, Stockholm School of Economics, Stockholm, Sweden

*Rickard Sandberg*, Department of Entrepreneurship, Innovation and Technology and Center for data Analytics, Stockholm School of Economics, Stockholm, Sweden

*Emma Sjöström*, Department of Marketing and Strategy, Stockholm School of Economics Institute for Research (SIR), and Mistra Center for Sustainable Markets (MISUM), Stockholm School of Economics, Stockholm, Sweden

*Abstract:* Sustainability is assumed to be an effective driver of value creation in the society. To measure sustainability is, thus, important. It is so to verify the present situation, but also to identify the processes to most efficiently improve sustainability performance.

Are the most widely applied measurement approaches good enough for those purposes, and in particular, are they accountable? Some important dimensions of the accountability concept are discussed in the paper. The measurements of actual versus stakeholder perceived sustainability are also discussed. Finally, a new approach for measuring sustainability, as perceived by stakeholders in the financial sector, is introduced.

[a] This is a working paper in a series of papers stemming from the project *"On Accountable Measurement of Corporate Sustainability; the Need for a Sustainability Index – a research project at the Stockholm School of Economics."* This project is funded by the Nasdaq Nordic foundations, and all authors gratefully acknowledged this financial support.

[b] Correspondence to: Anders Westlund, Department of Entrepreneurship, Innovation and Technology, and the Center for data Analytics, Stockholm School of Economics, P.O. Box 6501 (Sveavägen 65), 113 83 Stockholm, Sweden. E-mail: anders.westlund@hhs.se.

*1. Introduction*

Sustainability has established itself as a business megatrend. It emphasizes the need to conduct business with a holistic and long-term view so that present needs of society and the planet can be met without jeopardizing the ability of future generations to meet their needs. Today, most companies, and particularly those that are large-cap and publicly listed, are at least to some extent integrating sustainability dimensions in their business operations. This trend has been exacerbated by the establishment in 2015 of both Agenda 2030 with seventeen global sustainability goals, and the Paris Agreement, a legally binding international treaty on climate change. It is manifesting itself for example in that many companies (e.g., 90 percent of S&P 500 companies in 2019) are issuing annual sustainability reports to inform their stakeholders on past performance and future goals.[c] The EU has even mandated such disclosure for companies of a certain size.

In tandem with these developments, an increasing number of financial investors are including environmental, social, and governance (ESG) dimensions in their investment analysis and decision making, as evidence for example by the continuous growth of signatories to the UN backed initiative Principles for Responsible Investment (PRI). This has generated a growing investor demand for information on corporate ESG performance. As a consequence, several ESG ratings have been developed, either by specialised ESG rating agencies or by the established financial rating institutes. Such ratings will evaluate and score individual corporations on a range of ESG dimensions.

Not only do they provide the financial sector with information to ensure effective investment decisions. They could potentially also help business itself to assess its level of sustainability, and thus support the process of improving sustainability, on operational as well strategic levels. The prioritizing of improvement actions is then crucial.

With ESG ratings now being so influential in the financial sector, it spurs a number of important questions: How is sustainability in a business context typically measured today? Is the quality of such measurements good enough? And how should we define ”good enough quality”? Sustainability ratings for the financial sector is dominated by a small number of large

[c] https://www.ga-institute.com/research-reports/flash-reports/2020-sp-500-flash-report.html.

providers.[d] Some of the dominating actors include MSCI, Sustainalytics, RepRisk and ISS. These ratings are typically large scale and descriptive. They are based on a variety of ESG concepts, each one operationalized by many indicators. This is done for a large number of industries and companies. Finally, the measurements are weighed together to overall ESG ratings, used for the financial investment decisions.

The main purpose of these ESG ratings is to identify strong or weak sustainability performance. In that way they can help investors to prioritize between different investments. They can also help individual companies to benchmark themselves versus competitors. However, we find that they do not provide good enough support for prioritized improvement actions. It is so because ESG ratings typically do not provide effective and consistent information on the relative importance of alternative improvement alternatives.

It is important to verify the quality of the commonly used ESG ratings. It is after all guiding trillions of dollars. This is not an easy task, but for their purpose to provide extensive and descriptive investment support, our view is that they probably meet reasonable quality levels. However, as there are no commonly accepted standards for ESG ratings, the different providers take different paths to their overall ESG ratings. The differences concern *what* is measured (choice of concepts), and *how* they are measured (choice of indicators), and finally, how their choice of *weighting* algorithms is made. This situation reduces comparability between different ESG ratings. There is substantial divergence between them, that has been well documented in the academic literature (for an overview, see Sjöström et al, 2021). For example, Berg, et al. (2019) compare ESG ratings from six leading providers (KLD, Sustainalytics, Vigeo Eiris, RobecoSAM, Asset4/Refinitiv, and MSCI). They show that correlations between the ratings are in the range from 0.38 to 0.71 (with an average of 0.54). This signals that the users within the financial sector would do well to ask for more standards here. Berg, et al., (2019) try to identify the main enablers of this significant divergence (as manifested in very low correlations). They find that the two main enablers are (i) differences in what is measured, and (ii) how the concepts are measured, i.e., the choice of indicators. The different weighting procedures do not explain much of the observed divergence. The six different providers are, however, using very similar weighting procedures, not resulting in substantially different

[d] For an overview of the many mergers and acquisitions in the past decade, see SustainAbility (2020) https://www.sustainability.com/globalassets/sustainability.com/thinking/pdfs/sustainability-ratetheraters2020-report.pdf.

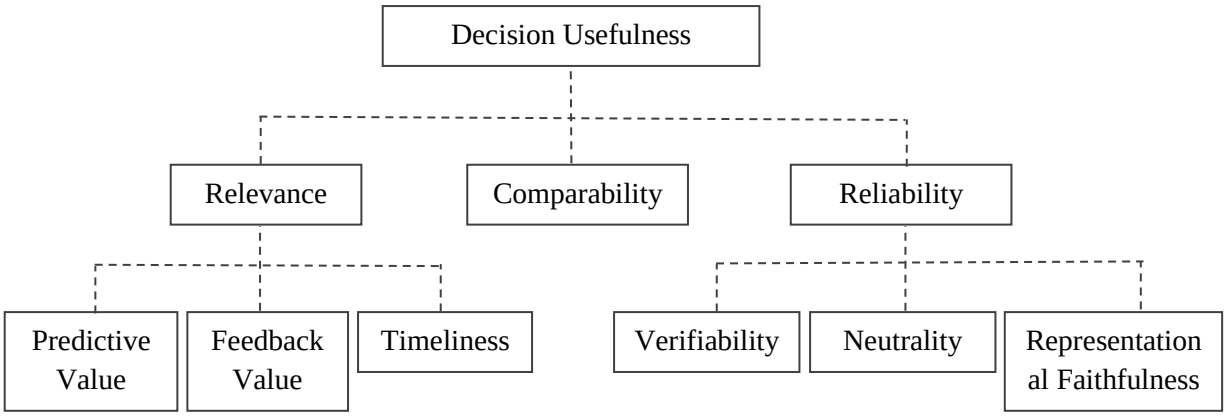
weights. That does not mean that the weights are correct, or even well decided. For example, the weightings do not consider what is more or less important to stakeholders to achieve their goals. Stakeholder perspectives are not at all considered.

Which quality dimensions in sustainability measurement are important, and what is to be meant by quality here? The Sustainability Accounting Standards Board (SASB) (year) provides metric characteristics and a taxonomy for concepts and indicators of accountability. These definitions are rather abstract and will allow for numerous different interpretations. Statistical operationalisation is probably needed. Not much is said, however, about the statistical considerations to achieve a high-quality standard. We discuss this in Section 2 of this paper. We also provide comments on whether, or not, the dominating providers for sustainability rankings meet with what we see as appropriate accountability requirements. Section 3 provides an example of a stakeholder based, econometric model for sustainability measurement, with focus on the financial sector. Finally, Section 4 gives some conclusions.

*2. Accountable Sustainability Measurement*

The characteristics of performance measures used in internal or external reports are, as already said, becoming more and more important. Most large and publicly listed companies also disclose non-financial information to their stakeholders. The Standards Committee of the American Accounting Association argues that non-financial performance measures, such as measures on sustainability, should be judged against the same criteria as financial performance measures (i.e., *relevance, reliability and comparability*; Maines, *et al.*, 2002). Although these criteria sometimes are difficult to qualify in order to assess the quality of sustainability information, we recommend to follow the conceptual framework of Financial Accounting Standards Board (FASB), (1980) - see *Figure 1* below - but operationalised according to statistical criteria. The SASB (year) standards provide the following metrics characteristics: Faithful, Complete, Neutral, Verifiable, Aligned, and Understandable. As seen, the overlap with the FASB framework is high.

*Figure 1:* FASB's qualitative characteristics of financial information.




It is also noticed that sustainability measurements should be implemented for two main reasons: (a) to manage and monitor internal activities related to the enhancement of non-financial and financial stakeholder assets, and (b) to report externally non-financial and financial performance. For (b) FASB (and SASB) provides the information quality requirements and standards given above. For (a) the same quality requirements help the process of effective stakeholder asset management (stakeholders could be customers, employees, society, financial investors, etc.). It is simply managerially sound to use the same high-quality requirements for (a) as for (b).

Sustainability measurement fulfilling the high-quality standards of FASB (and SASB) is here referred to as *Accountable Sustainability Measurement (ASM).* In order to implement these, sometimes rather abstract, quality requirements in practice, they should preferably be translated into statistical requirements and standards.

*2.1. Relevance*

Relevance presupposes Predictive (and Feedback) Value, and Timeliness. Predictive Value is first of all related to a possibility to link sustainability to future financial value. Studies trying to demonstrate a link between current non-financial measures and future financial information indicate that non-financial information should be useful to investors and creditors (see, e.g., Ittner and Larcker, 1998, Kristensen and Westlund, 2004, Westlund, *et al.*, 2005, and Bescos,

*et al*, 2007). This has shown to be true, also for sustainability information (see, e.g., Verheyden, et al., 2016). Here, relevance should be defined as whether it is possible to assess links between sustainability and future financial results. That is, there is a prediction and/or feedback value. Timeliness is a prerequisite for obtaining such values.

The Relevance dimension of ASM is more effectively met by measurement in a structural context. The measurement model should not just focus on individual stakeholder performance criteria, but also include their main enablers (such as ESG), and their main consequences (such as changes in risk assessments, trading behaviour, etc.). In particular, the measurement should determine the size of the interactions between enablers, key criteria, and the consequences. In that way, management will obtain information that identifies areas for improvement with high impact on stakeholder performance. These arguments are further discussed in Section 3. Here, statistical criteria, e.g., precision, lack of bias, and in particular robustness are crucial ingrediencies of ASM, and essential for sustainability excellence. If observed differences over time, or between different business units, are not statistically significant, the information value of sustainability is questionable. A key criterion for ASM is robustness. By that we mean the use of a methodology, the quality of which do not significantly deteriorate, even if measurement models are mis-specified, if there are survey design problems such as nonresponse, response errors, etc. One leading candidate for robust measurement will be structural modelling using Partial Least Squares; see, e.g., Cassel, et al., 1999, 2000, 2001.

### *2.2. Reliability*

Reliability is often defined as the quality of information that assures that information is reasonably free from error and bias, and faithfully represents what it purports to represent (Hermann and Thomas, 1997). According to the FASB framework, reliability is built from (i) Representational faithfulness, (ii) Verifiability, and (iii) Neutrality. Representational faithfulness is defined as a correspondence or agreement between a measure or description and the phenomenon it purports to represent (FASB, 1980). Verifiability is the ability through consensus among measures to ensure that information represents what it purports to represent, or that the chosen method of measurement has been used without error or bias. Finally, neutrality is defined as an absence of bias in reported information intended to attain a predetermined result or to introduce a particular mode of behavior.

Overall, however, from a statistical perspective, reliability refers to the quality of the measurement constructs of the model. Reliability is in general defined as

[1] $V(S2, True)/V(S2, Observed)$,

and typically estimated by studying correlations between two or more scores. Widyawati (2020) studies internal consistency reliability among scores and rankings for four agencies (Sustainalytics, Bloomberg, ASSET4, and MSCI KLD). Cronbach's alpha is used to estimate internal consistency of construct's factors. The results were here quite mixed, but typically not significant.

*2.3. Comparability*

Research indicates that different formats for reporting financial and non-financial (e.g., sustainability) performance measures can influence investors' use of that information by affecting the transparency of information and explicitness of the links between performance measures (Hirst, *et al.*, 2002). Then, studies suggest that investors' ability to use non-financial and financial information consistently across companies and time is impaired by non-comparability in measures or formats. That is clearly the case when measuring ESG, and ranking sustainability (see, e.g., Berg, et al., 2019, and Sjöström, et al., 2021). The non-comparability likely reduces the value of sustainability performance measures and may lead investors to focus primarily on financial measures for assessing performance. It is therefore important to ask providers of sustainability information for a clear explanation of the composition of their measures, their computation, and the original and underlying assumptions of the different elements of the calculation. Such explanations would permit comparability of different metrics or different definitions of identical measures. For the same reason of comparability, the requirement of consistency over time of the composition and the computation of indicators is necessary (COB, 2000).

*2.4. ESG Ratings Today - Accountable or Not?*

As said in our Introduction, the ESG ratings on the market for the financial sector are dominated by a small number of providers. Following the accountability requirements given by FASB and/or SASB, are those rankings accountable, or not? Due to a substantial lack of transparency

from those providers, it is impossible to make a specific evaluation here. However, some general observations are possible.

The lack of transparency is a problem in itself. That obstructs the possibility to verify *Comparability*, in particular from a methodological perspective. However, several empirical studies on differences between rankings have been made (see, e.g., Sjöström, et al., 2021). Most of these studies demonstrate substantial differences between rankings, which creates problems for users in the financial sector. As for *Reliability*, it is impossible to evaluate the existence of Neutrality and Representational Faithfulness. However, lack of Verifiability is a substantial problem, due to lack of transparency. As for *Relevance*, and with respect to the overall purpose of existing sustainability rankings, good Predictive and Feedback values, and high Timeliness, have been demonstrated in many studies; see, e.g., Verheyden, et al. (2016).

The relevance of ESG ratings as instruments to help the financial sector to take better investment decisions is clearly demonstrated. The *Decision Usefulness* could have been better with more focus on comparability and reliability. However, we question the role of ESG rankings as instruments for *managing sustainability improvements*. It is so because of lack of stakeholder focus when identifying what is important to prioritize to achieve sustainability improvements. In the next section we outline an alternative measurement, to some extent dealing with those challenges.

*3. Stakeholder Based ASM*

In Section 2 we outlined some criteria to obtain ASM (Accountable Sustainability Measurement). It is mostly about reliability and relevance. We argued that important prerequisites to reach reliability and relevance is to use a structural equation model, with a stakeholder focus.

Formally, such a measurement model has an inner structure:

[2] $Y = BY + CX + D,$

where $Y$ is a vector of latent endogenous variables, typically including a stakeholder Satisfaction Index (e.g., satisfaction of financial analysts), and changes in Stakeholder

Performances (e.g., risk assessments, trading behaviour, etc.). Furthermore, $X$ is a vector of latent exogenous variables, such as ESG concepts. The vector $D$ of error terms is assumed to fulfil the orthogonality condition $E(D|X) = 0$. $B$ and $C$ are weight matrices, estimated in the statistical measurement and analysis process.

The outer measurement structures specify relations between each latent variable, endogenous as well as exogenous, and a corresponding set of measurable manifests, e.g.:

[3] $$x = c(x)X + d(x)$$

and

[4] $$y = c(y) + d(y),$$

where $x$ and $y$ are vectors of measurable manifests, and $c(x)$ and $c(y)$ are matrices with a structure which makes sure that each indicator measures exactly one latent variable, i.e., the manifest variables are split into blocks, each being related to one latent variable. Often, the latent variables are scaled to have unit variance to avoid ambiguities in the presence of unknown terms. The measurement specifications [3] and [4] are called reflective, or outward directed, specifications. The measurable manifests are reflections of underlying latent variables. This situation is typical for the endogenous latents. For the exogenous latents (ESG), it will often be more appropriate to specify a relation where the latent is formed by the manifests, i.e.:

[5] $$X = \pi x + d(X),$$

where $\pi$ is another matrix of parameter weights. This specification is called formative or inward-directed. In each measurement application the interpretation of the specified relations will guide the choice between reflective and formative measurement models.

Finally, there are weight relations defining the estimated case values, or scores, of the latent variables:

[6] $\hat{Y} = w(Y)y$

[7] $\hat{X} = w(X)x.$

The weights *w* are determined depending on whether the measurement models are reflective, or formative.

The specifications [2] – [7] should be estimated with a robust algorithm, such as Partial Least Squares (PLS); see, e.g., Fornell and Cha (1994) and Cassel, *et al.* (1999). PLS is an iterative estimator with good statistical characteristics.

To illustrate the approach formally described above, we provide an example. The perhaps simplest version of the exogenous vector *X* in [2] could include three latent variables, representing Environmental Performance (E), Social Performance (S), and Governmental Performance (G). It could, however, easily be expanded to include further concepts (see Berg, *et al.*, 2019), representing various subdimensions or scopes of ESG. For example, *X* could represent scopes according to the taxonomy provided by Sustainability Accounting Standards (SASB). The choice of elements in *X* could be crucial (see Berg, *et al.*,2019).

The endogenous vector *Y* in [2] is typically divided into two blocks $Y(i), i = 1,2$. Following our suggestion of a stakeholder approach, $Y(1)$ includes the overall stakeholder perception of the quality of a firm's sustainability performance (e.g., overall stakeholder satisfaction). The second endogenous block, $Y(2)$, includes various dimensions representing the stakeholder behaviour, as a consequence of changes in a firm's sustainability performance, and the overall stakeholder satisfaction, it implies. For example, if the stakeholders are financial investors, $Y(1)$ could include risk judgements, intention to buy or sell stocks, etc.

Each latent in *X* and *Y* is measured by one or several manifests, *x* and *y*. As shown by Berg, *et al.* (2019), differences in the choice of manifests represent another important source of divergence between different ESG ratings.

Manifesting $Y(1)$ should, however, be standardized to secure comparability (another key dimension of accountability). Following the recommendations from stakeholder satisfaction

research, $Y(1)$ should be represented by three manifests; $y(11)$ = overall perceived satisfaction, $y(12)$ = overall satisfaction, subject to expectations, and $y(13)$ = overall satisfaction compared to ”best in class”.

The matrices of parameters *B* and *C* in [2], as well as loadings *c(x)* and *c(y)* in [3] and [4], and/or weights $\pi$ in [5] are preferably estimated by Partial Least Squares. This procedure gives a much more consistent weighting process than what is normally applied in ESG ratings. In particular, we could introduce restrictions saying that estimation is performed subject to maximum impact from $Y(1)$ to $Y(2)$. This secures relevance in that sense that $Y(1)$ is defined as the latent giving maximum impact on the stakeholder performance $Y(2)$. This procedure also guarantees consistent weighting through *c(x)* and/or $\pi$, i.e our weighting of manifests of ESG or other ESG scope factors. Normally, these weighting procedures are more or less *ad hoc* in common ESG rankings.

Sandberg, *et al.* (2021) provides an example of stakeholder based structural equation ASM. The stakeholders are investors in the Swedish financial sector. Data are collected through a web questionnaire for a random sample of these investors. Sustainability performance is evaluated for two companies from different industries.

*4. Conclusions*

We have argued that the major ESG ratings are reasonably well designed for their purpose of supporting the financial sector to take effective investment decisions. But with reference to FASB/SASB accountability criteria, the ESG rankings are not fully accountable. In particular, there is a lack of comparability (due to lack of transparency), mixed reliability (due to lack of verifiability), but a reasonable degree of relevance for their purpose outlined above. However, there is lack of relevance when it comes to supporting sustainability improvement processes. Effectiveness here requires information on *what is important* to prioritize, and that is often not just correlated to what is worse, or better, sustainability performance.

Furthermore, there must be more of a stakeholder perspective in such measurement processes. For that purpose, we have outlined an alternative measurement strategy, based on structural

modelling and PLS estimation. That procedure is empirically described in another paper (Sandberg, et al., 2021).

*References*

Berg, F., Koelbel, J.F., and Rigobon, R.; Aggregate Confusion: The Divergence of ESG Ratings, MIT Sloan School, Working Paper 5822-19 (2019)

Bescos, P-L, Cauvin, E., Decock-Good, C., and Westlund, A.H.; Characteristics of Performance Measures for External Reporting, *Total Quality Management, 18,* 1-20 (2007)

Cassel, C., Hackl, P., and Westlund, A.H.; PLS for Estimating Latent Variable Quality Structures: Finite Sample Robustness Properties, *Applied Statistics, 26,* 435-446 (1999)

Cassel, C., Hackl, P., and Westlund, A.H.; On Measurement of Intangible Assets: A Study of Robustness of Partial Least Squares, *Total Quality Management, 11,* 897-907 (2000)

Cassel, C., Hackl, P., and Westlund, A.H.; Structural Analysis and Measurement of Customer Perceptions, Assuming Measurement and Specification Errors, *Total Quality Management, 12,* 873-881 (2001)

COB; Creation de Valeur Actionnnariale et Communication Financiere, *Bulletin COB, 346,* 43-94 (2000)

Financial Accounting Standards Board (FASB), *Qualitative Characteristics of Accounting Information,* Statement of Financial Accounting Concepts, No 2, Stanford (Ct).

Fornell, C., and Cha, J.; Partial Least Squares, in R.P. Bagozzi (ed), *Advanced Methods of Marketing Research,* (Cambridge: Blackwell)

Herrman, D., and Thomas, W.B.; Reporting Disaggregated Information: a Critique Based on Concepts, *Accounting Horizons, 11,* 35-44 (1997)

Hirst, D.E., Hopkins, P., and Wahlen, J.; Fair Values: Performance Reporting, and Bank Analysts' Risk and Valuation Judgements, Working Paper, University of Texas, Austin

Ittner, C.D., and Larcker, D.F.; Are Non-Financial Measures Leading Indicators of Financial Performance? An Analysis of Customer Satisfaction, *Journal of Accounting Research, 36,* 1-36 (1998)

Kristensen, K. and Westlund, A.H.; Performance Measurement and Business Results, *Journal of Total Quality Management, 15,* 719-733 (2004)

Maines, L.A., Bartov, E., Fairfield, P.M., Hirst, D.E., Iannaconi, T.E., Mallet, R., Schrand, C.M., Skinner, D.J., and Vincent, L.; Recommendations on Disclosure of Non-Financial Performance Measures, *Accounting Horizons, 16,* 353-362 (2002)

Sandberg, R., Westlund, A.H., and Sjöström, E.; Multilayer Path Modelling for Accountable Measurement of Corporate Sustainability, forthcoming, (2021)

Sjöström, E., Sandberg, R., and Westlund, A.H.; ESG Ratings: Mixed Bag and its Implications, forthcoming, (2021)

Verheyden, T., Eccles, R.G., and Feiner, A.; ESG For All? The Impact of ESG Screening on Return, Risk, and Diversification, *Journal of Applied Corporate Finance, 28,* 47-55 (2016)

Westlund, A.H., Gustafsson, C., Lang, E., and Mattsson, B.; On Customer Satisfaction and Financial Results in the Swedish Real Estate Market, *Journal of Total Quality Management, 16,* 1149-1159 (2005)

Widyawati, L.; Measurement Concerns and Agreement of Environmental Social Governance Ratings, *Accounting and Finance,* (2020)