Bojan Radej, Forthcoming in Working papers of Slovenian Evaluation Society (9/2), Spring 2016 (ver. 6)

Aggregation Problem in Evaluation of Sustainable Development

»We used to think that if we knew one, we knew two, because one and one are two.
We are finding that we must learn a great deal more about 'and'.«
Arthur Eddington (1882–1944), English astronomer (in Mackay, 1977)

Abstract: Evaluators of large-scale policy interventions have an aggregation problem (Scriven) arising from incommensurability of assessed policy impacts. Leopold et al. recognised this and left evaluation results in a disaggregated form. However, they failed to observe that cross-sectional impacts are at least weakly commensurable. Ekins and Medhurst have acknowledged this, but did not implement the finding consistently. When inconsistency is removed, fragmented results are first partly aggregated into a square input-output matrix of impacts on the meso level, and then synthesised correlatively. The aggregation problem is further studied in a horizontal perspective by extending Dopfer et al.'s classification of Meso 1, 2 and 3 sub-levels of synthesis with Meso 2a and 2b. One of our conclusions is that a precondition for neutral evaluation of complex policy intervention is not only an objective analysis of policy impacts but also a consistent synthesis of poorly compatible and value laden detailed evidence.

Keywords: Aggregation, Policy Impact Evaluation, Social complexity, Meso level, Horizontal synthesis.

Aggregation Problem

A given complex social issue, such as sustainable development, needs to be researched simultaneously at multiple levels: vertically (micro and macro), and in multiple domains horizontally, for instance in terms of economic, social, and environmental sustainability. Results obtained on different levels and in different domains cannot be aggregated into uniform conclusions because they are incommensurable; measurable in incompatible units so they cannot be expressed quantitatively in terms of each other (Gutiérrez et al., 2013). They must be evaluated as qualitatively incompatible. Where no objective basis exists for comparison between domains of inquiry, there is no common denominator on which aggregation can be accomplished in a uniform way.

Arrow (1951) claimed there is a logical impossibility in devising a democratic and rational mechanism for the aggregation of results obtained from multiple contributions, because of the plurality of legitimate perspectives involved in researching collective issues that are circular in nature ('non-transitive') and thus not reducible to a single 'correct' view (Ravetz, 2003). Kuhn (1970) outlined that different sciences, in this case economics, sociology and environmental science, integrate information in different ways, and different theories weigh the appearances of the same world differently, so that different and even contradictory judgments are supported. In socially complex situations, incommensurable viewpoints about collective issues with different principles of legitimacy and social primacy must be reckoned with (Wacquant, 1997). As a result, even the most competent, honest and disinterested scientists may arrive at different problem framings and conclusions about sustainable development because of systematic differences in the way they collect, analyse and summarize available information (Mumpower et al., 1996). Yet without an aggregate measure of sustainable development, there is no straightforward answer to the question of how to comprehend and thus govern the sustainability of development, and how to evaluate the integrity of policy achievements.

Coleman traces the roots of the micro-macro problem to the philosophies of Hobbes, Smith, Locke, Rousseau, and Mill from the Seventeenth to the Nineteenth centuries (in Morçöl, 2012). 'In all these philosophies the central question was how purposeful action of individuals were connected to macrosocial phenomena. How do acts of individual actors lead to macro events? How dissatisfaction becomes revolution, individual fears into panic, individual aspirations into market demand or supply' (Morçöl, 2012)?

Methodological problems of aggregation in studying socially complex concerns are symptoms of an even broader challenge, one that invokes crucial questions about the methodology of social research in general, and about appropriate ways to sum up and make sense of what we know about complex social reality. Coleman and Parsons argue that the aggregation problem is the central issue in the social sciences (in Morçöl, 2012). This is especially emphasised in the forthcoming complexity turn in social sciences (Wallerstein, 2004). Increasing social complexity acts as a prominent driver for the highlighting of the aggregation problem (Simon, Ando, 1961; Foster, Potts, 2007) as one of the major methodological concerns in various walks of social sciences: in macro-economy (Keynes), social choice (Arrow), elections of political representatives, public opinion, social integration (Durkheim), and sustainable development (WCED, 1987) are only few among more prominent examples.

Despite a well-founded diagnosis of this challenge, explanations of laws that link the individual to the collective consideration level remains one of the least developed aspects of social sciences (Coleman in Åberg, 2000). All sciences reveal a micro-macro divide, and even the most advanced have not reconciled the two levels theoretically (Turner, 2006). List and Polak (2010) observe that the problem in general remains unresolved, as the literature has mainly concentrated on how to avoid it. The aggregation problem is also generally absent from conventional textbooks (Elsner, 2007). Social sciences 'ignored the problem and instead focused on individual actions or studied macro processes and structures only' (Morçöl, 2012). As a result, social research often lacks a big-picture vision about large processes that cut across diverse and essentially incompatible domains of inquiry, and instead pursues either a fragmented view of social reality or a too generalized view, both of which are misleadingly oversimplified.

One area of social research that is especially affected by this unresolved aggregation problem – and one that invokes a fully fledged paradigm crisis of the entire field – is policy impact evaluation (Virtanen, Uusikylä, 2004; Hertin et al., 2007). To evaluate (governmental) policy impacts is to collect detailed factual evidence of the performance of policy measures and to make a neutral judgment about the policy's overall worth or merit (Scriven, 1994).

Standard evaluation approaches are designed for the appraisal of homogeneous, sector-based policy interventions with commensurable impacts that are observable from very specific point of view (Elbers et al., 2007; Rotmans, 2002), such as assessing the economic impacts of economic policies. However, in the real world, governments usually intervene in situations that are not commensurable, since they have a number of unrelated goals (such as social and environmental, in addition to economic), all of which may hinder, support or reinforce one another. Furthermore, a democratic government needs to take account of all the divergent viewpoints from which its achievements are legitimately evaluated. Governmental inability of aggregation in complex conditions leads to inconsistent policy-making, particularly at the strategic level.

In situations that involve non-linear relations, where the whole is larger (or smaller) than sum of the parts, the summation procedure is far from trivial (Veen, Otter, 2002). The relevance of synthesis has long been recognized in policy impact evaluation (Rotmans, 2002; Weaver, Rotmans, 2006), but poorly dealt with. Several renowned studies have confirmed this. The Impact Assessment Board (IAB, 2009) estimates that the majority of impact-assessment studies provided to the European Commission supply the kind of information that fails to inform policy makers whether their global objectives are met. Policy evaluation has been in particularly failing in producing forms of knowledge that strategically inform actions intended to improve large-scale and cross-sector policies. Compatible conclusions are obtained by Hageboeck et al. (2013) in their meta-evaluation of USAID evaluation reports, and in Huitema et al. (2011), who assessed evaluation studies prepared for the needs of EU climate policy. Standard evaluation approaches in the EU in particular struggle with how to appropriately summarise incompatible policy

impacts on diverse evaluation criteria, such as the Strategic Impact Assessment (2001/42/EC), Impact Assessment Guidelines (SEC(2005)791), and the Territorial Impact Assessment (TIA; ESPON - 3.2, 2006).

Unsurprisingly, there is widespread recognition of the failure of policy impact evaluation to live up to its promise of improving public governance and more consistently contributing to social welfare where 'no one is left behind', as stated in new United Nations agenda for sustainable development (UN, 2015). The lack of explicit justification of the aggregation procedure is the Achilles heel of evaluation efforts (Scriven, 1994). This calls for review of the foundations of the aggregation methodology in evaluation (Scriven, 1994).

Evaluation methodologists have in fact contributed considerable efforts to include logic of complexity into synthetic evaluation of social issues. A wide range of methods have been developed for 'qualitative synthesis' of incommensurable contents, such as meta-narrative synthesis, critical interpretive synthesis, meta-study, meta-ethnography, thematic synthesis, textual narrative synthesis, and framework synthesis (Barnett et al., 2009; Gerrits, Verweij, 2015). These approaches are mainly aimed to narrow the aggregation problem by adopting various strategies of avoidance when they introduced new overarching or intermediating categories that recover the possibility of some sort of synthesis. In some approaches, this is achieved by introducing overarching meta-categories such as meta-narratives, synthetic constructs, frameworks of synthesis, and meta-summaries. On other occasions, the possibility of synthesis is recovered by instituting 'reciprocal translation algorithms', or by triangulation, by exploring contradictory claims in inputs to synthesis, by replacing quantitative synthesis with interpretative synthesis, or by replacing aggregating data with comparative understanding (Barnett et al., 2009).

Yet the challenge remains unresolved despite these contributions. How should one aggregate pieces of quantitative information that belong to qualitatively different domains of consideration? It is concluded with this case study that complex social issues can be evaluated consistently in a highly aggregated manner only when the horizontal and vertical axes of synthesis are organized in mesoscopic way, which is a less restrictive method compared to those that currently dominate the field.

The aggregation problem is studied through practical example by a comparative ex-ante evaluation of the sustainable impact of development programme for the Slovenian North-Eastern region Pomurje for 2007 – 2013 (RP; Radej, 2006). The Pomurje region is mainly agricultural and is the least developed Slovenian region (at Eurostat's NUTS 2 level)[1], covering 6.6% of the national territory, containing 5.5% of the national population, contributing 4.3% of the gross domestic product (GDP) and achieving less than 70% of national average GDP per capita. The region has a strong cultural and ecological identity – more than a third of its territory is protected as landscape parks or nature reserves, including the unique landscape along the River Mura. Its economic capital is fragile but has been improving since the mid-1990s. Its social capital is very frail and is further depleting. For half a century, the region had been surrounded by Cold War borders – from the West (Austria) as well as from East (Hungary) – and it is the only region of this kind in Central Europe.

Coinciding with the geostrategic realignments in Eastern Europe at the end of 1980s, the region suddenly found itself situated on the main European transport corridors, which exposed it, unprepared, to intensive international flows of capital, goods and people. The accession of Slovenia to the EU in 2004 further imposed a more stringent border regime between Pomurje and the Republic of Croatia (at that time not yet a member of the EU), borders which had been traditionally close up to that point. These shocks have further weakened the region's social and

1 NUTS - Nomenclature of territorial units for statistics.

human capital, leading to continued depopulation, brain drain, long-term unemployment, and prolonged health and social risks, so that the majority of the regional population is now officially classified as vulnerable.

Structural development lags have accumulated in regional social capital since the late 1980's despite the considerable inflow of investment from the national and EU budgets in the past two decades that was specially earmarked as stimulus for projects in less advanced regions, because not enough emphasis was placed on genuine local needs. Decision makers failed to address appropriately critical regional trends and trade-offs between the economic, social and environmental domains of regional sustainability. This case study therefore aims to evaluate ex-ante if the proposed Regional Programme is likely to contribute to a reversal of critical development trends, especially trends in social capital, and furthermore, if RP is likely to induce regional sustainable development in a consistent and synergetic way relative to inherited baseline conditions.

This case study is accomplished comparatively with three alternative approaches to the summation of detailed evaluation results: a disaggregated or micro level approach (with Leopold's matrix); a highly aggregated or macro level approach (with Ekins-Medhurst's matrix); and an intermediate or meso level approach (with application of square input-output matrix). The expectation is that different aggregation approaches will bring about dissimilar evaluative conclusions and lead to diverse and possibly even contradictory policy advice (Gutiérrez et al., 2013).

Setting the Scene

The purpose of synthesis is to render sensible a heterogeneous corpus of information (Encyclopaedia of Evaluation, 2004) from a more general viewpoint. Synthesis comprises a set of methods aiming at consolidating multiple values into a single value that represents the whole.[2]

Aggregation is the simplest method of direct micro to macro synthesis, aside from multiplication, standardization, correlation, and integration. Aggregation is accomplished through cumulative additions of similar elementary parts to a pile or stack that is acknowledged as a whole. It assumes linearity and conservation of the qualities in transition between the lower level of the elements to the higher level of the aggregate. Strong conditions of symmetry and homogeneity must be met in such a case. 'Being symmetric' in general usage is synonymous with 'being harmonious', and in technical usage, as in Euclid's Elements, it is synonymous with 'commensurable' (Lorenz, 2005), denoting that diverse parts share a common denominator which can be then applied, with the aim of aggregating diverse parts as substantively uniform manifestations of the same essence that is summarised into the aggregate.

When all social values are translated into one ultimate value that is reducible to a single quantum of measurement, then the demand of rationality seems clear: aggregate and maximize (Scharffs, 2001). Trust in the neutrality of simple aggregation logic has led many classical social scientists to constitute it in their 'Political arithmetic' (William Petty, 1690) as a main cohesive mechanism, such as with the majority principle in political decision-making or the maximisation of welfare in economics. The founder of modern utilitarianism, Jeremy Bentham, claimed that the interest of the community could be neither more nor less than 'the sum of the interests of members who compose it' (1789 in his 'The Principles of Morals and Legislation'). In accordance with the aggregative nature of his "greatest happiness" principle ('the greatest happiness of the greatest number that is the measure of right and wrong'[3]), the interests of individuals should be added up

2 https://docs.oracle.com/cd/B19306_01/olap.102/b14349.pdf, III 2013.
3 Wikipedia, #Jeremy Bentham, I 2015.

in a maximisation strategy in order to promote an underlying universal value system (James, 1981) for society as a whole.

A more contemporary example that represents the ultimate expression of the aggregative political ideal is the idea of an ‘e-government’, in which citizens participate directly through referenda conducted online on a day-to-day basis in majority-based decision-making.

Assumption of commensurability between parts as a precondition of forming wholes is generally true for physical entities in expressing their physical characteristics such as size, location or speed – but it may not be true for social characteristics (Scholes et al., 2013), which may have a far less deterministic structure in their defining characteristics because they are heavily ‘contaminated’ with incompatible value-based considerations.

Furthermore, the aggregate composed of essentially similar elements is merely a repetition of the elementary level, and in this sense tautological (Allport, 1928), and relatively uninteresting (Wimsatt, 2002). In this approach to synthesis, unity is already involved in the details – in the substantive essence of the parts that operate as the common denominator – so unity does not emerge by forming an aggregate. As a result, the aggregate is simply a repetition of the same details; it is not qualitatively different and thus cannot become independent of its parts. A new quality, such as ‘collective novelty’ (Perry, 1922) or a new wisdom about the whole cannot emerge from the aggregation of commensurable content (Allport, 1928).

For Isaiah Berlin all social values are reduced to commensurable relations, thus values are no more about universals but are simply a matter of calculation (in Scharffs, 2001), which cannot capture the ethical texture of social life (Mason, 2006). The unsolvable problem with aggregative maximizing strategies is that it will usually be impossible to settle on which value should be the one that is maximized (Scharffs, 2001), and which specific observed characteristic of reality should be taken as the commensurable representative for all others. In this regard, Schumpeter could not support Bentham’s maximisation doctrine, since there is no common good about which everyone agrees ‘because ultimate values – our conceptions of what life is and what society should be – are beyond the range of mere logic’ (in Coleman, 2005) and calculation.

Social reality is increasingly comprehensible to us as complex. There are different and incompatible valuations of what is to be counted as indispensable, and so all valuations of social facts cannot be unified on a common denominator (Funtowicz, Ravetz, 1994). For that reason, synthesis of heterogeneous contents needs to go beyond simple methodology in which detailed inputs to aggregation may not be routinely or fully aggregatable from a micro to a macro level.

The methodology of social research has offered a wide range of strategies for synthesis that aim to recover the possibilities of aggregation when commensurability of observed social facts is not secured. Three main lines of approaches can be distinguished for the narrow purposes of underlying inquiry: the non-aggregative, the aggregative and the intermediate or multi-criteria approach.

The first line of approaches covers methods, which do not permit cumulating detailed information into aggregates. Findings in one domain of research cannot be directly combined with findings in other domains. Each domain must stand on its own (Xu, Yang, 2001). A good example of this is the strong sustainability thesis (Pearce et al., 1990), which claims that trade-offs between economic, social and environmental capital are not permitted because capitals are of different natures. Aggregation can be accomplished in sustainability research only partially, within domains of sustainability, but full aggregation between domains from a micro to a macro level is beyond reach. As a result, research of large-scale social processes remains either very narrow in scope when focused on a single research domain, or poorly integrated, producing only relativist conclusions, when involving multiple but disconnected research domains.

Moreover, the family of aggregative approaches is furthered by compensatory or weighting methods (for 'non-compensatory method of aggregation' that for instance aggregates the corpus of information by relying only on one set of assessment results, such as best or worst achievements between multiple criteria; see Munda, 2004, 2012). A composite indicator may be obtained as an index derived from more detailed data for measuring the aggregated performance of a multi-dimensional issue, data that have no common units of measurement (Nardo et al., in Zhou, Ang, 2009). Under some additional restrictive conditions, composite indicators permit partial or full aggregation of detailed information. Substantively diverse issues are first translated into 'neutral form', such as into index number or standardized, and only then aggregated.

Another compensatory method relies on the application of social weights. For instance, higher weights are attached to environmental valuations because of the applying precautionary principle. In other cases, weights are proportional: each preference in collective choice should have a weight equal to the proportion of the population supporting the values represented by each alternative option. Economists have also developed different techniques for smoothing out conflicts in collective choice at a lower level in order to achieve maximal outcome at a higher level, such as by compensating losers for negative trade-offs (Pareto), or by weighting collective alternatives relative to willingness to pay (in Baumol, Oates, 1988). Thus, in compensatory approaches to aggregation, no specific value prevails simply by securing that all involved values in social inquiry, universal and particular alike, are taken as equally relative.

Methodological efforts aimed at shaping detailed analytical insights so that they are more aggregatable have offered valuable assistance in certain situations, such as when underlying incompatibilities in measured phenomena remain small and local; socially complex situations, on the other hand, involve radical diversity. Even when incompatibilities remain small, a large amount of additional parameters, such as trade-off coefficients, thresholds, and weights, must be constructed in order to enable aggregation, which may cause a loss of transparency and consistency in the aggregation model (Munda, 2004). In this way, substantive judgments may be easily replaced with purely technical and statistical concerns on how to collect ideally fitting data and how to calculate the most appropriate version of parameters. Furthermore, parameters may be very hard to establish scientifically, and are often selected ad hoc (Munda, 2004), thus becoming affected by the same type of incompatible value judgments that constrained implementation of simple aggregative methodology from the outset.

The third main intermediate family of approaches to synthesis that addresses a lack of commensurability between social facts consists of multi-criteria methods. Multi-criteria analysis is used to describe any structured approach to determining overall preferences among alternative options, where the options accomplish several objectives (DCLG, 2009). The methodology retains the idea of a single objective function and optimization, but emphasizes that analysis should more explicitly consider separate aspects of valuation in order to capture all horizontal domains of research indiscriminately (Söderbaum, 1998). Due to the application of independent horizontal criteria, criteria that may be in conflict, multi-criteria methods do not always have a conclusive or unique solution (Xu, Yang, 2001). Even if they do produce solutions, these soluctions consist of of a limited set of interrelated propositions aimed at understanding limited topics (Geels, 2007), so they cannot support definite judgments about that kind of holistic concerns that are outlined here.

Söderbaum (1998) analogously distinguishes between three sets of evaluation methods according to the degree of aggregation of detailed policy impacts assessment into an evaluative summary: disaggregated, aggregated and intermediate methods.

Disaggregated evaluation methods, e.g. monitoring, environmental impact assessments, or the Leopold assessment matrix, are multi-dimensional methods, so they do not capture only one

collective value or reduce diversified effects to a simplified scheme (Söderbaum, 1998). Leopold et al. (1971) proposed a detailed assessment method at the micro level from which a synthesis of results remains absent as a matter of principle. They claimed that detailed assessment results should be presented in fragments, leaving policy-makers with full responsibility for the synthesis of evaluation conclusions and for drawing its wider policy implications.

On the other hand, highly aggregated evaluation approaches, e.g. cost-benefit analysis or cost-effectiveness analysis, are useful in circumstances in which a consensus about 'specific valuation rules' (Söderbaum, 1998) has already been established. One of the first declared multi-domain macro-evaluation methods was the Strategic environmental assessment (SEA; Sadler, Verheem, 1996; SEA Directive, 2001/42/EC); but this method gives no indication about how to cumulate diverse economic impacts on various environmental criteria. The method's authors implied that strategic insight follows automatically and in a unified form of expression from all-encompassing assessments that are accomplished from a higher level of evaluative concerns. The missing horizontal element of both elemental (micro level) and structural multiplicity (meso level) in evaluative synthesis has been contributed by Ekins and Medhurst (2003, 2006), in their novel approach to assessment of the EU Structural Funds programs' impacts on regional sustainable development. Ekins previously proposed the Four-capitals model to serve as a conceptual framework for more structured evaluation of sustainable development (Ekins, 1992). The four-part evaluation framework can be traced back to the Brundtland report (WCED, 1987), and to the conference on sustainable development in Rio de Janeiro (UNCED, 1992), where Ekins and Munasinghe independently proposed analogous idea.

Ekins proposed a partly aggregated version of the Leopold matrix, which similarly evaluates programs' incommensurable impacts on economic, social, environmental, and human capital, bringing about four independent aggregated indicators of social-wide policy impact. This obtained extended matrix is named the Leopold-Ekins-Medhurst matrix, or LEM. LEM allows for partial aggregation of assessed impacts for all assessment criteria within each of four assessment domains that are placed in its columns. Policy impacts are aggregated for all criteria only within each impact domain, but not between them, appropriately accounting for their incommensurability. An analogous approach to synthesis is now accepted also in the EU's territorial impact assessment (ESPON – 3.2, 2006).

LEM has made an important contribution to cumulative evaluation methodology. Nevertheless, LEM's aggregation approach is inappropriate in its second step, when it allows for the summation of all policy impacts irrespective of the source of impact. Different policy measures affect given evaluation criteria in qualitatively different ways. Policy impacts may therefore be aggregated only partially.

By accounting for the incommensurability of policy impacts in evaluation, LEM must be reorganized into the square input-output matrix on an intermediate level of partially aggregated assessment results. This approach upgrades multi-criteria impact assessment with synthesis. The square matrix has very desirable properties for evaluation of complex social issues because it explicates the synergies and tensions between evaluation domains, as well as from directly observed effectiveness of policy measures only in primarily targeted areas of impact. The matrix presents a complex social system decomposed on its main sub-systems that can be evaluated partially as autonomous, in their principal targeted domains, and partially in interaction (Heylighen, 1989) or overlap, which results from cross-sectoral or secondary policy impacts.

The square matrix exists above the micro-level (Leopold matrix) because it is aggregated from it. At the same time, as a set of only partial aggregates, it exists below the macro level. The matrix therefore comprises an intermediate or meso level of policy impact evaluation. This intermediate evaluation method applies the meso-matrix, which is a relevant concept since the complex social

issues themselves are built upon meso units (Dopfer et al., 2004, Easterling, Kok, 2002) – in our case, evaluation domains. The meso-matrix enables the simultaneous evaluation of constitutive but divisive aspects of a given social matter together with marginal but integrative ones. This is sufficient to overcome the aggregation problem in evaluation. What is even more important, however, is that the meso level approach places synthesis at the centre of a new methodology of policy impact evaluation. Mesoscopic synthesis enhances evaluative and not only scientific insights into social reality. This suggests that in researching complex social issues, synthesis needs to be approached as an essentially evaluative undertaking.

Extreme Positions

First-generation methodologies relied on a disaggregated approach to matrical evaluation of large-scale policy interventions, with dispersed impacts assessed across multiple evaluation domains. Luna Leopold and colleagues (1971) developed their approach encompassing two incommensurable evaluation domains: economic and environmental. They accomplished evaluation only at the micro level, however, by assessing individual economic measures' impacts on a wide range of detailed environmental assessment criteria. The approach is nonetheless significant, because it goes beyond monitoring of implementation and assessment of internal criteria of sector-based effectiveness such as how a particular policy, such as an economic policy, can affect selected economic assessment criteria, such as value added growth, competitiveness or employment. Leopold et al. placed assessment of policy side effects or secondary impacts in the centre of evaluation concerns. This has been a substantial achievement. It goes beyond the scope of many prominent evaluation practices up to the present day (see Hageboeck et al., 2013; Huitema et al., 2011). Unfortunately, however, Leopold et al. decisively refused to capitalise on their groundbreaking achievement.

In the finest positivist tradition of analytical evaluation, Leopold et al. aim to assess economic policy performance through a pedantic description of its numerous individual environmental impacts. Their matrix presents 100 economic policy measures in rows and 88 evaluation criteria of environmental impact in columns, creating a matrix with 8,800 cells – each further divided into four sub-sections that describe every impact by its size (large/medium/small), direction (positive/negative/neutral), probability (high/low) and the assessment of risk involved in each individual impact (critical or not). In this way, policy impacts are assessed in sufficient detail to enable maximally informed policy decisions.

Appropriately recognizing the incommensurability between economic and environmental evaluation domains, they rejected the summation of detailed assessments into an aggregate impact indicator on how much all economic measures affect all evaluation criteria. For them, the task of the evaluator is to inform and comment, but not to synthesise or generalise. Refusal of aggregation is essential for neutral evaluation, argues Leopold, as it draws a demarcation line between evaluator and policy-maker to protect the former from value judgments and political interference (Kunseler, 2007). Since Leopold, evaluators have usually resisted calls to generalise, simplify, and formulate problem specifications or focus on holistic solutions to definite complex problems. This argument has been accepted as an evaluation standard for decades, and is followed in examples such as the EU's Impact Assessment Guidelines (SEC(2005)791), one of most authoritative reference documents for evaluators in the region.

Rejection of summation in evaluation and shifting this task to policy-makers, the media, and to politicians is sometimes dubious (Stiglitz et al., 2009). Detailed evaluation results make political decisions more informed but not necessarily easier (Diamond, 2005). Leaving it to users to make the synthesis supposes that users are equipped to do this consistently, which can be difficult to justify. According to Scriven (1994), it is 'letting the client down at exactly the moment they

need you most'. It is after all precisely the failure of politicians and scientists as social aggregators, recall Arrow's impossibility theorem (1951), which demanded the introduction of evaluation as indispensable constituent of the policy cycle in the first place.

Sanderson warns that cross-sectional evaluation, which seeks to isolate policy instruments and assess them separately from each other, will produce results of limited usefulness due to limited external validity. This problem is especially evident in the assessment of complex 'cross-cutting' or horizontal social concerns (Sanderson, 2000) such as sustainable development. Assessment that simply produces non-overlapping information for economic, social, human and environmental domains tends to underplay inherent system contradictions, legitimizing a disregard of legitimate stakeholders' concerns in the public realm (Stake, 2001). Without any explanation of how different domains of a policy or programme work together and interact, it is impossible to substantiate evaluation findings. Fragmented assessment results especially fail to satisfy information needs at the strategic level by producing banal answers to multi-dimensional questions (Virtanen, Uusikylä, 2004). This leaves interpretation of evaluation results fully exposed to political manipulation or at least to continued voluntarism that characterised policy-making before the introduction of policy impact evaluation.

Our case study illustrates the problem quite plainly. Table 1 presents detailed assessment results for RP. These resulsts were obtained in deliberative way, using experts' assessments of the possible positive, neutral or negative impacts of the 47 proposed RP measures on four sets of assessment criteria: social, economic, environmental, and human capital, each represented by a smaller number of priority criteria.

In an ideal case, assessment of individual impacts would be obtained analytically (such as in Radej et al., 2015). When this is not feasible, usually because of lack of data or time, ad hoc expert opinions may fill the gap. It is important to indicate that experts' assessments are funded on best available knowledge in a given context and that their expertise is sufficiently diverse to cover all evaluation domains indiscriminately.

Experts' opinions about the direction of the RP's impacts were considerably divergent in some instances. If reconciliation between disagreements could not be achieved, their assessments were averaged and in this way neutralized. The threat of a cancelling-out of incompatible assessments in itself encourages a cooperative attitude in a team of experts. Contrary to this solution, some evaluation approaches, such as CAF (Common Assessment Framework, 2006) suggest that disagreements among experts' assessments need to be discussed with the aim of reaching consensus. However, forcing consensus for every single assessment detail is risky because it may invoke existing asymmetries within the team, which can lead to a kind of closed, exclusive process (Connelly, Richardson, 2004) in which the dominant agent will prevail. Hence, if results of detailed assessments are conflicting but well founded, disagreement is irresolvable. This does not mean they cannot be aggregated, however. Expert assessments are not only equally well founded and legitimate, but are also only specific and partial claims about a given partial issue. Nevertheless, remaining disagreements, especially if systematic across a spectrum of connected issues, should be dealt with further in evaluation, separately from aggregation of assessment results.

The assessed detailed impacts are presented in Table 1. These would normally be summarised following three lines of descriptive interpretation. A prevalence of the programme's positive impacts would suggest that a majority of the RP's measures favourably contribute to regional sustainable development, which evidently supports endorsement of the programme. Negative impacts would focus policy-makers' attention on the weakest parts of the proposed programme, aspects that should be improved or at least neutralized in future. Absent impacts (0) are usually not taken as problematic for policy-makers and therefore may be skipped in RP's amendments.

Following standard lines of reasoning would suggest that the evaluators, and consequently policy-makers, should focus their attention on those areas of the programme that induce negative impacts. In these cases, changes to the programme are proposed, either by abandoning individual policy measures with the most damaging cross-sector impacts, or by providing corrective actions and compensation in order to offset voluntary victims.

Table 1: RP's impacts at micro level (Leopold's matrix)

Domains & Criteria of Evaluation / Programme Measures	Economic		Social		Environmental		Human	
	Income growth	Invest. Intensity	Unemp-loyment	Migration	Abatement Expenditure	Sewerage Connections	Students per capita	Ageing
1 Development lags	+	+	0	-	+	+	+	0
2 Competitiveness	+	+	-	-	-	+	+	0
3 Investment promotion	+	+	+	+	0	+	0	0
4 Endogenous advantages	+	+	+	+	+	+	+	+
5 Entrepreneurship	+	+	-	+	0	0	+	0
6 Regional tourist organization model	0	0	0	0	0	0	0	0
7 Pomurje as a tourist destination	+	0	0	+	0	0	+	0
8 Destination management	0	0	0	0	+	+	+	0
9 Destination marketing	+	0	0	+	0	0	0	0
10 Human resources in tourism	0	+	+	+	0	0	+	0
11 Quality management	0	+	0	0	+	+	+	0
12 Tourist infrastructure investment	+	+	0	0	+	+	0	0
13 R&D in tourism	+	+	0	0	0	0	+	0
14 Health inequality (criteria)	0	0	0	0	0	0	0	0
15 Health promotion network	+	+	+	+	0	0	+	0
16 Health inequality – regional	+	0	+	+	0	0	+	+
17 Health inequality–vulnerable groups	+	0	+	+	0	0	+	+
18 Quality, access to health	+	0	0	0	0	0	+	+
19 Healthy environment	0	+	0	0	+	+	0	0
20 Mental health	0	0	+	+	0	0	0	+
21 Agriculture modernisation	+	+	-	-	+	+	+	0
22 Environmental agriculture	+	+	+	+	+	+	+	0
23 Entrepreneurship in agriculture	+	0	+	0	0	0	+	0
24 Human development in agriculture	0	0	+	+	0	0	+	0
25 Value added growth	+	+	-	-	+	+	0	0
26 Products, services – farms	+	+	+	+	+	+	+	+
27 Products, services-agro industry	+	+	-	0	+	+	+	0
28 Marketing agro-products	+	+	0	0	0	0	0	0
29 Rural developm., products, services	+	+	+	+	0	0	+	+
30 Countryside development	+	+	+	+	+	+	+	+
31 Rural entrepreneurship	+	+	+	0	0	0	+	0
32 Rural stakeholders' co-operation	+	0	+	+	0	0	+	0
33 Water supply	+	+	+	0	+	+	0	0
34 Transport infrastructure	+	+	+	+	+	0	0	+
35 Alternative, local energy	+	+	+	+	+	0	+	0
36 Energy distribution network	+	+	0	0	0	0	0	0
37 Access to IT services	+	+	+	+	0	0	+	+
38 Waste waters treatment	+	+	+	0	+	+	0	0
39 Solid waste management	+	+	+	0	+	+	0	0
40 Communally equipped zones	+	+	+	+	+	+	+	0
41 Water quality	+	+	+	0	+	+	0	+
42 Revitalisation of hot-spots	-	-	0	0	+	+	0	0
43 Illegal land-filling, monitoring	+	+	+	0	+	+	0	0
44 Nature and culture conservation	-	+	-	0	+	0	+	0
45 Energy policy	+	-	+	0	+	0	+	0
46 Spatial planning	+	+	+	+	0	0	+	0
47 Communication strategies	+	+	0	0	0	0	+	0

Source: Radej, 2006. Scores: '+' positive impact, '0' absent or neutral impacts, '-'negative impact.

However, by concentrating attention in evaluation only on incidences of conflict between specific aspects, we lose sight of the programme as a whole, as well as of regional sustainability as an integral concept. Furthermore, absent impacts may not be irrelevant to evaluation since they may indicate an absence of cooperation and an overlap between elements of regional sustainability, in this way ignoring its synergetic essence. Neither may indicate the prevailing positive impact acceptable as authoritative evidence of satisfactory policy achievements. The policy proposals are routinely prepared by specialised and generally competent authorities. Policy proposals are furthermore carefully scrutinised, first by assisting external experts, and then painfully negotiated among various group interests before they are submitted for formal evaluation. This largely secures that policy-makers can prepare credible initiatives that guarantee a wide array of positive contributions to the resolution of given social problems. However, credibility should not be mixed with policy-makers' effectiveness in the provision of public goods for all legitimately concerned parties indiscriminately. Prevailing positive impacts do not guarantee society-wide effects if policy goals or assessment criteria are in conflict – a very common situation in policy-making. Moreover, policy impacts are sometimes assessed against criteria selected by a formally responsible implementation authority alone, not by end beneficiaries, and even less by a general public, which additionally calls into question the neutrality of assessed positive impacts. Even when this is not the case, positive impacts are assessed in relation to narrow evaluation criteria and thus in isolation from each other with no capacity to measure integral impact.

A prevalence of positive impacts in a Leopold matrix can inform policy-makers only about their effectiveness observed at atomistic (micro) level. Only when systematic evidence of positive impacts is obtained across all integral domains of evaluation can an evaluator assess the appropriateness of the regional programme as a whole. However, systematic evidence can be identified only when detailed assessment results are properly 'post-produced' with consistently designed algorithm of synthesis.

There are therefore several and dissimilar barriers to aggregating positive and negative impacts of a given policy measure on various assessment criteria or, analogously, impacts of different policy measures assessed against the same evaluation criteria. In aggregation, the evaluator has to take a position on the fundamental issue of compensability, i.e. the possibility of offsetting a negative impact of a given policy measure on one criterion by a positive impact on another (Munda, 2012). For instance, in Table 1, might negative impacts of entrepreneurship promotion on employment in Pomurje be outweighed by the positive impacts of entrepreneurship promotion on migration? Or, is it permitted in evaluation to cancel out additional tons of greenhouse emissions (a negative impact) with additional purchases of tradable pollution permits (a positive impact, because their proceedings pay for additional environmental protection measures at the permit seller's plant)? Greenhouse emissions trigger irreversible deterioration in the climate conditions, therefore the economic and climate domains of evaluation are not interchangeable but incommensurable. Thus, a trade-off between greenhouse gases and money is not adequate as a general principle in policy-making. However, climate change is not merely macroscopic, but is a complex phenomenon that must be observed at multiple levels with different evaluation principles in mind.

Trade-offs between different aspects of the public good may not be incommensurable locally, in every single case in which different valuations are confronted. Conflicts between valuation domains arise only when some thresholds are breached, but not within these limits where in fact a large majority of social interactions take place. To incorporate this peculiarity into the evaluation of policy impacts, system thresholds have been introduced – such as economic, ecological, social and human rights legislation and standards (for a survey of literature see

Muradian, 2001). This threshold marks a tipping point, beyond which a small quantitative change in one part of a system might have a disproportionately large effect on the entire system, involving high risk to its overall integrity.

The concept of system thresholds is closely linked to concept of 'social incommensurability' (Munda, 2004). System thresholds normatively define objective criteria against which trade-offs between domains are conditionally commensurable, depending exclusively on the specific considerations of those directly affected by trade-offs. Within the threshold limits, an agent either does not sense the qualitative difference between two distinct social conditions, or else refuses to declare a preference for one or the other (Luce in Munda, 2006), as in the case of minor environmental damage. Beyond these limits, any further trade-offs are ruled out, even if the victims' consent can be preserved, because this would negate building principles of co-existence in a given society.

Recognising system thresholds in policy impact evaluation is important, because it introduces a network of barriers to simple (micro to macro) aggregation that establish incompatible paths of only partial aggregation: for each domain separately according to its specific sector-based threshold constraints, and thus separately for each cross-sector consideration. Now Leopold's situation where detailed impacts are treated as strongly incommensurable can be abandoned and replaced by the claim that policy initiatives will conform to system thresholds, and are therefore, at least in principle, weakly commensurable and thus partly aggregatable. This suggests that detailed impacts from Leopold's matrix (Table 1) can be aggregated in two constrained ways: individual impacts for all evaluation criteria can be summarised only within each evaluation domain, and all policy measures' impacts are aggregated only within each main policy sector.

This newly obtained matrix of impacts was previously named LEM (Table 2). For simplicity, the original Four-capitals model by Ekins and Medhurst is provisionally reduced in LEM to only three capitals or evaluative domains because this is sufficient to discuss the aggregation problem in the evaluation of complex social issues. Later this simplification is removed, in order to study the model in its original four-leg design, because it outlines some important messages that demand further attention.

The number of rows in LEM is reduced from 47 programme measures to only six main sector-based policies. LEM presents impacts on a wider range of scores compared to Table 1, in order to increase its interpretative potential on an aggregate level: from the most robust positive impact with the highest score ('+++') to the most negative impact with the lowest score ('---'), with all five intermediate possibilities, including absent or negligible impact (0). When uncertain about how to round-off aggregate impact from Table 1 to Table 2, a decision was made based on comparison of the financial weight of the related measures (Radej, 2006).

Evaluation results from Table 1 can be now reinterpreted in aggregated form. Infrastructure development, as one of the six sectors involved in RP, will produce more positive effects than any other sector and therefore is assessed as the most welfare-enhancing, followed by rural development policy. The most problematic is the negative impact of the manufacturing sector on social capital, because it aims to increase cost efficiency in companies in a neo-liberal way, by lowering wages, which can result in higher unemployment and increased outward migration. The impacts of the health sector and tourism are largely absent – the reason being that the proposed measures generally do not relate to investments or to the provision of new services, but to the preparation of plans and regional organization structures that show no immediate or direct impact on regional sustainability.

The direction of the RP impacts on domains of regional sustainability, as presented in aggregate in the bottom row of Table 2, do not appear problematic. The programme will improve regional sustainability in all three integral domains, though a more positive impact is expected in the

economic (++, two plusses out of three maximally possible) than in social and environmental domains of sustainability (+). The summary impacts of the RP are feeble and unevenly positive, yet the differences are relatively small and do not defy regional strategy for sustainable development. Thus, this assessment concludes that the programme will have a rather weak but systematically positive effect on regional sustainability, so that RP could be recommended for adoption under the condition that considerable improvements are achieved in areas with negative impacts, which could strengthen the positive impact of RP on social capital that is regionally most vulnerable. Taken as a whole, RP is an acceptable development instrument for implementation in the Pomurje region.

Table 2: Macro level impact of the RP on domains of sustainability (LEM)

Evaluation Domains Policy Sectors	Economic (E)	Social (S)	Environment (N)
1 Manufacturing (rows 1-5)*	+++	–	+
2 Tourism (rows 6-13)	+	0	0
3 Health (rows 14-20)	0	+	0
4 Rural development (rows 21-32)	+++	+	++
5 Infrastructure (rows 33-40)	+++	++	++
6 Environment (rows 41-47)	+	+	+
Summary Impact	++	+	+

Source of data: Table 1. Note: * Summary of rows 1-5 in Table 1, etc. Scores: '+++' Very positive, '++' Moderately positive, '+' Weakly positive, '0' Absent or Neutral, '-' Weakly negative, '--' Moderately negative, '---' Very negative impact.

Macro evaluation of RP importantly upgrades the interpretative potential of evaluation compared to Table 1 by identifying the direction of impact on a strategic level of the programme in regional sustainability as a whole, summarized in Table 2. However, a very high level of aggregation of three compound indicators of RP impact does not allow for sufficiently structured evaluation of relations between three incommensurable capitals, neither in terms of conflicts and synergies between them nor in terms of how they horizontally overlap. These crucial questions about policy synergy are indispensable in evaluation of sustainable development, and for any other complex social concern. However, macro evaluation presents obtained results for each domain separately, as if they were independent constituents of regional sustainability.

Standard methodologies have proven to be either too individualistic or too summative for evaluation of complex social concerns. If detailed assessment results are not aggregated, as in Leopold's approach, the evaluation produces findings that are too fragmented for holistic comprehension of complex social issues. In contrast, full aggregation results are already too unified for structured judgment at a strategic level of considerations. In both cases, evaluation is neither effective nor efficient, since extensive analytical effort is invested in obtaining detailed impact assessments that remain poorly exploited in highlighting macro level integrity of policy interventions. A new approach is therefore needed for the summative organization of detailed evaluative findings, an approach that can negotiate a compromise between microscopic aversion and macroscopic passion for aggregative synthesis.

Resolution in the Middle

The origin of troubles in the aggregation procedure introduced by Ekins and Medhurst is too narrow an understanding of otherwise appropriately observed incommensurability between evaluation domains. Ekins and Medhurst did not acknowledge that the assessed impacts are vertically not fully aggregatable in LEM, by all RP measures' impacts on a given evaluation domain. The column aggregation in Table 2 assumes the homogeneity of policy impacts – a

highly dubious conjecture. Many studies indeed demonstrate that a given policy does not influence all relevant areas of impact in the same way (Schnellenbach, 2005). Policy impacts are by their nature either: (i) direct, needed when pursuing their stated goals vertically, or (ii) indirect side effects or trade-offs, operating horizontally and bringing up secondary meanings in evaluation because they fall under the jurisdiction of some other policy sector (Rotmans, 2006) with divergent goals and evaluation criteria. Discriminative direct and indirect sectoral impacts are confirmed even for those policies that had previously been taken as most neutral, for monetary reasons (Lucas, 1972) and for tax policy (Leith, Thadden, 2006).

When policy measures are focused and sector-based, they should always be analysed not only in relation to specific goals, but also, and more importantly, in relation to the general interest they are meant to serve (Donzelot, 1991 in Burchell et al.). Sectoral specialisation and the non-holistic nature of policy-making means that policy evaluation should take secondary issues as indispensable. Yet secondary impacts routinely fade out in evaluation because it is assumed that 'they are too complex' (Morçöl, 2011) and impossible to track. The recent meta-evaluation of 340 USAID evaluation studies found that only 15% reported on unplanned effects, and only 10% discussed secondary causes that might be contributing to results (Hageboeck et al., 2013). Huitema arrived at a similar conclusion, finding that some 60-80% of 260 evaluation studies prepared for the needs of EU climate policy either avoid or attempt to diminish the complexity of the evaluated objects (Huitema et al., 2011), by focusing specifically on directly observable issues while side-lining indirect links.

Systematic disregard for extensive secondary impacts in policy-making might explain why good, sector-based, vertically designed policies, founded on strong values and even on common sense, often lead to disappointing overall results (Chapman, 2004). In response to these negative trends, a new generation of evaluation approaches (Guba, Lincoln, 1989) or new wave of evaluation studies (Vedung, 2010) has emerged, emphasizing the complexity of evaluative challenges. In particular, these new approaches outline the importance of secondary impacts in policy evaluation by highlighting 'horizontal themes', such as gender equality or sustainability of development, that span across all sector-based policy concerns.

Accounting for secondary impacts in evaluation calls for a specific aggregation procedure in two regards: firstly, summarizing secondary impacts separately from direct ones; and secondly, aggregating different types of secondary impacts independently. For example, if we choose to differentiate three domains of evaluation – economic, social and environmental – then economic and social policy's impacts on environmental criteria will not be not fully commensurable and therefore must be aggregated separately; for example, economic impacts on the natural environment must be aggregated separately from social impacts on natural environment.

The partial aggregation rule initially seems at odds with the strong version of the incommensurability thesis. Some authors even argue against a generalizing concept of strong incommensurability (Morgan, 2007; Nola, Sankey, 2000; Thagard, Zhu, 2002). They propose instead making a distinction between relations of strong and weak incommensurability. Martinez-Alier et al. (1998) point out that in situations when there is an irreducible value conflict in public affairs, we can only search for weak comparability as a facilitator of collective discourse.

Policy impact will be conceptualised here as 'weakly commensurable' when specific limitations are imposed in aggregation so that it can be accomplished only partially, within borders set by evaluation domains that order a network of system thresholds into a matrix of value-based oppositions that are involved in evaluation of socially complex issues. Furthermore, impacts that are weakly commensurable in two or more incommensurable domains of the evaluation, such as for hybrid socio-economic impacts, in addition to purely social and purely economic impacts, are conceptualized as 'weakly incommensurable'.

The difference between strong and weak incommensurability can be illustrated by the following "juicy" example that shows how it is possible to enforce integrity by forwarding co-existence between partial compatibility and stringent incompatibility. It is entirely possible to mix apples and oranges because their juice, as their secondary content, is only weakly incommensurable and can be tastefully combined under certain threshold conditions that can be provided, for instance, by a bartender's recipe.[4] Despite the obvious possibility of a perfectly harmonious mix of their secondary contents, however, it will still never be possible to grow an apple tree from the seed of an orange, because the seeds of apples and oranges are strongly incommensurable in their primary essence. Strong incommensurability therefore relates only to core concerns in evaluation (apple and orange seeds), whereas weak incommensurability relates only to marginally overlapping concerns (apple and orange juices).

Weakly incommensurable impacts are evaluated against two otherwise incompatible sets of criteria. These impacts have a hybrid character, which allows for a delicate possibility of translation between domains; they are therefore the main drivers of evaluative synthesis. By identifying weak incommensurability between many policy impacts, aggregation can now be accomplished in a consistent way for any socially complex consideration.

An evaluator can easily capitalise on the distinction between strong and weak in/commensurability in order to enhance synthesis. To account for weak commensurability of many assessed impacts, one needs to regroup all rows of policy measures in Tables 1 and 2 in the same way that impact areas are grouped in columns – by three incommensurable domains. This divides the Leopold matrix (and LEM) into three sections vertically and three sections horizontally, resulting in nine sub-sections. A partial aggregate in each sub-section is obtained by summarising detailed impacts by a source (row) and area of impact (column). These sub-aggregates are organised as fields of a square input-output matrix (Table 3).

A square matrical evaluation enhances two separate but integral views of how a regional programme can impact selected criteria of sustainability. The direct view, presented on the diagonal of the matrix, is accomplished by assessing the policy relative to achieving its primary goals; the indirect view, located on the non-diagonal fields, describes trade-offs as cross-sectional overlaps between policy domains (inputs, in rows) and evaluation domains (outputs, in columns). An overlap is denoted with the sign '∩' (from set theory). For example, the economic policy impact on the social domain is denoted as E∩S (E overlaps S), while the social policy impact on the economic domain is denoted as S∩E. Obviously, in policy impact evaluation E∩S is not identical to S∩E, as they would be mathematically. In set theory, an overlap consists of shared characteristics (membership) between the simultaneous elements of two sets, equally belonging to both sets. In a complex evaluation such as the ones presented here, one needs to account for the incommensurable relation between S and E.

A mesoscopic perspective of RP's impacts on three domains of regional sustainability is shown in Table 3.

Table 3: The meso level impacts of RP (input-output matrix)

Evaluation domains / Policy domains	E	S	N
E*	E∩E (+++)	E∩S (–)	E∩N (+)
S**	S∩E (+++)	S∩S (++)	S∩N (++)
N***	N∩E (+)	N∩S (+)	N∩N (+)

Source of data: Table 2.

Notes:

4 See, for example : http://www.welchs.com/products/essentials-juice-cocktails/essentials-orange-pineapple-apple-juice-cocktail or http://www.absolutdrinks.com/en/drinks/virgin-sangria/, IX 2015.

(1) Aggregation of scores: * Aggregated from rows 1 and 2 from Table 2; ** Rows 3 and 4 from Table 2; *** Rows 5 and 6 from Table 2.

(2) Range of scores: '+++' Very positive impact, '++' Moderately positive impact, '+' Weakly positive impact, '0' Absent or Neutral impact, '-'Weakly negative impact, '--' Moderately negative impact, '---' Very negative impact.

The diagonal elements in Table 3, read from top left to bottom right, present the intended impacts of RP on targeted areas of regional development (E∩E, S∩S, and N∩N). These exhibit a vertical perspective of RP's effectiveness at the meso level of evaluation. The diagonal elements ss strongly incommensurable cannot be aggregated any further and must be interpreted as they are. They suggest that implementation of RP would be very successful in promoting economic capital (maximum, three pluses, instead of two in Table 2), moderately successful in enhancing social capital (two pluses, instead of one in Table 2) and only weakly effective in achieving primary environmental goals (one plus in both Tables). Table 3 suggests that three domains of sustainability are handled in RP in an unbalanced way. This observation does not match with the one previously obtained from LEM. Note however, that the summary row elements in Table 2 and the diagonal elements in Table 3 do not represent the same content, despite being treated in both cases in the same way, as the program's overall impacts on three main pillars of sustainability. The former is a compound indicator of the total impacts (direct and indirect) of all policy measures on a given evaluation domain, while the latter more appropriately presents only the aggregate of direct intra-domain impacts.

The secondary impacts, or RP's side effects (E∩N etc.), are located below and above the diagonal in Table 3. They disclose the RP's crosscutting impacts that must be evaluated horizontally. Each cross-sectional relation in matrix, such as E∩N, has its inversely symmetric counterpart, N∩E, and when they are correlated, they establish an overlap between two domains, providing insight into RP's synergies.

The three domains of evaluation in this case study produce three bilateral overlaps. They are exhibited in Table 4: SE, for the 'socio-economic' overlap between S∩E and E∩S; NS, for the overlap between the Natural Environment and Social sustainability; and NE, for the overlap between Natural Environment and Social domains.

As far as RP's indirect impacts are concerned, the overlap in SE is strong but damaging for the social domain, which indicates a socially obstructive impact of economic measures (Table 4). The moderate overlap in NS is linked to a mismatch between nature conservation measures and the urgent needs of social capital due to the lack of enhanced opportunities for the local population in landscape parks, by imposing conservative measures only normatively and institutionally in a largely passive manner. Here we recall the previous finding (shown in Table 3), that environment protection and nature conservation policy will not be very effective in pursuing their primary goals; thus the RP will, at least in relative terms, impose additional social constraints for rather very poor environmental improvements. Additionally, the economic domain integrates very poorly with S and with N; this can also be observed in the feeble secondary impacts of N on E and on S. RP's poor overall impact on S points to a continuation of unfavourable trends in the Pomurje region. It will moderately improve because of RP's direct impacts, but will systematically deteriorate further in relative terms. This finding points to a situation in which evaluative conclusions significantly diverge from the assessment of direct or indirect impacts.

Table 4 outlines some additional asocial features of RP that become visible only when indirect impacts are evaluated in their overlap. This aspect is absent from the assessment results presented in Tables 1 and 2 because their construction is too simple to be able to identify it. The aggregate conclusion from meso-level evaluation is that the RP is only inconsistently contributing to regional sustainability, and in fact it fails to promise a reversing of critical development trends in

Pomurje. This is again a more decisive and certainly a less optimistic evaluation of RP compared to the conclusions obtained from Tables 1, 2 and 3.

Table 4: The RP's correlation matrix of synergy between scopes

	E	S	N
E	EE, Economic Sustainability (+++): Very effective	SE, Socio-Economic Sustainability (+++ , -): Strong overlap but involving negative trade-offs in favour of E	NE, Environmental-Economic Sustainability (+, +): Weak overlap, balanced between N and E
S	-	SS, Social Sustainability (++): Moderately effective	NS, Environmental-Social Sustainability (++, +): Moderate overlap, unbalanced in favour of N
N	-	-	NN, Environmental Sustainability (+): Poor effectiveness

Source of data: Table 3.

Diverging or even incompatible evaluation conclusions derived from Tables 1, 2, 3 and 4 are, of course, not due to different detailed impact assessments at the micro level. They arise solely from the summative endeavour. The standard micro and macro methodologies are premised on strong commensurability or strong incommensurability of detailed impacts, while a complex methodology is established on the distinction between strong and weak incommensurability of evaluation domains on a meso level while admitting only a weak commensurability of detailed impacts on micro level.

This case study also demonstrates that different definitions of the relations between assessed impacts on micro, macro and meso levels require different aggregation algorithms, leading to diverging evaluation conclusions and policy recommendations. A policy's failure to implement more integrative interventions into regional development is therefore not necessarily a result of intentional policy-makers' biases, but can be a straightforward outcome of an inadequate understanding of policy challenges in the presence of social complexity. This validates the premise presented in the introduction that inconsistent policy intervention may be caused by methodologically inadequate handling of socially complex phenomena in policy impact evaluation.

Social processes unravel both on the micro and on the macro levels, but descriptions on the two levels cannot be directly linked and synthesized. Microscopic issues can only be approached in a fragmented way, and macroscopic issues only in highly compacted way. In both cases, a uniform and highly divisive way of thinking is furthered. The evaluation of complex social issues requires instead an accounting for all main domains of judgment simultaneously, both in their overlapping and non-overlapping contents. This can only be accomplished by the evaluative synthesis that begins from an intermediate or meso level (Dopfer et al., 2004).

Horizontal Extension

An intermediate goal of this case study is achieved by explaining the mesoscopic aggregation procedure in evaluation of complex social issue for the three-part version of Ekins's model. Thus we can return to the original four-part model in order to identify some further implications of horizontal synthesis. Picture 1 schematises both models in the form of a three- and a four-part Venn diagram.

A brief theoretical introduction will be helpful here in order to broaden understanding of the meso level as a hierarchical structure with many sub-levels. We rely on Dopfer, Potts and Foster (DFP; 2004), who developed a theoretical basis for the hierarchical organization of meso logic. DFP do not understand meso as a single level but divide it into three sub-levels that describe

three core processes in a trajectory of a given social system between two successive macro states, such as that of a given region before and after implementation of the RP. They distinguished the three-phase process of a system's trajectory as Meso 1, 2 and 3.

Meso 1 describes the micro to meso emergence of novelty by its early affirmation between initial followers based on a local affinity or similarity that connects them in relation to a novelty. In our case, Meso 1 is involved in the partial aggregation of detailed and weakly commensurable impacts into domains of input-output table. Meso 2 consists of a meso to macro (Dopfer et al., 2004) correlative process in which novelty is integrated or disseminated into a broader context through mixing and blending – these were represented in the case study by correlating impacts in three overlaps between the non-diagonal fields of the input-output matrix, SN, NE and ES. The novelty is finally integrated into the system and established as a new macro order in Meso 3. With the concept of Meso 3, DFP account for the meta-correlation between the emergent products of Meso 2.

DFP's classification of three meso sub-levels is an important achievement because it develops deeper insight into the emergent processes that take place in micro-meso-macro synthesis. Meso 1 forwards an analytical perspective in constructivist objectifications of social reality. Meso 2 is different in its character compared to Meso 1; it is evaluative, studying the overlaps between constructs of Meso 1 in the case study between pairs of sustainability domains. The character of Meso 3 is interpretative, aiming to identify a sense of triadic (or higher order) overlap between dual overlaps. The outcome of ordering on the Meso 2 level is circular due to the non-transitivity (see Arrow, 1951) of triangulated binary overlaps. Overall conclusions cannot be automatically evident from correlated assessment results, because, as only binary and rational expressions of complex phenomenon, they partly contradict each other as symmetrically opposite (such as E∩S vs S∩E, etc). To simplify the assessment of complex configurations, we usually prefer a binary (rational and relational) explanation of its constituents. However, the preference for simplicity has a price of throwing the researcher into a set of contradictions involved in overlapping contents. Interpretation is then needed in synthesising knowledge about complex social issues in order to deal with these contradictions. Interpretation is synthesising as much as it is able to form a coherent understanding of partly mismatching overlaps between incommensurable valuations imposed on scientific, objectivized, but only partly valid descriptions. The upshot from the cyclical construction of mesoscopic conclusions is that synthesis is not narrowly scientific, but is an evaluative and interpretative undertaking. Synthesis uses scientific results as its base, but goes beyond factual claims by explaining broader meaning of constituents in relation to their diversified value considerations.

Inspired by DFP, we aim to extend the hierarchy of meso levels one step further. As classical evolutionists of Schumpeterian school, DFP were not sufficiently equipped to accomplish this step. They produced a hierarchical theory of the meso level in the vertical direction of complexity only. The horizontal axis is not explicitly elaborated upon in their model and this curtailed their prospects to capitalize fully on their achievement.

Dopfer et al.'s triadic classification is illustrated in Picture 1a, while Picture 1b schematises the hierarchy of meso levels, as is implied in Ekins's (Munasinghe's) original Four-capital model by the methodology of complexity.

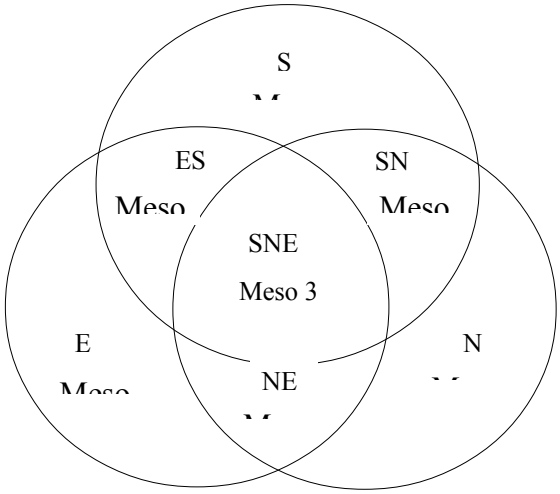
By adding the fourth domain (H, for Human capital) to the evaluation model, Meso 2 further decomposes into the sub-levels Meso 2a and Meso 2b. As a result, the mesoscopic procedure of synthesis is lengthened for one full cycle of correlations. RP's secondary impacts (Meso 2a) are now presented by six dual overlaps (ES, EN, HE, SN, HS, HN), where previously they had been only presented by three (Table 4). In addition, four triple overlaps are obtained (HSE, HNE, SNE, HSE; Meso 2b), where previously there was only one (SNE, or DFP's three-part Meso 3 in

Picture 1a). In the Four-capital model, Meso 3 is shifted upwards from triple to a quadruple overlap (HNES, Picture 1b). Multiplication of horizontal domains therefore extends mesoscopic reasoning 'into its own middle' (Prigogine, Stenger, 1982). By adding new domains, complex analysis becomes increasingly deeper, instead of increasingly detailed, as it is pursued in the standard positivist methodology.

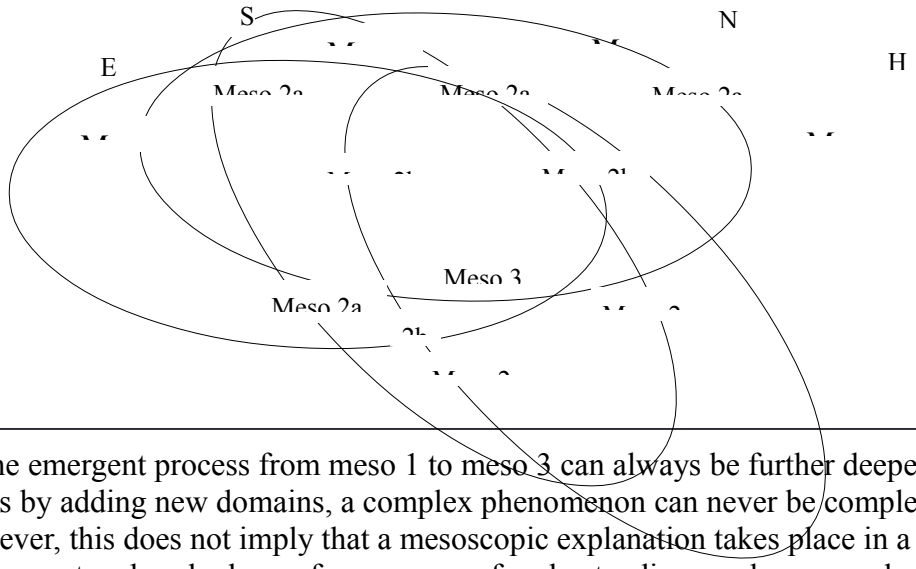
The possibility of a hierarchical mesoscopic emergence of meaning with cyclical evaluative interpretative synthesis suggests that the initial triadic micro–meso–macro explanatory approach would be more consistently framed as a meso–meso–meso approach. In the latter, meso is not seen merely as an intermediating phase in the transition between micro and macro, but arises as an overarching perspective for integral inquiry into complex social issues.

Picture 1: Venn diagram of sustainable development

Picture 1a: Venn diagram of three meso sub-levels of sustainable development



Picture 1b: Venn diagram of overlaps between four evaluation domains of sustainability (the Four-capital model)



As the emergent process from meso 1 to meso 3 can always be further deepened on meso 2 sub-levels by adding new domains, a complex phenomenon can never be completely explained. However, this does not imply that a mesoscopic explanation takes place in a black box, since each aspect and each phase of emergence of understanding can be comprehended from a unified mesoscopic perspective.

The mesoscopic algorithm of synthesis for socially complex situations in the case of four domains of sustainability can be described with a few procedural steps. Synthesis formally starts

with the definition of the quadruple overlap between domains in direct relation between the Meso 3 and the Meso 1 sub-levels:

$$HNES = H \cap N \cap E \cap S.$$

HNES can be more appropriately rewritten as the overlap between four triple overlaps of Meso 2b as it appears in the Venn diagram, Picture 1b:

$$HNES = HSE \cap HSN \cap HNE \cap SNE. \quad (1)$$

This is a practical reformulation because it translates a four-part correlation matrix into four sub-matrices of the third order, which can be solved the same as in Table 4 (non-diagonal fields of correlation matrix) on the Meso 2a level, with a set of dual overlaps:

$$HSN = H \cap S \cap N = (H \cap S) \cap (H \cap N) \cap (S \cap N), \text{ etc.}$$

If further simplified by replacing $H \cap S = HS$, etc. we obtain:

$$HSN = HS \cap HN \cap SN. \quad (2)$$

The formula of quadruple overlap (1) converts through (2) into the relationship between the bilateral overlaps, which is useful since they can now be easily analysed through the correlation between dual overlaps (Table 4):

$$HSNE = (HS \cap HN \cap SN) \cap (HS \cap HE \cap SE) \cap (HN \cap HE \cap NE) \cap (SN \cap SE \cap NE). \quad (3)$$

Due to the idempotency of overlapping sets, we can simplify, $HS \cap HS = HS$, etc., so that equation (3) can be rewritten shorter:

$$HSNE = (HS \cap NE) \cap (HE \cap SN) \cap (SE \cap HN).$$

The area of the Meso 3 is therefore obtained as three meta-overlaps between dual overlaps at the Meso 2b sub-level:

$$(HS \cap NE) \rightarrow \text{Meta-overlap } A,$$

$$(HE \cap SN) \rightarrow \text{Meta-overlap } B,$$

$$(SE \cap HN) \rightarrow \text{Meta-overlap } C.$$

Three meta-overlaps constitute three heterogeneous indicators of compound impact, which present in a circular way three panoramic views of RP's overall effects on regional sustainability. Even though they are distinctive and independent views, they closely reinforce one another. Circular evaluation of social issues is relevant because the nature of society is also circular. Circularity ensures that the interpretation of evaluation results remains to a certain extent contradictory and irrational, but because of this it is always open to deeper considerations. Such an algorithm is exactly consistent with the complex nature of social matters as evolving; they can never be fully revealed to any observer in their entirety, nor as fixed and final.

Each meta-overlap identifies one integral evaluation perspective and together they bring about a holistic understanding of complex social issues on Meso 3. Note that mesoscopic synthesis brings forward a radical perspective of wholes. In a socially complex setting, what counts as a whole is not counted as one, but as many, where a whole on a higher level emerges from a smaller number of partial wholes on an immediately lower level of inquiry (instead of in a conventional aggregative way, from many commensurable particles).

This case study has been helpful previously in illustrating conceptual achievements; now it appears to be time to call upon this case study once again. RP's numerous and diversified impacts from Table 1 are now aggregated into four evaluation domains, presented in Table 5, first in a square matrix (5a) and then in a matrix of overlaps (5b).

Table 5: RP's impacts on sustainability of Pomurje region – four-part presentation

Table 5a: Four-part input-output matrix of RP's impacts

	E	H	S	N
E	E∩E (+++)	E∩H (+)	E∩S (-)	E∩N (+)
H	H∩E (0)	H∩H (+)	H∩S (+)	H∩N (0)
S	S∩E (+++)	S∩H (+)	S∩S (++)	S∩N (++)
N	N∩E (+)	N∩H (+)	N∩S (+,)	N∩N (+)

Table 5b: Four-part correlation matrix of RP's overlapped impacts

	E	H	S	N
E	EE (+++): Very effective	HE (0, +): Very weak overlap, imbalanced in favour of H	SE (+++ , -): Strong overlap, imbalanced in favour of E	NE (+, +): Weak overlap, balanced
H	-	HH (+): Poor effectiveness	HS (+, +): Weak overlap, balanced	HN (0, +): Very weak overlap, balanced
S	-	-	SS (++): Moderately effective	SN (++, +): Moderate overlap, imbalanced in favour of N
N	-	-	-	NN (+): Poor effectiveness

Source of data: Table 1. Scores in Table 5a: '+++' Very positive, '++' Moderately positive, '+' Weakly positive, '0' Absent or Neutral, '-' Weakly negative, '--' Moderately negative, '---' Very negative impact.

The first meta-overlap, denoted by *A*, presents the interaction between material (N, E) and non-material aspects of regional sustainability (H, such as knowledge and creativity; and S, such as trust and cohesion). In the correlation matrix (Table 5b) it is evaluated as: (NE (+, +): Weak overlap; and HS (+, +): Weak overlap). RP's contribution to this specific aspect of regional sustainability is balanced at its lowest positive effects. The weak connection within a non-material aspect of sustainability amid RP's feeble positive effects on H and S suggests that RP cannot contribute significantly to diminishing the presently high vulnerability of the regional population. Another observation that is also striking is how poorly RP performs in the materialist aspect (NE), despite being clearly prioritised in RP. RP is materialistic only when prioritising the economic domain, leaving N relatively marginalized (Table 5a). In addition to this, the relationship between the two materialist aspects of sustainability is also only a weak relationship. This demonstrates that the materialist emphasis of RP is further narrowed only to focus on the direct economic effects and to the internal effectiveness of the economic policy domain.

The meta-correlate *B* describes the interaction between progressive factors of sustainability that are the most responsive to policy stimulus on the short run (H, E), and conservation factors (S, N) that are slow in response and mainly forward long-term goals. The overlap (HE ∩ SN) has been assessed as only weak in RP, while emphasizing conservation factors of regional sustainability. This is in line with the identified passive attitude of regional policy-makers to progressive initiatives in the field of sustainability of regional development. RP can be in this regard again seen as an instrument of inertia, instead of producing qualitative change in regional baseline conditions.

The third meta-overlap *C* describes the nature of the interactions between the produced factors of regional sustainability (S, E, which are created and processed) and non-produced factors (H and N, which are enhanced as they are; not socially created). RP induces medium strong overlap between (SE ∩ HN). It both positively and negatively affects produced factors significantly more than non-produced factors. This seems to indicate that RP forwards a specific theory of change. It is more process-oriented (input-output) and incremental rather than transformative (old-new), so that RP contribution to breakthroughs in regional development trends are also not very likely.

An interpretation of the four dimensional evaluation results on Meso 3 can be condensed from the contents that converge in all three meta-correlates from the point of view of sustainability of regional development: RP attributes privilege to economic goals in a primary as well as in a secondary aspect, but not consistently. Inconsistency is also identified in RP's conservative attitude: despite an aversion to external impositions, preconditions for endogenous development will be further diminished with RP by poor cohesiveness between regional factors of sustainability and by a too weak emphasis on non-produced capitals. RP acts as a regional instrument of defiance to progress and deep changes in particular. The Program could only contribute to regional sustainable development in non-ambitious, logically inappropriate and even contradictory ways. Regional authority therefore should reconsider not only a considerable number of measures in the context of their absent or negative side-effects (discovered on more detailed levels of assessment), but also main premises should be re-evaluated, such as the internal logical consistency of program's theory of change.

Horizontal extension allows for a more in-depth interpretation of evaluation results. Hypothetically, if a fifth horizontal element, for instance culture, were added to the model of sustainable development, its meso structure and its interpretative potential would deepen further. However, extension beyond four domains may not necessarily provide a better understanding of regional sustainable development, due to the rapid complication that accompanies multiplication of overlaps. How would one, for instance, sensibly explain the overlap of nature, human and culture, versus the overlap of culture, society and the economy? It is not that these overlaps do not exist, but rather that they might not be structured in policy-making and in our understanding in similar complicated ways, so these overlaps might operate only as artificial constructions of the evaluation model. For that reason, such constructed overlaps are of questionable worth. Instead of adding more and more horizontal evaluation domains of sustainability, evaluators and programme designers would do much better to form new triadic or quadruple structures of complexity on higher levels of generality, such as the overlap of culture and sustainable development, etc.

This case study suggests that mesoscopic exploration of complex social issues needs to remain constrained to only a moderate number of co-existing deep horizontal (and vertical) oppositions. This usually will suffice for resolving the aggregation problem in the evaluation of complex social issues. The mesoscopic methodology is not meant to cover all at once every important substantive matter for understanding a complex social issue. It is not holistic due to being all-encompassing, but rather due to its ability to indiscriminately account for all main multi-dimensional oppositions (at a meso level) that constitute the evaluation of complex social issues for a specific collective context of considerations.

Evaluative basis of synthesis

Complexity is challenging standard methodologies of policy evaluation and of social research in general by forcing them to enter a new 'Age of Synthesis' (Belloc in Wilhelmsen, Bret, 1970) as a precondition for successfully coping with social complexity. If we sidestep the problem of social complexity, aggregation does not make much sense in the social sciences and vice versa (Kirman et al., 2009). Ritzer and Smart (2003) outlined that algorithm of synthesis is one of the central issues in the methodology of social research, because it is about standards of consistent understanding of social reality that are always constituted on culturally specific formulae of intellectual coherence (Prigogine, Stengers, 1979). Geels (2002, 2007) emphasised that aggregation is not simply a cognitive process of drawing general lessons from local projects; one can also show how power, negotiation, and framing play a role.

Every epoch of social history forwards a different logic of synthesis, and these different logics are always an integral part of the ruling political arithmetic (Ritzer, Smart, 2003). 'The type of mathematics found in any major Culture is a clue, or key, to the distinctive character of the Culture taken as a whole' (Wilder).[5] Bentham synthesised social facts with the aim of maximisation that is induced by a dominant value system. Arrow broadened the scope of synthesis by inquiring into the formal conditions for synthesis that are not only rational in dominant values but also democratic from the aspect of diversity of values involved in social choice.

Dominant reductionist models of commensurable aggregation are highly restrictive, with obvious asocial consequences, because they rest on an extensive exclusion of all atypical inputs to synthesis that are not directly compatible with arbitrarily selected common denominators. Standard algorithms of synthesis are thus not cohesive. For a methodology of social research, this deductive and divisive character of the standard aggregative procedure poses a serious problem and invokes questions about the appropriate rational of the standard aggregation algorithm, which is after all meant to be a synthesising tool. For gaining an ability to understand complex social issues integrally, social research first needs to develop and implement a more connective approach to synthesis. We have therefore synthesised our object of concern from the meso aspect of circular synergies, with which synthesis is cohesive because it is thought of as a process – it is never completed, but it nevertheless enhances its own preconditions for more holistic inquiry.

The aggregation problem has arisen from ignoring the horizontal aspect of synthesis that is the main reason for the aggregator's inability to identify weakly commensurable and weakly incommensurable impacts. Even though synthesis is an eminently vertical, micro to macro procedure, it cannot be accomplished directly without denying the essential diversity of translating mechanisms operating in the middle world. Vertical synthesis that ignores horizontal differentiation is trivial. Already in the Fifth century CE, the Macedonian Stobaeus said that 'things that were alike and of the same kind had no need of harmony, but those that were unlike and not of the same kind and of unequal order'.[6]

The horizontal axis of synthesis, as the main novelty introduced by mesoscopic methodology, is proposed as an indispensable precondition for providing a connective and creative character of vertical aggregation in social research. Complex synthesis is not accomplished any more from extreme positions (micro, macro), but it takes place from the middle, where the potential of synthesis is the largest. This case study has clarified that complex social issues can be comprehended very differently when evaluated from a commensurable, incommensurable or from a weakly in/commensurable viewpoint. All three are relevant, and mesoscopic methodology organizes them according to their synthesising potential, which is the slimmest when aggregation rests on the assumption of commensurable valuation of social facts.

This experiment in mesoscopic evaluation demonstrates that social incommensurability is not an irresolvable obstacle to more holistic reasoning about collective concerns. In fact just the opposite is true! The imperative of incommensurability is established as a safety mechanism, reminding the evaluator that social issues are complex and cannot be holistically explained from any specific point of view, even a point of view that is seen as vital and resting on universal justification. However, strong and principal distinctions are crucial in evaluations for only a small number of the principal concerns. Even though contemporary societies are built on incommensurable oppositions, which of course cause strong social fractures, the majority of issues important for the understanding and cohering of everyday social life are not principal but

---

5 Wilder R.L. The Cultural Basis Of Mathematics. http://www-history.mcs.st-andrews.ac.uk/Extras/Cultural_Basis_I.html; II 2011

6 Wikipedia, #Stobaeus, I 2011.

are weakly in/commensurable. This distinction is crucial, because it is precisely sufficient for the type of soft synthesis that is needed when an object of evaluation is defined as complex.

The lesson to be learned here is that any complex social matter must be evaluated between an explanation of its primary contents – contents that are of course constitutive for it, but in a deeply dividing way – and an explanation of its secondary contents that enable synthesis only in matters that are not of primary importance to anybody (Radej et al., 2012). What is important to note, because it is usually ignored, is that emphasis on primary (or secondary) impacts is for the evaluator of complex social issues never forwarded in isolation; rather, it consistently appears in relation to secondary (or primary) concerns. To remain neutral, an evaluation must consider equally a large number of local synergies and a small number of deep and irresolvable oppositions. For Foucault, neither difference nor unity can be seen as primary, but need to be kept in balance (in Olssen, 2002) so that they can be dealt with on the same meso plane (Althusser in Levačić, 2009)[7] of inter-paradigmatic standards (Kordig, 1973).

We conclude that a precondition for a neutral evaluation of a complex policy intervention is not only an objective analysis of policy impacts, but also a consistent evaluative synthesis of findings that can only be holistically interpreted when addressed from the middle level of evaluation. Neutrality is not characteristic of fragments of knowledge, but we nevertheless seek to achieve it in holistic understanding (Kuhn, 1970). Synthesis must therefore be seen by a researcher of complex social matters as a creative act of inventing and operationalizing the middle-world from which connective meanings can emerge between radically diversified contributions.

7 Levačić M. 2009. Materialna eksistenca ideologije. Air Beletrina, http://www.airbeletrina.si/index.php?option=com_content&task=view&id=1713&Itemid=82

Bibliography

Åberg Y. Individual Social Action and Macro Level Dynamics: A Formal Theoretical Model. Sage, Acta Sociologica, 43/3(2000):193-205.

Allport F. H. "Group" and "Institution" as Concepts in a Natural Science of Social Phenomena. Publications of the American Sociological Society, 23(1928):83-99.

Armitage D. 2014. How history forgot its role in public debate. Aeon Essay, 2 October.

Arrow K. 1951. Social Choice and Individual Values. New Haven: Yale University Press, 2nd ed., 1963.

Baumol W.J, W.E.Oates. 1988. The Theory of environmental policy. Cambridge: Cambridge University Press, 2nd ed.

Barnett E., P. Thomas , J. Thomas. Methods for the synthesis of qualitative research: a critical review. London: Institute of Education, ESRC National Centre for Research Methods. NCRM Working Paper Series. No. 1/2009.

Burchell G., C. Gordon, P. Miller (eds). 1991. The Foucault Effect: Studies in governmentality. Hertfordshire: Harvester Wheatsheaf.

Chapman J. 2004. System failure – Why governments must learn to think differently. Second edition, London: Demos.

Coleman J. Foundations of Social Theory. Revue française de sociologie, 46(2005), Supplement.

Collins P. H. 2000. Black feminist thought: Knowledge, consciousness, and the politics of empowerment (2nd ed.). New York: Routledge.

Common Assessment Framework. 2006. Brussels: EIPA – European Institute of Public Administration.

Connelly S., T. Richardson. Exclusion: The necessary difference between ideal and practical consensus. Journal of Environmental Planning and Management 47/1(January 2004):3-17.

DCLG. 2009. Multi-criteria analysis: a manual. London: Department for Communities and Local Government

Diamond J. Establishing a Performance Management Framework for Government. Instituto de Estudios Fiscales, Presupuesto y Gasto Público, Working Paper 40(2005):159-83.

Dopfer K. 2006. The Origins of Meso Economics Schumpeter's Legacy. Jena: Max Planck Institute of Economics, Evolutionary Economics Group, The Papers on Economics and Evolution No. 610.

Dopfer K., J. Foster, J. Potts. Micro–meso–macro. Springer, Journal of Evolutionary Economy 14/3(2004):263–79.

Easterling W.E., K. Kok. Emergent Properties of Scale in Global Environmental Modelling – Are There Any? Integrated Assessment 3/2–3(2002):233–46.

Ekins P. 1992. A Four-Capital Model of Wealth Creation in Real-Life Economics: Understanding Wealth Creation in Ekins P., M. Max-Neef, (eds.), London, New York: Routledge.

Ekins P., J. Medhurst . 2003. Evaluating the Contribution of the European Structural Funds to Sustainable Development. Presented at the 5th European Conference on Evaluation of Structural Funds, Budapest, June 26-7.

Ekins P., Medhurst J. The European Structural Funds and Sustainable Development: A Methodology and Indicator Framework for Evaluation. Sage: Evaluation, 12/4(October 2006):474-95.

Elbers C., J. W. Gunning, K. de Hoop. 2007. Assessing Budget Support with Statistical Impact Evaluation: a Methodological Proposal. Amsterdam: Tinbergen Institute, Discussion Paper 075/2.

Elsner W. Why Meso? On 'Aggregation' and 'Emergence' and Why and How the Meso Level is Essential in Social Economics. Springer, Forum for Social Economics, 36/1(2007):1-16.

Encyclopaedia of Evaluation. 2004. 'Synthesis'. Sage Publications.

ESPON - 3.2, Vol. 5. 2006. Spatial Scenarios and Orientations in Relation to the ESDP and Cohesion Policy – Territorial Impact Assessment, Final Report.

European Commission. 2001. SEA Directive 2001/42/EC on the assessment of the effects of certain plans and programmes on the environment.

European Commission. 2006. Impact Assessment Guidelines. SEC(2005)791.

Faris E. Sociology and Human Welfare. Social Forces 18(1939):1-9.

Foster J., Potts J. 2007. A micro-meso-macro perspective on the methodology of evolutionary economics: integrating history, simulation and econometrics. The University of Queensland, School of Economics, Discussion Paper No. 343.

Funtowicz S. O., J. R. Ravetz. The worth of a songbird: ecological economics as a post-normal science. Ecological Economics 10(1994):197–207.

Geels F. W. Technological transitions as evolutionary reconfiguration processes: A multi-level perspective and a case-study. Elsevier, Research Policy 31/8-9(2002):1257-74.

Geels F.W. Feelings of Discontent and the Promise of Middle Range Theory for STS Examples from Technology Dynamics. Sage, Science, Technology, & Human Values 32/ 6(November 2007):627-51.

Gerrits L., S. Verweij. Taking stock of complexity in evaluation: A discussion of three recent publications. Sage: Evaluation, 21/4(2015):481–91.

Guba E., Y. S. Lincoln. 2000. Competing paradigms in qualitative research, in Denzin, Lincoln (ed.). Handbook of Qualitative Research, 2nd ed., London, New Delhi: Sage.

Gutiérrez F., D. Buitrago, A. González. Aggregating Political Dimensions: Of the Feasibility of Political Indicators. Springer Science, Social Indicators Research, An International and Interdisciplinary Journal for Quality-of-Life Measurement, 110/1(2013):305-26.

Hageboeck, M., M., Frumkin, S. Monschein, S. 2013. Meta-Evaluation of Quality and Coverage of USAID Evaluations 2009-2012, Management Systems International under subcontract to DevTech Systems, Inc. for USAID Contract No. AID-OAA-M-11-00026.

Hertin J., A. Jordan, M. Nilsson, B. Nykvist, D. Russel, J. Turnpenny. 2007. The practice of policy assessment in Europe: An institutional and political analysis. EU/FP6 Project MATISSE Working Paper 6.

Heylighen F. 1989. Self-Organization, Emergence and the Architecture of Complexity, in: Proceedings of the 1st European Conference on System Science, AFCET, Paris.

Huitema D., A. Jordan, E. Massey, T. Rayner, H. van Asselt, C. Haug, Roger Hildingsson, S. Monni, J. Stripple. The evaluation of climate policy: theory and emerging practice in Europe. Springer: Policy Science, 44/2(2011):179-98.

IAB. 2009. Impact Assessment Board report for 2008. Third strategic review of Better Regulation in the European Union. SEC(2009)55.

James M. Public Interest and Majority Rule in Bentham's Democratic Theory. Sage Publications: Political Theory, 9/1(Feb., 1981):49-64.

Kordig C. R. Discussion: Observational Invariance. Philosophy of Science, 40(1973):558-69.

Kuhn T. S. 1970. The Structure of Scientific Revolutions. Chicago: The University of Chicago Press.

Kunseler E. 2007. Towards a new paradigm of science in scientific policy advising. Environment and Health Department. NUSAP.net.

Leith C., L. von Thadden. 2006. Monetary and fiscal policy interactions in a new Keynesian model with capital accumulation and non-Ricardian consumers. Working Paper Series No. 649.

Leopold L. B., F. E. Clarke, B. B. Hanshaw, J. R. Balsley. 1971. A procedure for evaluating environmental impact. Washington: Geological Survey Circular 645.

List C., B. Polak. 2010. Introduction to judgment aggregation. Yale University, Cowles Foundation Discussion Paper No. 1753.

Lorenz K. Reflections on the concept of symmetry. European Review, 13/2(2005):3–11

Lucas R. E. Jr. Expectations and the neutrality of money. Blackwell, Journal of Economic Theory, 4/2(1972):103-24.

Mackay A. L. 1977. The Harvest of a Quiet Eye Quotes. http://www.quotes.net/ Mackay, XII 2011.

Martinez-Alier J., G. Munda, J. O'Neill. Weak comparability of values as a foundation for ecological economics. Elsevier, Ecological Economics 26(1998):277-86.

Mason E. 2006. Value Pluralism; entry in Stanford Encyclopedia of Philosophy (in Zalta N.E, ed.), Stanford: Centre for the Study of Language and Information, Stanford University.

Morçöl G. 2011. Why Complexity Theory? What Kind of a Theory Is It? Presented at the Governance of complex systems conference, Rotterdam, June 23-25.

Morçöl G. 2012. Coplexity theory for Public Policies. Routlege.

Morgan D. L. Paradigms Lost and Pragmatism Regained: Methodological Implications of Combining Qualitative and Quantitative Methods. Sage, Journal of Mixed Methods Research, 1/1(2007):48-76.

Mumpower J. L., T. R. Stewart. Expert Judgement and Expert Disagreement. Taylor and Francis, Thinking and Reasoning, 2/2-3(1996):191-211.

Munasinghe M. 1992. Environmental Economics and Sustainable Development. Paper presented at the UN Earth Summit, Rio de Janeiro, Washington: World Bank, Environment Paper No. 3.

Munda G. 2006. A NAIADE based approach for sustainability benchmarking. Inderscience, International Journal of Environmental Technology and Management, 6/1-2(2006):65–78.

Munda G. Choosing Aggregation Rules for Composite Indicators. Social Indicators Research - An International and Interdisciplinary Journal for Quality-of-Life Measurement, Springer, 109/3(2012):337-54.

Munda G. Social multi-criteria evaluation: Methodological foundations and operational consequences. European Journal of Operational Research, 158/3(2004):662-77.

Muradian R. Ecological thresholds: a survey. Elsevier, Ecological Economics, 38/1(2001):7–24.

Nola R., H. Sankey (eds). 2000. After Popper, Kuhn & Feyerabend: Issues in Theories of Scientific Method. Australasian Studies in History and Philosophy of Science, Kluwer.

Olssen M. Michel Foucault as "Thin" Communitarian: Difference, Community, Democracy. Cultural Studies, 2/4(2002):483-513.

Pearce D.W., E.B. Barbier, A. Markandya. 1990. Sustainable development: economics and environment in the third world. Hants: Edward Elgar.

Perry R.B. Is There a Social Mind? American Journal of Sociology 27(1922):561-572

Prigogine I., I. Stengers. 1979. La nouvelle alliance: Métamorphose de la science, Paris: Gallimard.

Radej B. 2006. Assessment of structural funds effectiveness on the sustainable development: Pomurje regional case study. Final report, 2006, 6th Framework program – Framework project SDRTOOLS: Regional sustainability Appraisal: Developing evaluation methods and sustainable policies (Contract SCS8-CT-2004-502485).

Radej B., K. Žvokelj Jazbinšek, M. Dolinšek. Measuring Smartness of Innovation Policy. New Delhi: UN's Economic and Social Commission for Asia and the Pacific, Asian And Pacific Centre for Transfer of Technology, Asia-Pacific Tech Monitor, 32/3(Jul-Sep 2015):28-36.

Radej B., M.Golobič, M.Macur. 2012. Kompleksna družba v radikalni sredini. Ljubljana: Oddelek za krajinsko arhitekturo Biotehnične fakultete Univerze v Ljubljani, foreword, S. Dragoš.

Ravetz J. 2003. Notes for the lecture: Pluralistic uncertainty management: social scientific basis. EU Advanced Summer Course in Integrated Assessment. ICIS, University of Maastricht, September.

Ritzer G., B. Smart. 2003. Introduction: Theorists, Theories and Theorizing, in Ritzer G., B. Smart (ed.). Handbook of Social Theory. Sage.

Rotmans J. Scaling in Integrated Assessment: Problem or Challenge? Swets & Zeitlinger, Integrated Assessment, 3/2–3(2002):266–79.

Rotmans J. Tools for Integrated Sustainability Assessment: A two-track approach. Vancouver: University of British Columbia, The Integrated Assessment Journal, 6/4(2006):35–57.

Sadler B., R. Verheem. 1996. Strategic environmental assessment: Status, challenges and future directions. Hague: Ministry of Housing, Spatial Planning, the Environment.

Sanderson I. Evaluation in complex policy systems. Sage, Evaluation, 6/4(2000):433–54.

Scharffs B.G. Adjudication and the problems of incommensurability. William and Mary Law Review, 42/4(April, 2001):1367-435.

Schnellenbach J. The Dahrendorf hypothesis and its implications for (the theory of) economic policy-making. Cambridge Journal of Economics, 29/6(2005):997-1009.

Scholes R.J., B. Reyers, R. Biggs, M.J. Spierenburg, A. Duriappah. Multi-scale and cross-scale assessments of social–ecological systems and their ecosystem services. Elsevier, Current Opinion in Environmental Sustainability 5/1(March 2013):16–25.

Scriven M. The Final Synthesis. Sage, American Journal of Evaluation, 15/3(1994):367-82.

Simon H.A, A. Ando. Aggregation of variables in dynamic systems. Econometrica, 29/2(1961):111-38.

Söderbaum P. 1998. Economics and Ecological Sustainability. An Actor-network Approach to Evaluation, in N. Lichfield, A. Barbanente, D. Borri, A. Khakee and A. Prat (eds) Evaluation in Planning: Facing the Challenge of Complexity, pp. 51–72. Dordrecht: Kluwer.

Stake R. E. A Problematic Heading. American Journal of Evaluation; 22/3(2001):349–54.

Stiglitz J.E, A. Sen, J.-P. Fitoussi. 2009. Report by the Commission on the Measurement of Economic Performance and Social Progress, Paris: CMEPSP - The Commission on the Measurement of Economic Performance and Social Progress.

Thagard P., J. Zhu. Emotion and action. Philosophical Psychology 15/1(2002):19–36.

Turner J. H. (ed.). 2006. Handbook of Sociological Theory. Springer.

UN. 2015. Transforming our world: the 2030 Agenda for Sustainable Development. A/RES/70/1.

UNCED. 1992. Agenda 21. United Nation Conference on Environment and Development.

Vedung E. Four Waves of Evaluation Diffusion. Sage, Evaluation, 16/3(July 2010):263-77.

Veen A. Van Der, H. S. Otter. Scales in Space. Swets&Zeitlinger, Integrated Assessment, 3/2–3(2002):160–6.

Virtanen P., P. Uusikylä. Exploring the Missing Links between Cause and Effect. A Conceptual Framework for Understanding Micro–Macro Conversions in Programme. Sage, Evaluation, 10/1(2004):77–91.

Wacquant L. 1997. Reading Bourdieu's "Capital". Foreword to the English-language translation of The State Nobility. Cambridge: Polity Press.

Wallerstein I. 2004. The Uncertainties of Knowledge. Philadelphia: Temple University Press.

WCED. 1987. Our Common Future. Oxford and New York: Oxford University Press, World Commission on Environment and Development – WCED.

Weaver P. M., J. Rotmans. 2006. Integrated Sustainability Assessment: What? Why? How? EU/FP6 Project MATISSE, Working Paper 1 (October).

Wilhelmsen F. D., J. Bret. 1970. The War in Man, University of Georgia Press.

Wimsatt W.C. 2002Emergence as non-aggregativity and the biases of reductionisms. University of Chicago Department of Philosophy.

Xu L., J.-B. Yang. 2001. Introduction to Multi-Criteria Decision Making and the Evidential Reasoning Approach. Manchester School of Management, Working Paper No. 106.

Zhou P., B. W. Ang. Comparing MCDA Aggregation Methods in Constructing Composite Indicators Using the Shannon-Spearman Measure. Springer, An International and Interdisciplinary Journal for Quality-of-Life Measurement, 94/1(October 2009):83-96.