SPES
Sustainability performances, evidence & scenarios

# SPES Focus - Work Package #3
# Assessing measurement challenges and sensitivity in transition performance indicators

## June, 2024

Funded by European Union's Horizon Europe Programme under Grant Agreement No. 101094551

# Authors

**Amaia Palencia-Esteban** - London School of Economics
**Pedro Salas-Rojo** - London School of Economics
**Paolo Brunori** - London School of Economics
**Luca Lodi** - Università degli Studi di Firenze

### Contributors and peer reviewers:

Manlio Calzaroni, ASviS: András Gábos, TARKI; Mario Biggeri, University of Florence; Emilia Rocco, University of Florence.

### Aknowledgements:

The authors would like to thank all SPES partners for their inputs provided in several SPES meetings and communications.

### Cite as:

Palencia-Esteban, A., Salas-Rojo, P., Brunori, P., Lodi, L. (2024). Assessing measurement challenges and sensitivity in transition performance indicators. SPES Focus - Work Package #3, SPES project – Sustainability Performances, Evidence and Scenarios. Florence: University of Florence. Available at: https://www.sustainabilityperformances.eu/publications-deliverables/

# Disclaimer

# Contents

# Abstract

We explore measurement challenges associated with five composite indicators broadly used by international policy institutions to capture well-being, sustainable development, and economic transition. We perform several sensitivity and robustness analyses after formally discussing their construction and checking data requirements. Departing from a baseline scenario drawn for the European Union in 2019, we estimate the sensitivity of the selected indicators to perturbations in their components across countries. We find indicators insensitive to changes in most components, and rather robust to perturbations in the data. The main drivers of this insensitivity are the countries' relative position within the component range of variation and their dispersion around the mean value. We propose the need to rethink the construction of transition performance indices given their limited capacity to capture socioeconomic changes and, especially, assess sustainable development.

# 1 Introduction

The use of aggregate or composite indicators has expanded in recent decades (Stiglitz et al., 2009). Contrary to one-dimensional measures such as gross national product or income-based inequality and poverty measures, or subjective well-being composite indicators aim at accounting for multidimensional phenomena. Most of these indexes aggregate information across diverse dimensions such as pollution, democracy proxies, gender gaps, inequality, or energy use, collapsing their information into just one value. This summarizing ability has promoted them as a relevant tool to capture the complexity of well-being, sustainable development, and economic transition. Some of the most relevant international institutions, such as the United Nations (UN), the European Commission (EC) or the Organisation for Economic Co-operation and Development (OECD) have developed their own measures, which are used to rank countries and assess transition and development performance (European Commission, 2022; Sachs et al., 2019).[1]

Composite indicators summarize multi-dimensional concepts, aggregating the complexity of socioeconomic or environmental phenomena in an aggregate measure. These indexes facilitate cross-country comparisons or analyzing the country performance over time.[2] The relevance of these indexes bears shortcomings. Aggregating different dimensions, such as air pollution, life satisfaction, and voter turnout often deliver results hard to interpret or disentangle. Although their intuitiveness makes them accessible to the general public and policymakers, the lack of a concrete interpretation questions their utility for policy advising beyond the elaboration of country ranks (Fleurbaey, 2009). The construction of aggregate indexes also implies relevant decisions on weighting, variable selection, normalization, and aggregation, implying potentially undesirable properties and misleading results (Costanza et al., 2004; Kubiszewski et al., 2013; Nardo et al., 2005; Ravallion, 2012)

Economies in the European Union are expected to face major challenges in the coming years. The long-lasting effects of the Great Recession, the COVID-19 crisis, the notorious effects of global warming, the rising prices of energy and consumer prices, and the rising international insecurity pose severe threats to social, democratic, and economic performance (European Commission, 2022; UNDP, 2022). Ambitious investment programs like *InvestEU*, focused on energy transition, *Horizon2020*, promoting research and innovation, or the *NextGenerationEU* funds, are recent examples of the diverse financial efforts made towards sustainable development in this increasingly complex context. Although the evaluation of specific policy interventions is nec-

[1] Respectively, these institutions have developed the Human Development Index, the Transition Performance Index, and the Better Life Index. All addressed in this work.

[2] Cultural and social norm disparities may limit cross-country comparisons. For example, countries with more social recognition of gender disparities may be more prone to denounce gender violence, thus artificially rising the associated metrics.

essary for an efficient and effective allocation of resources, measures aimed at capturing the overall ability of countries to transition towards sustainability in all dimensions is key.

The literature has already presented a myriad of transition performance indicators (see Hoekstra (2019) and Gábos et al. (2023) for a review). Although assessments have been performed in some indexes (European Commission, 2022; Galotto et al., 2020; OECD, 2022)), a comprehensive exploration of the sensitivity of aggregate indicators to changes in their components poses crucial relevance. If economies enhance their performance in one or more dimensions associated with greater well-being or sustainability and the indicator fails to reflect this, its utility and significance may be questioned (Biggeri et al., 2024). The interpretability of indicators and their ability to aggregate complex aspects of economic development should not come at the cost of ignoring relevant dimensions or highlighting components due to undesirable properties.

Our contribution explores this extent. We focus on the aggregate indicators selected by Gábos et al. (2023). The five measures that reflect different approaches to transition performance towards sustainability are the Planetary Pressure Adjusted Human Development Index, the Transition Performance Index, the Better Life Index, the Green Growth Index, and the Sustainable Development Goals Index. With a focus on the EU-27 in 2019, we first compute the values for all indexes and define a baseline scenario. Then, we run several Monte Carlo simulations to address the reaction of the aggregate baseline indicators to perturbations in the individual components. A set of robustness checks is performed to relax some assumptions in the simulation exercise.

Our main results raise doubts about the performance of these five indicators in their current form. Overall, they are found rather insensitive to changes in their components and unable to capture improvements or worsening in several transition or sustainability dimensions. We find that, if the component's values are highly clustered for a set of countries, a small change in a country's component value may drastically alter its relative position. The normalization process may exacerbate this effect, so countries receiving a low normalized score in a component could turn into a high score, and vice versa. This change in the score has a latter effect on the aggregate indicator. This way, these measures are more reactive to changes in components characterized by a low coefficient of variation, this is, very clustered around the mean. This distributional property of the components leads to undesirable results. For instance, the Green Growth Index is quite insensitive to changes in forest areas, air pollution, or gas emissions, but very sensitive to changes in the gender gap. Several robustness checks and the computation of standard errors confirm the reliability of these findings.

The remainder of the article is schemed as follows. First, Section 2 reviews the literature and addresses how the selected measures have been evaluated. Then, Section 3 describes the indicators, explains the data limitations and presents the sensitivity evaluation methods. Section 4 explores our main results and robustness checks. Section 5 concludes.

# 2 Literature review

This work aims to analyzing the measurement challenges and the sensitivity of five indicators selected through a rigorous process described in Gábos et al. (2023). First, these authors evaluated an initial list of 44 composite indicators and dashboards, measuring well-being, human development, sustainability, and transition performance. Then, they refined the initial pool to 15 indicators, focusing on the effective coverage of the four pillars of human development: productivity, equity, environmental sustainability, and citizen democratic participation. Finally, five indicators were shortlisted: those considered the most appropriate and complete to measure countries' well-being and human development and evaluate sustainable transition performances. The five indicators are the following:[3]

- The **Planetary Pressure Adjusted Human Development Index (PHDI)** from the United Nations Development Programme (UNDP, 2022), measures the level of human development adjusted by carbon dioxide emissions (based on production) and material footprint per capita (based on consumption), thus considering the disproportionate human impact on the planet.

- The **Transition Performance Index (TPI)** from the *European Commission* (European Commission, 2022) encompasses a scoreboard that evaluates and ranks countries based on their progress in achieving sustainable prosperity through four types of transitions: *i)* economic (education, wealth, labor productivity, research and development intensity, industrial base), *ii)* social (healthy life, work, and inclusion, leisure time, equality), *iii)* environmental (greenhouse gas emissions reduction, biodiversity, material use, energy productivity), and *iv)* governance (fundamental rights, security, transparency, sound public finances). This framework consolidates the foundation of a new prosperity model emphasizing resilience, inclusiveness, and sustainability, aligning with the *EU's 2022 Annual Sustainable Growth Strategy*.

- The **Better Life Index (BLI)** (OECD, 2022) from the *Organisation for Economic Co-operation and Development (OECD)* covers 11 topics: housing, income, jobs, community, education, environment, civic engagement, health, life satisfaction, safety, and work-life balance. Each topic is represented by up to three equally weighted indicators, selected according to relevance and data quality. These indicators effectively measure well-being in cross-country comparisons and are set to expand gradually, enriching the index scope.

- The **Green Growth Index (GGI)** from the *Global Green Growth Institute* (Acosta et al., 2019) evaluates the performance of countries in meeting sustainability objectives and encom-

[3] We refer to Section 3.1 to a complete formulation of each indicator.

passing the Sustainable Development Goals, the Paris Climate Agreement, and the Aichi Biodiversity Targets. The four dimensions of green growth highlight efficient and sustainable resource use, protection of natural capital, fostering green economic opportunities, and promoting social inclusion.

- The **Sustainable Development Goals (SDG) Index** of the *SDG Transformation Center* (Sachs et al., 2023) measures and ranks countries according to their performance in the 17 Sustainable Development Goals of the Agenda 2030.

The remainder of the section reviews, first, the literature evaluating composite indicators, and then schemes previous analyses on the five proposed indicators.

The construction of composite indicators involves multiple steps (European Commission, 2008). Decisions made by the indicator designers and practitioners are not trivial, as they often have remarkable effects on the outcomes. Providing valid and reliable composite indicators goes beyond the technical dimension, and a strong theoretical framework is also needed (Fleurbaey, 2009). Indeed, if poorly constructed or misinterpreted, these measures can convey misleading or contradictory policy messages. For instance, Freudenberg (2003) addresses the inclusion or exclusion of different variables, alterations in weights, the application of diverse standardization methods, and the choice of different base years. Saisana et al. (2005) recommends confirming the accuracy of models and components used in aggregate indicators.

Beyond these and other early contributions (see Ebert and Welsch (2004) and Chowdhury and Squire (2006)), several international institutions formulated a Handbook discussing how to create reliable composite indicators (European Commission, 2008). This seminal publication aimed at offering a comprehensive set of guidelines to assist indicator designers in improving the quality of their outputs. It discusses ten steps required to build a proper measure of the phenomenon of interest, from the development of the theoretical framework to the technical details of variable selection, including a multivariate analysis to understand the structure of the indicators or the feasibility of the data. The methodology to construct the index shall also be justified, including the imputation of missing data, and the normalization, weighting, and aggregation techniques. Overall, the Handbook suggests that it is necessary to self-assess the robustness and sensitivity of the outcomes and check the external validity by testing links to other variables capturing similar information to the indicator. In the same line, Mazziotta and Pareto (2013) broadens the set of guidelines for selecting individual indicators, addressing aggregation methods and comparability, and finally acknowledging that there's often no unique and unequivocal solution.

While general guidelines are essential to the construction of composite indicators, the second pivotal aspect in the literature consists on the evaluation of the indicators performance. One key dimension of the evaluation process lies on the selection of the appropriate weighting and aggregation methods. These are determined either through theoretical constructs or by consulting experts and stakeholders on the relevance of specific dimensions. One early example can be found in Chowdhury and Squire (2006), which scrutinized the "equal weights" approach employed in the Human Development Index (HDI) (see also Ravallion (2012)). Through an opinion survey targeting researchers, they suggest that a simple weighting scheme does not significantly differ from more complex methodologies for these indicators. Subsequently, Cherchye et al. (2007) discussed the *benefit of the doubt* methodology, addressing how to construct composite indicators amidst this uncertainty. While not being a cure-for-all, their research offers a framework that could mitigate methodological disputes undermining the credibility of composite measures.

In the same vein, Becker et al. (2017) proposed a statistical-based approach to evaluate the reliability of weights. They suggested a series of steps, namely: *a)* employing a non-linear Pearson correlation ratio, estimated via Bayesian Gaussian processes, to determine the relevance of components to the aggregate indicator; *b)* isolating and examining the effect of each variable through regression analysis; and *c)* implementing an optimization procedure to align weights with pre-established importance values. A more recent contribution to post-evaluation methods is Greco et al. (2019). The authors critically analyzed the methodologies underpinning composite indicators, emphasizing the weighting and aggregation and underscoring the importance of robustness analysis. They proposed the usage of Stochastic Multi-criteria Acceptability Analysis to reflect a diverse range of individual weighting preferences, challenging the dependence on a representative weight vector.[4] The authors advised caution when interpreting composite indicators due to inherent construction flaws. They also engage in a discussion on the continuum between subjective and objective weighting approaches, including the compromises inherent in aggregation methods.

The quality and robustness of composite indicators have also been addressed by institutions such as the *European Commission's Competence Centre on Composite Indicators and Scoreboards* (COIN). Recently, the *Joint Research Center* (JRC) published a set of detailed reports for auditing various indicators. Two of our five selected indicators are addressed: the SDG index, and the Transition Performance Index (TPI).

First, the audit of the SDG Index (Papadimitriou et al., 2019) is based on uncertainty and sensitivity analysis. As in Saisana et al. (2005), the authors address whether such an indicator represents a significant endeavor to consolidate the 17 adopted SDGs into a singular measure. The country ranks are quite stable, as they have limited uncertainty due to the aggregation meth-

[4] Stochastic Multi-criteria Acceptability Analysis is used in decision-making processes that evaluate various alternatives based on multiple criteria. It incorporates uncertainty and variability in the data (ordinal and cardinal) and weighting approaches, providing a probabilistic assessment of the acceptability of each alternative (Lahdelma et al., 1998).

ods and the set of chosen indicators. Regarding the TPI report (European Commission, 2022) the evaluation includes *(i)* a principal component analysis to evaluate the degree to which the statistical method validates the underlying conceptual structure, *(ii)* Monte Carlo simulations using randomly assigned weights, and *(iii)* the imputation of missing data employing the nearest neighbor technique. The statistical audit suggests that the indicator is appropriate and informative even though its structure is not entirely unidimensional.[5] Regarding its components, the environmental pillar seems to operate somewhat independently compared to the other three, resulting in its lesser contribution to the overall index. As noted in the report, it would be beneficial for the developers to conduct further analyses to emphasize the distinct scores (and rankings) stemming from the environmental pillar in contrast to the combined effect of the other three pillars.

Acosta et al. (2019) and Acosta et al. (2022) have also explored the robustness of the GGI, dealing with the uncertainty and sensitivity of the indicator. The first report (Acosta et al., 2019) explores the sensitivity through two principal sources of uncertainty of the input factors: the indicators (values and set) and the sustainability targets. The uncertainty analysis tests the assumptions made and methods used to build the model of the index. In particular, they evaluate for aggregation, normalization, weights, and outlier effects. This first self-assessment of the GGI reveals that changes in input values affect mostly countries with higher ranks than countries at the bottom. Conversely, under the GGI framework, different assumptions affect countries with lower ranks more than countries at the top of the ranking. This is mainly due to the greater deviation across indicators in countries at the lower bound. Beyond this preliminary analysis, (Acosta et al., 2022) evaluates the explanatory power of the GGI through correlation and regression analyses that estimate how much the indicator's variance explains the GGI's scores. This analysis aims at pinpointing the variation within the index and determining the significance of each indicator. While the GGI is found relatively robust to changes in the values of the indicators (especially for countries at the top and bottom of the ranking), missing values/data are found to affect the ranks.

These contributions reflect that the validation of composite indicators involves a multifaceted approach. On the one hand, researchers should explain the selection and handling of the raw data, highlighting potential biases arising from the sampling errors of the survey used to build composite indicators (Mauro et al., 2018a). This step should also involve an assessment of

[5] A composite indicator should serve as a unidimensional measure of the phenomenon studied, capturing the essence of the different variables required to construct it. In the case of the TPI, the principal components analysis carried out by the JRC was aimed at verifying the existence of a singular statistical dimension across the four TPI pillars. From a technical standpoint, they expected to find a sole principal component with an eigenvalue exceeding 1, or accounting for over 70% of the variance, meaning that the indicator captures a unique dimension that underpins the four pillars. Still, their results show a likely bi-dimensional structure of the indicator. The environmental pillar behaves differently concerning the other three pillars (European Commission, 2022)

the weighting and aggregation schemes, acknowledging before hand the consequences of the methodological decisions performed. On the other hand, a statistical and sensitivity exploration of the variables (or sub-indicators) should ensure that the composite indicator accurately measures the intended phenomenon without undue influence from irrelevant sub-indicators or interactions (European Commission, 2008). Following this binary scheme, Section 3.2 explores the data availability and limitations of the selected indicators, while Section 3.1 and Section 3.3 respectively described the technical aspects and the sensitivity tests.

# 3 Methods and Data

## 3.1 Normalization and aggregation

Although every composite indicator has its particularities they are all constructed following two consecutive steps: a normalization of its sub-indicators or components, and their aggregation.

**Normalization:** This step must be conducted before aggregating data since the components within a dataset typically vary in their measurement units. Aggregating diverse components such as health, income, education, pollution, and other social dimensions, makes it impossible to use regular units or levels. As reviewed in (Nardo et al., 2005), several normalization methods were developed including ranking, standardization, re-scaling, and measuring the Euclidean distance to a reference or categorical scale, among others.

Choosing a particular method requires careful consideration (Ebert and Welsch, 2004). In general, re-scaling normalizes components within a range, which in general is defined between 0 and 1. The five aggregate indicators under scrutiny in this study follow a min-max normalization, which performs a linear transformation on the original data to get the scaled data in the above-mentioned range. In doing so, the minimum and maximum values can be defined theoretically or empirically.[6] The normalization process turns country levels into relative positions, enabling cross-component comparisons and rankings. Countries performing well in a given component will usually present high normalized values, close to 1, while countries performing badly in a component will show small normalized values, close to 0.

The selection of in-sample or theoretical thresholds in the normalization process is not trivial. Indexes using sample thresholds are dependent on the sample composition, so the individual country scores will be affected depending on what are the countries they are being compared to. The value of the thresholds can change because a new country is added, because the value of the top/bottom country varies, or because a country over/under performs and reshuffles the rank, setting the new maximum/minimum. This way, when a new sample-based thresholds is defined, the final score of all countries will also change even if their components remain the same.

Moreover, if all countries improve by the same rate, including those defining the two limits of the normalization range, the composite indicator will remain the same for all countries instead of signaling a generalized improvement. Theoretical thresholds are more sensitive to changes in

[6] When the maximum and minimum values are defined empirically, this is, when they come from the highest and lowest values found in the sample, the thresholds and final scores may vary depending on the countries selected.

component levels. However, if these thresholds are too broad, or too narrow, they may also lead to distorted results. In the first case, they would not capture small advances in one component, while in the other they could exacerbate small changes.[7]

Components can be interpreted as "positive", measuring a dimension associated with an improvement in general welfare (i.e., higher education, life expectancy, more green zones in the cities...), while others may be thought of as having a "negative" effect (i.e. higher levels of pollution, higher material or organic waste, wider gender gaps...). Necessarily, the normalization process must consider the interpretation of the variable, ensuring that a high value corresponds to a positive outcome and a low value corresponds to a negative outcome. With this in mind, the normalization for component $x$, when interpreted positively, is obtained as a simple rank standardization:

$$x_{normalized} = \frac{x - min(x)}{max(x) - min(x)} \tag{1}$$

Analogously, normalization for component $x$ when higher values are interpreted with a negative connotation is obtained as:

$$x_{normalized} = \frac{max(x) - x}{max(x) - min(x)} \tag{2}$$

**Aggregation:** Once the normalization is executed, the different components must be aggregated into one measure that summarizes the performance in each country. The five selected indicators use standard aggregation techniques, based on arithmetic or geometric averages of the normalized component values. Choosing one or another approach is not trivial either. The geometric mean is used to reflect the cumulative effect of multiple factors. When considering deprivations, the geometric mean ensures that the overall measure is more sensitive to disparities. In other words, a high level of deprivation in one dimension can significantly impact the overall score, reflecting the idea that deprivations in different areas can compound and have a more severe impact on wellbeing. This approach contrasts with the arithmetic mean, where components are assumed to be perfectly substitutes in generating wellbeing. The geometric mean results in smaller values than the arithmetic mean (see Ravallion (2012) for a discussion), so the levels in the final composite indicators will necessarily be smaller. Multiplying components assigns a higher implicit weight to lower values, hence highlighting the bad effect of cu-

[7] Theoretical thresholds are set according to theoretical or empirical criteria. For example, the upper bound of the Gross National Income per capita in the PHDI is derived from Kahneman and Deaton (2010), who argue that there is no significant gain in human development and wellbeing from annual incomes above $75,000 (per capita).

mulative deprivations when bad scores obtained across components. Moreover, the geometric mean is restricted to positive values and is less sensitive to outliers, making it more robust when the component data is volatile. However, since component values are ranked beforehand, these properties have a minor effect on the composite indicators.

After normalization and aggregation, countries performing well in most components will occupy higher relative positions in the normalized component and will also score higher in the aggregate indicator, while the opposite will happen with countries performing poorly in many components.[8] Following this binomial scheme (normalization and aggregation) we proceed to illustrate the construction of the indicators.

The **PHDI** collects information on the three components conforming to the HDI (income, education, and health) plus two extra dimensions related to climate change (carbon dioxide emissions and material footprint per capita).[9] The index is built aggregating 7 components, one per dimension and two for education. The normalization follows Equation (1) and Equation (2) and use theoretical threshold values (UNDP, 2022). The maximum values are set to 75,000 USD for GDP per capita, 85 years for life expectancy at birth, 15 years for the mean schooling, 18 years for the expected schooling, 68.72 tonnes for carbon dioxide emissions per capita, and 107.42 tonnes for footprint per capita. Similarly, the minimum values are set to 100 USD for GDP per capita, 20 years for life expectancy at birth, 0 years for the mean and expected schooling, and 0 tonnes for carbon dioxide emissions and footprint per capita. After normalizing, for each country, the PHDI is estimated as follows:

$$PHDI = (LEI * EI * II)^{1/3} * \frac{C + MF}{2} \tag{3}$$

Where LEI stands for life expectancy, EI for the education index, II for income-related index, C for Carbon dioxide emissions, and MF for material footprint.[10] Note that the combination of the geometric mean of HDI components and the arithmetic average of the climate variables effectively act as weights, making the sensitivity of PHDI different for both sets of components. The latter component will have a higher relevance, provoking a strong re-ranking in countries with respect to the standard HDI index, as further discussed in Section 4.

The **TPI** ranks 28 components across four transition dimensions: economic, social, environmental, and governance. Normalization is performed according to Equation (1) and Equation (2),

[8] Both, the normalization and aggregation consist of trivial mathematical operations that pose no software or hardware limitations.

[9] See Ravallion (2012) for a broad review of the HDI index. Its main advantages and limitations also apply to the PHDI.

[10] The education index EI is computed as the arithmetic average between the average years of schooling and the expected years of schooling normalized values.

multiplying the resulting rank scores by 100 and using the theoretical thresholds defined in European Commission (2022). The aggregation method, follows a weighted arithmetic average, where economic and social-related components weigh 0.2, environmental 0.35, and governance 0.25. Weights were decided with expert opinions, and then corrected to enhance robustness. After normalizing components, the TPI is computed as follows:

$$TPI = \sum_{d=1}^{4} w_d \sum_{i=1}^{N_{id}} \frac{1}{N_{id}} x_{id} \tag{4}$$

Where $N_{id}$ is the number of components belonging to dimension $d$, $x_{id}$ is the score of component $i$ under dimension $d$, and $w_d$ represents the weight assigned to each dimension.

The **BLI** aggregates 24 components across 11 dimensions of well-being, ranging from subjective life satisfaction to life expectancy at birth. The normalization is performed according to empirical values defined in the sample of countries and the aggregation consists of a simple arithmetic average (European Commission, 2022). For each normalized component $i$ belonging to dimension $d$ ($x_{id}$), the BLI is computed as:

$$BLI = \sum_{d=1}^{11} \frac{1}{11} \sum_{i=1}^{N_{id}} \frac{1}{N_{id}} x_{id} \tag{5}$$

The **GGI** collects information on 36 components aggregated into 16 categories and four main dimensions or goals: efficient and sustainable resource use, natural capital protection, green economic opportunities, and social inclusion. The normalization process follows Equation (1) and Equation (2) and multiplies the scores by 100. Thresholds are defined empirically, based on the sample (Acosta et al., 2019). After normalizing, the components ($i$) are aggregated using an arithmetic mean by categories ($c$) ($x_{ic}$).

$$GGI_{category} = \frac{1}{N_{ic}} \sum_{i=1}^{N_{ic}} x_{ic} \tag{6}$$

In a second and third level, a geometric aggregation is applied to the 16 categories and the 4 dimensions, in sequence:

$$GGI = (\prod_{1}^{4}(\prod_{1}^{16} GGI_{category})^{1/16})^{1/4} \tag{7}$$

Similarly, the **SDG** aggregates across 114 components (85 global and 29 specific for OECD economies) grouped into 17 sustainable development goals. The normalization is performed according to Equation (1) and Equation (2), multiplying the resulting rank scores by 100 and taking empirically defined lower and upper bounds, the latter being defined after several theoretical steps (Sachs et al., 2019). Once more, the SDG first aggregates normalized components $i$ within each dimension or goal $d$ ($x_{id}$) using an unweighted average mean. Then, the 17 goals are averaged using an arithmetic mean to conform the final SDG.

$$SDG = \sum_{d=1}^{17} \frac{1}{17} \sum_{i=1}^{N_{id}} \frac{1}{N_{id}} x_{id} \tag{8}$$

Table 3.1 reviews the main features of the five selected indicators. The first column presents its components and how they are aggregated into broader dimensions. The second column summarizes the normalization method employed, and the third column describes the aggregation method, geometric or arithmetic, and whether weights are used.

Excluding the PHDI, all indicators aggregate their components using an arithmetic mean, hence implying certain assumptions. The arithmetic mean implicitly compensates between indicators that, in principle, should not be substituted (Mazziotta and Pareto, 2013), hence ignoring the heterogeneity between variables. For example, a low value in the "Life Expectancy" component in the BLI can be compensated by a high value in the "Voter Turnout". Concerned with this effect, the UNDP decided to update the HDI in 2010 and substitute the arithmetic by the geometric mean (see a complete discussion in Ravallion (2012)). As explained before, the geometric mean also bears its assumptions and shortcomings (Mauro et al., 2018b).

The TPI is the only indicator including a specific weighting function, thus assigning a different level of importance to its components. The other indicators assume equal weights. As noted in Nardo et al. (2005), equal weighting does not imply "no weights", but rather assumes that all components are equally important. Nevertheless, when the components are grouped into different dimensions and further aggregated into the composite indicator, equal weighting may become unequal, as dimensions capturing a larger number of components will implicitly receive a higher weight. As in the case of the PHDI, this design may lead to an unbalanced structure of the indicator. Employing equal weights may risk double counting when two or more collinear indicators are aggregated without adjusting their weights for this effect, resulting in an indicator being implicitly weighted higher than intended (Freudenberg, 2003; Greco et al., 2019). Indeed, equal weighting disregards implicit effects arising from mutual dependence across variables in the indicator structure.

Table 3.1: Summary of Indicators and Aggregation Methods

| Indicator | Areas-Components | Normalization | Aggregation |
|---|---|---|---|
| PHDI | HDI: 4 components aggregated into 3 dimensions. PP: 2 components | Min-max transformation with fixed values | HDI geometric mean adjusted by the PP arithmetic mean |
| TPI | 28 components aggregated into 4 dimensions | Min-max transformation with fixed values | Dimension-specific weights with arithmetic averaging across dimensions |
| BLI | 24 components aggregated into 11 dimensions | Min-max transformation based on sample values | Equal weights with arithmetic aggregation |
| GGI | 36 components aggregated into 16 indicator categories, further aggregated into 4 dimensions | Min-max transformation based on sample values | Equal weights with arithmetic aggregation of normalized components and geometric aggregation of indicator categories and dimensions |
| SDG | 114 components aggregated into 17 SDGs (85 global and 29 specifically for OECD countries) | Min-max transformation with fixed values | Equal weights with arithmetic mean of components for each goal and average scores across all 17 goals |

Note: Own elaboration.

## 3.2 Data availability

Now we move to appraising the data accessibility associated with the five indicators. Given our emphasis on the European Union (EU), our discussion focuses on the availability of data within the context of the 27 countries currently conforming to the EU-27, taking 2019 as the reference year.[11] We have chosen 2019 because it is the most recent year with wide data availability before the onset of the COVID-19 pandemic. This global event possibly had an uneven effect on several components. Take the PHDI as an example. The low geographical mobility may have smoothed carbon emissions, but at the same time, the high mortality might have also affected life expectancy. Bearing this in mind, to avoid spurious results, we try to focus on the previous year. For completeness, data availability has been addressed for a decade (2011-2021).

[11] The BLI is only computed on those countries from the EU-27 belonging to the OECD.

The necessary data to compute the PHDI is provided by the United Nations Human Development Reports (UNHDR) repository.[12] For our selected sample of countries all information is complete between 1995 and 2021, so the index can be easily computed after downloading the data.

The European Commission (EC) provides the data needed to construct the "The Transitions Performance Index" (TPI).[13] The dataset is rather complete for the period of analysis. Whenever data was missing, the developer imputed the values following diverse procedures (European Commission, 2022). All data is publicly available, except the "Energy productivity (2015 PPP per kilogram of oil equivalent)". This proprietary component is obtained from the International Energy Agency (IEA), and the European Commission is therefore unable to share it.[14]

Raw data for the "Better Life Index" (BLI) is available at OECD Statistics, although it poses several limitations.[15] First, the data cannot be compared across time because each edition of the BLI is a snapshot of the most updated values at the time of the release. Past editions are not revised backward to account for revisions in the time series. Second, the website to download the data is not the same across editions, because the BLI database does not feature past editions but only the most recent ones. Third, the BLI does not cover EU-27 countries not belonging to the OECD, so there is no information for Bulgaria, Cyprus, Croatia, Malta, and Romania. Fourth, values have been imputed for a small number of observations but, while editions 2013-2017 include the imputation estimates, the 2019 and 2020 raw data sets still contain missing values. For instance, the component "Time devoted to leisure and personal care" presents missing values in 9 out of the 22 countries over which the data is available.[16]

The Green Growth Index is produced by the Global Green Growth Institute (GGGI).[17] Some components capture different concepts across editions. For instance, "AB3" stands for "Internet broadband and mobile cellular subscriptions" in 2019, but it represents "Universal access to sustainable transport" in 2022. All input data for the index, including imputed values, can be easily downloaded. Still, we find recurring missing values across the available waves (2000-2020). Out of its 36 components, 3 indicators have one or two missing values. Most remarkably, the variables "Tourism and recreation in coastal and marine areas (CV2)" and "Proportion of urban

[12] Data: https://hdr.undp.org/data-center/documentation-and-downloads.

[13] Data: https://research-and-innovation.ec.europa.eu/strategy/support-policy-making/support-national-research-and-innovation-policy-making/transitions-performance-index-tpi_en.

[14] The necessary data to compute the TPI in this report was provided by the IEA.

[15] Editions 2013 to 2017 can be found here: https://stats.oecd.org/Index.aspx?DataSetCode=BLI2017. Edition 2019 in https://www.oecd-ilibrary.org/social-issues-migration-health/data/oecd-social-and-welfare-statistics/better-life-index-edition-2019-1_74ade212-en. The last edition (2020): https://stats.oecd.org/index.aspx?DataSetCode=BLI or https://www.oecdbetterlifeindex.org/.

[16] For 2020, the imputed values cannot be downloaded, but the imputation values can be obtained when checking the scores country by country by visiting the webpage (https://www.oecdbetterlifeindex.org/).

[17] Data: https://ggindex-simtool.gggi.org/SimulationDashBoard/downloads.

population living in slums (SP3)" are not available for 5 and 24 countries, respectively. Data gaps for each country are presented in Table A1.14 of the GGI Technical Report.[18] The Green Growth Performance Measurement (GGPM) team capped the values in several indicators to avoid outliers (see the above-mentioned report for details). The data does not include these corrections, so the researcher should consider this step when building the index.[19]

Finally, the data used to construct the Sustainable Development Goals (SDG) is freely downloadable at the SDG datasite.[20] To minimize biases from missing observations each edition of the SDG only includes countries with valid data for at least 80% of the indicators, that have been available in previous editions and have data for at least 75% of the indicators. The raw time series data (2000-2023) is rather incomplete, and information for all components is only available for the specific year/edition of download. Additionally, the components have varied over the years, hindering its comparability over time. Due to the lack of widely accepted statistical models for imputing country-level data for many SDG priorities, the data managers do not impute or model missing data beyond a few exceptional circumstances. As a result, various indicators are missing for several countries even in the complete databases, making this data the most incomplete source. For instance, in 2019, "Nitrate in groundwater" data was absent in 11 EU-27 countries, and "Access to justice" data was missing in 8 countries.[21] Missing observations may reduce statistical power, bias estimation of parameters, reducing the representativeness of the samples and increasing the complexity of analysis.

## 3.3 Sensitivity analysis

Appropriate one-dimensional inequality, poverty, and welfare indicators should be sensitive to the distribution of values in the variable they measure. For instance, the most widely used inequality measures, the Gini and the General Entropy Indexes fulfill the principle of transfers (Cowell (2011)); the index rises when a poorer individual transfers resources to a richer individual. Similarly many poverty measures, such as the Foster-Greer-Thorbecke, generally include a focus or anonymity axiom and are sensitive only to income changes among the poor (Foster et al. (2010)). As such, measuring the sensitiveness of these indexes is a task that involves perturbations at different parts of the distribution and checking their reaction. In contrast, measuring the sensitivity of multidimensional indicators, which by construction aggregate information on

[18] From Acosta et al. (2019), the data is provided in https://greengrowthindex.gggi.org/wp-content/uploads/2019/12/Green-Growth-Index-Technical-Report_20191213.pdf.

[19] We have excluded the variable SE2 -Urban-rural access electricity- from the analysis because all countries in our sample had the maximum value, so results from simulation exercise proposed in Section 3.3 could not be properly interpreted.

[20] The most recent data can be found in: https://dashboards.sdgindex.org/downloads.

[21] Full Data 2019: https://www.sustainabledevelopment.report/reports/2019-europe-sustainable-development-report/

different aspects, is not so straightforward.

As a matter of example take the BLI, which includes up to 24 different components. Measuring the impact of distributional changes within each component and its subsequent impact on the composite indicator (BLI) would probably be too exhaustive and not easy to expose. Besides, they may not be informative when distributional changes do not affect the final indicator, for instance, if changes in the distributional tails in two components compensate each other due to the arithmetic aggregation process. Because composite indicators focus on the final aggregated score and neglect within-distributional aspects, addressing their sensitivity involves measuring how they react to changes in the component values themselves (Freudenberg, 2003).[22]

Taking the partial derivative of each composite indicator with respect to each component is a feasible idea (Ravallion, 2012). As a rule, the elasticity $\epsilon$ of an indicator $I$ with respect to component $x_i$ could be derived as:

$$\epsilon_{I,x_i} = \frac{\partial I}{\partial x_i} \times \frac{x_i}{I} \tag{9}$$

These indicators are obtained from simple geometric or arithmetic aggregations, often unweighted, so the partial derivatives will be identical across components. Thus, $\epsilon_{I,x_i}$ would only depend on the specific $I$ and $x_i$ values. Besides, $x_i$ is introduced in $I$ after normalizing, so $\epsilon_{I,x_i}$ captures the elasticity of the indicator to changes in the normalized value, instead of capturing the sensibility of the indicator to changes in the component absolute value before normalization. This way, estimating the elasticity following Equation (9) would hinder a complete analysis of the association between the (non-normalized) component values and changes in $I$.

Bearing these limitations in mind, instead of computing the standard elasticity $\epsilon_{I,x_i}$, we perform a simulation exercise similar to that proposed in Saisana et al. (2005), Papadimitriou et al. (2019) and Acosta et al. (2022). In a nutshell, we simulate the sensitivity of each indicator $I$ to changes on a given component $x_i$, first, by implementing a modification in each component, and then, by addressing the associated change in the indicator. Thus, the sensitivity of the indicator I to variations in the component $x_i$ ($\triangle_{I,x_i}$) can be expressed as:

$$\triangle_{I,x_i} = \frac{I' - I}{I} \tag{10}$$

With $I'$ being the value of the indicator *after* the perturbation, and $I$ being the value of the indicator *before* the perturbation.[23] Our exercise is very simple. Departing from the first component,

[22] Indeed, this could be thought of as a limitation of these indicators. For instance, the PHDI and the TPI employ the GDPpc, such that a rich country with high inequality levels, ceteris paribus, would rank higher than a not-so-rich country with much lower inequality levels (Ravallion, 2012).

[23] Similarly, $\triangle x_i = x' - x$, where $x'$ is the value of the component *after* the perturbation, and $x$ is the value of

we increase the observed value by some percentage points $p$ in a given country.[24] All values are then normalized and used to estimate the new composite indicator. Repeating this exercise for every country delivers 27 new final measures, that can be used to estimate their relative increase/decrease with respect to the original value. This relative change captures the direct indicator change when a given component is increased. The average values of these 27 relative changes would reflect the mean sensitivity of the composite indicator to some percentage point changes in the selected component. The exercise is then repeated for every single component of the composite indicator. The algorithm set to find these sensitivity values for each component can be schemed in this way. We depart from an indicator *I* with components *I*$(x_1, x_2, .., x_n)$. Each component is a vector containing country *(c)* specific values, $x_1(x_{1c_1}, x_{1c_2}, ..., x_{1c_s})$. For each $x_i \in$ *I*:

1. Define a p% increase. In our case, *p=0.02*.

2. Define maximum and minimum values of component $x_i$: (min $x_i$, max $x_i$).[25]

3. Take a country and apply the change $(x'_{ic1} = x_{ic1} * (1 + p))$.[26]

4. Normalize $x'_i$ using (min $x_i$, max $x_i$) defined in step 2.

5. Estimate the new indicator *I'* and calculate the sensitivity measure as the relative change with respect to the original indicator *I* as in Equation (10). [27]

Our approach bears some limitations. First, it only attends to partial direct effects and ignores associations across components. For instance, consider two components in the PHDI, education (EI) and income levels (II). Indeed it seems reasonable that a higher education index provoked, for instance, by an educational reform, could also lead to a higher income-related index. In such a case, a rise in EI would increase the PHDI in two ways: a direct effect (EI rises, so PHDI also rises) and an indirect effect (EI rises, making the II rise too, such that it also increases

the component *before* the perturbation.

[24] As explained in Section 2 some components have a "positive" connotation, while others are interpreted as "negative". Naturally, rising some percentage points will increase the value of the indicator in the former case, and decrease the value of the indicator in the latter case.

[25] Depending on the indicator, these values are based on the original sample of countries or theoretical thresholds, see Table 3.1.

[26] When the new non-normalized value exceeds the theoretical or sample-based maximum, we assign the maximum value.

[27] When the original value of the non-normalized component equals the theoretical or sample-based maximum, the normalized version receives the maximum score (1 or 100, depending on the component). In these cases, the indicator will not change when we increase its component because the rank associated with the new normalized value remains the same.

the PHDI value).[28] Our approach only considers the former effect, ignoring the latter. Analyzing complex relations across the many components in the selected indicators is beyond the scope of this paper, as we simply focus on the statistical association of changes that can be produced by real changes in the values, but also by measurement errors.

The value of the perturbation $p$ is the same in all simulations, thus reducing the sensitivity of the indicators to components where country ranks are more scattered. We considered using an endogenous value of $p$, for instance, setting $p = share * SD(x_i)$, this is, taking a share (like 10%) of the standard deviation of the values in component $x_i$. This option was disregarded because sensitivity values were extremely sensitive to components affected by outliers (like the GDPpc in the PHDI or net wealth in the BLI), thus bearing unrealistic results.[29]

The sensitivity levels heavily depend on $p$. For simplicity, we focus our main results on $p = 0.02$, thus showing the reaction of the indicator to a $2\%$ change in its components. We run simulation exercises with many other $p$ values, and the main ideas remain.[30] We consider 2% to be a realistic value. Take, once more, the PHDI. In Europe, sustained rises in life expectancy or education rates beyond 2% are hardly possible. Aimed at strengthening our main conclusions, Section 4.3 relaxes the $p = 2\%$ assumption and repeats the exercise applying bootstrapped random values to $p$. Note that this robustness analysis somewhat deals with the former limitation of the $p$ values being the same across components.

[28] This indirect effect may also include lagged associations between components, making the analysis even more cumbersome.

[29] An example can be found in Figure A.17, where the sensitivity for the BLI indicator is estimated with 10% of each component-specific standard deviation. Income and wealth-related variables have a much larger standard deviation than rates or shares, by definition bounded between 0 and 1, or 100. The distinct nature of the variables makes their statistics hardly comparable in this exercise. We repeated the complete exercise including Turkey, a (bottom) outlier, and our main results remained similar. Since our sample already includes some of the richest and most developed countries in the world, we believe it is unrealistic to include outliers in the top tail. We thank Jorge Davalos Chacon and Luca Tiberti for this insight.

[30] The association is simple: higher $p$ values lead to higher sensitivity. These robustness exercises are available upon request.

# 4 Results

## 4.1 Baseline analysis

To benchmark the analysis and explore our baseline data, Figure 1 deploys six maps with the results of the five indicators plus the standard HDI index, estimated with 2019 data for the EU-27 countries (see Section 3.2). Including the HDI index, widely used in the well-being literature, allows us to compare the similarities and differences across the score values in the five main indicators under scrutiny.[31]

To begin with, the PHDI demarcates two well-separated regions. On the one hand, relatively high values are found in Western Europe, especially in Spain, Greece, Sweden, and Italy (ranging between 79 and 83 points). These values contrast with the Eastern European countries, who joined the EU in the middle of the 2000s and early 2010s. In these economies, the scores move between 0.65 and 0.75. Notable exceptions to this geographical division are found in Ireland, Netherlands, Finland, and Belgium, which exert values closer to the latter group rather than to the more similar Western economies.

The gradient is more clear for the remaining indicators. The Nordic countries (Finland, Sweden, Denmark) tend to perform better than the rest. The next step in the ladder is conformed by Central-European Western countries (Germany, Netherlands, Belgium, Luxembourg, and Austria). The third group includes the western Mediterranean countries (Spain, Portugal, France, Italy) and Ireland, with scores often lying between 50 and 65 in the TPI, GGI, and SDG, or 0.55 and 0.60 in the BLI. Finally, among the EU-27, Eastern European economies and Greece tend to exert the worst values.

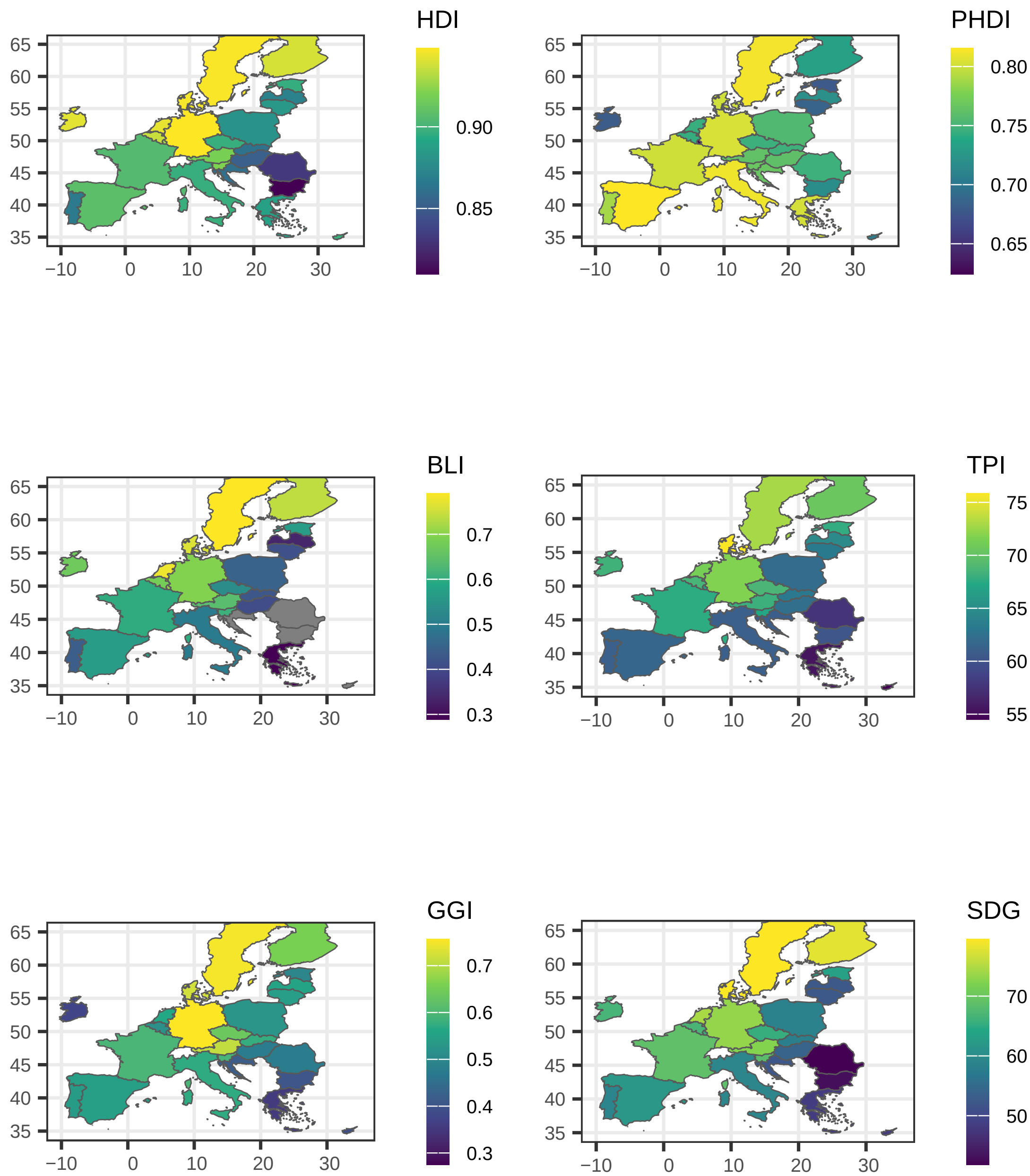
The different geographical division obtained with the PHDI is better understood when looking at the original HDI. Note that the HDI pattern highly resembles the one found for the GGI, SDG, BLI, and TPI, with the Nordic and center-European Western countries experiencing high scores, the Mediterranean presenting middle scores, and the Eastern showing relatively low scores. The difference in the ranking of the PHDI can thus be attributed to the climate-related variables, which unevenly push the scores down, partially due to the differences between the geometric and arithmetic averaging explained in Section 3.1.

Overall, the scores align with the conventional rationale and do not yield surprising results in terms of country ranks and levels. While the scores in the other four indexes studied, without taking into account the HDI, range between 0.5 (or 50) and 0.75-0.8 (or 75-80), the scores in BLI

[31] Its estimation simply requires assigning a value of 1 to "carbon dioxide emission (C)" and "material footprint (MF)" components in Equation (3).

range between 0.35 and 0.7, with the highest scores being achieved in Sweden and the lowest in Greece and Latvia. Note that, as discussed in Section 3.2, non-OECD countries -colored in grey- are excluded in this component.

Figure 1: Indicator by country




Source: Authors' calculations. Data: UNHDR (PHDI, HDI), GGGI (GGI), EC (TPI), OECD (BLI), SDG (SDG).

## 4.2 Sensitivity analysis

### 4.2.1 The Planetary Pressure Adjusted Human Development Index - PHDI

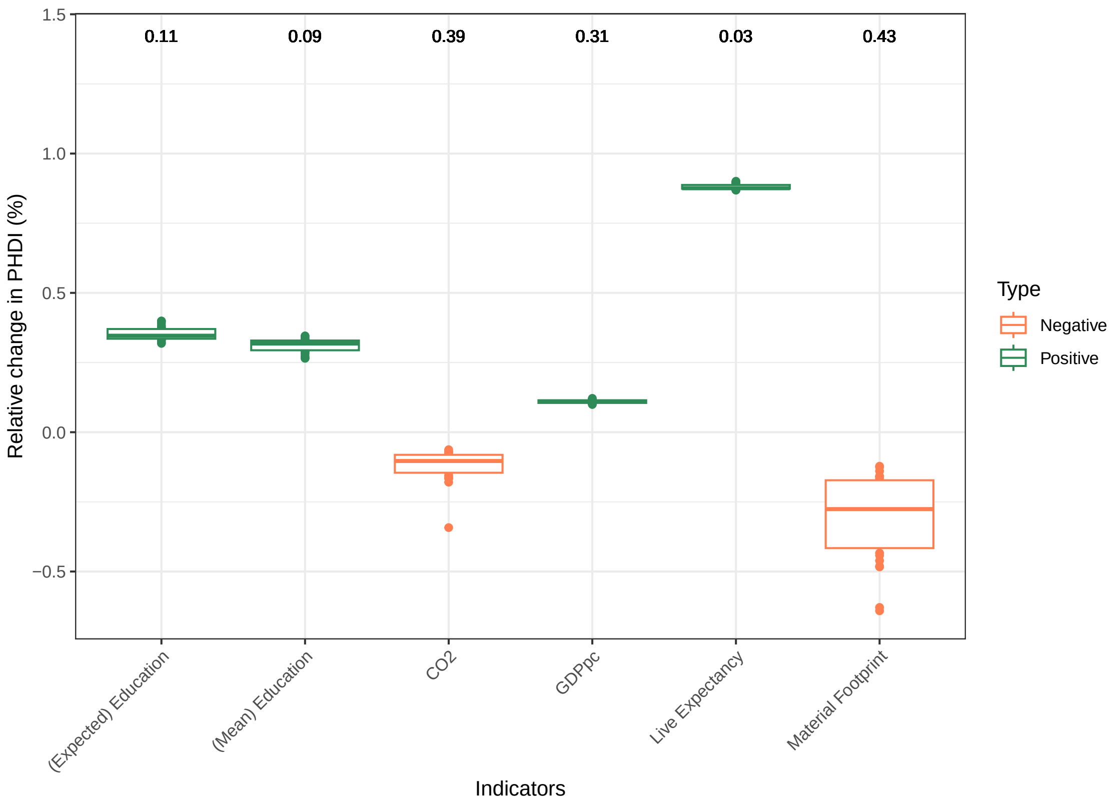
Figure 2 explores the sensitivity of the PHDI and shows the relative change in the score after increasing by 2% the values in each component. Since the value is changed for each country separately, we obtain 27 new PHDI scores for each component under this exercise. The 27 sensitivity values are shown as box plots, with the interquartile range delimiting the upper and lower sides and the average sensitivity being expressed as a horizontal line within the box. Variables associated with positive outcomes, such as education, income, and life expectancy, are colored in green, and have a positive sensitivity, meaning that a 2% increase in the component is associated with a rise in the indicator score. Analogously, variables associated with negative outcomes, such as carbon emissions and material footprint, show negative values and are colored in red.

We find the PHDI to be rather insensitive to changes in the scores of the components, with the relative change being often below |0.5%|. For instance, a 2% rise in per capita GDP is associated with a 0.2% increase in the PHDI, with all values being very clustered around the mean. The map deployed in Figure 1 would remain unchanged after a perturbation in the component. Interestingly, all positive components show a very low dispersion, while negative elements, associated with climate-related variables, are more scattered.

The highest sensitivity is associated with life expectancy (0.85%). To further investigate the possible causes, we refer to Equation (3). The aggregation method weights education, health, and income equally, so the construction of the index is not responsible for different sensitivity levels among these components. Thus, we use the coefficient of variation (CV) to explore the dispersion in the values before normalization.[32] The CV values are shown in all figures right over the corresponding boxplots, such that in Figure 2 the expected education has a CV of 0.11.

[32] The CV of a variable x is estimated as $CV = \frac{Sd(x)}{\bar{x}}$, with $Sd$ being the standard deviation and $\bar{x}$ the sample mean of x.

Figure 2: Planetary Pressures–Adjusted Human Development Index




Source: Authors' calculations. Data from UNHDR.

The CV of life expectancy is very low, merely reaching 0.03. This means that the values in the non-normalized variable are very clustered around the mean, such that the ratio between the standard deviation and the mean is tiny. In these cases, a small change in a value may provoke a big change in the country's ranking. If the theoretical values delimiting the normalization are narrow enough, the new normalized values will be very different when compared to the baseline normalized values. Countries with originally low baseline ranks could escalate up to high positions, thus affecting their country rank and, after aggregation, the indicator scores. Ultimately, this sensitivity will be reflected in higher sensitivities.

As a matter of example, consider two components $\{x, y\}$ measuring two dimensions in five countries $\{A, B, C, D, E\}$. The first component is assigned to countries such that $x = \{1, 2, 3, 4, 5\}$, while the second is $y = \{1001, 1002, 1003, 1004, 1005\}$. Note that $\bar{x} = 3$, $Sd_x = 1.58$, and $CV_x = 0.53$, while $\bar{y} = 1003$, $Sd_y = 1.58$, and $CV_y = 0.002$. Although both components have the same standard deviation, the means are entirely different, so the CV is much higher in the former than in the latter. For component $x$, a 2% change in the values does not alter the ordering: if country A raises its value by 2%, the new set is $x' = \{1.02, 2, 3, 4, 5\}$. The rank remains the same, and thus the index remains unaffected after normalizing and aggregating. However, if country A rises by 2% its value in the $y$ component, the new set is $y' = \{1021, 1002, 1003, 1004, 1005\}$. Regardless of how thresholds in the normalization are defined, the rank of countries within the component is heavily affected, with the indicator reflecting a higher sensitivity.

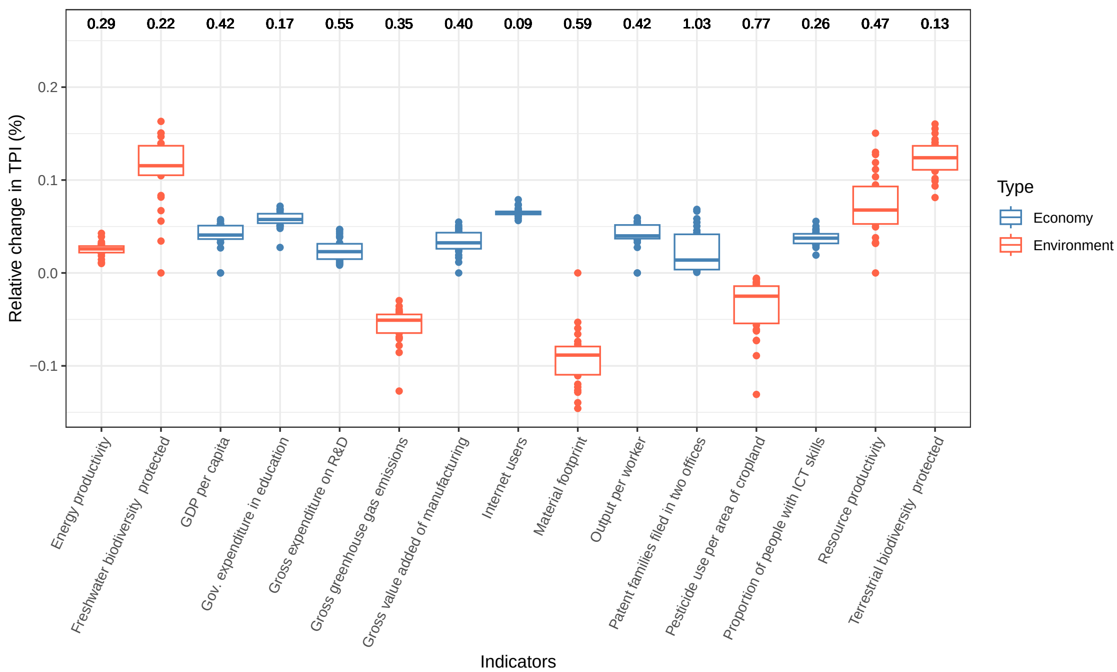
The association between the dispersion of non-normalized component values and the sensitivity of the final indicator prevails in all five measures. Low CV values are associated with changes in the ranks during the normalization and a higher sensitivity in the indicator scores. This poses a significant shortcoming to these measures, as the indicator sensitivity is not driven by the theoretical or wellbeing relevance of the components, or their level, but rather by their ability to change country rankings. This dimension is affected by a variety of factors, such as the thresholds or the measurement unit over which the normalization is applied. If these factors are not carefully considered in the definition of the index, improvements in the levels of components that, in principle, may be relevant for the indicator, would have a limited impact if they do not alter country ranks.

### 4.2.2 The Transition Performance Index - TPI

Figure 3 and Figure 4 show the sensitivity values for the TPI index. The first graph displays results for those components composing the Economic and Environmental dimensions, and the second includes the components that build the Governance and Social dimensions. Once again, the sensitivity values are small, evidencing the insensitivity of this index to changes in its components. Results and country orderings shown in Figure 1 would be barely affected. The greatest sensitivity is found when the component "Healthy life expectancy at birth" increases by 2%, which with a CV of 0.03 poses an average sensitivity of 0.2%. Other components, such as "the homicide rate" or "the employment ratio gender gap for those aged above 25" do not alter the TPI, the CV values being 0.78 and 0.28, respectively. Remember that each dimension is weighted differently in the TPI: economic and social-related components weigh 0.2, environmental 0.35, and governance 0.25. Thus, although small CV values are associated with more sensitivity, components with the smallest CV do not always present the largest changes in the TPI. Focusing on Figure 3, "Internet users" has the smallest CV (0.09), but there are other environmental variables with a higher CV but larger sensitivity to changes in the component, like "Material footprint" or "Greenhouse emissions".

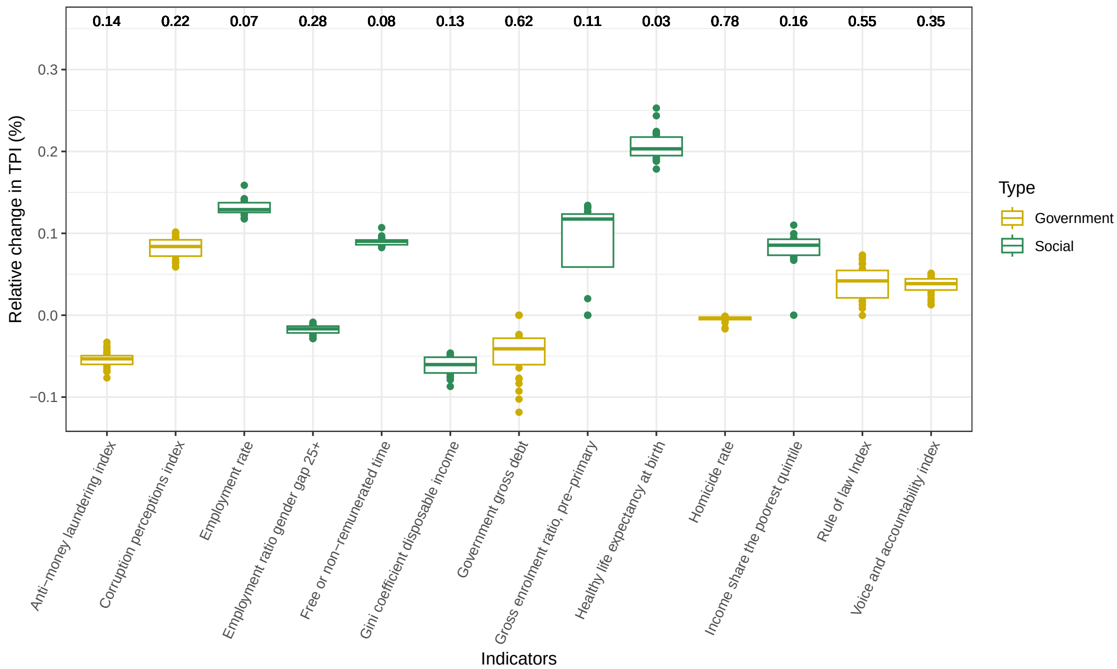
As already mentioned in Section 3.3, the relative change equals 0 when the original value of the non-normalized component equals the theoretical or sample-based maximum. Since the normalized version has the maximum score (1 or 100, depending on the component) before and after applying the 2% change, the indicator will remain unchanged. For instance, in 2019, GDP per capita (PPP $) was 120,490 dollars in Luxembourg and 91,812 dollars in Ireland, values well above the theoretical maximum of 75,000 dollars, so both countries obtained the maximum value after normalization. Given that the variable was already above the maximum threshold, increasing GDP pc 2% will not change the normalized score nor the indicator. In fact, the point-taking value 0 in Figure 3 corresponds to these two countries.

Figure 3: Transition Performance Index (Economic and Environment)




Source: Authors' calculations. Data from EC.

Figure 4: Transition Performance Index (Government and Social)



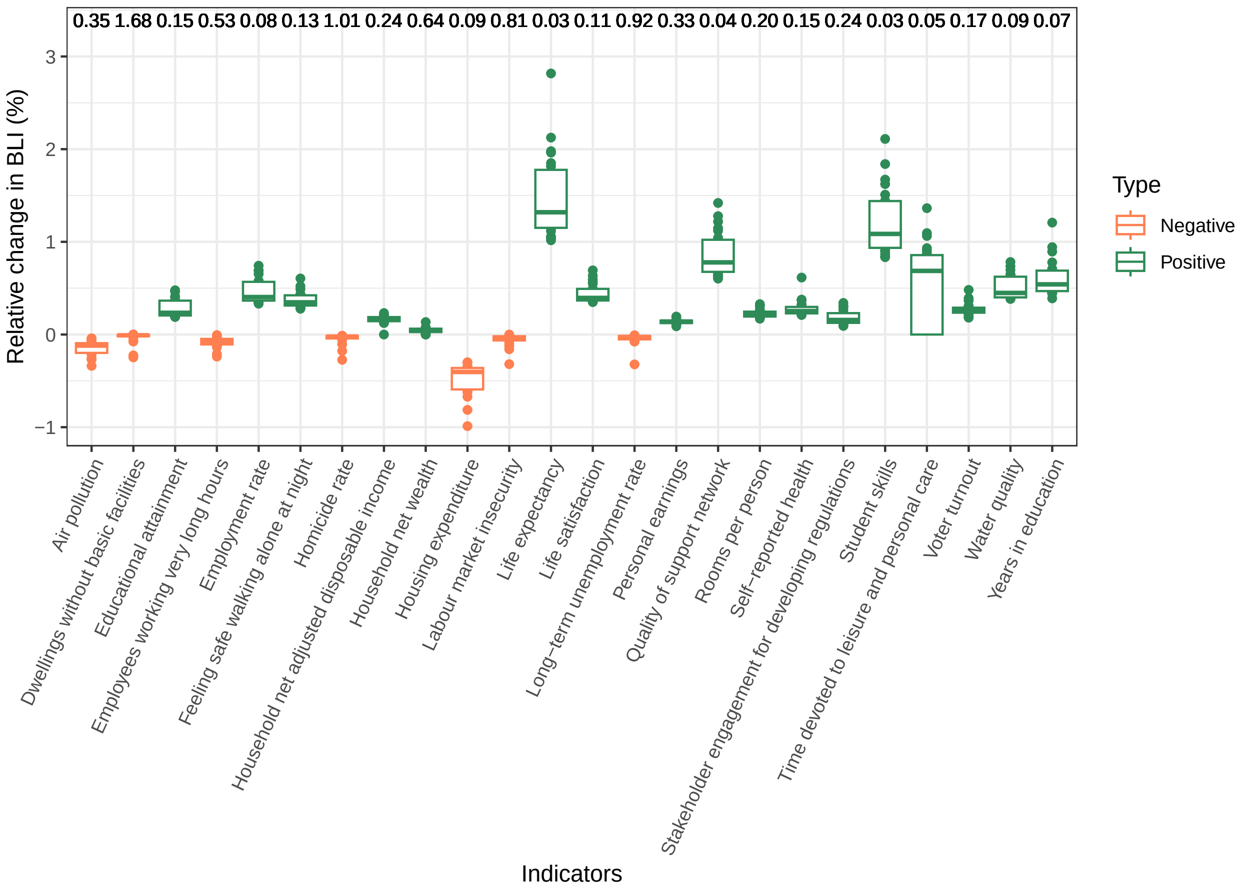

Source: Authors' calculations. Data from EC.

### 4.2.3 The Better Life Index - BLI

Figure 5 presents the relative change in the BLI. Overall, sensitivity values are higher than in other indexes, most of them ranging between 0.4% and 1.5%. Low CVs in "Life Expectancy" and "Student Skills" (0.03 and 0.05, respectively), together with "Time devoted to leisure and personal care" (0.05) are associated with high sensitivity values. Components such as "Homicide rates", "Household net wealth", "Disposable income", "Long-term unemployment rates", or "Air pollution", which one could in principle consider as relevant indicators for a "Better Life", are characterized by a high CV and thus merely affect the BLI.[33]

[33] Since the BLI normalizes the components taking the maximum and minimum values from the sample, we have checked the robustness of the results by including a somewhat different country. We have included Turkey because the BLI only considers OECD countries and is one of the few available countries whose values differ from European standards, and may thus alter sample-based maximum and minimum. We find the results unchanged, suggesting that neither the normalization nor the indices are affected by the new min-max values defined with Turkey. Results are available upon request.

Figure 5: Better Life Index

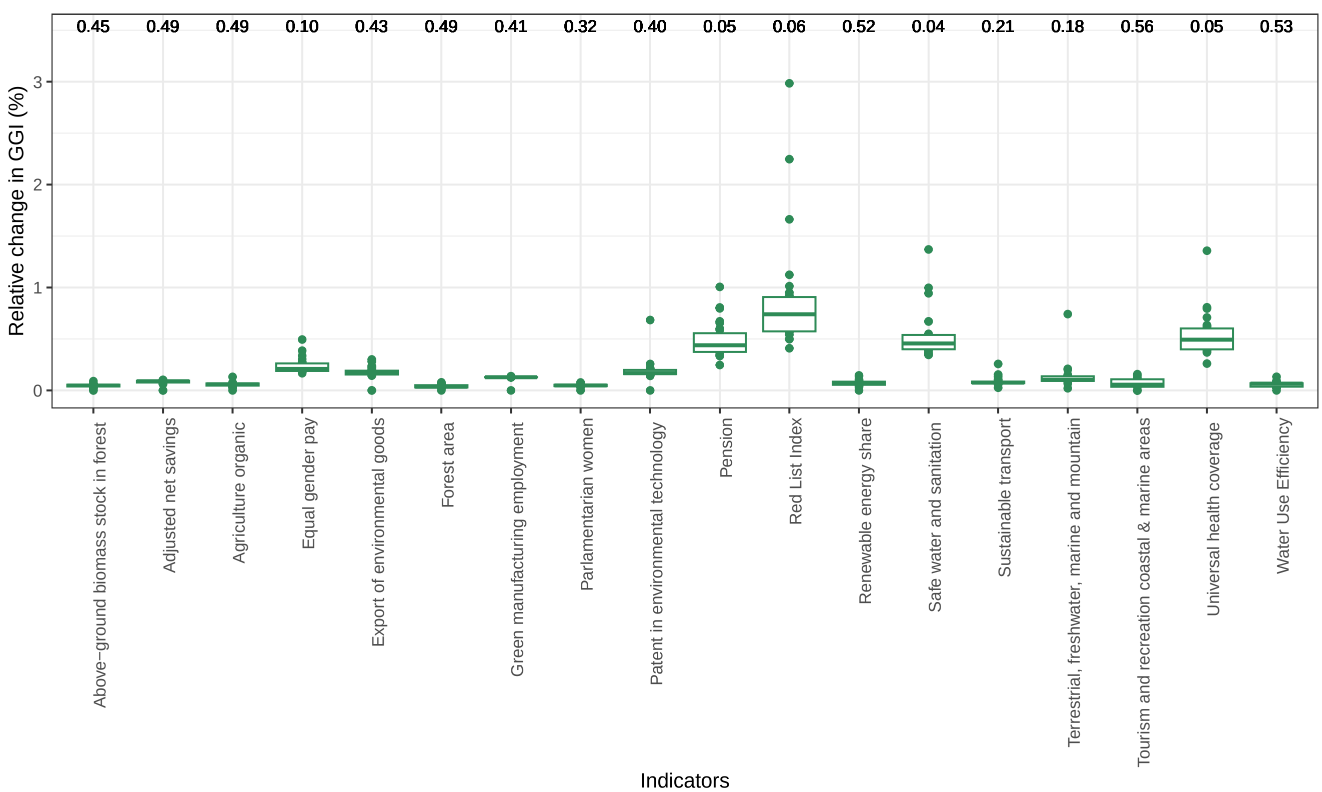



Source: Authors' calculations. Data from OECD.

### 4.2.4 The Green Growth Index GGI

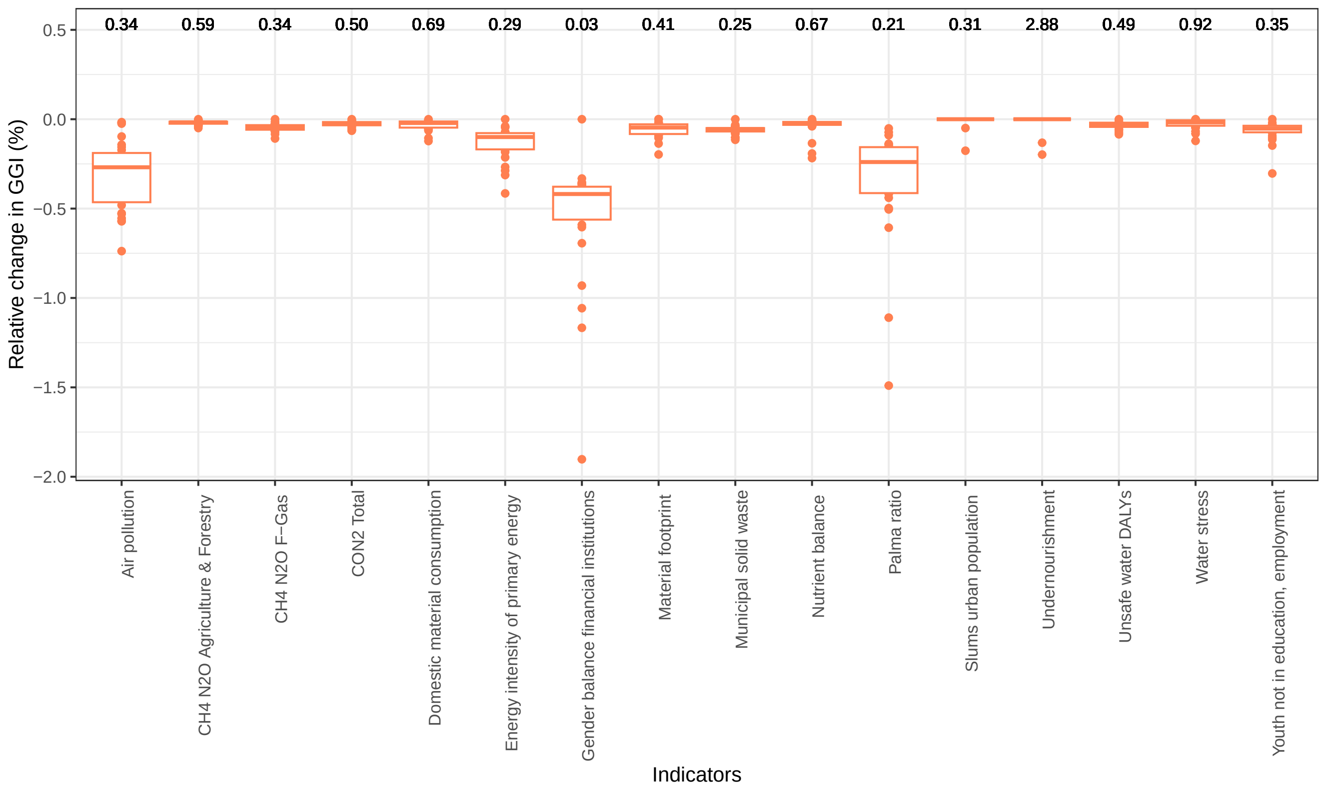
Figure 6 displays the results for components that positively contribute to the GGI. Overall, this index shows a higher sensitivity when compared to the TPI and PHDI, with all sensitivity mean values ranging between 0.2% and 1%. Again, components with CV values below 0.1, such as "Pension", "Red List Index", "Safe water and sanitation", and "Universal Health Coverage", are associated with higher sensitivity values, with some outliers rising the index to relatively high values. On the contrary, components expected to be more related to green growth, like "Water use efficiency" or "Forest areas" present low sensitivity values and are respectively associated with CVs of 0.53 and 0.49. This rationale also applies to the negative components. As shown in Figure 7 all sensitivity values are very close to zero except the "Gender gap in financial account ownership", which has the smallest CV (0.03). It is remarkable how insensitive is the Green Growth Index to components related to climate, pollution, and resource use.

Figure 6: Green Growth Index (Positive)




Source: Authors' calculations. Data from GGGI.

Figure 7: Green Growth Index (Negative)



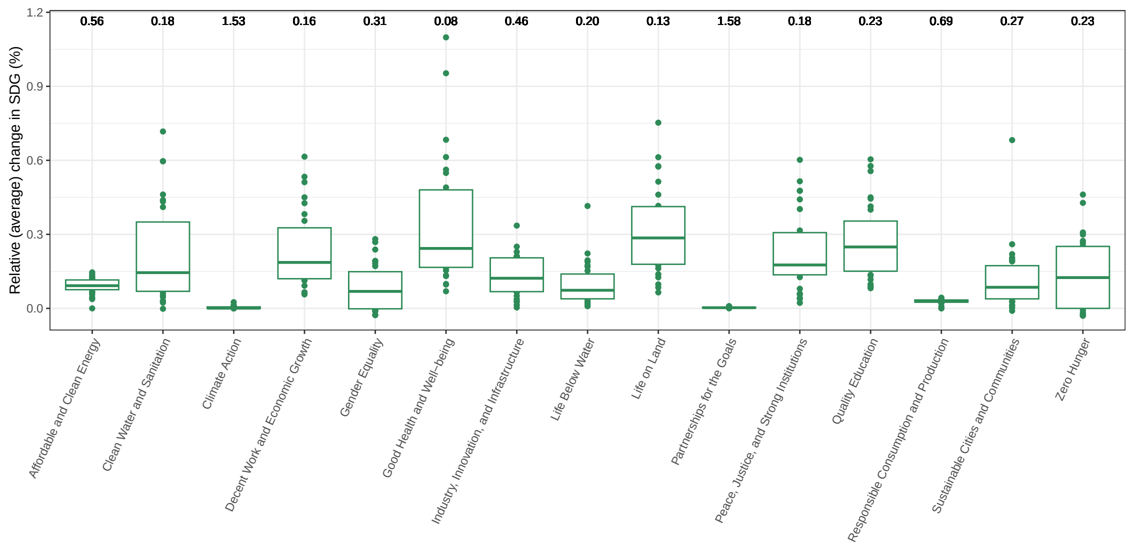

Source: Authors' calculations. Data from GGGI.

### 4.2.5 The Sustainable Development Goals Index - SDG

Finally, Figure 8 and Figure 9 respectively address the sensitivity values for positive and negative variables in the SDG Index. As explained, this indicator is formed by 114 different components grouped into 17 categories referring to Sustainable Development Goals. The comparison of the SDG and the other indexes would only be direct if we showed the sensitivity of the index with respect to the 114 components. However, to simplify the exposition, we have averaged and grouped individual component sensitivity values into these 17 dimensions.

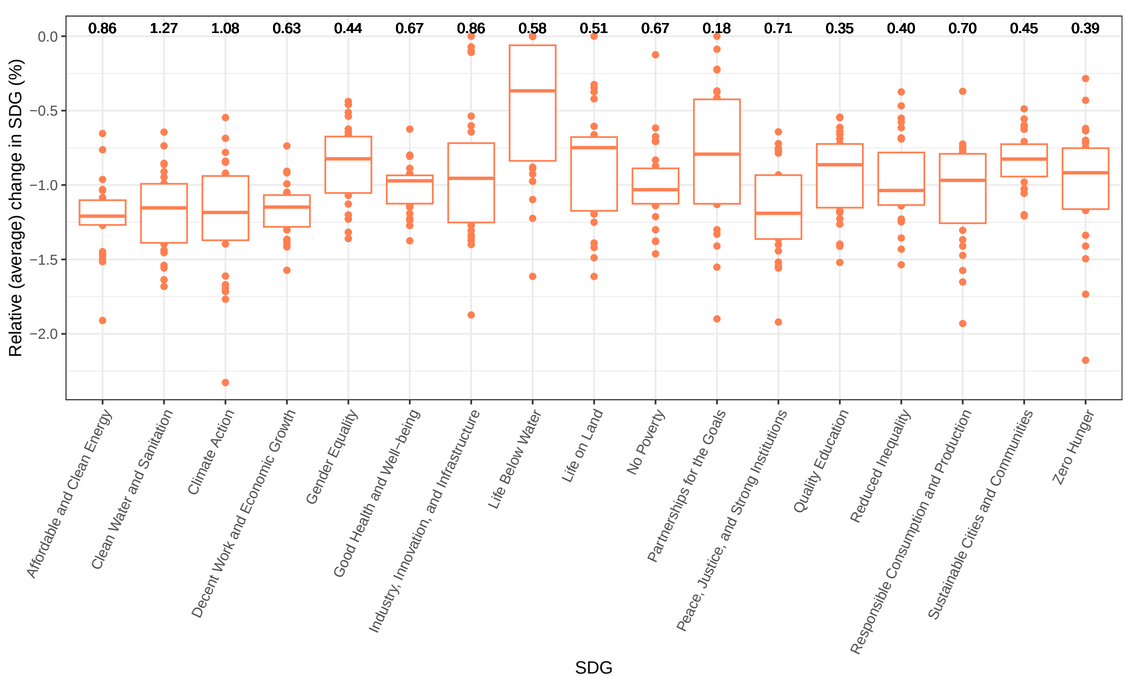
Sensitivity values barely surpass 1.2% for "positive" components, with most being located below 0.5% and not altering the aggregate picture deployed in Figure 1. Interestingly, negative components of the SDG do show higher sensitivity, with the average value being located at around -1%. Due to the fact that we are averaging across several "inner" components within each goal, we cannot comment further on the CV effect.

Figure 8: Sustainable Development Growth (Positive)




Source: Authors' calculations. Data from SDG.

Figure 9: Sustainable Development Growth (Negative)




Source: Authors' calculations. Data from SDG.

Table 4.1 overviews the main results. The first column presents the range of sensitivity values by indicator, presenting the spectrum between the smallest and highest values observed along the Y-axis in the preceding graphs. We separate results based on the sign or contribution of the individual components. The maximum value of the positive components is larger than the minimum value of the negative, so our selected indicators seem to respond more to changes in positive components. This is especially relevant for the BLI, whose maximum mean positive relative change reaches 2.8%, while the minimum negative sensitivity merely reaches 1%. The only exception is the SDG, whose results are obtained after aggregating many subcomponents. There is not association between the number of components in the indicator and the sensitivity estimated.

Table 4.1: Sensitivity overview by indicators

| **Indicator** | **Range Sensitivity (in %)** | **Mean Sensitivity (sd)** | **Pearson Correlation (p-values)** |
|---|---|---|---|
| PHDI | Positive: (0.1 ; 0.9)<br>Negative: (-0.64 ; -0.06) | Positive: 0.41 (0.29)<br>Negative: -0.21 (0.14) | -0.92 (0.01) |
| TPI | Positive: (0 ; 0.25)<br>Negative: (-0.15 ; 0) | Positive: 0.07 (0.05)<br>Negative: -0.05 (0.03) | -0.49 (0.01) |
| BLI | Positive: (0 ; 2.8)<br>Negative: (-0.99 ; 0) | Positive: 0.48 (0.43)<br>Negative: -0.10 (0.19) | -0.21 (0.23) |
| GGI | Positive: (0 ; 2.98)<br>Negative: (-1.90 ; 0) | Positive: 0.21 (0.29)<br>Negative: -0.11 (0.20) | -0.18 (0.01) |
| SDG | Positive: (-0.19 ; 2.18)<br>Negative: (-2.39 ; 0) | Positive: 0.19 (0.26)<br>Negative: -1.01 (0.50) | -0.50 (0.00) |

Note: Own elaboration. The Pearson Correlation shows the p-values of the correlation between the CV and the sensitivity for each indicator.

The second column captures the mean sensitivity value, showing the sd in parenthesis. The highest average (and standard deviation) sensitivity values are found for the PHDI, the BLI, and the GGI, while the SDG and the TPI seem less reactive. The last column presents Pearson's product-moment correlation coefficient between the CVs and sensitivity values, with the p-value of the statistical test being displayed in parenthesis. The correlation is negative for all indicators, hence confirming that a bigger coefficient of variation in the non-normalized variables is associated with a smaller sensitivity. The correlation is significant at 99% in the PHDI, TPI, and GGI. As previously discussed, although the BLI indicator reaches the highest sensitivity with high CV values, some low CVs are also associated with higher sensitivity values. Finally, the low sensitivity of the SDG, with most values being close to zero, makes the correlation negligible.

The five indices exhibit considerable insensitivity to a 2% variation in their components. For instance, the highest sensitivity observed registers a value of 2.8 and is associated with the "Life Expectancy" component of the BLI in Greece. Considering the original variable at 81.5 years, a 2% increase would raise its value to 83.1 years. Despite this considerable shift in life expectancy by 1.6 years — representing a significant and somewhat unrealistic sudden event — the BLI for Greece would remain largely unchanged, shifting from 0.29 to 0.3. This observation underscores the sensitivity of these indicators, undermining their capacity to address change in relevant dimensions of transition towards sustainability.

## 4.3 Robustness and statistical sensitivity

We perform two complementary analyses. First, we address the robustness of our choice of $p = 0.02$ and confirm that it does not drive our conclusions. Second, we check statistical robustness with confidence bounds around the indicator values.

As acknowledged at the end of Section 3, our main results might be affected by the arbitrary setting of $p = 2\%$. Choosing higher $p$ values would lead to higher sensitivity values at the potential cost of reaching unrealistic results, such as an increase of 8% in life expectancy in the PHDI. Having a small sample size of only 27 countries hampers the analysis of the statistical significance of the simulated sensitivity values. We do not have enough observations to check whether a 2% increase in a component leads to a variation in the index statistically different from zero.

We repeat the sensitivity analysis relaxing the assumption of a 2% variation. Following Acosta et al. (2022), for each component and country we draw 200 random percentage points ($p$ values) from a uniform distribution $U(\mu = 0, \sigma = 0.1)$, and estimate the change in the indicator as schemed in Section 3.3 for all options.[34] This Monte Carlo approach allows us to average sensitivity values across repetitions so we obtain the point estimate of the mean sensitivity and $95\%$ confidence intervals.

We find two main advantages in this approach. First, relaxing the baseline 2% sensitivity setting helps us reflect a more plausible scenario, where shocks of different magnitudes affect differently the components. Second, obtaining 200 bootstrapped shocks for each country and component allows us to get confidence intervals and dig into the statistical sensitivity of these measures. This comes at the cost of losing tractability on the specific $p$ leading to a change in the sensitivity.

[34] The average $p$ after 200 random extractions from the uniform distribution is 0.0475, while the standard deviation reaches 0.0289.

Results from this second simulation exercise are shown in Appendix A and confirm the main findings in Section 4.2. Since the average $p$ is 4.75, which more than doubles our baseline value of $p = 2$, the sensitivity values are higher than those exposed in the main results. For instance, the average relative change in the PHDI (Figure A.1) now reaches 2% when we increase Life Expectancy, and equals 3.1% in the BLI also for Life Expectancy (Figure A.4). In any case, the rank of sensitivity values remains with higher values being once more associated with lower CV values. The TPI is the exception, as components with the smallest CV do not always present the largest changes. The TPI is indeed the only indicator including weights, which ultimately show how giving different importance to each dimension can help overcome the limitations attached to the normalization process.

After 200 bootstrapped repetitions, most confidence intervals cross the zero intercept, suggesting that changes in components are not significant at a 95% significance. As explained, higher sensitivity values are associated with smaller CV values. In these cases, country ranks are more volatile and more sensitive to changes in $p$, so the confidence intervals widen and make them surpass the zero line. Opposed, components with smaller sensitivity values are more robust and exert smaller confidence intervals.

We have focused on the sensitivity of indicators to changes in the components due to shocks, such that $p$ was assumed to be higher than zero. Assuming $p$ to be smaller than zero would have delivered (close to) symmetric results, but mixing positive and negative values in $p$ may have made the sensitivity values compensate each other. Aimed at analyzing the statistical robustness of the indicators, we relax this assumption.

Measurement errors can indeed under or overestimate the values of the components, and shocks can be either positive or negative. Our final exercise repeats the sensitivity analysis, drawing for each component and country 200 random percentage points ($p$ values) from a normal distribution $\mathcal{N}(0, 1)$.[35] For each draw, we estimate the change in the indicator as schemed in Section 3.3.

The results are shown in Appendix A, from Figure A.9 to Figure A.16. Point estimates in the vast majority of plots lie over zero, ensuring that the number of iterations is enough for the exercise. The magnitude of the 95% confidence intervals aligns with the main previous findings. They are found quite robust, with most relative changes being small or negligible. As argued, indicators are more sensitive to very homogeneous variables, such as life expectancy in the PHDI or the BLI, or the gender balance in financial institutions in the GGI, thus supporting our main findings.

[35] We have no information about how estimation errors or shocks are distributed across components, so we arguably assume they follow a Gaussian distribution.

# 5 Conclusions

Composite indicators contribute positively to the policy debate by informing and assisting policymakers in monitoring and evaluating processes. Besides assessing multidimensional socio-economic phenomena, they enable cross-country comparison and are accessible to the general public (Bandura, 2008; Stiglitz et al., 2009). However, the construction of a composite index requires many steps, and researchers' decisions are not trivial (Becker et al., 2017; Freudenberg, 2003; Mazziotta and Pareto, 2013). If poorly constructed or misinterpreted they can convey misleading policy messages and hide relevant dimensions of wellbeing and sustainability (Fleurbaey, 2009). Understanding the sensitivity of aggregated indicators to changes in their components is key to interpreting time variations and exploring further their economic and social implications.

In this paper we provide statistical robustness analyses for five indicators selected through a rigorous process discussed in Gábos et al. (2023), namely, the Planetary Pressure Adjusted Human Development Index (PHDI), the Transition Performance Index (TPI), the Better Life Index (BLI), the Green Growth Index (GGI), and the Sustainable Development Goals (SDG).

These indicators are found, overall, rather insensitive to changes in their components. The sensitivity is higher when the component values are very clustered around the mean, thus exerting a low coefficient of variation. In these cases, small changes in the components values are associated with big changes in the country's relative position within the countries distribution, so the effect on the indicator is higher. We argue that this association is problematic, because the effect of a change in the component over the indicator is not driven by its theoretical relevance, but by the distribution of the components' values. This provokes that, in some cases, components presenting higher sensitivity values are not necessarily considered "green" variables. As an example, the GGI is quite affected by changes in the gender gap, but rather insensitive to rises in pollution or material footprint.

These results undermine the reliability of composite indicators as sole measures of economic and social transition towards sustainability. Countries improving in one or more dimensions or components may not find these changes reflected in the aggregate indicator if other economies also improve their situation by a similar rate. Besides, inappropriate thresholds, -whether sample-based, or theoretically set too broadly or narrowly- can either smooth out or amplify the impact of changes in the indicator through the normalization process.

Practitioners should always acknowledge the indicators' limitations before compelling policymakers to interpret them. It is also mandatory to propose alternatives to enhance their ability to assess wellbeing. One possible way to move forward could imply developing weights across multiple components (see Decancq and Lugo (2013) for a discussion). Despite weighting schemes have limitations, such as the difficulty of weighting different dimensions, recent advances in the literature propose techniques to improve their evaluation (Becker et al., 2017). Well-developed weighting schemes should highlight components with more theoretical relevance and compensate for the country-rank effect caused by the normalization, making indicators more sensitive to changes in transition-related components. These weights could also counterpart the perfect substitutability inherent to arithmetic aggregations. Avoiding in-sample thresholds, which exacerbate the effect of the analysis sample as well as the country's relative position in the range of variation, and establishing theoretical limits, subject to continuous updates, would also contribute to reflecting changes in specific components.

All in all, composite indicators are aimed at aggregating across different dimensions and they will never - and are not intended to - capture details on the specific role of components. As in European Commission (2022) or UNDP (2022), their exposition should always be complemented with one-dimensional measures, such as greenhouse gas emissions, income or wealth inequality, life expectancy, gender gaps, poverty rates, water supply, or biodiversity proxies, among others. An example of good practices can be found in the European Commission targets for a more social Europe by 2030, including a 78% employment rate or a 60% of adults participating in training every year.[36] Well-defined target thresholds and values could serve to acknowledge the performance of countries beyond ranks, and get a deeper understanding of their performance. The big picture allows for a broad understanding of reality, but the devil - and the angels - are always in the details.

[36] Country-specific targets can be found here: https://ec.europa.eu/social/BlobServlet?docId=25728&langId=en.

## References


Acosta, L., Maharjan, P., Peyriere, H., Galotto, L., Mamiit, R., Ho, C., Flores, B., and Anastasia, O. (2019). Green growth index: Concepts, methods and applications. https://greengrowthindex.gggi.org/wp-content/uploads/2019/12/Green-Growth-Index-Technical-Report_20191213.pdf.

Acosta, L., Nzimenyera, I., Jr., R. S., Munezero, R., Nantulya, A., Shula, K., Quiñones, S., Luchtenbelt, H., Czvetkó, T., Lee, S., and Adams, G. (2022). Green growth index 2022 – measuring performance in achieving sdg targets. https://greengrowthindex.gggi.org/wp-content/uploads/2023/02/2022-Green-Growth-Index-1.pdf.

Bandura, R. (2008). A survey of composite indices measuring country performance: 2008 update. *New York: Office of Development Studies, United Nations Development Programme (UNDP)*.

Becker, W., Saisana, M., Paruolo, P., and Vandecasteele, I. (2017). Weights and importance in composite indicators: Closing the gap. *Ecological indicators*, 80:12–22.

Biggeri, M., Ferranini, A., L.Lodi, and J. Cammeo, A. F. (2024). The "winds of change": the spes framework on sustainable human development. *SPES Project Deliverable*, 2(1).

Cherchye, L., Moesen, W., Rogge, N., and Puyenbroeck, T. V. (2007). An introduction to 'benefit of the doubt'composite indicators. *Social indicators research*, 82:111–145.

Chowdhury, S. and Squire, L. (2006). Setting weights for aggregate indices: An application to the commitment to development index and human development index. *The Journal of Development Studies*, 42(5):761–771.

Costanza, R., Erickson, J., Fligger, K., Adams, A., Adams, C., Altschuler, B., Balter, S., Fisher, B., Hike, J., Kelly, J., et al. (2004). Estimates of the genuine progress indicator (gpi) for vermont, chittenden county and burlington, from 1950 to 2000. *Ecological Economics*, 51(1-2):139–155.

Cowell, F. A. (2011). *Measuring inequality*. Oxford University Press.

Decancq, K. and Lugo, M. A. (2013). Weights in multidimensional indices of wellbeing: An overview. *Econometric Reviews*, 32(1):7–34.

Ebert, U. and Welsch, H. (2004). Meaningful environmental indices: a social choice approach. *Journal of Environmental Economics and Management*, 47:270–286.

European Commission, D. G. R. I. (2022). Transitions performance index 2021 – towards fair and prosperous sustainability.

European Commission, J. R. C. (2008). *Handbook on constructing composite indicators: methodology and user guide*. OECD publishing.

Fleurbaey, M. (2009). Beyond gdp: The quest for a measure of social welfare. *Journal of Economic literature*, 47(4):1029–1075.

Foster, J., Greer, J., and Thorbecke, E. (2010). The foster–greer–thorbecke (fgt) poverty measures: 25 years later. *The Journal of Economic Inequality*, 8:491–524.

Freudenberg, M. (2003). Composite indicators of country performance: A critical assessment. OECD Science, Technology and Industry Working Papers 2003/16, OECD Publishing.

Galotto, L., Gerrard, S., and Acosta, L. (2020). Gggi's concept for the green growth index: Comparative assessment of relevant global green growth indices. GGPM Working Paper 1, Global Green Growth Institute, Seoul, Republic of Korea.

Greco, S., Ishizaka, A., Tasiou, M., and Torrisi, G. (2019). On the methodological framework of composite indices: A review of the issues of weighting, aggregation, and robustness. *Social indicators research*, 141:61–94.

Gábos, A., Lelkes, O., and Cammeo, J. (2023). Report on mapping indicators and composite indices relevant to measure transition performances. Spes report no. 3.1, University of Florence.

Hoekstra, R. (2019). *Replacing GDP by 2030: Towards a common language for the well-being and sustainability community*. Cambridge University Press.

Kahneman, D. and Deaton, A. (2010). High income improves evaluation of life but not emotional well-being. *Proceedings of the national academy of sciences*, 107(38):16489–16493.

Kubiszewski, I., Costanza, R., Franco, C., Lawn, P., Talberth, J., Jackson, T., and Aylmer, C. (2013). Beyond gdp: Measuring and achieving global genuine progress. *Ecological economics*, 93:57–68.

Lahdelma, R., Hokkanen, J., and Salminen, P. (1998). Smaa-stochastic multiobjective acceptability analysis. *European Journal of Operational Research*, 106(1):137–143.

Mauro, V., Biggeri, M., and Maggino, F. (2018a). Measuring and monitoring poverty and well-being: A new approach for the synthesis of multidimensionality. *Social Indicators Research*, 135:75–89.

Mauro, V., Biggeri, M., and Maggino, F. (2018b). Measuring and monitoring poverty and well-being: A new approach for the synthesis of multidimensionality. *Social Indicators Research*, 135:75–89.

Mazziotta, M. and Pareto, A. (2013). Methods for constructing composite indices: One for all or all for one. *Rivista Italiana di Economia Demografia e Statistica*, 67(2):67–80.

Nardo, M., Saisana, M., Saltelli, A., Tarantola, S., Hoffman, A., and Giovannini, E. (2005). Handbook on constructing composite indicators. https://www.oecd-ilibrary.org/content/paper/533411815016.

OECD (2022). Better life index: Definitions and metadata. https://www.oecd.org/wise/OECD-Better-Life-Index-definitions-2021.pdf.

Papadimitriou, E., Neves, A. R., and Becker, W. (2019). Jrc statistical audit of the sustainable development goals index and dashboards. Technical report, EUR 29776 EN. JRC116857.

Ravallion, M. (2012). Troubling tradeoffs in the human development index. *Journal of development economics*, 99(2):201–209.

Sachs, J., Lafortune, G., Fuller, G., and Drumm, E. (2023). Sustainable development report 2023. https://sdgtransformationcenter.org/reports/sustainable-development-report-2019.

Sachs, J., Schmidt-Traub, G., Kroll, C., Lafortune, G., and Fuller, G. (2019). Sustainable development report 2019. https://sdgtransformationcenter.org/reports/sustainable-development-report-2019.

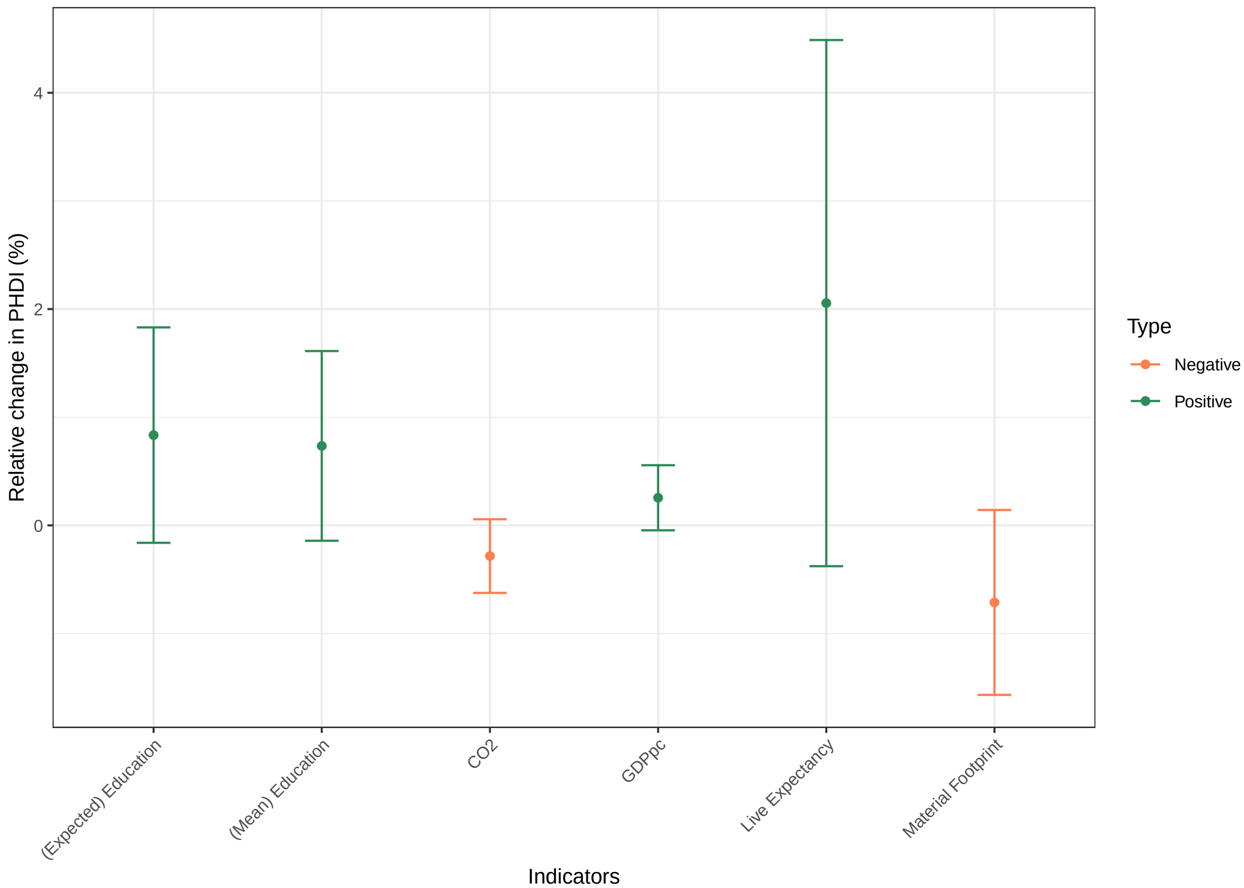
Saisana, M., Saltelli, A., and Tarantola, S. (2005). Uncertainty and sensitivity analysis techniques as tools for the quality assessment of composite indicators. *Journal of the Royal Statistical Society Series A*, 168(2):307–323.

Stiglitz, J. E., Sen, A., Fitoussi, J.-P., et al. (2009). Report by the commission on the measurement of economic performance and social progress.

UNDP (2022). Human development report 2021-22. http://report.hdr.undp.org.

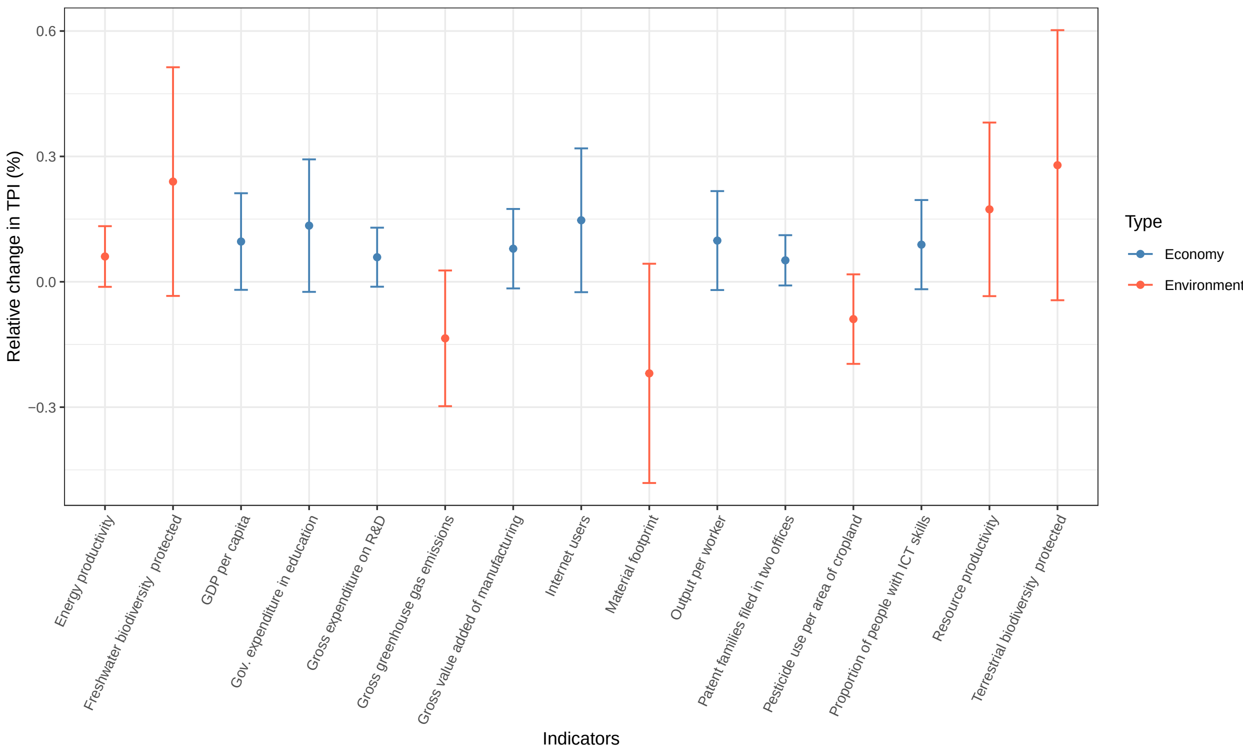
# A Figure Appendix

Figure A.1: Planetary pressures-adjusted Human Development Index (Random Component Shock)



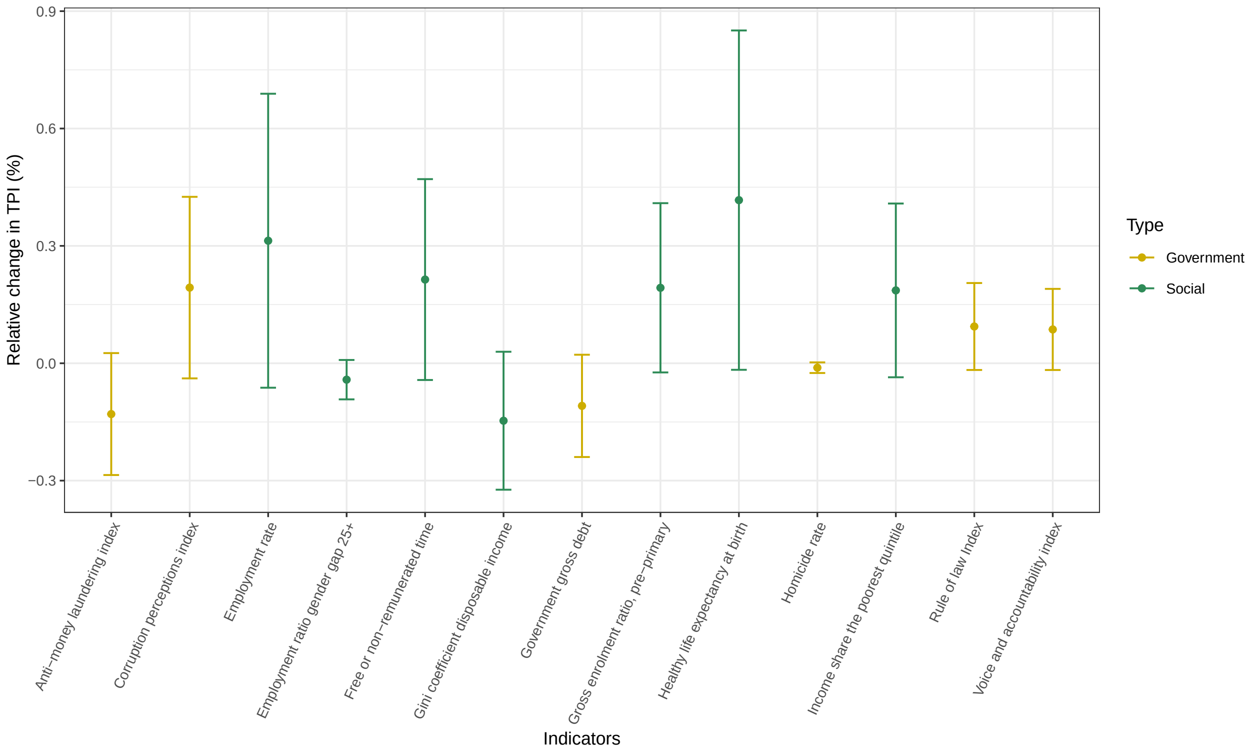

Source: Authors' calculations. Data from UNHDR. Confidence intervals at $95\%$ are estimated with 200 bootstrapped repetitions.

Figure A.2: Transition Performance Index (Economic and Environment, Random Component Shock)



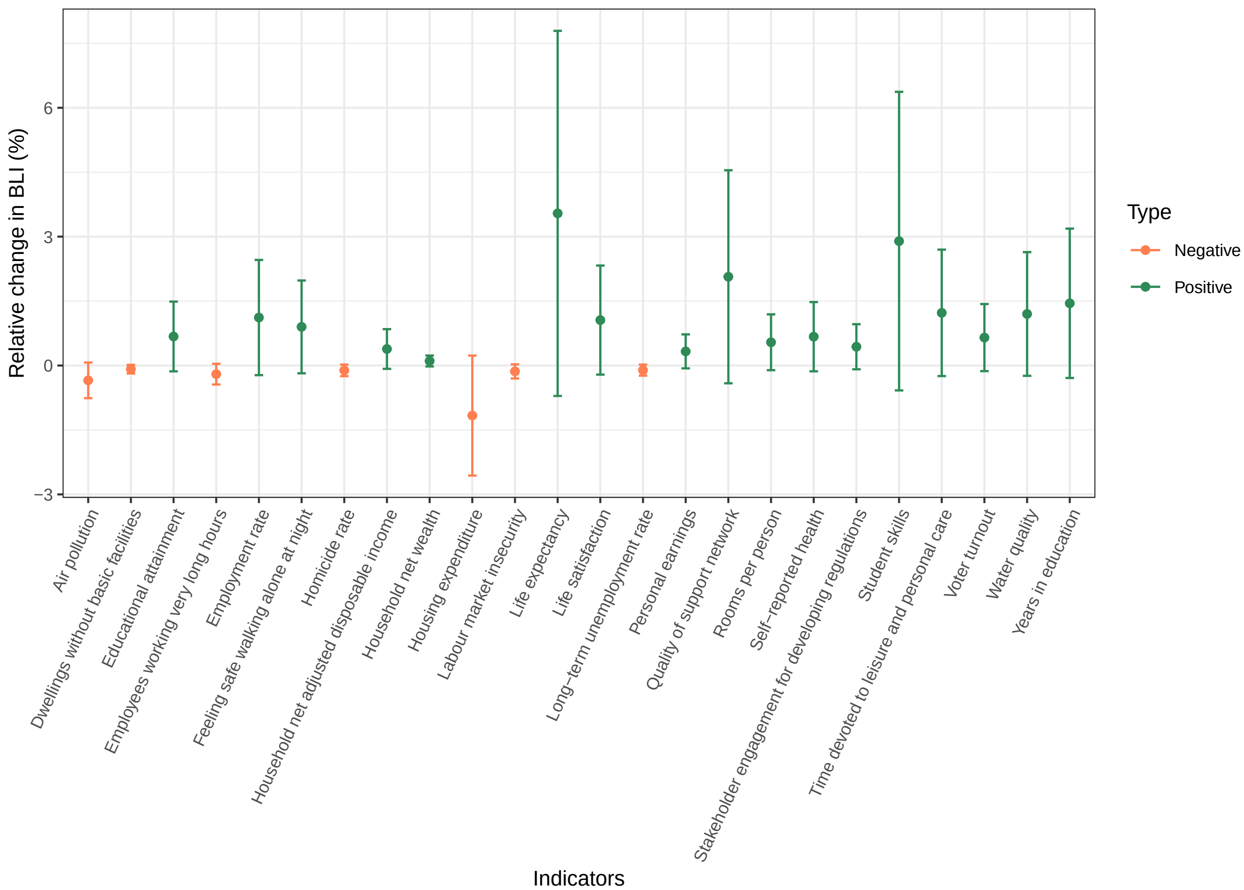

Source: Authors' calculations. Data from EC. Confidence intervals at $95\%$ are estimated with 200 bootstrapped repetitions.

Figure A.3: Transition Performance Index (Government and Social, Random Component Shock)



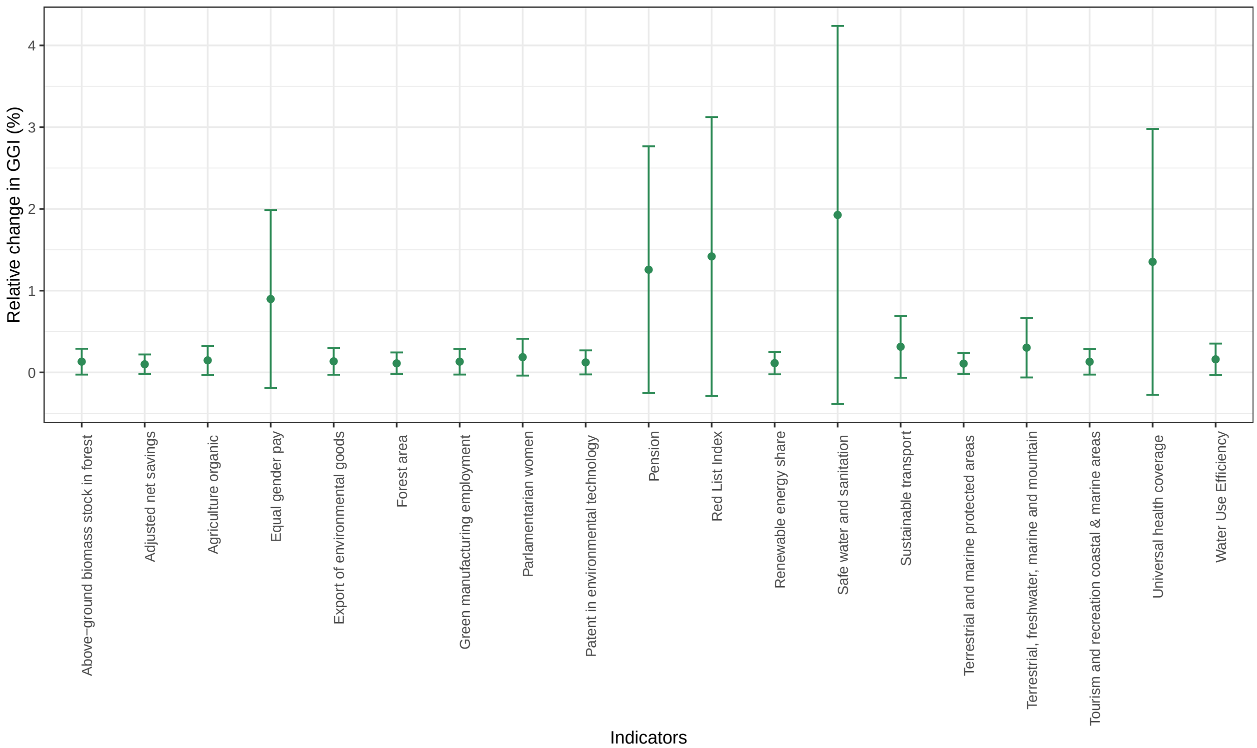

Source: Authors' calculations. Data from EC. Confidence intervals at $95\%$ are estimated with 200 bootstrapped repetitions.

Figure A.4: Better Life Index (Random Component Shock)



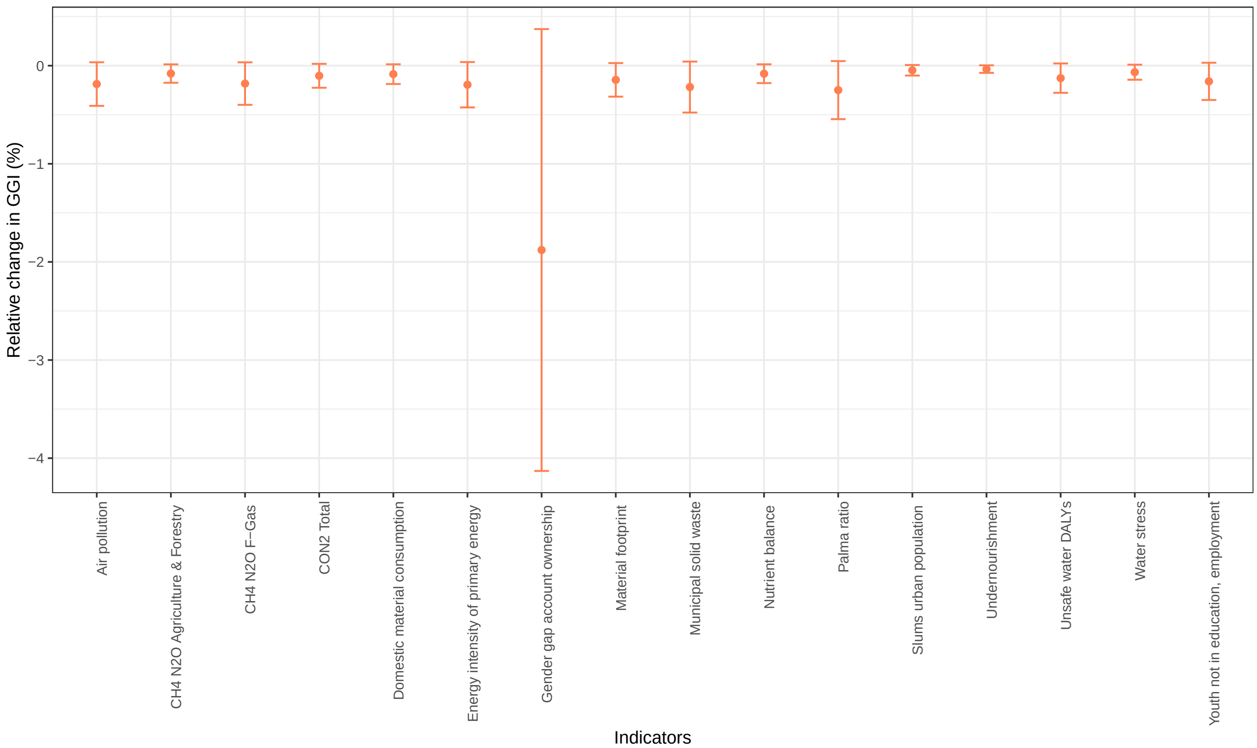

Source: Authors' calculations. Data from OECD. Confidence intervals at $95\%$ are estimated with 200 bootstrapped repetitions.

Figure A.5: Green Growth Index (Positive, Random Component Shock)



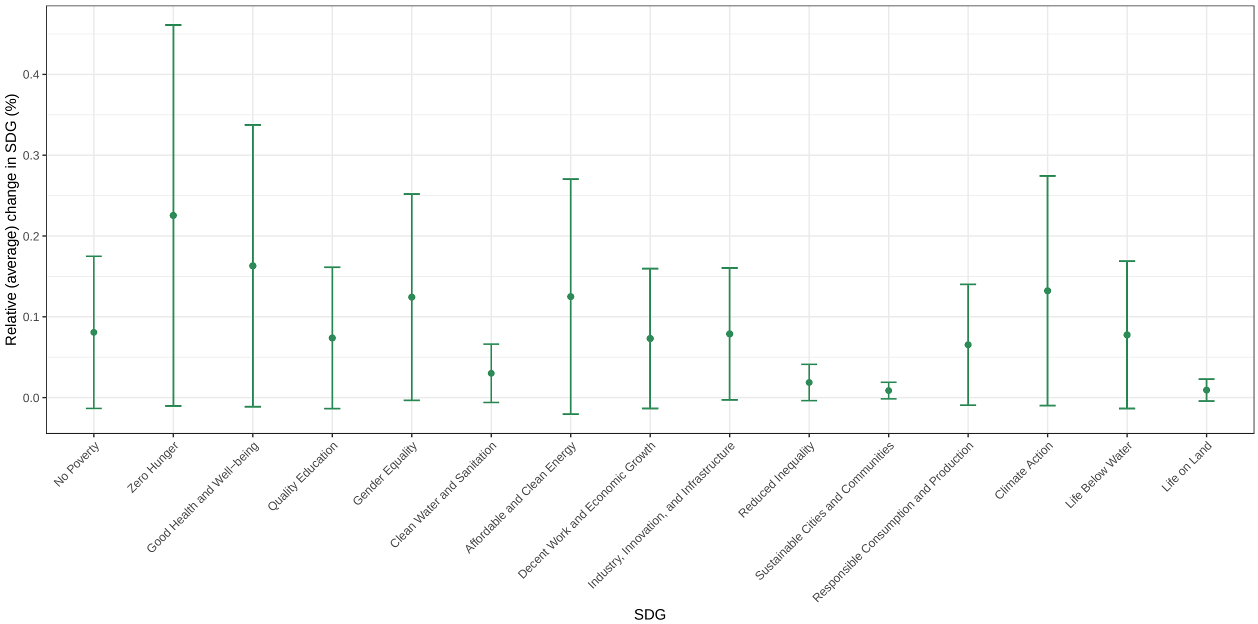

Source: Authors' calculations. Data from GGI. Confidence intervals at $95\%$ are estimated with 200 bootstrapped repetitions.

Figure A.6: Green Growth Index (Negative, Random Component Shock)



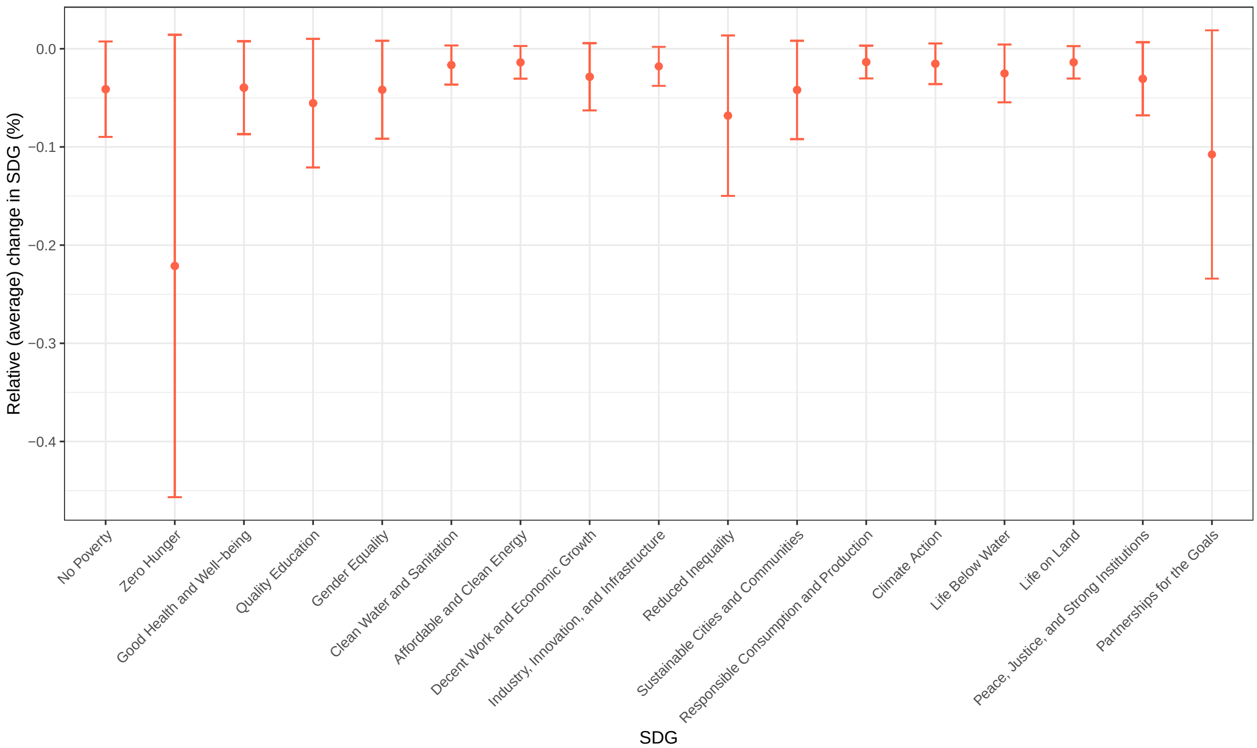

Source: Authors' calculations. Data from GGI. Confidence intervals at $95\%$ are estimated with 200 bootstrapped repetitions.

Figure A.7: Sustainable Development Growth (Positive, Random Component Shock)



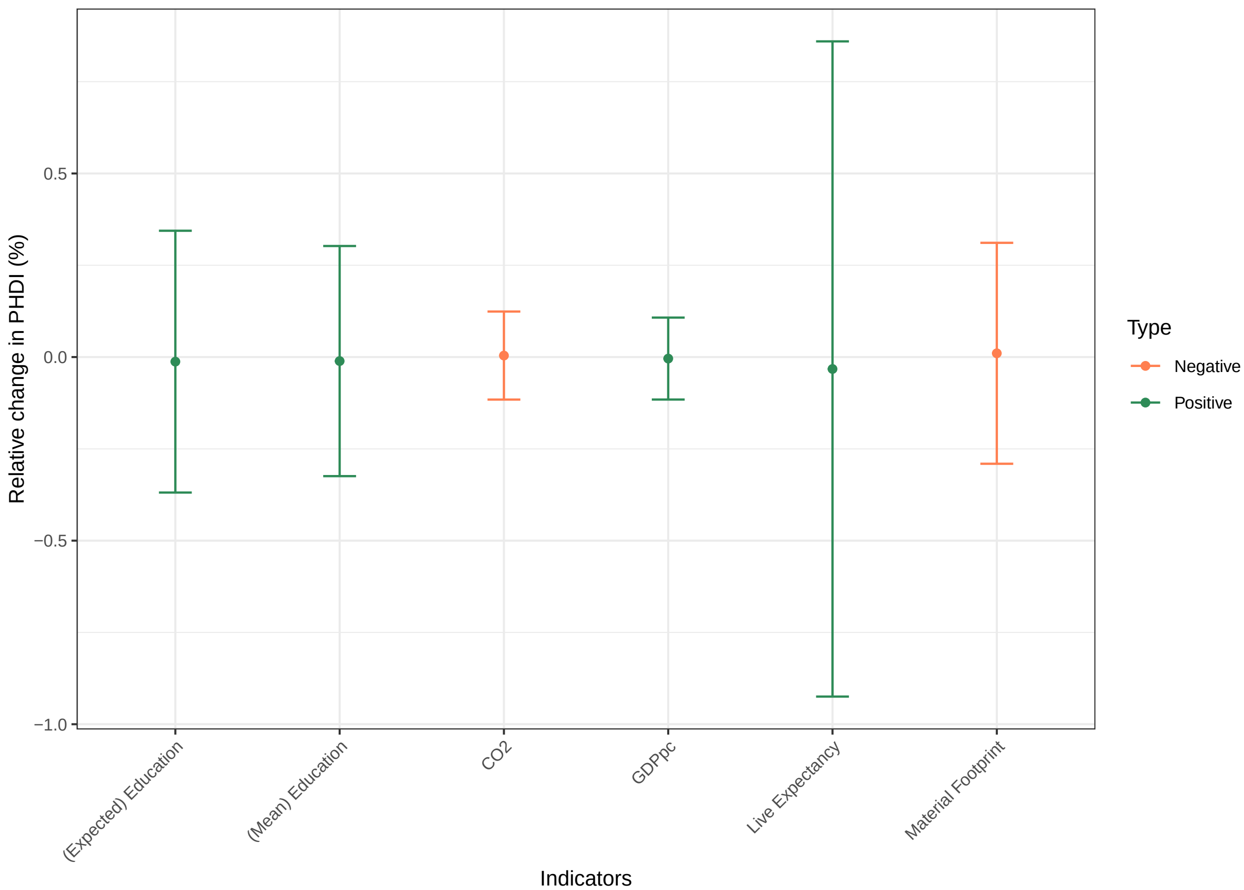

Source: Authors' calculations. Data from SDG. Confidence intervals at $95\%$ are estimated with 200 bootstrapped repetitions.

Figure A.8: Sustainable Development Growth (Negative, Random Component Shock)



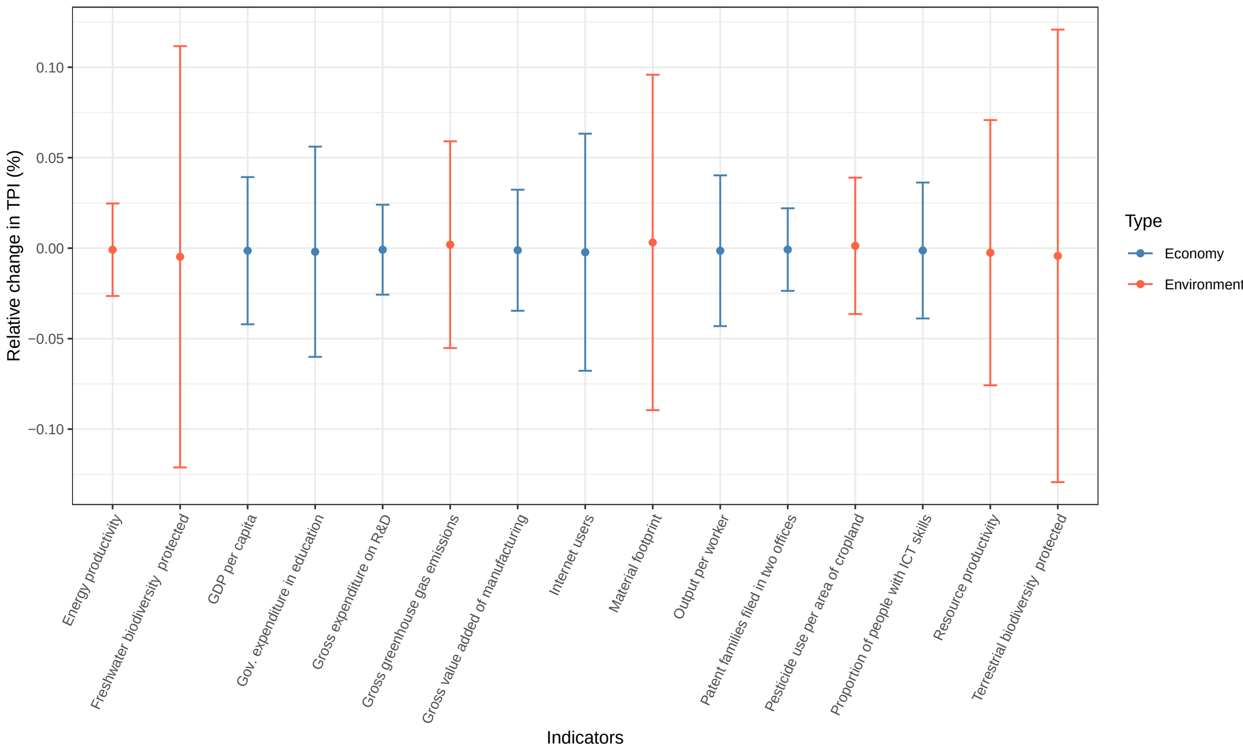

Source: Authors' calculations. Data from SDG. Confidence intervals at $95\%$ are estimated with 200 bootstrapped repetitions.

Figure A.9: Planetary pressures-adjusted Human Development Index (Standard Error)



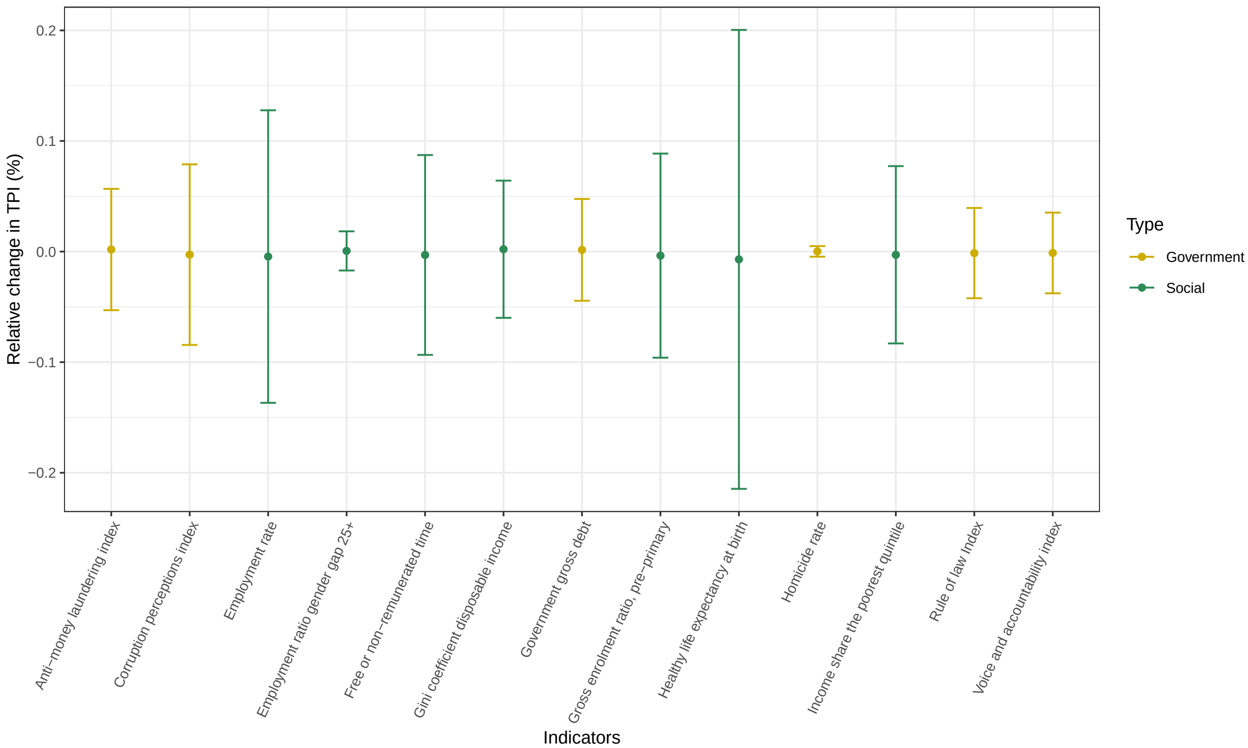

Source: Authors' calculations. Data from UNHDR. Confidence intervals at $95\%$ are estimated with 200 bootstrapped repetitions.

Figure A.10: Transition Performance Index (Economic and Environment, Standard Error)



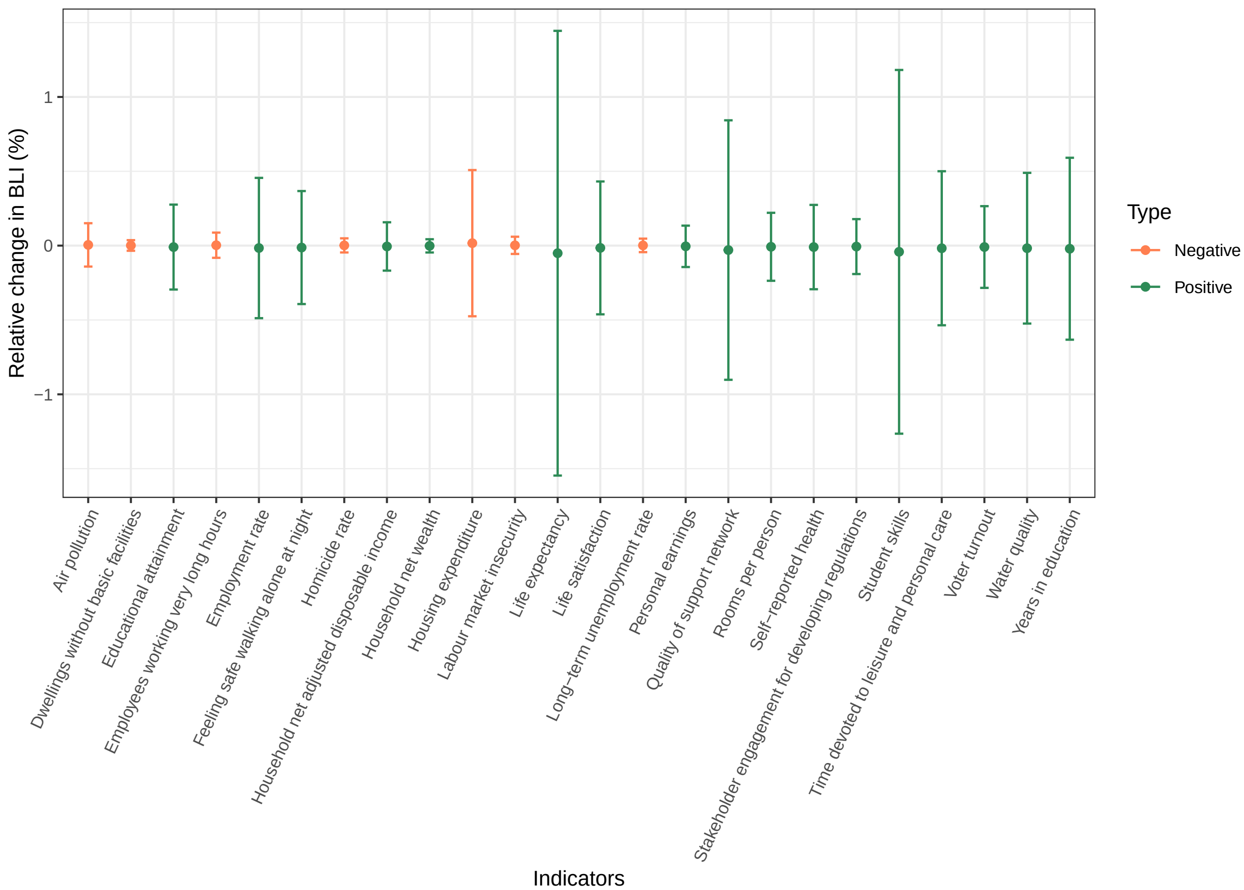

Source: Authors' calculations. Data from EC. Confidence intervals at $95\%$ are estimated with 200 bootstrapped repetitions.

Figure A.11: Transition Performance Index (Government and Social, Standard Error)



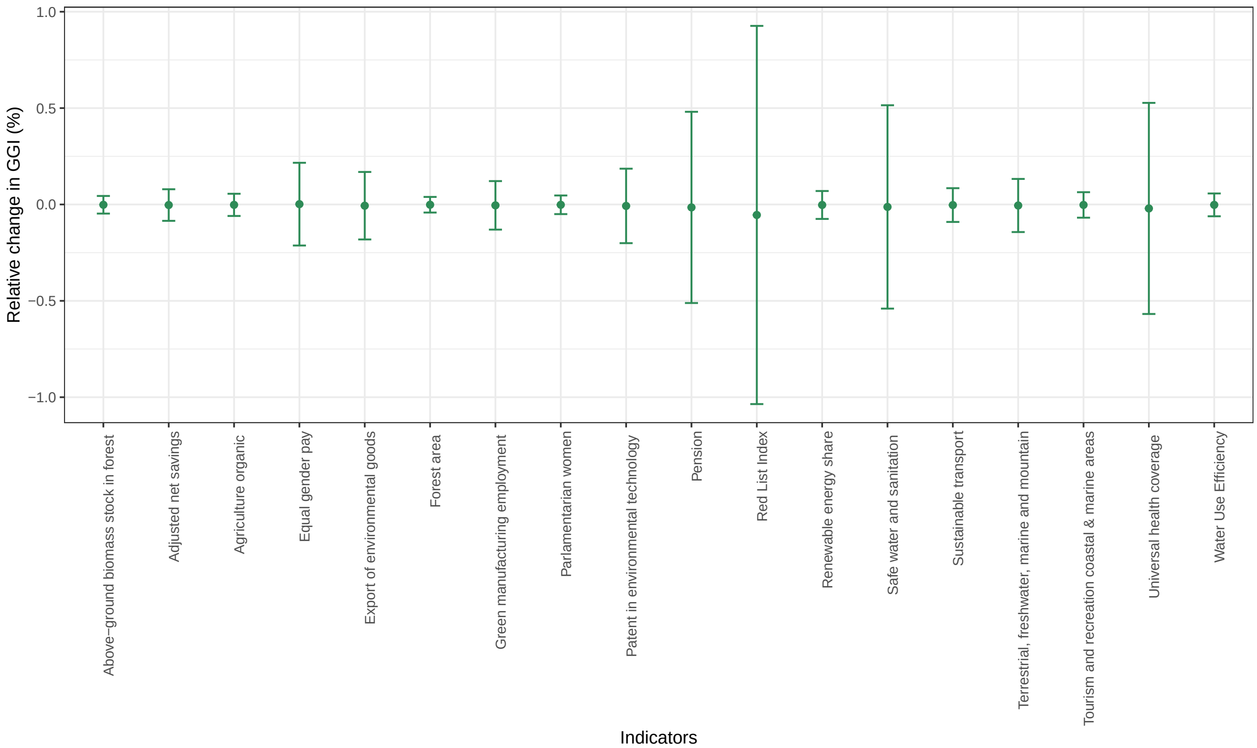

Source: Authors' calculations. Data from EC. Confidence intervals at $95\%$ are estimated with 200 bootstrapped repetitions.

Figure A.12: Better Life Index (Standard Error)



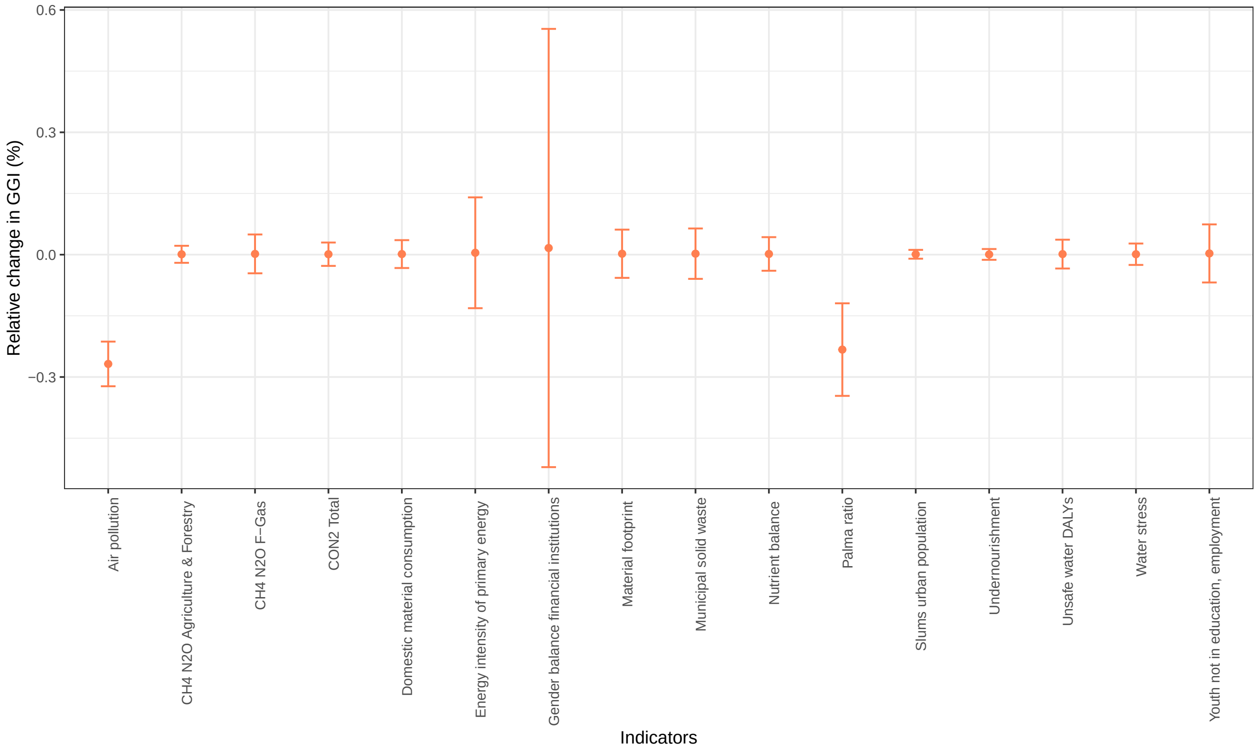

Source: Authors' calculations. Data from OECD. Confidence intervals at $95\%$ are estimated with 200 bootstrapped repetitions.

Figure A.13: Green Growth Index (Positive, Standard Error)



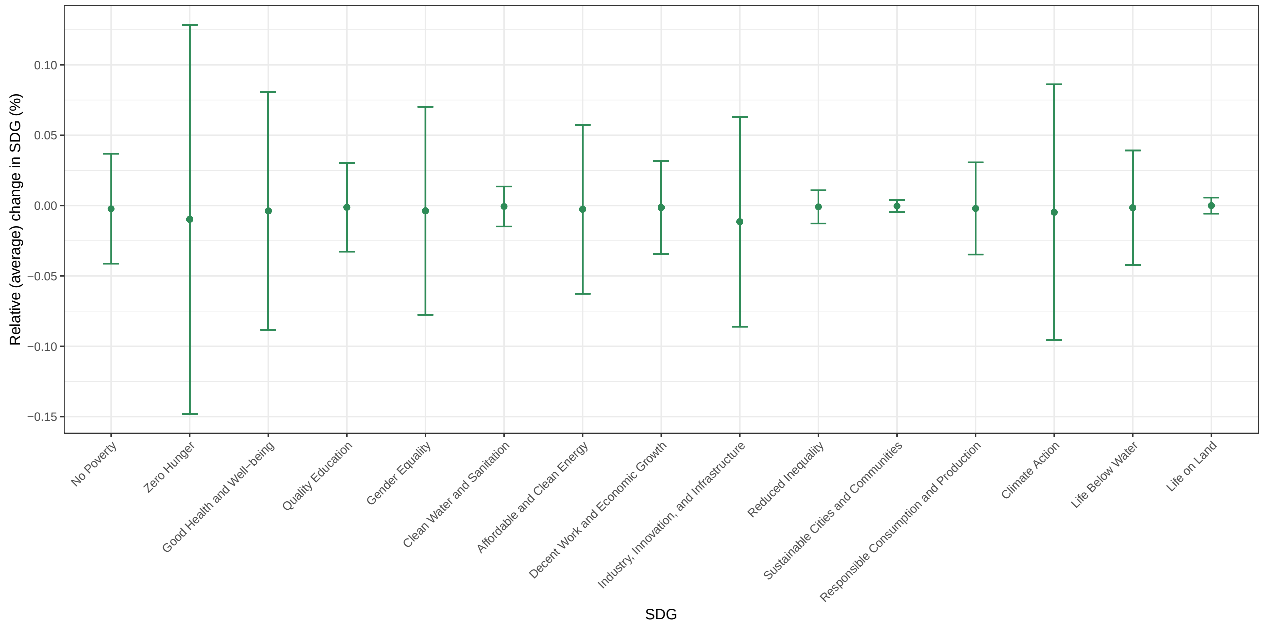

Source: Authors' calculations. Data from GGGI. Confidence intervals at $95\%$ are estimated with 200 bootstrapped repetitions.

Figure A.14: Green Growth Index (Negative, Standard Error)



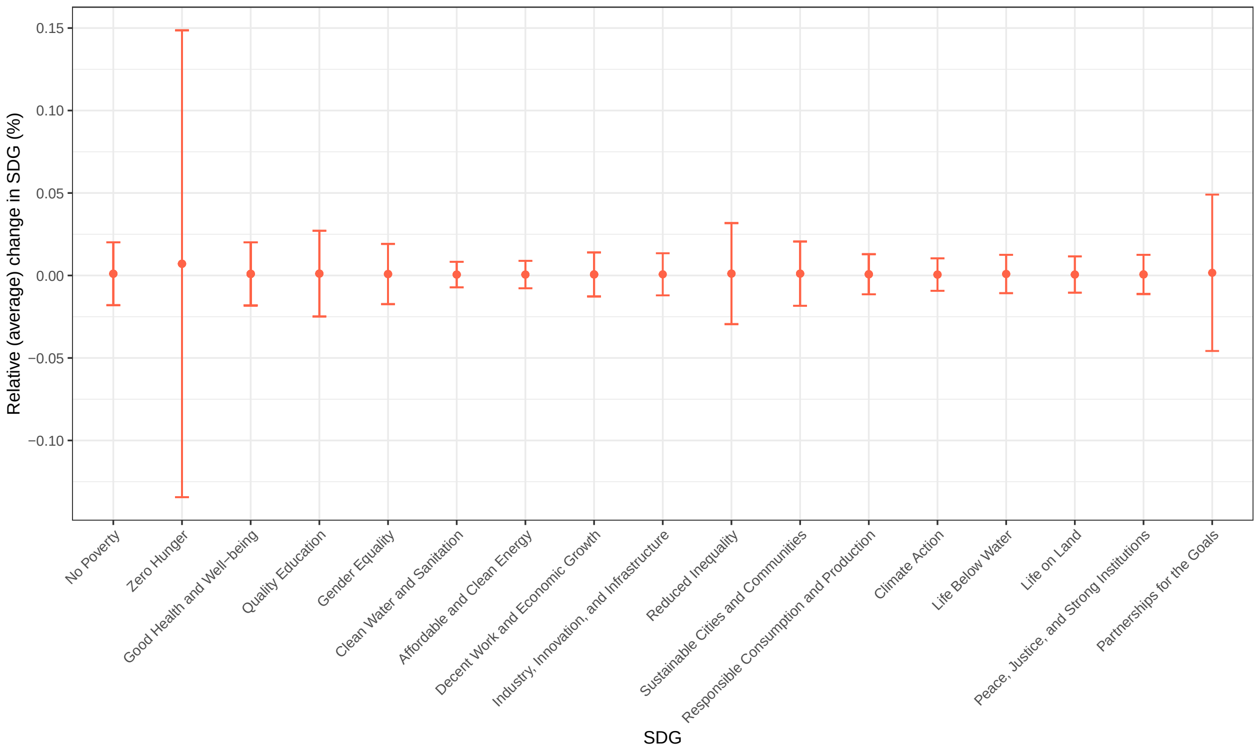

Source: Authors' calculations. Data from GGGI. Confidence intervals at $95\%$ are estimated with 200 bootstrapped repetitions.

Figure A.15: Sustainable Development Growth (Positive, Standard Error)



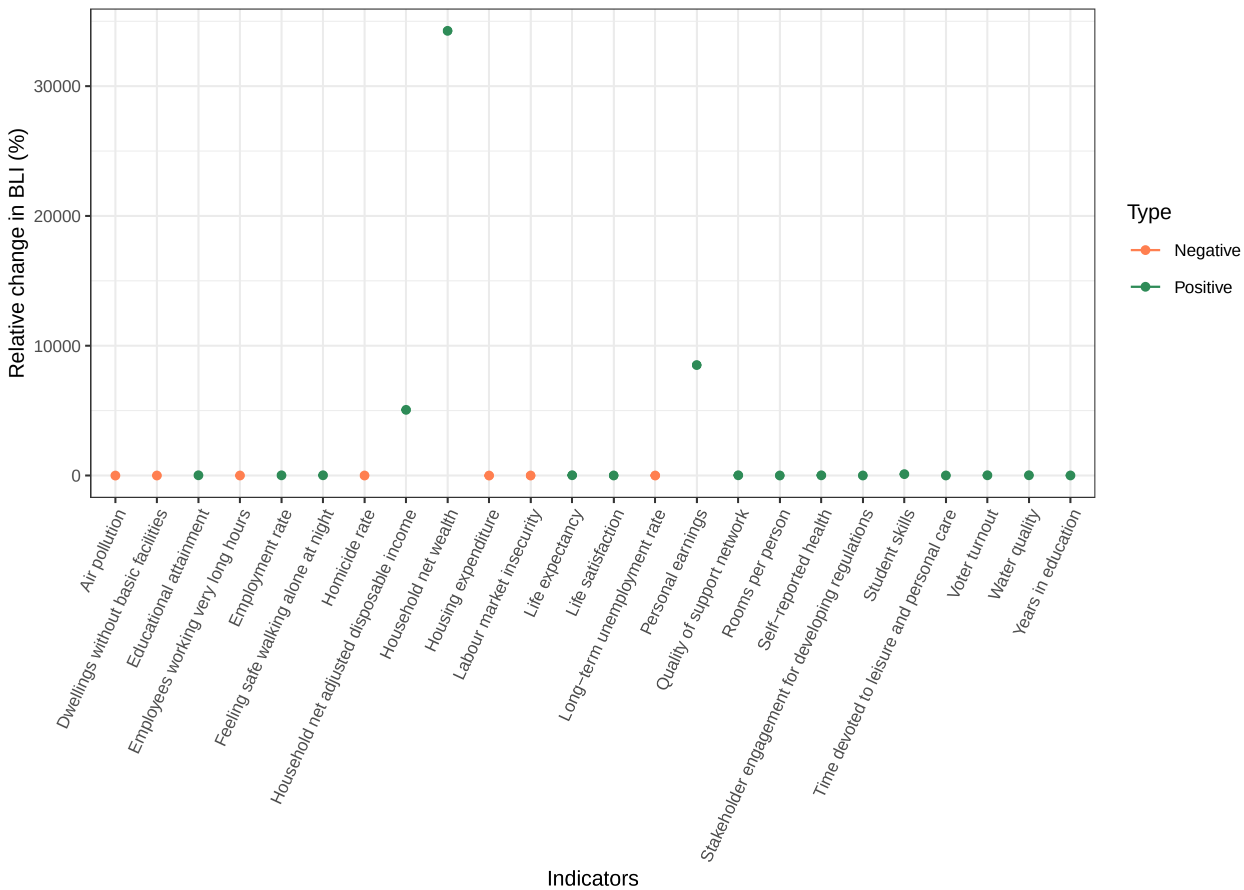

Source: Authors' calculations. Data from SDG. Confidence intervals at $95\%$ are estimated with 200 bootstrapped repetitions.

Figure A.16: Sustainable Development Growth (Negative, Standard Error)




Source: Authors' calculations. Data from SDG. Confidence intervals at $95\%$ are estimated with 200 bootstrapped repetitions.

Figure A.17: Better Life Index (Sensitivity using the Standard of the specific variables)




Source: Authors' calculations. Data from OECD. Confidence intervals at $95\%$ are estimated with 200 bootstrapped repetitions.