



Image: DeCe - https://stock.adobe.com/be_en/images/group-of-different-pine-cones-isolated-on-white/238264138?prev_url=-detail and cloud7days - https://stock.adobe.com/be_en/images/pine-cones-isolated-on-white-background/212650363?prev_url=detail

Icon in page corners: Conifer cone icon by arloenl evinniev, from the Noun Project (<https://thenounproject.com/search/?q=pine%20cones&i=2213049>)

The Use of Quantitative Methods in the Policy Cycle

Main authors: Giuseppe Munda, Daniel Albrecht, William Becker, Enkelejda Havari, Giulia Listorti, Nicole Ostlaender, Paolo Paruolo, Michaela Saisana

Contributors: Marcos-Dominguez-Torreiro, Leandro Elia, d'Artis Kancs, Sven Langedijk, Rossana Rosati, Eckehard Rosenbaum, Paul Smits, Daniel Vertesy, Ian Vollbracht

Key takeaways:

- The chapter presents a set of quantitative modelling approaches, connected to various steps of the policy cycle, that aim at being learning tools facilitating the dialogue among policymakers, scientists and different social actors.
- Quantitative policy modelling cannot provide exact answers, but it can help policymakers and all social actors involved by providing a scientific sound framework for a systematic, coherent and transparent analysis.
- Practical guidelines for structuring policy problems by using uncertainty and sensitivity analysis (SA), multicriteria decision analysis (MCDA), composite indicators and ex-post impact evaluation are provided.
- Uncertainty analysis (UA) and SA are quantitative 'Model Quality Assurance' techniques that provide information to understand uncertainty and ultimately enhance the quality of policy modelling.
- MCDA is the most widespread multidimensional approach to ex-ante impact assessment. The basic idea is that in comparing policy options, a plurality of technical dimensions and social perspectives is needed.
- Composite indicators are powerful practical tools that can help policymakers in summarizing complex and interdependent phenomena that are not directly observable and help the monitoring of their progress over time.
- Ex-post impact evaluation is used to measure the impact of policy changes on the ground. It is based on the 'what works?' approach to the policy cycle, in which the design of policy measures is guided by careful analysis that generates clear evidence of what has worked in the past.

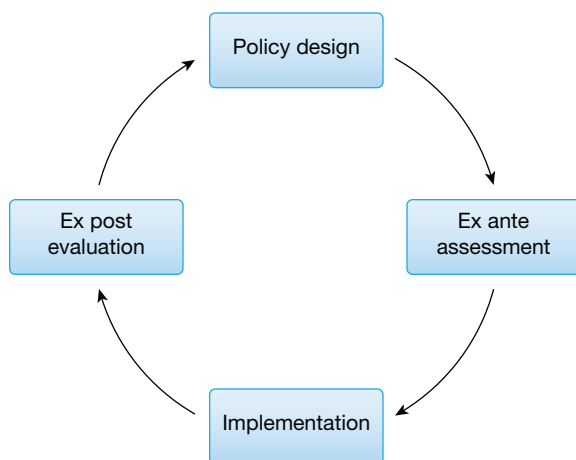


FIGURE 1 A simplified scheme of the policy cycle.

The real world is characterized by deep complexity. This apparently unremarkable observation has important implications for the manner in which policy problems are framed. A system is complex when the relevant aspects of a particular problem cannot be captured using one single perspective, thus any representation of a complex system reflects only a subset of its possible representations. This implies that for any policy modelling exercise,¹ there is a need to decide what is important for different social actors as well as what is relevant for the representation of the real-world entity described in the model (see, e.g., [Frame and O'Connor, 2011](#); [Jerrim and de Vries, 2017](#)).

When a public administration wishes to implement policies, there is a previous need of comparing different options to assess their social attractiveness. This is the traditional scope of welfare economics and decision theory which is grounded in the use of theoretical models helping the definition of the intended effect of policy interventions in question (see, e.g., [Dasgupta, 2001](#)). However, theoretical models have their limits and can never *fully* capture reality in all of its complexity. Ultimately, policymaking is bounded by a reality check expressed by the simple question: *what worked and for whom?* To answer this question, the design of policy measures should be guided by careful analysis that generates clear evidence of what has worked in the past. Quantitative impact evaluation aims at measuring the impact of policy changes on the ground.

Of course, policy assessment and evaluation is not a one-shot activity. On the contrary, it takes place as a learning process that is usually highly dynamic, so that judgments regarding the political relevance of items, alternatives or impacts may present sudden changes, hence requiring a policy analysis to be flexible and adaptive in nature. This is the reason why assessment/evaluation processes have a cyclic nature. By this is meant the possible adaptation of elements of the process due to continuous feedback loops among the various steps and consultations among the actors involved. The policy cycle is usually described as the circular phases of policy design, ex-ante impact assessment, implementation and ex-post impact evaluation (see [Fig. 1](#)).

Obviously, the learning process that takes place while analyzing the issue and defining policies will itself influence perceptions and alter significantly the decisional space in which alternative strategies are chosen. At the other end, institutional and cultural representations of the same system, while also legitimate, are on their own insufficient to define what should be done in any particular case.

Funtowicz and Ravetz have shown that these concerns were not considered very relevant by scientific research as long as time was considered an infinite resource. On the other hand, the new nature of the problems faced in this third millennium implies that, when dealing with problems that may have long-term consequences, policy analysis and modelling is confronting issues 'where facts are uncertain, values in dispute, stakes high and decisions urgent'

¹Throughout this chapter the term 'model' is referring to any quantitative approach to policy analysis.



(Funtowicz and Ravetz, 1990, 1991). Clearly ethical connotations are there; this implies that people, social scientists and governments differ significantly on what they consider to be an important policy issue, dimension, etc. As a consequence, values by which a society wishes to live cannot be relinquished to the preferences of an elite but need to be subject of extended dialogue with all constituencies (Benessia and Guimarães Pereira, 2015). Therefore, there is a need to invite the relevant constituencies into conversations about meanings and representations of policies, and the ethics and social practices that need to be fostered.

This chapter presents a set of quantitative modelling approaches that are designed and used within European Union (EU) policymaking processes, and beyond, with the objective of helping real-world policy processes without the ambition of being exhaustive; on the contrary, they aim at being learning tools to be used to facilitate the dialogue among policymakers, scientists and different social actors. Although they cover a wide spectrum of scientific approaches and can be used in different steps of the policy cycle, all of them have the common characteristic of being tools for boosting effectiveness, consistency and transparency of the policy process. Section 2 focuses on the issue of model transparency, Section 3 introduces quantitative approaches related to uncertainty and SA, Section 4 illustrates MCDA, Section 5 presents composite indicators, Section 6 is devoted to ex-post impact evaluation and finally in Section 7 some conclusions are drawn.

Model Transparency

Transparency is defined here as the extent to which external actors are afforded access to information about the way public organizations operate (Cucciniello et al., 2017: 36). The transparency and reproducibility of scientific evidence underpinning policy is crucial to build and retain trust. With respect to models, transparency serves to improve the understanding of the workings of a model and how it behaves in different contexts and for different policy measures. But it may also help to validate a model and test its behaviour. In addition, greater transparency of appropriately designed and tested models may also encourage a more widespread use of impact assessments at later stages of the policymaking process and thereby contribute to better regulation.

Maintaining an overview of ongoing modelling activities is an elementary first step for a more transparent and coherent use of models in the policy cycle. It is, however, also a major challenge: For example, in the European Commission, more than 200 models are applied within various policy departments and services with hundreds of scientific and administrative users in various countries. The list of domains ranges from greenhouse gas (GHG) emissions, energy consumption and economy to structural integrity assessment, to name but a few. In addition, the majority of these models run in combination with other models. Thus, they form networks of interaction, further complicated by the related input datasets and assumptions. It is clear that capturing this knowledge, and communicating it in an understandable manner, will not only foster reproducibility and transparency of model use but also enable collaborative, interdisciplinary research that serves cross-policy issues.

“

The transparency and reproducibility of scientific evidence underpinning policy is crucial to build and retain trust.

”



In order to address some of these challenges, the corporate modelling inventory and knowledge management system, MIDAS contains descriptions of models previously or currently in use in support of the EU policy cycle (Ostlaender et al., 2015). One of the starting points of MIDAS was a particular case, where the question was raised whether models used in combination to answer a specific policy question were actually based on the same underlying datasets. As it turned out, answering this question was not that straightforward. In order to better understand how models are used, MIDAS therefore puts these models in their scientific and policy context, by asking the experts to link them to other models, to data, to domain experts, and publications, and to supported policies. In this manner, we are creating a network of thousands of relationships, documenting and preserving model knowledge and use.

To effectively communicate the resulting complex network to a nontechnical audience, MIDAS relies on ‘the simple power of good visualization tools to help untangle complexity and just encourage you to ask questions you didn’t think of before’ (Berlow, 2010).² Especially the latter we frequently experienced over the last years: Scientists and policymakers that are confronted with MIDAS visual aids, start to ask questions about the bigger picture of model use, enquire on relationships between models and data and discover interesting facts that can be found in second or third order relationships, and that are usually hidden to the human eye (Fig. 2).

All of these enhance the transparency of models, the traceability of model results and our understanding of the ongoing support within and across different domains. This makes MIDAS an important corporate tool to use, reuse and document models in a proper way.

Uncertainty and Sensitivity Analysis

In the modelling process, the modellers generally use model assumptions and take input data, from different sources such as sound theory, experts and stakeholders opinions, experience, intuition surveys, measurements, rule of thumb or merely a guess (lack of knowledge). At each step of this process, uncertainty comes into play and may be very significant.³ UA and SA are statistical analysis methods taking this uncertainty into account. Intrinsically, the UA and SA information are of quantitative nature; UA quantifies the uncertainty at the inputs and the output(s) of a model, while SA identifies the factors (assumptions, variables, data, etc.) that influence the most the output(s) and by how much (see e.g., Cacuci et al., 2005; Helton et al., 2006; Saltelli et al., 2008). The steps involved are the following ones:

1. The modellers, the experts, the stakeholders and the UA–SA practitioners identify which assumptions and parameters to vary, and from experiments, estimations, physical bounds, developers, expert and stakeholders opinion, characterise the uncertainty for each selected input by assigning a value, a range, a probability distribution etc.,

²Berlow, E. (2010). Simplifying Complexity, TEDGlobal 2010, at http://www.ted.com/talks/eric_berlow_how_complexity_leads_to_simplicity, accessed 12 November 2018.

³Recently, Silberzahn et al. (2018) showed a study where 61 analysts used the same dataset to address the same research question. The study concluded that a significant variation in the results of analyses of complex data may be difficult to avoid, even by experts with honest intentions; as a consequence, transparency of key assumptions used is always needed.

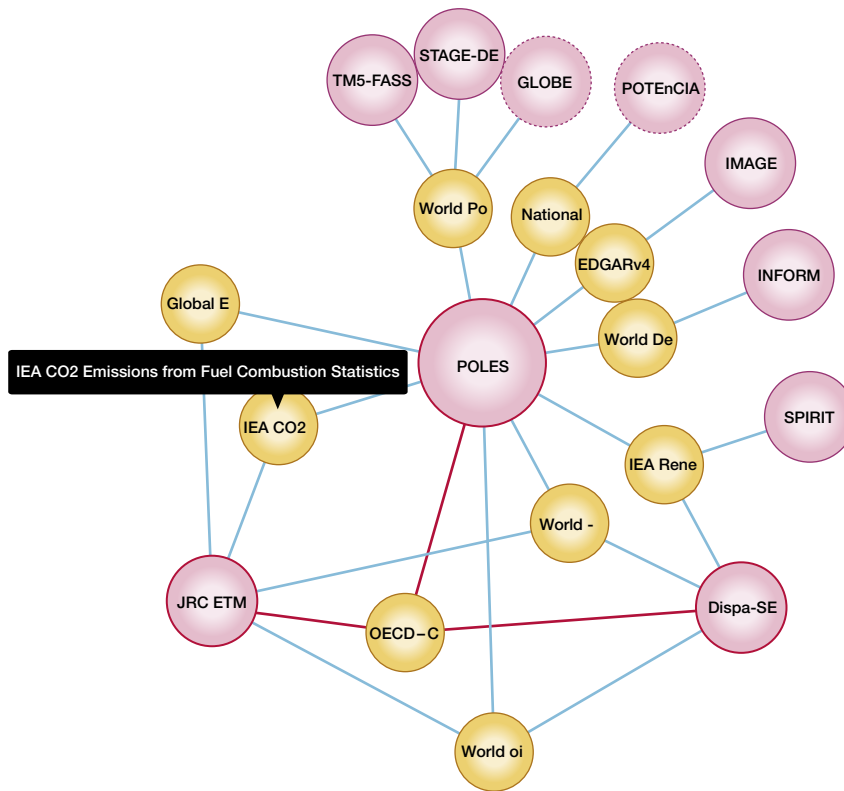


FIGURE 2 Example of datasets (in orange) which the model POLES (in the centre) shares with other models (in magenta).

2. The modellers, the experts, the stakeholders and the UA-SA practitioners define the output of interest of the model and the objectives of the SA,
3. SA practitioners generate randomly the input samples from the previously defined probability distributions (Monte Carlo⁴). The samples are generated so as to explore the whole space of uncertainty,
4. Modellers repeatedly run the model (Monte Carlo), using the values of the input variables dictated by the generated samples (scenario). For each run, they record the value of the output variable of interest,
5. SA practitioners calculate measures of uncertainty on the model output and the different sensitivity measures with respect to the input(s),

⁴Monte Carlo methods are based on reiterated random input generations. The random inputs are run through the model, and the associated output is recorded and analysed. From this uncertainty analysis, it is possible to infer valuable information about the behaviour of the model.



6. SA practitioners identify any major factor (assumption, variable, coefficient, etc.) and examine the impact that these factors might have, namely their respective influence on the model output (ranking),
7. SA practitioners report the conclusions/inferences to the modellers, the experts and the stakeholders, for model improvement (model, data),
8. Together with the modellers, the experts, the stakeholders and the UA–SA practitioners will report to the policymaker on the quality of the model, and the credibility of the policy options and the related policy impact.

Overall, UA and SA give an increased understanding of the inner workings of the model, for example, the sensitivity measures can identify particularly sensitive assumptions and sometimes the uncertainty in these assumptions can be reduced, or an alternative model that does not rely on such assumptions can be built. However, in carrying out these practical steps, various caveats have to be considered, and above all, one has to remember that UA and SA concern with the internal validity of a model. Focusing on models' internal validity (e.g., the quality and quantity of the input parameters), they have little to say about the veracity of models. Veracity and validity are two different concepts; the former referring to the truth value of a model with reference to the real world and the latter referring to its logical consistency. A recent application of UA and SA to impact assessment modelling can be found in [Pisoni et al. \(2018\)](#).

At empirical level, UA and SA have proven to be essential when the quality of a model is an issue. Not only UA–SA techniques are strongly recommended by the Commission Better Regulations [SWD(2017), 350] as tools for modelling quality assurance but also the same techniques are adopted and included in the modelling guidelines of international agencies like the Intergovernmental Panel on Climate Change. The full potential of UA–SA is still little exploited by modellers and model users. This is often true in the field of impact assessment where the impact of uncertainty on the model output, the presumed impact of uncertainty on a policy option and thus its ranking may turn out to be significant particularly at the EU level.

Multicriteria Decision Analysis

The traditional welfare economics approach to ex-ante impact assessment is cost-benefit analysis that is grounded on market mechanisms; this implies that only the behaviour of individuals as consumers in one single institution, i.e., market, is considered. Is this fully acceptable in public policy? The Nobel Prize laureate A. Sen (2009, p. 239) gives the following answer: *'... there is such a long tradition in parts of economics and political philosophy of treating one allegedly homogeneous feature (such as income or utility) as the sole 'good thing' that could be effortlessly maximized (the more the merrier), that there is some nervousness in facing a problem of valuation involving heterogeneous objects, ... And yet any serious problem of social judgement can hardly escape accommodating pluralities of values, ... We cannot reduce all the things we have reason to value into one homogeneous magnitude'*.

The most widespread multidimensional approach to ex-ante impact assessment is MCDA (see, e.g., [Arrow and Raynaud, 1986](#); [Ishizaka and Nemery, 2013](#); [Keeney and Raiffa, 1976](#); [Roy, 1996](#)). The basic methodological foundation of MCDA is incommensurability, i.e., the notion that in comparing options, a plurality of technical dimensions and social perspectives is needed



(Munda, 2016). MCDA builds on formal modelling techniques serving the purposes of decision and policymaking; its basic idea is that in assessment problems, one has to first establish objectives, i.e., the direction of the desired changes of the world (e.g., maximize profits, minimize environmental impact, minimize social exclusion, etc.) and then find useful practical criteria, which indicate the consistency between a policy option and a given objective. As a consequence, MCDA and in particular social multicriteria evaluation (SMCE), which has been explicitly designed for public policy, are very useful methodological and operational frameworks (Munda, 2008). In this framework, mathematical models aim at guaranteeing consistency between assumptions used and results obtained. This is a key success factor since multicriteria mathematics does answer to the standard objection that the aggregation of apples and oranges is impossible in a definitive way.

As a tool for policy assessment and conflict management, SMCE has demonstrated its usefulness in many real-world problems in various geographical and cultural contexts (see, e.g., Figueira et al., 2016). SMCE accomplishes the goals of being inter/multidisciplinary (with respect to the research team), participatory (with respect to the community) and transparent (since all criteria are presented in their original form without any transformations in money, energy or whatever common measurement rod) (see Fig. 3). In operational terms, the application of an SMCE framework involves the following main steps:

1. Description of the relevant social actors. For example, institutional analysis may be performed on historical, legislative and administrative documents to provide a map of the relevant social actors.
2. Definition of social actors' values, desires and preferences. In an SMCE framework, the pitfalls of the technocratic approach can be overcome by applying different methods of sociological research.
3. Generation of policy options and selection of assessment criteria as a process of cocreation resulting from a dialogue between analysts and social actors. In this way, criteria become a technical translation of social actors' needs, preferences and desires.
4. Construction of the multicriteria impact matrix synthesizing the scores of all criteria for all policy alternatives, i.e., the performance of each option according to each criterion.
5. Construction of an equity impact matrix, including all the distributional consequences of each single option on the various social actors.
6. Application of a mathematical procedure in order to aggregate criterion scores and obtain a final ranking of the available alternatives.

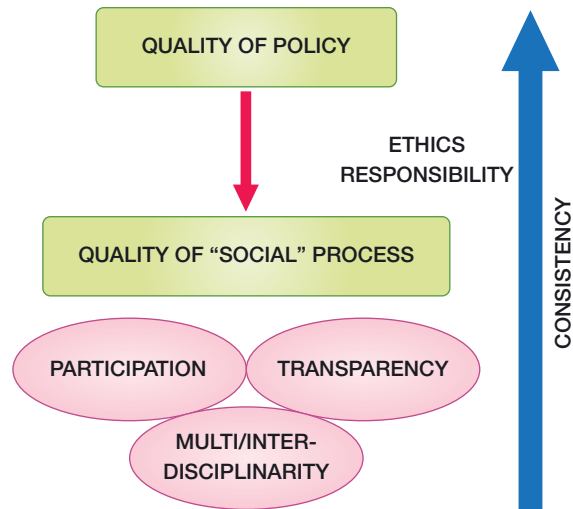


FIGURE 3 Main characteristics of SMCE.



7. Finally, sensitivity and robustness analysis look at the sensitivity of results to the exclusion/inclusion of different criteria, criterion weights and dimensions. While such analysis may look very technical, in reality a social component is always present too. That is, inclusion/exclusion of a given dimension, or set of criteria, normally involves a long story of social, political and scientific controversy and involves social values and social actors.

These steps are not rigid; problem structuring may vary a lot across different real-world problems.

Monitoring Progress Through Composite Indicators and Scoreboards

Hardly any newspaper can resist the temptation of including a reference to an international country ranking. With the same ‘apparent’ ease and simplicity that football teams are ranked according to their performance, countries are ranked according to their ability to provide a high standard of living, competitiveness, and an attractive business environment. The rankings’ use by global institutions (World Bank, EU, World Economic Forum, OECD, International Monetary Fund) has further captured the attention of the media and policymakers around the globe, and their number of applications has surged ever since. The appeal of the reported rankings lies in their simplicity and in their power to act as advocacy tools.

Composite indicators can be powerful practical tools that can help policymakers summarise complex and interdependent phenomena that are not directly observable, such as competitiveness, freedom of press or climate hazards. They provide the big picture, are easy to interpret, easy to communicate and attractive for the public. The resulting figures facilitate country, region or city comparisons and benchmarking. They help monitoring progress over time and evaluate ex-ante policy options based on multicriteria analysis. Scoreboards of indicators have, to some extent, similar objectives to composite indicators, yet they do not consist of a mathematical aggregation of data.

Indexes and scoreboards are also drivers of behaviour and of change by forcing institutions and governments to question their standards (Kelley and Simmons, 2015). The literature often describes the example of the Human Development Index (HDI), which has received a vast amount of criticism since its creation due to the arbitrariness of its methodological framework. However, it is the most well-known composite index to date. Moreover, it led the 1998 Nobel Prizewinning economist A. K. Sen, once one of the main opponents of aggregation, to change his position due to the

attention that the HDI attracted and the debate that it fostered afterwards. A. K. Sen characterised it as a ‘success’ on drawing attention to human development, beyond looking at GDP alone, that would not have happened in the case of nonaggregation. Another similar example is the development of the Global Slavery Index by the Walk Free Foundation, which aims at drawing attention to a hidden and everyday problem: the confronting reality that even in the present day, men, women and children all over the world remain victims of modern slavery.



Indexes and scoreboards are also drivers of behaviour and of change by forcing institutions and governments to question their standards





The increase of composite indicators over the past 20 years is exponential, and the number of yearly publications shows no sign of a decline. Nevertheless, this increase in the index popularity has underlined a key issue: the need for the index interpretation and consolidation. An example of the criticism surrounding the use of composite indicators comes from the Economist. Performance indicators are popular, influential and, if not flawed, then at least subjective and easy to manipulate: get rid of them. This was the main message of two articles in the Economist back in 2014.⁵ The truth is that we are still far from settling the disputes and the criticism concerning the stages of the construction process. This is natural, as there are many stages in the development of a composite indicator and criticism could grow simultaneously regarding each of them. Moreover, if the procedure followed is not clear and not reasonably justified to everyone, there is considerable room for manipulation of the outcome.

One cannot disagree that there are composite indicators out there that are not of sufficient quality. But this should not imply that composite indicators should altogether be dropped from informing policy and public debates. Undoubtedly, the one-number-per country practice may send misleading or partial policy messages because of poor construction or misinterpretation (Hoyland and Moene, 2012). This is why it is important that one **systematically assesses the relevance, added-value and quality of existing (or new) composite indicators**. In this respect, there exist some principles of good practice that would allow the users of these ‘statistical artefacts’ to gauge their quality:

(a) a handbook on composite indicators (OECD/JRC, 2008) offering a ten-step process guide that aims to enhance the transparency and the soundness of the process; (b) SA offering tools to assess the extent to which a composite indicator is volatile to the specificities of its underlying assumptions (see e.g., Becker et al., 2017; Saisana et al., 2005); (c) sensitivity auditing assessing whether and to what extent models (including composite indicators) conform to the scientific standards of accessibility, transparency and reproducibility (Saltelli et al., 2013) and (d) a plethora of audits of international indices conducted upon request of the index developers, and many inspiring articles offering methodological advances in the field. Taken together, these tools, best practices and lessons learnt at least facilitate a public debate based on explicit reasoning in which composite indicators and scoreboards play a role.

The 10-step guide and checklist starts with the development of a conceptual framework that is relevant, easy to remember and apply. In steps 2 and 3, guidance is offered on how to select suitable indicators that meet key criteria and how to conduct analysis to inspect, clean, transform and model data. Step 5 is about normalization that aims at mapping data on a common scale. Steps 5 and 6 are about weighting indicators to allow for more balanced decisions and about aggregating data to reliably synthesize the maximum amount of information. Step 7 measures the statistical coherence of the framework through statistical control tests that help to ensure that information is not lost in aggregation. Step 8 relates to assessing the impact of the assumptions/decisions on the final outcome by means of robustness and SA. Step 9 zooms into the underlying data in order to help interpreting the final results. Finally, step 10 offers suggestions on how to visualise, interpret and communicate the results. Fig. 4 illustrates the 10 steps and an example regarding steps 7 and 8.

The main pros and cons of using composite indicators are the following (see Box 1).

⁵<http://www.economist.com/news/international/21631039-international-comparisons-are-popular-influential-and-sometimes-flawed-ranking-rankings>; <http://www.economist.com/news/international/21631025-I-earn-ruses-international-country-rankings-how-lie-indices>.



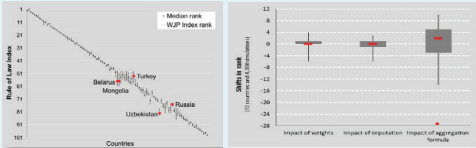
1. Framework How to develop a conceptual framework that is relevant, easy to remember and apply	2. Data selection How to select suitable indicators that meet key criteria																																																
3. Data treatment How to conduct analysis to inspect, clean, transform and model data	4. Normalisation How to map data on a common scale																																																
5. Weighting How to weigh data to make more balanced decisions	6. Aggregation How to aggregate data to reliably synthesise the maximum amount of information																																																
7. Statistical coherence How to use quality control tests to ensure that information is not lost in aggregation	8. Robustness and sensitivity How to assess the impact of the assumptions																																																
<table><thead><tr><th>SSI categories</th><th>SSI Indicators</th><th>I</th><th>II</th><th>III</th></tr></thead><tbody><tr><td rowspan="3">I. Basic Needs</td><td>Sufficient food</td><td>0.82</td><td>0.71</td><td>0.59</td></tr><tr><td>Sufficient to drink</td><td>0.90</td><td>0.72</td><td>0.58</td></tr><tr><td>Safe sanitation</td><td>0.97</td><td>0.79</td><td>0.64</td></tr><tr><td rowspan="3">II. Health</td><td>Healthy life</td><td>0.87</td><td>0.87</td><td>0.70</td></tr><tr><td>Clean air</td><td>0.75</td><td>0.87</td><td>0.62</td></tr><tr><td>Clean water</td><td>0.40</td><td>0.68</td><td>0.52</td></tr><tr><td rowspan="4">III. Personal & Social development</td><td>Education</td><td>0.80</td><td>0.82</td><td>0.71</td></tr><tr><td>Gender equality</td><td>0.47</td><td>0.57</td><td>0.52</td></tr><tr><td>Income distribution</td><td>0.35</td><td>0.37</td><td>0.81</td></tr><tr><td>Good governance</td><td>0.62</td><td>0.76</td><td>0.76</td></tr></tbody></table>	SSI categories	SSI Indicators	I	II	III	I. Basic Needs	Sufficient food	0.82	0.71	0.59	Sufficient to drink	0.90	0.72	0.58	Safe sanitation	0.97	0.79	0.64	II. Health	Healthy life	0.87	0.87	0.70	Clean air	0.75	0.87	0.62	Clean water	0.40	0.68	0.52	III. Personal & Social development	Education	0.80	0.82	0.71	Gender equality	0.47	0.57	0.52	Income distribution	0.35	0.37	0.81	Good governance	0.62	0.76	0.76	
SSI categories	SSI Indicators	I	II	III																																													
I. Basic Needs	Sufficient food	0.82	0.71	0.59																																													
	Sufficient to drink	0.90	0.72	0.58																																													
	Safe sanitation	0.97	0.79	0.64																																													
II. Health	Healthy life	0.87	0.87	0.70																																													
	Clean air	0.75	0.87	0.62																																													
	Clean water	0.40	0.68	0.52																																													
III. Personal & Social development	Education	0.80	0.82	0.71																																													
	Gender equality	0.47	0.57	0.52																																													
	Income distribution	0.35	0.37	0.81																																													
	Good governance	0.62	0.76	0.76																																													
✓ Assess if an indicator dominates the framework [(correlation between indicator and index) > 0.95] ✓ Assess if an indicator behaves as “noise” in the framework [-0.3 < (correlation between indicator and aggregates) < 0.3] ✓ Assess if an indicator is lost in the aggregation [(correlation between indicator and index) < 0.3] ✓ Assess if indicators behave as trade-offs in the index [(correlation between indicator and index) < -0.3]	✓ Identify the sources of uncertainty in the index development and assess their impact to the index scores or ranks ✓ Complement aggregate scores/ranks with confidence intervals and consider policy guidance based on country groups than exact ranks ✓ Identify which uncertainties are more crucial in determining the results (why certain countries notably improve or deteriorate their performance given the assumptions) ✓ Do not confuse robustness with quality																																																
9. Back to the data How to make the most of an index by looking back at the data	10. Visualisation and Communication How to visualise, interpret and communicate results																																																

FIGURE 4 The 10-step guide for constructing a composite indicator.

Ex-Post Evaluation

Ex-post impact evaluation aims at answering questions of the type: ‘What is the added value of the policy intervention?’ This question cannot be fully answered by qualitative, process or monitoring evaluations because these approaches lack the ability to identify a causal effect; answering this type of question requires specific methods and data.



Box 1: Pros and Cons of Composite Indicators

Pros	Cons
• Can summarise complex, multidimensional realities with a view to supporting decision-makers. • Are easier to interpret than a battery of many separate indicators. • Can assess progress of countries over time. • Reduce the visible size of a set of indicators without dropping the underlying information base. • Thus make it possible to include more information within the existing size limit. • Place issues of country performance and progress at the centre of the policy arena. • Facilitate communication with general public (i.e., citizens, media, etc.) and promote accountability. • Help to construct/underpin narratives for lay and literate audiences. • Enable users to compare complex dimensions effectively.	• May send misleading policy messages if poorly constructed or misinterpreted. • May invite simplistic policy conclusions. • May be misused, e.g., to support a desired policy, if the construction process is not transparent and/or lacks sound statistical or conceptual principles. • The selection of indicators and weights could be the subject of political dispute. • May disguise serious failings in some dimensions and increase the difficulty of identifying proper remedial action, if the construction process is not transparent. • May lead to inappropriate policies if dimensions of performance that are difficult to measure are ignored.

For instance, to quantify the impact of a training programme on employment, one can compare the employment status of participants to the training programme, the treated group, with the one of a group similar in all respects to the participants, but which was not exposed to the training, the so-called control group. In controlled experiments, units are either assigned to treatment or given a placebo at random. The comparison between treated and control groups is useful to disentangle the causal effect of the policy from additional factors that may occur in parallel and influence the outcome of interest, but in a manner that is not due to the policy.

The comparison with the control group is an example of how to construct a *counterfactual* outcome, which is defined as the outcome that *would have occurred without the policy intervention*. This type of evaluation, called counterfactual impact evaluation (CIE), is used to measure the causal effects of an intervention on a given outcome.

Causal evaluations require data of the appropriate granularity and appropriate methodology. In fact, in order to find a valid counterfactual, data on single individuals, firms or local level of government are usually needed; these data are called microdata, as opposed to macrodata. Data of different granularities are good for different policy evaluation purposes. Microdata are especially fit for finding evidence of policy impacts at the individual level, while aggregate data are useful for studying macroeconomic (or general equilibrium) effects.

Microdata can come from surveys or from other sources, like administrative records. These are usually collected for administrative purposes by governments or other public



administration agencies in the course of their regular activities; examples include data on income collected in tax declarations, data on wages and employment collected by social security, data on education collected by Ministry of Education, etc. During the past 20 years, interest in administrative data for social research and policy evaluation has increased exponentially (see e.g., [Poel et al., 2015](#); Card et al., 2010).

Microdata from different sources can be linked together; for instance, income data on a single individual can be linked to her/his education record coming from the registry in the Ministry of Education. Linked microdata offer several advantages: they are relatively inexpensive, they are typically of good quality and they usually provide information on large sets of individuals (units) for which statistical inference works well, see [Crato and Paruolo \(2019\)](#). Full respect for the legal safeguards for handling personal data is essential in this area of research and analysis. The General Data Protection Regulation, which came into force in May 2018, sets out the legal rules in force in the EU.

The major CIE techniques in current use are difference-in-differences (DiD), regression discontinuity design (RDD), instrumental variables (IVs) and propensity score matching. They are briefly reviewed below; see [Imbens and Wooldridge \(2009\)](#) and [Imbens and Rubin \(2015\)](#) for more details.

DiD: This CIE technique estimates the average treatment effect by comparing the changes in the outcome variable for the treated group with those for the control group, possibly controlling for other observable determinants of the outcome variables. Because it compares the changes and not the attained levels of the outcome variable, this technique eliminates the effect of the potentially different starting points of the two groups. A basic assumption of DiD is the common trends assumption, namely that treated and control groups would show the same trends across time in the absence of policy intervention. Hence, the change in the outcome variable for the control group can be used as an estimate of the counterfactual change in the outcome variable for the treated group ([Box 2](#)).

RDD: This CIE technique allows one to estimate the impact of a programme when participation in the programme depends on a rule or a threshold in any observed characteristics (e.g., age, income). Consider, for instance, a training programme that targets individuals below a given age. In this setting, to assess the impact of the programme on the employment status or wages, one can compare the outcomes of the marginal beneficiaries and nonbeneficiaries on either side of the threshold. RDD assumes that individuals on either side of the threshold but close to it are similar in terms of observed and unobserved characteristics. The size of the jump in the outcome of interest determines the impact of a programme. Since the observations that contribute to identifying the causal effect are those around the threshold, RDD may require large sample sizes, so it is not always easy to implement it. Furthermore, RDD identifies a local effect around the threshold, and in order to extrapolate the effect away from the threshold, stronger assumptions are needed to detect the effect ([Box 3](#)).

IVs: This CIE technique, known as IVs uses information from an observable variable, called the instrument, which predicts the assignment of units to the policy intervention (determines participation) but does not have other direct impacts on the outcome of interest, i.e., the instrument impacts the outcome variable only through its influence on the treatment status ([Box 4](#)).

Propensity score matching: This CIE technique has become a standard tool when evaluating the impact of participating in Active Labour Market Policies on participants' employment



Box 2: Example of an application of the DiD

Consider an impact evaluation of some regulations aimed at reducing the level of airport charges. Assume that for some reasons only one group of airports (treated group) was exposed to the regulation, while a second group of airports (control group) was not. One should not look at the average level of airport charges once the regulation is in place because the starting values for the airport charges in the two groups may be different. The DiD approach instead compares the magnitudes of the *changes* in the average level of airport charges in the two groups of firms. This gives an estimate of the impact of the regulation.

Fig. 5, taken from Conti et al. (2019), illustrates the effect of the implementation of the Airport Charges Directive. What emerges is that after the adoption, there is evidence of a significant decrease of the level of airport charges in treated airports compared to control ones.

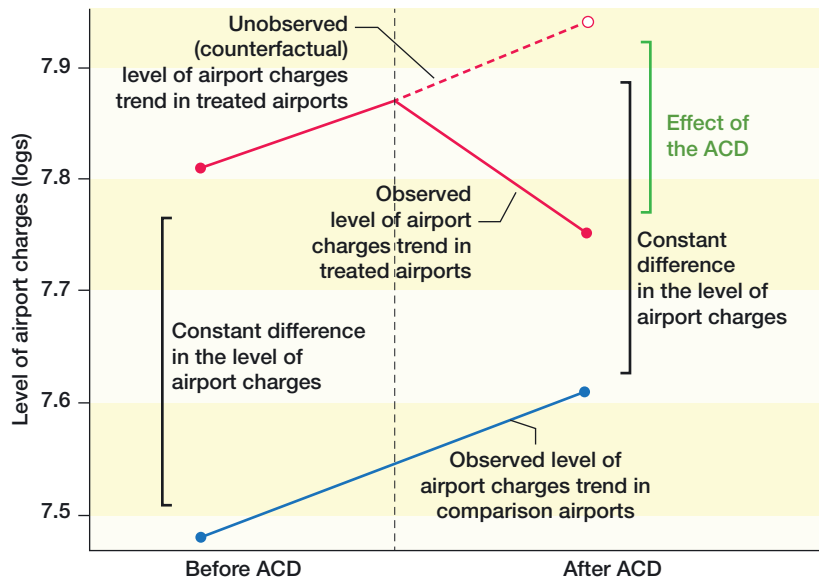


FIGURE 5 Illustration of the estimated Average Treatment Effect via DiD of the Airport Charges Directive in Conti et al. (2019).

prospect. The idea is to find a match for each participant in the programme (treated group) from the sample of individuals that did not participate in the programme. Matching occurs on basis of observed characteristics, such as age, gender, education, etc.

After creating all the matched couples, the effect is estimated by comparing the average of the outcome variable between the treated and control groups. Usually, individuals in the two groups are matched on the so called propensity score, which is the predicted probability of participation in the programme given the set of observed characteristics. This allows to reduce the complexity of the exercise, as one may run out of matches when increasing the number of observed characteristics used for the matching (Box 5).



Box 3: Example of an application of the regression discontinuity design

Consider the Youth Guarantee (YG) programme in Europe, launched in 2014 to cope with the emergency of high unemployment rates among youth of age 15–29. For some programmes, priority to participation was given to registered unemployed youth aged below 25. In this setting, one can analyse the impact of a given programme financed under the YG by comparing the outcome (e.g., employment status) of the participants (those younger than 25) and nonparticipants (those older than 25).

For example, Bratti et al. (2018) evaluate the effectiveness of a vocational training (VT) programme on employment outcomes within the YG scheme in Latvia. Fig. 6, taken from this study, shows the estimated RDD impact of age on the probability to participate, with a clear discontinuity in the probability of participation at this training programme around age 25, normalised as age 0. RDD assumes a local randomisation around age 25. (Note that people cannot manipulate their age in order to participate in the programme.) The study then looks at the impact of participation in the VT programme on employment by comparing the employment status of individuals younger than 25 and those older than 25 exploiting rich administrative data.

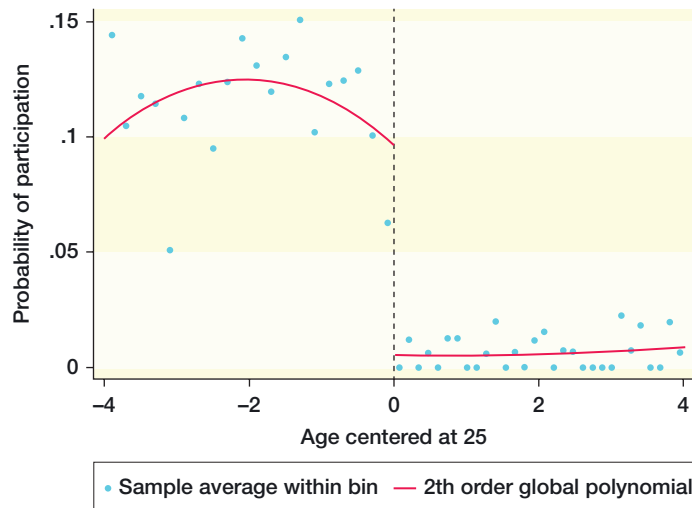


FIGURE 6 RDD estimate of the effect of age on the probability of participation to vocational training programme financed by the Youth Guarantee in Latvia, taken from Bratti et al. (2018). They study this effect as the first stage in the analysis of the effects of the programme on employment status.

In summary, counterfactual techniques typically provide causal inference by comparing a treated and a control group. When these groups are formed by random assignment or by other appropriate statistical techniques, the average impact of many extraneous factors, which might undermine the robustness of other assessment methodologies, is controlled for in the analysis. Counterfactual techniques provide a higher level of rigour to impact evaluation than other more descriptive analytical techniques. Counterfactual techniques can be combined with other econometric approaches to investigate which causal channels have been activated by the policy intervention, see Heckman (2010) and Heckman and Vytlacil (2007a,b).



Box 4: Example of the application of instrumental variables

A study by [Martelli et al. \(2018\)](#) uses IVs to understand whether the political commitment to reduce emissions at the municipal level undertaken by an incumbent mayor has an impact on the share of votes in the subsequent elections for the mayor's second term. In this study, political commitment is represented by the participation in the EU covenant of mayors (CoM), which is a mainstream European movement involving local authorities that decide to participate on a voluntary basis to meet the European target of GHG emissions under their mandate (–20% by 2020).

Because participation in the CoM is voluntary, it is not possible to assume that participation to the programme is random, and one needs to deal with *selection bias* in programme participation. To solve this problem, the authors use an IV strategy and exploit the exogenous variation in the incentives to participate in the CoM given by the creation of a Covenant Territorial Coordinator (CTC), which is an entity offered by some provinces and regions. This is considered a good instrument because the creation of a CTC is expected to impact the probability to participate in the CoM by reducing its nonfinancial costs, but it should not affect the electoral outcomes in any other way. The authors find no reduction in the share of votes of the incumbent mayor who signed up for CoM.

Box 5: Example of an application of the propensity score matching

A recent study (Ghirelli et al., 2019) used this technique to evaluate the impact of an ESF-funded programme on employability and quality of the job (type of contract). The programme was a job training for registered unemployed graduates residing in the Italian region of Umbria. To estimate the impact of the programme, the study relies on matching graduates participating in the training with similar graduates that did not participate in the programme. To carry out the impact evaluation, the authors relied on administrative data on the whole population of registered unemployed graduates in the region.

The availability of data of a sufficiently granular level of detail is essential for the use of all of the counterfactual techniques described in this section. Where such data are not available, one may need to resort to other more aggregate (macro) techniques; however, this may render the effects of specific policies invisible to the analysis. On the other hand, macroanalysis may be useful to investigate general equilibrium effects.

Conclusions

In a recent Financial Times debate on the future of economics,⁶ the main question was '*has economics failed?*' In general terms, everybody agreed that economics (and all quantitative policy analysis approaches) has developed many important tools to help real-world policymaking

⁶May 2018, more than 300 readers commented the initial debate between Tom Clark, editor of Prospect Magazine, and Chris Giles, the FT's economics editor.



although it is not an exact science, obviously. Quantitative modelling of policy problems cannot provide exact answers, but it can help policymakers and all social actors involved by providing a scientific sound framework for a systematic and transparent analysis.

In our opinion, in this context, 'evidence-based policy' does not mean that public policy is actually made on scientific evidence to the exclusion of all other considerations. It means that evidence from science and rigorous analysis should be used to the maximum extent possible, but of course within the wider constraints of legitimacy, politics and so on. Amartya Sen has raised the point that democracy to be effective in practice and needs a shared language of the public sphere, and this cannot be supplied by any specific single science.

Given that a complex real-world problem cannot be completely described by a model, since descriptive completeness and formal consistency are incompatible, it is essential to remember that a model always depends on the assumptions introduced to make the model relevant for a given policy problem, i.e., which representation of reality is used. Fairness in the policy process can be seen as an ethical obligation to take a plurality of social values, perspectives and interests into account in a coherent and transparent manner.

The responsible use of policy models requires some level of understanding of their limitations as well as their strengths. It is imperative that modellers devote efforts to explaining the assumptions underlying their work, margins of potential error and possible inconsistencies across different policy objectives. But it is also crucial that those formulating policy based on modelling inputs do so in an appropriate manner. Models and model results can of course be 'appropriated' by policymakers to support preordained objectives. In light of this, a two-way dialogue between modellers and model output users would seem to offer the greatest scope for responsible use of models in the policy process.