

COST ACTION 356 - WG2
Task 2.2 REPORT

Criteria and methods for indicator assessment and validation

- a review of general and sustainable transport related
indicator criteria and how to apply them

Background Report for Chapter 4 in COST Action 356
Scientific Report

Author:

Henrik Gudmundsson, DTU, Denmark

With contributions from:

Aud Tennøy, TØI, Norway

Robert Joumard, INRETS, France

Patrick Waeger, EMPA, Switzerland

Lennart Folkesson, VTI, Sweden

Holger Dalkmann, TRL, UK

Thomas Fischer, University of Liverpool, UK

Enrique Calderon, UPM, Spain

Ana Paula Ramos, TIS, Portugal

COMPLETE VERSION Rev. 06
April, 2010

[blank]

Contents

1	Introduction and overview	7
1.1	Purpose of this report in the context of COST Action 356	7
1.2	The need for criteria for EST indicators	7
1.3	The content of the report	8
2	Initial framing of indicator criteria in COST Action 356	9
2.1	Environmentally Sustainable Transport indicators: What to indicate?	9
2.2	Developing an initial list of criteria	10
3	Literature about criteria for indicators	14
3.1	Overview of the literature search	14
3.2	Criteria for ‘good’ indicators at the three levels of indicator application	15
3.3	Criteria in transport applications	27
4	Developing indicator criteria for COST 356 and EST	36
4.1	Correspondence of criteria: literature compared with Section 2 lists	36
4.2	Correspondence of criteria: Section 2 lists compared with literature	38
4.3	Reorganizing and rewording indicator criteria	39
4.4	Proposal for an intermediate revised set of criteria	41
4.5	An internal trial exercise with indicator criteria	44
4.6	Further refinement of the criteria list	49
5	Frameworks and methods for assessing indicators	57
5.1	Introduction	57
5.2	Validation frameworks and selection procedures	58
5.3	Summary discussion of criteria methods and frameworks	64
5.4	Discussion of criteria and ‘joint consideration’ of indicators	65
6	Proposed approach and recommendations	70
6.1	General approaches for the assessment of EST indicators	70
6.2	Approach and guidelines for subsequent work within COST Action 356	71
7	Conclusion	73
	References	74

[blank]

List of tables

Table 1 . COST Action 356. Starting points for indicator assessment	9
Table 2. Initial list of criteria and descriptions	10
Table 3 Intermediate structure for indicator criteria (to be revised)	12
Table 4. Level 1 - 'Scientific measurement' criteria.....	16
Table 5 . Level 2 - 'Monitoring' system criteria.....	18
Table 6. Level 3 - 'Policy/management' criteria	19
Table 7. NCHOD 2005. Comprehensive list of Clinical Health indicator criteria.....	21
Table 8. Rice & Rochet 2005. Fisheries Management - ICES.....	22
Table 9. Niemeijer & de Groot (2008) Environmental indicator sets.....	23
Table 10 'Acronym' criteria listings.....	24
Table 11. Heterogeneity of criteria groupings across various references (examples).....	25
Table 12. COST 350 criteria (Goger et al 2006).....	27
Table 13. General data quality for traffic measurement (Batalle et al 2004)	28
Table 14. Traffic Safety Indicators at EU level (Farchi et al (2006)	29
Table 15. Environmental Indicators for US transportation (US EPA (1999)	30
Table 16. Sustainable urban infrastructure in Europe (Lahti et al 2006)	31
Table 17. (Sustainable transportation indicators for Canada (Gilbert et al 2002).....	32
Table 18. STPI topics and questions (Gilbert et al 2002)	32
Table 19. Zietsman & Rilett (2002).....	33
Table 20 Criteria from selected references selected by Marsden et al (2005)	34
Table 21. Revised intermediate list of potential indicator criteria	41
Table 22 . Number of respondents.....	45
Table 23 . Exercise responses concerning criteria (Trial 1).....	46
Table 24. Trial assessment of candidate indicators	47
Table 25. Proposal of revised criteria from R Joumard (tentative).....	50
Table 26. (a) Proposed criteria with regard to representation.....	54
Table 27. Framework for selecting indicators (Rice & Rochet 2005).....	62
Table 28. Types of evidence to consider quality of candidate indicators (Rice and Rochet 2005)	62
Table 29 Quality criteria for Multicriteria methods (adapted from DeMontis et al (2004).....	68

List of Figures

Figure 1. General version of the assessment matrix applied in COST Action C8 (Lahti et al 2006)...	31
Figure 2 Decision tree as defined by Bockstaller & Girardin (2003)	59
Figure 3. '3S' methodology for indicator validation as proposed by Cloquell-Ballester et al (2006).....	61

1 Introduction and overview

1.1 Purpose of this report in the context of COST Action 356

The general purpose of this report is to help establish criteria and methods to assist in the identification and selection of appropriate indicators for Environmentally Sustainable Transport (EST).

The report is part of the work in COST Action 356 “Towards a measurable Environmentally Sustainable Transport (EST)”. More specifically it is the extended output from *task 2.2* of COST Action 356 Working Group 2, which deals with ‘indicators as measurement tools’ for the assessment of transport’s environmental impacts.

According to the Work Program of the COST Action the purpose of task 2.2. has been:

‘To identify operational quality criteria needed for assessing usability of indicators (representativity, simplicity, transparency etc) based on available literature – thus forming the basis for task 2.3.’

The work has thus been intended for use in the subsequent *Task 2.3* of the Action which aims to construct or select ‘*indicators per environmental impact*’ using criteria and methods as identified in the present report.

Meanwhile it has become clear that indicator criteria also relates strongly to other parts of the COST Action 356 including Working Group 3, which has dealt with ‘*indicators as decision making tools*’, especially task 3.1 ‘*EST indicators from the planning and decision making point of view*’ and task 3.2 looking at ‘*Options for integrating EST indicators*’. Those aspects are also addressed in this report, although with less detail.

Furthermore it has become clear that criteria alone do not provide sufficient guidance for the identification and selection of indicators. Guidance on how to apply the criteria, as also reported in literature, is another aspect that is addressed in this report.

In terms of the *Scientific Report* of COST Action 356 this report will have the following functions:

- It *provides the main content of Chapter 4* of the Scientific Report with the working title ‘Criteria and methods for indicator assessment and selection’.
- It *informs work that will go into writing mainly Chapter 5* ‘Assessment of some indicators within impact’, partly Chapter 6 ‘Methods for joint consideration of indicators’ and partly Chapter 3 ‘The dimensions and context of transport decision making’.

It has been decided to finalize this report as a background document, because it conveys the full internal process in the work on indicator criteria in COST Action 356 as well as its results. Documenting the process and results together may have value for COST Action 356 participants and potentially also for others who intend to embark on a process of indicator review and selection using criteria.

1.2 The need for criteria for EST indicators

An important background for the work of COST Action 356 is a recognized need for appropriate indicators to measure the environmental impacts and sustainability of transport systems and policies (‘EST indicators’).

But what are appropriate EST indicators, and how to identify them?

Indicators proposed or already used in transport assessment today is one potential source. However, as noted by several (Goger et al 2005; May et al 2007, Litman 2006, Jeon & Amekudzi 2005) such indicators often reflect practical compromises and are not necessarily comprehensive or accurate reflections of transport impact on environmental sustainability. The literature in areas like environmental monitoring, resource management and urban planning, may provide a great number of other potential indicators on the environment. However such measures may also have limitations, for example in their ability to support assessment of the impact of transport systems or policies. Finally, new concepts and measures may be introduced where existing sources such as the above do not provide adequate information. Even in that case yardsticks are needed to gauge the adequacy of the input for assessing impacts of transport on the environment

In short there is a need for *criteria* to assess the relevance, quality and sufficiency of measures to be proposed as EST indicators, both existing, and potential, new ones. The aim of this report is to review and establish such criteria and to discuss how to apply them.

The task is pursued along two directions. The one direction is a top-down approach, where the literature about indicator assessment and selection is compiled. This has been done in two steps:

- reviewing and systematizing literature about *criteria* for indicator quality, appropriateness etc
- review of literature about *procedures and methods* for applying such criteria in practice

The other direction is a bottom-up approach where specific requirements regarding indicator criteria for the area of EST indicators are identified. The latter also involves two elements:

- outlining key questions that indicators must help answer in the EST and COST 356 areas,
- involving COST 356 Action members in steps along the way to help identify, review and apply criteria and approaches thought to be of particular relevance for EST.

The *principal* emphasis in the report is on the top-down approach (literature review based), while the bottom-up approach provides necessary framing and adaptations to the present context.

1.3 The content of the report

The report contains the following sections:

Section 2 reports on the initial work in the task where the questions to be addressed by EST indicators is considered, and the first steps to identify indicator criteria of relevance for COST Action 356 are taken in a 'bottom-up' process, with only limited guidance from the literature.

Section 3 categorises and reviews key literature on indicator quality and selection criteria, including both general indicator criteria literature, and previous work on sustainable transport indicator criteria, such as work in COST Action 350. A broad range of possible criteria is identified.

Section 4 further develops the first tentative list of EST criteria defined in Section 2 in the light of the literature findings of section 3. A revised, consolidated, list of potential criteria and associated tentative definitions is developed and then subsequently tested with regard to usability etc in the context of COST Action 356. A 'final' consolidated list of criteria is emerging.

Section 5 addresses the question of how to apply the indicator assessment criteria in practice. Hence, the section reviews methods proposed in the literature for criteria based indicator selection procedures, and discusses them in the context of the tasks of COST Action 356.

Section 6 summarises key points of the report and give recommendations for the following work within and beyond COST Action 356.

Section 7 gives the conclusions.

2 Initial framing of indicator criteria in COST Action 356

2.1 Environmentally Sustainable Transport indicators: What to indicate?

An important observation made by Lenz et al (2000), US EPA (2006) Jeon & Amekudzi (2005) and others is that *indicator selection should primarily be driven by the questions that the indicators are supposed to answer*. Some consideration of context is necessary since indicators are always supposed to be indicators of something, for something

In the present context the overall questions that indicators should help answer are given by the focus on environmental sustainability impacts of transport systems and policies in general, which include, for example:

- What are the environmental impacts of transport systems and flows,
- how are potential or actual transport policies and decisions influencing such impacts in a positive or negative way,
- how significant are the environmental impacts of transport with regard to sustainability or other standards for acceptability.

Further directions for the work on indicator criteria of COST Action 356 has been extracted from the Memorandum of Understanding, the Work Program and subsequent decisions on the purpose and scope of the action, as summarised in Table 1.

<p>Table 1 . COST Action 356. Starting points for indicator assessment</p> <p>The purpose of COST 356 is stated in the Memorandum of Understanding (MoU):</p> <p>“The purpose is to design harmonised and scientifically sound methods to build better environmental indices (or indicators) by using existing European indices, and to build methods to be applied to the decision-making process of the transport sector in the different European countries (...) the whole range of impacts is necessary to ensure that sustainability takes into account environmental issues to a satisfactory degree”.</p>	
<p>List of environmental impacts considered in COST 356:</p>	
<ul style="list-style-type: none"> • Noise and vibration • Local air quality • Regional air quality • Quality and use of water • Protected areas • Waste • Loss of biodiversity • Light pollution 	<ul style="list-style-type: none"> • Technological hazards • Landscape, cultural and built heritage • Land use • Non-renewable resource use • Ozone depletion • Climate change • Safety of transport users and residents
<p>• Transport is defined as including Infrastructure building, vehicle production, transport energy production and distribution, traffic, vehicle and infrastructure destruction. It has been decided to consider the <i>full life cycle</i> of transport impacts, and to address <i>all modes of transport</i>.</p> <p>The definition of indicators used in COST 356:</p> <ul style="list-style-type: none"> • “An indicator is a variable, based on measurements, representing as accurately as possible and necessary a phenomenon of interest to human beings. • An environmental impact indicator is a variable based on measurements, representing an impact of human activity on the environment, as accurately as possible and necessary. • An indicator of environmentally sustainable transport is a variable, based on measurements, representing potential or actual impacts on the environment, or factors that may cause such impacts, due to transport systems, flows or policies, as accurately as possible and necessary”. 	

It is clear that COST Action 356 does not itself aim to establish concrete indicators for a particular types of environmental impact or transport system. The aim is to help develop general *scientific methods* for the identification and use of EST indicators for all types of environmental impacts, transport systems, and aspects of transport decision making. Key aspects of the work include,

- a concern to ensure a comprehensive and sufficient consideration of the environmental impacts
- focus on both measurement and decision making aspects of indicators
- interest in both indicators and indices
- addressing both existing and potential new indicators

This suggests a *broad approach* to review criteria and methods for indicator assessment and use, rather than focus on requirements for specific scientific disciplines, compartments of the environment, transport modes, policy making contexts or, or development of knowledge and methodology.

2.2 Developing an initial list of criteria

The work to identify indicator criteria of potential relevance for COST Action 356 was initiated at the 1st WG2 meeting in Lisbon in March 2007. A group of WG members held a brainstorm where a first list of indicator ‘quality’ criteria of potential interest for selecting indicators in the context of COST Action 356 were drafted. The list was based on the members own previous personal and professional experiences with sustainable transport assessments, and the subsequent dialogue.

The list was consolidated in a subsequent process over email where an extended group of WG members was involved in providing descriptions of the content of each criterion and its relevance for the Action. The (unedited) results of this first round is shown in Table 2.

Criterion	Action member descriptions
Aggregatability	How easy and to which degree indicators can be aggregated, to higher geographical levels, with other indicators etc.
Replicability	One should be able to reproduce measures about the relationship between the indicator(s) used to measure a phenomenon, and the phenomenon in focus (like fragmentation and loss of biodiversity) in later studies and by other scientists. This implies among others transparency:
Transparency	To which degree it is described in an understandable way how the indicator is constructed, and how it varies with what it represent (the phenomenon in focus). This implies that input data, assumptions, methods, models and theories involved are described. The reasons why these particular data, assumptions, models, methods and theories are used should be explained, as should the implications of the choices made. It is important to point out that other conclusions might have been reached if another set of input data, assumptions, methods, and models were chosen.
Representativity	To which degree the indicator is representing the phenomenon it is developed for.
Preciseness	How precise the indicator can be measured (accuracy, reliability...) and/ or how precise the indicator is showing development of the phenomenon it is developed for.
Theoretical (foundations)	The extent to which an explicit theoretical basis exists, that provides an explanation why the indicator represents something else and/ or something more than themselves, which is relevant for the topic the indicator is developed for (examples: CO2 as indicator for global warming, fragmentation as indicator for loss of biodiversity)
Measurability (included also forecast ability – will it change with time),	Data required to figure out the indicators should be reliable. E.g. air quality measures should be taken through consistent procedures and using standard equipment. Likewise, population affected by different levels of pollution should be objectively calculated. When forecasting present values of magnitude, technically sound models should be preferable. Subjective assessment of significance is highly variable in time and space.

Table 2 Continued	
Criterion	Action member descriptions
Common definition exists	Indicators defined and used frequently in other situations are preferable, as are less subject to dispute. However, significance should be assessed on a case-by-case basis
Data availability in terms of quality, quantity and timeliness (on time, how long does it take to produce an indicator from the data)	Indicators that can be accessible in time series and on a cross-geographical basis should be preferred. Decisions, mostly at the local level have to be based on local data and it is no use to recommend indicators that cannot be documented numerically or through generally agreed (undisputable) information. Detected lack/gaps of information at individual locations can, however, serve as basis for policy instructions to start data collection when particular indicators are generally used elsewhere. This criterion is related to one on space transferability, i.e. indicators that can be adequately measured and forecast in different locations should be preferable.
How frequently are data updated	See previous description. The validity of extrapolated values is ruled by statistical significance. To establish trends a minimum number of values are required. Social surveys at regular intervals may be required to highlight changes in perceptions. Again, significance may vary locally in addition to temporally.
Robustness	This refers not only to technical issues but likewise to time sustainability. An indicator sustained by mathematical measures and models is preferable. Moreover, the significance of an indicator should be sustained over time, in other words, not subject to local and temporal circumstances
Certainty (monitoring and predictions)	Description can be derived from previous comments. Measures taken with reliable instruments and using internationally accepted procedures are less subject to challenge and can be used for comparisons. Robustness of forecasting models is essential. Beware of indicators based on subjective perceptions (e.g. value of scenery) and of allocated values of significance.
Discountability	Discounting influences people's assessment and evaluation of impacts that will be perceived in different moments of time, as well as trade-offs with other effects characterized in other moments and through other indicators. Discounting factors are affected not only but subjective perceptions but, likewise, by changes in technology and by people becoming used to situations
Appropriate time series	Indicators should have allocated the correct time series in order to be reliable. The time series adoption should be very rigorous to represent as much as possible the scenario under evaluation.
Independence from each other	Indicators should be as much as possible independent from each other.
Causality	This criteria denotes the logical relationship between one physical event (cause) and another physical event (effect) being the direct consequence (result) of the first event. Causality simply means (by definition) that the effect is the consequence (result) of the cause.
Reliability	The ability of an indicator to perform its pre-defined functions in routine circumstances, as well as hostile or unexpected circumstances. The IEEE defines it as ". . . the ability of a system or component to perform its required functions under stated conditions for a specified period of time." Reliability of an indicator may also be 'the idea that something is fit for purpose with respect to time'.
Transferability (comparability and usefulness across borders)	This means the capability that an indicator has to be used in other similar contexts in order to compare different scenarios. This characteristic is useful in the case of cross-border issues.
Simplicity (1)	Condition, or quality of an indicator be simple or un-combined. This characteristic in some situations is better to turn easier the explanation of certain things than complicated ones.
Simplicity (2)	How easy it is to understand the indicator: how it is constructed, how it is related to and varies with the main phenomenon etc, and/ or how easy it is to measure and calculate the indicator

This brainstorm allowed a broad range of possible criteria to be brought forward, taking into account what is considered important for selecting EST indicators. The further development into the list in Table 2 reflects the WG2 member's combined understanding of what would be relevant to consider in terms of indicator quality needs of COST 356. In this way the list provided a valuable starting point for the continued work. However it could be noted that the unedited list suffers from a number of deficits:

- Overlaps and possible redundancies among criteria
- Partly intuitive descriptions (not official, consolidated definitions)
- No prioritisation among criteria
- No order in terms of where or when in an indicator selection process each criterion would apply

To the second WG2 meeting in Stockholm in June 2007 a first attempt to structure the list was proposed. The aim was to reduce overlaps, and to distinguish 'Measurement' related criteria (WG2) from more 'Decision making' related criteria (WG3) (See Table 3).

The structure with four categories of criteria was inspired by two contributions from the research literature on indicator selection criteria, namely Innes (1990) work on validation of policy indicators and the work of Kurtz et al (2001) on criteria for environmental management indicators.

Table 3 Intermediate structure for indicator criteria (to be revised)		
Criterion category	Proposed Criteria	Subsumed criteria in the original 'brainstorm' list (Table 3)
Conceptual and theoretical aspects	Representativity /Validity	Representativity Robustness Preciseness
	Theoretical Foundation	Theoretical Foundation Causality Independency from each other
	Transparency	Transparency Common definition exists
Measurement aspects	Reliability	Reliability Replicability Appropriate time series
	Measurability	Measurability
	Data availability	Data availability How frequently are data updated
Data structuring aspects	Aggregatability without loss of representativeness	Aggregatability
	Discountability	Discountability
'User' aspects	Transferability	Transferability
	Simplicity	Simplicity 1+2

The following explanation to this structure – still intuitive, but partly informed by literature, can be given.

Criteria in the *first category* (Conceptual and theoretical aspects) are supposed to be the ones that are always of importance for the measurement qualities of any indicator in any general function. According to the encyclopaedic article about indicators given by Bollen (2004), "...(t)he most critical and most neglected aspect (...) is providing a clear theoretical definition of the concept that a researcher seeks to measure." If there is no clear accepted theoretical foundation, the indicator may measure anything or nothing, depending on point of view. Further aspects in this domain include concepts such as 'representativity', 'validity', and 'transparency'. 'Validity' involves several technical subcategories, as we will return to again later.

Criteria in the *second* category refer to the more practical preconditions for indicators to actually operate as measurement tools, including (obvious) criteria of measurability (do methods exist to measure or calculate from measurements the problem of interest?), reliability (do similar measurements reproduce similar results?), and data availability. The distinction between category 1 and 2 is partly arbitrary, since there is an intricate link between for example valid theoretical conceptions and verified reliability of an assumed casual relations. 'Data availability', however is clearly distinct from, and secondary to, the criteria referring the establishing of causal relations and representativity

Some criteria included in Table 2 such as 'aggregatability' and 'discountability' are put in a *third* category 'data structuring' with reference to Innes 1990 pp 214 ff) : such criteria may be important for allowing data analysis and advanced presentations, including aggregation of indicators *across several impacts*, while they seem less directly related to conceptual or measurement aspects. Still aggregation and discounting should of course build on conceptual, theoretical justification (or be avoided) and on appropriate sound measurement and calculation principles.

Finally, 'Simplicity' and 'transferability' may especially be important for the use of indicators e.g. in decision making and policy, but are not directly relevant for conceptual or measurement aspects. They are here put under the *fourth* category 'user aspects'.

In this resulting structure seen in Table 3 the number of criteria have been tentatively reduced and collapsed from Table 2, to avoid a number of partial overlaps stemming from the mostly intuitive definitions given. This may on the other mean that some criteria aspects will be lost in the streamlining. The list is therefore to be consulted again following the literature review in Section 3, and some items may reappear or be re-categorized, if they are found to be salient.

Most importantly, the revised list in Table 3 does not provide any recognized or 'official' definitions of the criteria nor any specifications as to how to apply them.

This will be addressed through the literature review in Section 3 and the and subsequent developments in Sections 4 and 5.

3 Literature about criteria for indicators

3.1 Overview of the literature search

A continuous search and review of literature about criteria for indicator assessment has been conducted with varying intensity through the period from March 2007 to March 2009.

The work has involved several electronic searches using specific search terms such as 'indicators' 'criteria' 'selection', 'assessment' etc at a number of journal websites, journal databases and academic search engines (including Google Scholar, SCIRUS, Ingenta, Web of Science, EBSCO-Host, ScienceDirect and others).

The review has considered both general literature of indicator selection, and more specific reviews of indicator studies in various areas connected to the COST 356 field, such as environmental sciences, ecosystem management, sustainability assessments, and health studies. Special attention has been given to reports on indicator criteria for the area of Environmentally Sustainable Transport, such as work in COST Action 350 (Goger et al 2005), the UK DESTILLATE project (Marsden et al 2007), and others.

A rather large number of publications (books, papers, reports and guidelines) about indicator quality, desirable properties of indicators, criteria lists, and methods for selecting indicators have been identified.

The majority of the references are ones that contain a *list of criteria* for the selection of indicators in various more or less specific domains such as environmental assessment or sustainability (e.g. Dale & Beyeler, 2001; Eyles & Furgal 2000; OECD 2003; WHO 2006; Pact 2005). Some focus on criteria as such, while other also report the use and application of the criteria in a particular case. It varies to what extent these references provide actual *definitions* of the criteria, sometimes they are operational ad hoc formulations while only in a few cases more rigid definitions drawing from basic scientific literature is drawn upon.

A smaller group of these references propose *methods or procedures* for how to apply criteria in actual indicator development and selection processes (e.g. Bockstaller & Girardin 2002, Cloquell-Ballester et al 2006; Hardi & DeSouza-Huletey 2000, Jackson et al 2000; Innes 1978). These references are typically more valuable for the present work than the simple 'list and apply' type. Only very few references provide accounts or evaluations of the actual procedures that have been followed in the selection of particular indicator sets (See especially Rochet & Rice 2005, as we shall return to).

A few meta-reviews of the indicator criteria literature also exist, as for example found in Boyle (1998), Niemeijer & de Groot (2008), and NCHOD (2005). In addition a few encyclopaedic articles provide short overall conceptual reviews of indicator criteria (e.g. Bollen 2004; Leviton 2001). Even these meta-references are typically concerned with criteria for indicators in a certain *domain*, such as 'environment' or 'health' or 'social reporting' or 'Management'. No completely universal review of indicator criteria literatures was identified.

Finally there are as mentioned a number of publications dealing specifically with indicator criteria for **transport and/or sustainable transport** (e.g. Marsden et al 2005; Dobranskyte-Niskota et al 2007; STPI 2003, US EPA 1999; Farchi et al 2006, apart from the work in COST Action 350 (Goger et al 2006). These references mostly belong to the first group mentioned above, basically suggesting the same types of criteria as proposed in other fields. However, some specific considerations for the selection of transport and sustainable transport indicators can be found.

It is characteristic for almost all of the indicator criteria references that they designate different *types* of criteria, divided into categories such as 'scientific' versus 'policy related' criteria, with the

groups, however, often defined in a way unique to the particular reference. Very few references deal solely with one category of criteria alone, meaning that *there are not major literatures* on for example ‘measurement’ related criteria for indicators only. As suggested also by Turnhout et al (2007) it seems that the notion of indicator selection *inherently* suggests a need to consider both measurement and management aspects, in contrast to the more fundamental scientific literature, where management concerns are more typically absent.

In the following review of the literature we do therefore not follow a procedure grouping each *reference* uniquely into a certain category, but seek to organize results according to different *types of criteria* found in the literature across references. Each reference typically suggest criteria for more than one category.

We distinguish between *three different levels of criteria* that reflect different intended functions of the indicators:

- Level 1: Indicators treated as **scientific units** measuring particular system properties or endpoints (subsection 3.2).
- Level 2: Indicators considered as reporting units in **monitoring programs** (subsection 3.3)
- Level 3: Indicators treated as decision making units in **policy or management strategies** (subsection 3.4)

As we shall see different aspects (and hence criteria) are typically considered relevant for each level of function. Level 1 (measurement) is a basis for the two others. A common critique of level 2 and 3 approaches is however that “...management and monitoring programs often lack scientific rigor because of their failure to use a defined protocol for identifying ecological indicators” (Dale and Beyeler 2001). In other words, level 1 criteria would be considered basic level criteria from an ‘indicators as measurement tools’ point of view, while level 2 and 3 could be considered *added levels* (not *replacing* level 1) if the indicators are to be used for monitoring or management purposes. Level 2 and 3 are necessary because measurement is not the final purpose of the indicators. We will return to this distinction again later

The literature about criteria used in the assessment of transport indicators is reviewed separately (in subsection 3.5).

The review of literature did not specifically address the topic of criteria for aggregate indicators, indices, composites, or other forms of joint consideration of indicators. Some references do touch upon various aspects of this topic and a few consider criteria to assess aggregates, (e.g. de Montis et al 2000). The topic is discussed briefly in Section 5,4 of the present report. The more substantial analysis of methodological issues regarding joint consideration of indicators has been addressed elsewhere in the work and is covered in Chapter 6 of the COST 356 Scientific Report.

3.2 Criteria for ‘good’ indicators at the three levels of indicator application

3.2.1 Level 1 - ‘scientific measurement’ of endpoint criteria

Approaches concerning criteria for *indicators as units of scientific measurement* (level 1) typically emphasize how to ensure that an indicator validly and reliably represents key properties in a particular system or endpoint of interest (for example how to select appropriate indicators to describe the state of eutrophication of a lake ecosystem). Examples of references adopting this approach include e.g. Cameron et al 1998 (for soil quality); Breckenridge et al 1995 (for rangelands); Franceschini 2005 (for air quality indices), and Babisch 2006 (for noise).

Table 4 below cite three typical sources that each attempt to summarize which ‘scientific’ indicator criteria are most important for work in the different areas (here: ecosystem and human health).

Jørgensen et al (2005) is a large scientific compendium over indicators and indices for measuring ‘ecosystem health’. Surprisingly the five general ‘scientific’ criteria (Table 4) proposed are not further detailed in the book, but still suggest an interesting summary of what ecosystem health

scientists should be most concerned with when selecting indicators for their analytic work. *Eyles and Furgal (1998)* propose a set of criteria to select indicators of human health effects of ecosystem changes. They distinguish between ‘scientific’ criteria (Table 4) and *use – based* criteria (not shown here). The criteria have been established in a consensus process, and are widely cited by others. The proposed criteria for ‘indicator validity’ consist of elements that have been established mainly in psychology and social sciences (Crocker 2001) (see later in section 3).

The *World Health Organization (WHO)* has several indicator programs for health monitoring. In WHO (2006) indicators for reproductive health are established. In Table 4 the criteria proposed in this report for selecting indicators that are ‘scientifically robust’ are cited. Additional criteria related to the other levels (monitoring and management) follow in Table 5.

As can be seen from the three example there is some overlap in concerns, but not a full consensus about what the ‘scientific’ measurement criteria for indicators are. A basic problems is that indicators in many cases are substitutes for actual scientific models or methods. Hence their ‘scientificness’ will always have some limitation; the ‘validity’, ‘reliability’ etc can typically not be established with the same rigor as in a fully developed scientific model.

Jørgensen et al (2005) (Ecosystem health)	Eyles & Furgal (2000) (Human health in ecosystems)	WHO (2006) (Reproductive health)
<ul style="list-style-type: none"> • Ease of handling • Independence of reference states • Sensibility to small variations of environmental stress • Applicability in extensive geographical areas • Possible quantification 	<ul style="list-style-type: none"> • Data availability, suitability and representativeness - with respect to sampling of populations. • Indicator validity: <ul style="list-style-type: none"> -‘face validity’ (is it reasonable?) -‘construct validity’ (does it behave as expected?) -‘predictive validity’ (does it predict outcomes?) - ‘convergent validity’ (different measures react in same way?) • Reliability (repeatability across times and sources). • Responsiveness to change • Disaggregation capability - across personal and community characteristics. • Comparability -across populations and jurisdictions. • Indicator representativeness - Coverage of important dimensions of concern 	<ul style="list-style-type: none"> • Valid. An indicator must actually measure the issue or factor it is supposed to measure. • Reliable. An indicator must give the same value if its measurement were repeated in the same way on the same population and at almost the same time • Sensitive. An indicator must be able to reveal important changes in the factor of interest • Specific. An indicator must reflect only changes in the issue or factor under consideration.

3.2.2 Level 2 - 'Monitoring' system criteria'

References to *indicators as elements in monitoring systems* (level 2) often do include some level 1 aspects, and then adds various operational criteria related to actually collecting, continuously monitoring, and communicating indicators in a monitoring context. Is it feasible to monitor the indicator? Are data available or can they be obtained? Is it cost-effective?

Another typical concern at this level is how to ensure a comprehensive and non-redundant *suite of indicators* for monitoring a whole system or domain of interest (such as 'sustainability' or 'ecosystem health'). (Niemeijer & de Groot 2008; Dale and Beyeler 2001; Jørgensen et al 2005; Bossel 1996). In this way monitoring criteria highlight an important aspect of relevance for comprehensive assessment (completeness) - although the same concern is actually also implied in some of the individual measurement related criteria (e.g. 'representativeness').

The three sources cited in Table 5 each attempt to summarize which indicator criteria are important in a monitoring context.

Boyle (1998) did a large study about different indicators sets and systems for monitoring the state of ecosystems in Canada. The study is remarkable because it is based on an extensive literature review over ecosystem indicator methodologies, including a many-page appendix section covering the indicator selection criteria literature rather broadly. The entries in Table 5 are those monitoring concerns that Boyle conclusively believe should guide criteria application to indicators. The three first entries of Boyle all deal with the need to devise an appropriately comprehensive *framework*, not individual indicators.

Dale & Beyeler (2003) propose a procedure for selecting indicators for comprehensive monitoring of ecosystems in the US. They note: "In general, ecological indicators need to capture the complexities of the ecosystem yet remain simple enough to be easily and routinely monitored". (Dale & Beyeler 2003, p 6). Here the 'integration' criterion is the one that addresses the need for an appropriate suite of indicators. The Dale & Beyeler criteria are rather widely cited by other references.

WHO (2006) Is the same source as in Table 4 above, but here the other proposed criteria beyond 'scientific robustness' are included, such as 'accessibility' and 'ethics'.

These examples illustrate especially how practical and communication issues enter when the purpose shifts from basic measurement issues to regular monitoring programs. Still measurement aspects are also present, albeit with less detailed specifications than the examples in Table 4.

Table 5 . Level 2 - 'Monitoring' system criteria		
Boyle 1998 (Ecosystem monitoring)	Dale & Beyeler 2003 (Ecosystem monitoring)	WHO 2006 (Monitoring Reproductive health)
<ul style="list-style-type: none"> • Sustainable Management goals and objectives: provide information that is timely • Conceptual model of the system: clearly relate to a specific societal or environmental concern • Issues framework: be clearly relevant to articulated goals and objectives • Knowledge base; be scientifically valid, statistically and analytically sound ,demonstrated to be practical through case studies • Data: use data that are available and accessible, accurate, comparable over time, complete with historical information and covering sufficient geographic area • Reporting: provide information that is understandable to potential users, unambiguous, easy to use; provide information that is at the appropriate scale for decision making 	<ul style="list-style-type: none"> • Be easily measured: The indicator should be straight-forward and inexpensive to measure. • Be anticipatory, i.e. signify an impending change in key characteristics of the ecological system: Change in the indicator should be measurable before substantial change in ecological system integrity occurs. • Predict changes that can be averted by management actions: The value of the indicator depends on its relationship to management actions. • Are integrative: the full suite of indicators provides a measure of coverage of the key gradients across the ecological systems • Have a known response to disturbances, anthropogenic stresses, and changes over time: The indicator should have a well-documented reaction to both natural disturbance and to anthropogenic stresses • Have low variability in response: Indicators that have a small range in response to particular stresses allow for changes in the response value to be better distinguished from background variability. 	<ul style="list-style-type: none"> • Scientifically robust (Valid, Reliable, Sensitive, Specific - see Table 4 above) • Useful: At national level, an indicator must be able to act as a "marker of progress"... the data should also be useful locally, i.e. follow-on action should be immediately apparent • Representative. An indicator must adequately encompass all the issues or population groups it is expected to cover • Understandable. An indicator must be simple to define and its value must be easy to interpret • Accessible. The data required should be available or relatively easy to acquire by feasible data collection methods that have been validated in field trials • Ethical. An indicator must be seen to comply with basic human rights and must require only data that are consistent with morals, beliefs or values of the population.

3.2.3 Level 3 - Policy/management criteria

Publications about *indicators as elements in policy or management strategies* (level 3) usually include some level 1 and 2 aspects, but emphasize in addition criteria related to communication aspects and in particular *to what extent indicators address policy relevant issues*, and to what extent they allow an *assessment of policy responses* or management interventions (OECD 2003; EEA 2004, Kusek & Rist 2004; Segnestam 1999). Hence the concern is less with the role of indicators to provide a comprehensive description of a 'natural system', and more with how they describe targets or measures of relevance in policy and decision making

The three sources cited in Table 6 are widely used or cited as standards for selection of useful indicators for policy or management in the area of environment.

OECD 2003 (Environmental Performance)	EEA 2004 (Environmental Performance)	Segnestam 1999 (Environmental Performance)
<p><u>Policy relevant and useful indicators should:</u></p> <ul style="list-style-type: none"> • provide a representative picture • be simple, easy to interpret and able to show trends over time • be responsive to changes • provide a basis for international comparisons • have a threshold or reference value against which to compare it <p><u>Analytically sound indicators should:</u></p> <ul style="list-style-type: none"> • be theoretically well founded in technical and scientific terms • be based on international standards and international consensus about its validity • lend itself to being linked to economic models, forecasting and information systems <p><u>Measurable indicators are based on data that should:</u></p> <ul style="list-style-type: none"> • be readily available or made available at a reasonable cost/benefit ratio • be adequately documented and of known quality • be updated at regular intervals in accordance with reliable procedures 	<ul style="list-style-type: none"> • Be policy relevant - support EU policies' priority issues • Monitor progress toward the quantified targets • Be based on ready available and routinely collected data within specified timescale at reasonable cost-benefit ratio • Be consistent in space coverage and cover all or most of EEA countries • Time coverage – sufficient/insufficient time trends • Primarily be national in scale and representative for countries • Be understandable and simple • Be conceptually and methodologically well founded and representative; and based on consultation with countries • EEA priorities in management plan • Be timely (be produced in reasonable and "useful" time) • Indicator well documented and of known quality 	<ul style="list-style-type: none"> • Direct relevance to project objectives • Limitation in number. It is most effective to be selective and use smaller sets of well-chosen indicators • Clarity in design. • It is important that the indicator is clearly defined to avoid confusion in the development or interpretation • Realistic collection or development costs • Clear identification of causal links • High quality and reliability • Appropriate spatial and temporal scale • Targets and baselines. To measure the environmental problem at three points in time: before the project begins, during project implementation, and after the project has ended

The *OECD (2003)* criteria in the first column have been used for more than a decade in connection with assessment of environmental policy performance in OECD member states. It is among the most well known criteria sets. It does cover scientific, monitoring as well as management aspects (all levels), although in a somewhat peculiar mix (where 'provide a representative picture' for example is listed under policy relevance rather than 'analytical soundness' while 'linked to economic models' etc is listed under 'analytical soundness, rather than 'policy relevance')

The *European Environment Agency (EEA 2004)* used the criteria listed in the second column to establish its 'core set of indicators' for reporting on the European environment. The purpose was to identify the *best available* indicators for a number of key issues, taken from several large indicator sets already existing ones. Hence there is less of a need to emphasize basic conceptual measurement issues such as 'validity' etc, since the 'core' ones are to be selected only from already well established basic ones; this is illustrated by the criterion 'indicator well documented and of known quality.' There is a strong concern for applicability at the European level.

Segnestam (1999), proposed the criteria listed in the third column for use in the World Bank's assessment of the environmental performance of projects in developing countries. The role of the

indicators in project assessment is strongly highlighted, e.g. measuring 'fulfilment of project objectives', while some measurement aspects are also considered important for this function (e.g. 'design clarity' and 'reliability'). The need to limit the range of an indicator set to make it workable in a management context is also noted.

3.2.4 Comprehensive lists of criteria

Some sources establish particularly comprehensive lists of criteria for use in indicator assessment, covering all of the three levels of criteria. Tables 8, 9 and 10 each illustrates such an example. The three examples refer to application in three different areas (health, fisheries, environment).

NCHOD 2005, see Annex 12, 9 427 ff) (Table 7) is a comprehensive list of criteria used to quality assess indicators by the National Health Service in the UK. The criteria were derived from 18 independent sources, and grouped into 4 categories *scientific criteria*; *policy criteria*; *methodological criteria*; and *statistical criteria* (categories not shown here, but see Table 11). The review of the 18 references show that the 7 most frequently applied criteria are: 'validity', 'policy-relevance', 'measurability', 'comparability', 'data quality', 'data reliability', and 'interpretability' (mentioned by more than 10 of the 18 sources ≥ 10). 'Scientific soundness', 'actionability', 'explicit methodology', 'timeliness', 'frequency', 'sensitivity to change', and 'representativeness' were listed by ≥ 5 sources. Other criteria were less frequently mentioned. Methods for how best to apply each criterion in selecting actual indicators is also briefly proposed (e.g. expert assessment, statistical verification etc), and a case example is used to demonstrate how the assessment may work out (case is 'Hospital Admissions for children with respiratory tract infections').

Rice & Rochet (2005) (Table 8) propose criteria for selecting indicators for the international management of fisheries in the International Council for the Exploration of the Sea (ICES). There are 'only' 9 criteria cited in the article, but for each main criterion a set of subcriteria for assessment apply. This reference and criteria list is particularly interesting, because it specifies detailed operational questions to assess candidate indicator quality for each criterion (albeit detailed and specific for fisheries management). An accompanying article (Rochet & Rice 2005) even reviews to what extent the criteria based approach to select indicators has been helpful for fisheries management.

Niemeijer & de Groot (2008) (Table 9) presents the most comprehensive list of criteria in any one table we have found in the literature. There are 34 criteria in total, divided into 6 categories, 'Scientific dimension', 'Policy and management', 'Systemic dimension', 'Intrinsic dimension', 'Historic dimension' and 'Financial and practical dimensions'. Like NCHOD (2005) the list derives from a review of a range of other references, albeit different ones (here environmental) and only half as many (nine). The most frequently cited criteria are here 'analytical soundness', 'time-bound', 'measurability', 'resource demand', and 'relevance'. The individual criteria are slightly less developed here than in the previous two sources. The main emphasis is on proposing a general procedure for devising sets of indicators for so-called 'causal networks'.

We can note that many criteria overlap, but with slightly or widely different definitions in each list. Also not even the most comprehensive list (Niemeijer & de Groot 2008) contains all of the criteria proposed by the two other comprehensive ones. We will return to the procedural aspects of criteria application proposed by these comprehensive references in Section 5 of this report.

Policy-relevance	Does the phenomenon under measurement represent significant public interest, disease burden or cost?
Actionability	Can the factors which influence the phenomenon be positively influenced to induce a future health / cost benefit?
Perverse incentives	Will the measurement process encourage undesired behaviours by those under measurement?
Explicit definition	Is the indicator explicitly defined by appropriate statistical units of measurement and clinical terminology?
Indicator validity	Will the indicator measure the phenomenon it purports to measure i.e. does it make sense both logically and clinically?
Scientific soundness	How scientific is the evidence / selection process (systematic / non-systematic) to support the validity of the indicator?
Explicit methodology	Are measurement tools / procedures explicitly defined, understood and monitored?
Attributability	Are the factors which influence the phenomenon likely to be identified e.g. patient risk factors, practitioner procedure etc?
Timeliness	What is the average time (months) between measurement and results?
Frequency	What is the average time (months) between reporting of results?
Sensitivity to change	Do the measurement tools and timing of results allow changes to be observed over time?
Confounding	What is the risk that variations between organisations and changes over time may be influenced by confounding factors?
Acceptability	What percentage of stakeholders accept the process of measurement and the reasons for it?
Measurability	Is the measurement process possible within the available budget and resources?
Cost-effectiveness	Does the likely output represent a cost-effective use of budget/resources?
Explicit methodology	Are measurement tools / procedures explicitly defined, understood and monitored?
Specificity	Does the measurement appropriately capture the level of detail required e.g. sub-group analyses, accurate diagnosis?
Comparability	Is the measure comparable between relevant sub-groups e.g. are age/sex/geography-specific data standardised and consistent?
Representativeness	Are sample sizes representative across all required sub-groups
Data quality	Data quality % of the information missing from the records?
Data reliability	% agreement (kappa coefficient) between measured records and those collected by an independent source?
Uncertainty	Have appropriate techniques been selected to demonstrate the effects of variation, dispersion and uncertainty
Interpretability	Can understandable, meaningful and communicable conclusions be drawn from the results?

Table 8. Rice & Rochet 2005. Fisheries Management - ICES	
(IND = Indicator)	
Concreteness	<ul style="list-style-type: none"> • Concrete property of physical/biological world, or abstract concept ? • Units measurable in the real world , or arbitrary scaling factor ? • Direct observations , or interpretation through model ?
Theoretical basis (competing theories to allow contrast is important)	<ul style="list-style-type: none"> • (i) Not contested among professionals ; (ii) basis credible, but debated - can account for patterns in many data sets; (iii) credible, but competing theories have adherents and empirical support is mixed; (iv) adherents, but key components untested or not generally accepted • If IND derived from empirical observations: <ul style="list-style-type: none"> (i) concepts readily reconciled with established theory (ii) concepts not inconsistent with, but not accounted for by ecological theory (iii) concepts difficult to reconcile with ecological theory • Theory allows calculation of reference point associated with serious harm
Public awareness	<ul style="list-style-type: none"> • Is it a property with a high or low public awareness outside the use as an IND? • Does public understanding correspond well or poorly with technical meaning of IND? • If awareness high, is public likely to demand action that is: (i) proportional to IND value as determined by experts ; (ii) disproportionately severe ; (iii) largely indifferent • Does the nature of what constitutes "serious harm" (used to define a reference point) depend on values that are widely shared or vary widely across interest groups ? • Internationally binding agreements, national or regional legislation require that a specific IND be reported at regular intervals , to agreements/legislation require environmental status reporting, but IND not specified to no such requirements
Cost	<ul style="list-style-type: none"> • Uses measurement tools that are widely available and inexpensive to use , to needs new, costly, dedicated, and complex instrumentation
Measurement	<p>Can variance and bias of IND be estimated? Yes ; No</p> <ul style="list-style-type: none"> • If variance can be estimated, is variance low to high • If bias can be estimated, is bias low to high ? • If IND biased, is direction usually towards overestimating risk , or towards underestimating risk If both can be estimated, have variance and bias been consistent over time , or have they varied substantially • Probability that IND value exceeds reference point can be estimated with accuracy and precision , to coarsely or not at all • IND measured using tools with known accuracy and precision , to unknown or poor/ inconsistent • Value obtained for indicator unaffected by sampling gear , to sampling methods can be calibrated , to calibration difficult or not done • Seasonal variation unlikely or highly systematic to irregular • Geographic variation irrelevant or stable and well quantified , through random to systematic on scales inconsistent with feasible sampling • Taxonomic representivity: IND reflects status of all taxa sampled/modelled (High), through ecologically predictable subset of species , to only specific species with no identifiable pattern of representivity
Historical data	<ul style="list-style-type: none"> • Necessary data are available for: periods of several decades to only relatively recent period , to opportunistic or none available • Necessary data are: from the full area of interest , to restricted but consistent sampling sites (Moderate), to opportunistic and inconsistent sources, or none ** • Necessary data have high contrast, including periods of harm and recovery , to high contrast but without known periods of harm and recovery , to uninformative about range of variation expected (Low) • The quality of the data and archiving is known and good , to data scattered with reliability but not systematically certified, and archives not maintained • Data sets are freely available to research community , to private or commercial Holdings
Sensitivity (length of time-series used for testing important)	<ul style="list-style-type: none"> • IND responds to fishing in ways that are: (i) smooth, monotonic, and with high slope; (ii) smooth, monotonic, and with low slope ; (iii) smooth, monotonic over a restricted range of effort characteristics ; (iv) unreliable; depending on when it fails to inform about fishing effects); (v) insensitive or irregular. Magnitude of response does not depend on magnitude of signal in effort
Responsiveness (length of time-series for testing important)	<ul style="list-style-type: none"> • IND changes within 1-3 years of implementation of measures , or IND only reflects system responses to management on decadal scales or longer
Specificity (contrast in data set for testing important)	<ul style="list-style-type: none"> • Is impact of environmental forcing on IND known, and small or strong ? • If environmental forcing affects IND, effect systematic and known , to irregular or poorly understood • Relative to other factors, IND: (i) known to be unresponsive ; (ii) responds to specific factors in known ways ; (iii) thought to be unresponsive (F); (iv) responds to many factors in only partly understood ways

Analytically soundness	Strong scientific and conceptual basis
Credible	Scientifically credible
Integrative	The full suite of indicators should cover key aspects/components/gradients
General importance	Bear on a fundamental process or widespread change
Historical record	Existing historical record of comparative data
Reliability	Proven track record
Anticipatory	Signify an impending change in key characteristics of the system
Predictable	Respond in a predictable manner to changes and stresses
Robustness	Be relatively insensitive to expected source of interference
Sensitive to stresses	Sensitive to stresses on the system
Space-bound	Sensitive to changes in space
Time-bound	Sensitive to changes within policy time frames
Uncertainty about level	High uncertainty about the level of the indicator means we can gain something from studying it
Measurability	Measurable in qualitative or quantitative terms
Portability	Be repeatable and reproducible in different contexts
Specificity	Clearly and unambiguously defined
Statistical properties	Have excellent statistical properties that allow unambiguous interpretation
Universality	Applicable to many areas, situations, and scales
Costs, benefits and cost-effectiveness	Benefits of the information provided by the indicator should outweigh costs of usage
Data requirements and availability	Manageable data requirements (collection) or good availability of existing data
Necessary skills	Not require excessive data collection skills
Operationally simplicity	Simple to measure, manage and analyze
Resource demand	Achievable in terms of the available resources
Time demand	Achievable in the available time
Comprehensible	Simply and easily understood by target audience
International compatibility	Be compatible with indicators developed and used in other regions
Linkable to societal dimension	Linkable to socio-economic developments and societal indicators
Links with management	Well established links with specific management practice or interventions
Progress towards targets	Links to quantitative or qualitative targets set in policy documents
Quantified	Information should be quantified in such a way that its significance is apparent
Relevance	Relevance for the issue and target audience at hand
Spatial and temporal scales of applicability	Provide information at the right spatial and temporal scales
Thresholds	Thresholds that can be used to determine when to take action
User-driven	User-driven to be relevant to target-audience

3.2.5 Popularized ‘acronym’ lists of criteria

Quite a number of sources try to provide comprehensive but condensed lists of indicator criteria with seemingly intelligent catchword acronyms such as SMART (Broughton & Hampshire 1997), CREAM (Kusek & Rist (2001), SUMIR (Cameron et al 1998), or PICABUE (Mitchell et al 1996). Three examples are illustrated in Table 10.

The acronyms themselves seem to be intended mainly as memory assistance for supposed indicator users. The potential added value of these condensed lists beyond that would be their efforts to boil down many criteria to fewer ‘key’ ones, but unfortunately the references are usually not informative about the basis for this selection. Some ‘acronym’ lists may even confuse more than they assist. For example the item in the ‘SMART’ list ‘Attainable’ would hardly refer to an indicator as such, but more to any target set as a benchmark to measure an indicator against. All in all the acronym lists should be considered mostly as well—intended curiosa.

	Broughton & Hampshire (1997)		Kusek & Rist (2001)		Cameron et al (1998)
S	Specific. Key indicators need to be specific and should relate to the conditions the project seeks to change	C	Clear. Precise and unambiguous	S	Sensitivity. Sensitivity of indicator to degradation or remediation process
M	Measurable. Each indicator should be measurable and hence requires a precise definition	R	Relevant. Appropriate to the subject at hand	U	Understanding. Ease of understanding of indicator value
A	Attainable. The indicator must be attainable at reasonable cost using an appropriate collection method	E	Economic. Available at a reasonable cost	M	Measurement. Ease and/or cost effectiveness of measurement of indicator
R	Relevant. Indicators should be relevant to the management information needs of the people who will use the data	A	Adequate. Provide a sufficient basis to assess performance	I	Influence. Predictable influence of properties on soil, plant and animal health, and productivity
T	Timely. An indicator needs to be collected and reported at the right time to influence many management decisions	M	Monitorable. Amenable to independent validation	R	Relationship. Relationship to ecosystem processes (especially those reflecting wider aspects of environmental quality and sustainability).

[gap on top of next page is not intended – caused by some automatic MS Word function]

NCHOD 2005 (Clinical Health)	Niemeijer & de Groot 2008 (environment)	Jackson et al 2000 (ecosystems)	OECD 2003 (env. policy)
Scientific criteria	Scientific dimension	Conceptual Relevance	Analytically sound
<ul style="list-style-type: none"> • Explicit definition • Indicator validity • Scientific soundness 	<ul style="list-style-type: none"> • General importance • Credible • Analytically soundness • Integrative 	<ul style="list-style-type: none"> • Relevance to the Assessment • Relevance to Ecological Function 	<ul style="list-style-type: none"> • Theoretically well founded • Based on international standards and consensus • Linkable to economic models, forecasting etc
Policy Criteria	Policy and management	Feasibility of Implementation	Policy relevant and useful
<ul style="list-style-type: none"> • Policy relevance • Actionability • Perverse incentives 	<ul style="list-style-type: none"> • Relevance • Comprehensible • International compatibility • Linkable to societal dimension • Links with management • Progress towards targets • Quantified • Relevance • Spatial and temporal • Thresholds • User-driven 	<ul style="list-style-type: none"> • Data Collection Methods • Logistics • Information Management • Quality Assurance 	<ul style="list-style-type: none"> • Representative • Simple, easy to interpret • Responsive • International comparison • Threshold or reference value
Methodological criteria	Systemic dimension	Response Variability	Measurable
<ul style="list-style-type: none"> • Explicit methodology • Attributability • Timeliness • Frequency • Sensitivity to change • Confounding • Acceptability • Measurability • Cost-effectiveness • Explicit methodology 	<ul style="list-style-type: none"> • Anticipatory • Predictable • Robustness • Sensitive to stresses • Space-bound • Time-bound • Uncertainty about level 	<ul style="list-style-type: none"> • Estimation of Measurement Error • Temporal Variability - Within Season • Temporal Variability - Across Years • Spatial Variability • Discriminatory Ability 	<ul style="list-style-type: none"> • Available at reasonable cost/ • Documentation • Updated/ reliable procedures
Statistical criteria	Intrinsic dimension	Interpretation and Utility	
<ul style="list-style-type: none"> • Specificity • Comparability • Representativeness • Data quality • Data reliability • Uncertainty • Interpretability 	<ul style="list-style-type: none"> • Measurability • Portability • Specificity • Statistical properties • Universality • Measurability 	<ul style="list-style-type: none"> • Data Quality Objectives • Assessment Thresholds • Linkage to Management 	
	Historic dimension		
	<ul style="list-style-type: none"> • Historical record • Reliability 		
	Financial and practical dimensions		
	<ul style="list-style-type: none"> • Costs, benefits and cost-effectiveness • Data requirements and availability • Necessary skills • Operationally simplicity • Resource demand • Time demand 		

3.2.6 Summary of the general criteria review (levels 1- 3)

We can make the following observations from the literature reviewed in the above sections:

A large number of possible indicator selection criteria have been proposed in the literature. There are many similarities and repeated items with regard to indicator selection criteria across the different references (different levels, different domains). The definitions of the criteria are often quite brief, while methods to assess indicators based on the criteria are sometimes (but often not) specified in the references. It is not uncommon that the same criteria are defined in different ways

Some of the most frequently mentioned criteria with regard to 'indicators as measurement tools' include the following:

- Validity (is the right thing being measured with regard to concept, theory, domain, scope, use?)
- Reliability (is it measured in the right way?)
- Representativity (does indicator 'represent' the object or the problem? Often with same meaning as validity; sometimes refers to a range of indicators)
- Theoretical basis (is there a theory behind the indicator? Sometime element part of 'validity')
- Sensitivity (is the indicator responsive to stress on the entity to be indicated?)
- Explicitness and consistency of methodologies (are the methods used comprehensible, reproducible, comparable, transparent?)

'Data availability', 'cost-effectiveness', 'timeliness' and 'understandability' are also frequently mentioned criteria of relevance for measurement, but they are typically considered more as a practical (operational) concerns in a context of monitoring or management. 'Policy relevance' and 'target-linked' also appear frequently, and are clearly related to user and decision making aspects.

We will return to discuss and suggest more specific definitions for the criteria to be used in section 4 of this report.

It seems evident that various criteria may cater more to some situations than others. Some criteria would for example be particularly relevant for the outset of an indicator process ('level 1' in section 2), where basic measures have to be defined, taking into account conceptual and scientific considerations of appropriate representation, via measurement, calculation, modelling etc. Other criteria refer to situations where monitoring or policy programs need criteria (levels 2 and 3) to select among potential indicators that have already been defined, calculated and produced at conceptual and measurement levels.

As shown in Table 11 there is however only limited consensus about which *categories* to use to classify or typologise indicator selection criteria. For example the same criterion, 'responsiveness/sensitivity' is classified under completely different categories, like in this example under 'policy relevance' (OECD 2003); 'systemic dimension' (Niemeijer & De Groot 2008); 'methodological' (NCHOD 2005) or its own category 'Response variability' (Jackson et al 2000).

We will return to review and discuss methods for grouping, application, and ranking of criteria in section 5.

3.3 Criteria in transport applications

Below we illustrate some sets of indicator criteria that have been proposed or applied in various *transport assessment* publications, including some dealing with sustainable transport indicators specifically. The references here often cut across the three levels, although many of the transport references are primarily concerned with criteria that they see as important for policy and management related functions (the third level).

3.3.1 Selection criteria defined in COST Action 350

It is natural to start with the work in COST 350, which is a predecessor and direct inspiration for COST 356. Meanwhile it must also be considered that COST 350 we confined to indicators useful for Strategic Environmental Assessment of transport policies and projects, while 356 has a much broader scope as well as a more methodological aim.

COST 350 defined nine overall criteria to be used in their selection of indicators for Integrated assessment of environmental impact of traffic and transport infrastructure (Goger et al 2006). Five of the criteria were defined as 'general', while four were defined as 'strictly linked to the goals of COST 350'. In Table 12, the criteria are listed, together with the definitions or explanations of each indicator, as described in (Goger et al 2006).

Criterion	Descriptions
General	
Significance	<ul style="list-style-type: none"> • The importance of the indicator • How good it is to provide the basis for the evaluation of actions and plans • How well it provides an early warning of potential problems • How well it demonstrates a move towards or away from sustainability • How well it covers the targets • How well it is to give prognosis; the ability to evaluate long term effects of the plan
Completeness	<ul style="list-style-type: none"> • How well the indicator covers the different parameters of the DSIPR framework • How well the whole set of indicators issues the impact pressure of the project • What relation exists between the different indicators (non-redundancy)
Simplicity and applicability	<ul style="list-style-type: none"> • How well the indicator can be calculated using easy tools • How well it can be calculated, during updating in the years, using easy tools • How is the number of indicators relatives to same topic (lower is better) • How well it can be calculated using simple data that are easily achievable in the terms of money and time and, above all, that are at a raw level (non elaborated)
Scientific validity	<ul style="list-style-type: none"> • How well the indicator will describe the impacts effectively • How well will it describe the impacts precisely • How big the consensus on the validity of the indicator is • How well it can be calculated, avoiding errors due to the calculation methods; reliability in avoiding bias
Transferability	<ul style="list-style-type: none"> • In time • How well the indicator can be used in different time periods (past, present, short and long term future) • How well it performs to provide a basis for comparison across time • In space • How well it can be used in different geographical areas maintaining its performance • How well it can be used in a standardised way at different geographical scales
Strictly linked to the goals of COST 350	
European rules-oriented	<ul style="list-style-type: none"> • How well the indicator follows the European rules and how well it covers the targets.
Transport oriented	<ul style="list-style-type: none"> • How is the responsibility of the transport sector in the considered impact evaluated by the indicator • How well the indicator shows the contribution of the transport sector in the considered impact evaluated by the indicator
SEA-oriented	<ul style="list-style-type: none"> • How good the indicator is to provide a basis for actions and plans • How well the indicator assesses the environment on the strategic level

Decision making oriented	<ul style="list-style-type: none"> • How useful the indicator is for the end-users (decision makers) • How well it is comprehensible to the public/decision makers
--------------------------	--

The list represents a mix of measurement related issues, policy making issues, and practical considerations. Some criteria refer to individual indicators, others (e.g. 'Completeness') refer more to sets of indicators.

For each example some comments are offered to highlight aspects that may be of particular relevance for COST 356, or where there are some specific issues that are not fully addressed in the general criteria sets.

Criteria such as 'European Rules-related' or 'SEA-related' are not directly relevant for COST 356, even if they are of course useful to identify indicators for such specific applications.

'Transport-oriented' on the other hand, is highly relevant, and we will return to this in the summary of section 3.3 and further on.

3.3.2 General traffic measurement indicator criteria

The quality criteria for traffic measurements in Table 13 (Batalle et al 2004) are from the US. They are not intended to apply directly to indicators, but to identify good quality traffic measurement data in general. This is why there is no reference to any particular problem or concept that is to be indicated (for example 'optimal mobility' or 'sustainable transport'). Hence there is not so much explicit consideration of for example conceptual aspects.

Anyhow, most of these measurement criteria are clearly described and defined, providing some more detail than e.g. the COST 350 list in Table 12. The criteria seem relevant for measurement aspects of sustainable transport indicators.

Table 13. General data quality for traffic measurement (Batalle et al 2004)
Accuracy – The measure or degree of agreement between a data value or set of values and a source assumed to be correct. It is also defined as a qualitative assessment of freedom from error, with a high assessment corresponding to a small error.
Completeness (also referred to as availability) – The degree to which data values are present in the attributes (e.g., volume and speed are attributes of traffic) that require them. Completeness is typically described in terms of percentages or number of data values.
Validity – The degree to which data values satisfy acceptance requirements of the validation criteria or fall within the respective domain of acceptable values. Data validity can be expressed in numerous ways. One common way is to indicate the percentage of data values that either pass or fail data validity checks.'
Timeliness – The degree to which data values or a set of values are provided at the time Required or specified. Timeliness can be expressed in absolute or relative terms.
Coverage – The degree to which data values in a sample accurately represent the whole of that which is to be measured. As with other measures, coverage can be expressed in absolute or relative units.
Accessibility (also referred to as usability) – The relative ease with which data can be retrieved and manipulated by data consumers to meet their needs. Accessibility can be expressed in qualitative or quantitative terms.

3.3.3 Traffic Safety Indicator criteria

The criteria in Table 14 (Farchi et al 2006) have been applied by the so-called *European Road Accident Indicator Working Group* in a project to identify appropriate indicators to track the health effects of traffic.

Table 14. Traffic Safety Indicators at EU level (Farchi et al (2006))
A clear and commonly accepted Definition
Association with other Public Health indicators
Relevance
Power of discernment (ability to detect small changes in the phenomenon)
Sensitivity (depending on the source: % of detected cases on total existing cases)
Comparability in time
Comparability between countries
Timeliness (time elapsed from the event to the publication of the indicator)
Availability of information
Stability (how much is influenced by other factors, not regarding road accident field?)
Continuity (how long are the historical series for the indicator available?)
Cost effectiveness
Theoretical validity (how well the indicator represents the subject of interest)
Reliability (depending on the source: how good and valid is the figure given by the indicator)
Interpretability
Coverage (is the indicator available for all countries?)

Some of the criteria are self-explanatory, while others are less so (for example what is meant by 'Association' and 'Interpretability' is not quite clear). The source paper does not offer additional explanations to each criterion. Reportedly the Working group has used the criteria in their indicator selection process. It can be noted traffic safety indicators presumably are more widely adopted than sustainable transport ones, and that lessons from that area may be learned.

The criteria are quite similar to more general ones in section 4.2 and from COST 350 above.

'Reliability' and 'Sensitivity' are mentioned. 'Validity' and 'Transparency' are not explicitly mentioned. The criteria 'Power of discernment' (ability to detect small changes in the phenomenon) and 'Stability' (how much is influenced by other factors, not regarding road accident field) may be critical for distinguishing transport impacts as a part of overall impacts.

3.3.4 Environmental Indicators for transportation in the US

The criteria in Table 15 (US EPA 1999) stem from a report about potential indicators for transport and environment in the US, produced for the US EPA.

The report presents a large number of tables and graphs with existing actual data for a wide range of transport impacts, which according to the report could potentially be used to measure and monitor transport and environment at the national level in the US. However, there is no final selection of indicators, and no follow up in terms of monitoring is reported.

The criteria do not directly include the key general ones, validity, reliability, sensitivity, transparency, etc. This may be because the report focus on existing data sets, where it may be assumed

that general indicator criteria have already been applied, or because the report, as said, has not been the basis for defining an actual indicator system.

Instead there are some other important considerations of specific relevance for sustainable transport indicators.

First, there is the emphasis on the need to *link transport with end results or impacts (outcomes)*. It is considered more environmentally relevant to look at indicators for impacts than at for example outputs (or pressures) such as emissions. This corresponds to the idea in COST 356 to focus on impacts as the end point.

Secondly the criteria highlight the need to single out *transport's contribution to the problems*, as part of overall impacts. This is also highly relevant for COST 356. However, no general method for this is proposed.

It stand out as like a significant challenge to combine those two criteria ('outcome focus' + 'transport specificity' = 'transport specific outcomes') since the causal chain from transport to impact is sometimes long and indirect. The report illustrates that most of the available data/potential indicators in the US actually did not accomplish this combination. They are either about impact in general or about transport shares of outputs (pressures) only.

The US EPA report does not resolve this issue, but points to an important aspect to consider for COST 356,,: how to make those ends meets; how to find 'compromises'.

Otherwise, the proposed criteria (Table 15) are not very elaborate.

Table 15. Environmental Indicators for US transportation (US EPA (1999))
Focus on end results . Information should be provided on outcomes, such as number of illnesses caused, not outputs or activities that cause outcomes
Isolate transportation's share of the impact — The indicator should identify the effect of transportation rather than providing an estimate of environmental quality that may depend on numerous sources
Be reasonably certain — Although modelling may be necessary to estimate the national effect, the indicator should be generally agreed upon as reasonably accurate and reliable
Be stated in meaningful units — The indicator should be presented in units that allow comparison to other sources of a problem or to a goal

3.3.5 Sustainable urban infrastructure

The report referred to in Table 16 and Figure 1 (Lahti et al 2006) presents a set of indicators that was used to characterize a range of urban infrastructure case studies with regard to their sustainability performance (from COST Action C8). It is not just about transport but transport is part of the project. Table 16 lists the indicator criteria used in the selection.

The criteria were used in a process to establish indicators that were then used to characterise and assess the individual project examples or cases. For each indicator is given a direction and a colour code for easy interpretation (down-red= negative trend, green-up =positive trend).

Table 16. Sustainable urban infrastructure in Europe (Lahti et al 2006)
Dimension (economic, social, environmental)
Geographic coverage (local, regional, global)
Time frame (short term effects, long term effects)
Linkage; (direct, primary, indirect, secondary)
Data Availability (Yes, No)
Sector (general, sectoral)
Relevance to EU policy (Yes, No)

Most of the indicator selection criteria in Table 16 are not *evaluative* but *descriptive* at the level of the individual indicators. This means that those criteria cannot serve directly to *select* among particular indicators, but could rather help to ensure that an assessment as a whole (a range or set of indicators) will cover all of the desired aspects and dimensions. Hence the criteria might be most useful in COST Action 356 WG3 concerning with policy relevance.













ECOLOGY		ECONOMY		SOCIAL ASPECTS	
Are emissions to air, water and soil within the restrictions set locally and internationally? Are the emissions decreasing?		Is the cost/ effectiveness/ and or cost/ benefits of the system reasonable compared to other systems? Compared to other needs in the city and to political goals?		Has the planning and decision-making for the infrastructure system been done in a democratic and participative way?	
Is the use of natural resources reasonable compared to other comparable systems? Is the use decreasing?		Are the citizens willing to pay for the services offered? Are the services affordable to all citizens?		Is the function and the consequences of the system transparent to and accepted by the citizens? Is the system promoting responsible behaviour by the citizens?	
Is the system allowing a reasonable bio-diversity with regard to the kind of area studied? Is the bio-diversity increasing?		Is the organisation(s) that finance maintain and operate the system effective?		Is the system safe to use for the citizens? (hazards, health, well-being)	
Is the system more or less sustainable than a conventional system regarding ecology?		Is the system more or less sustainable than a conventional system regarding economy?		Is the system more or less sustainable than a conventional system regarding social aspects?	

Figure 1. General version of the assessment matrix applied in COST Action C8 (Lahti et al 2006)

3.3.6 Sustainable transportation indicators for Canada

Table 17 shows selection criteria used in a project to define sustainable transport indicators for Canada, undertaken by the Centre for Sustainable Transportation (CST) (Gilbert et al 2002).

The project was aimed at practical monitoring purposes (with the 'Transport Canada' branch of the Canadian government as the client). The report is therefore mostly concerned with the selection of indicators that are policy relevant and have a sufficient coverage and representativity for national monitoring.

Rather than specifying any technical measurement criteria, there is a recommendation to use 'reputable' sources of data, where such criteria have supposedly already been applied. In this respect the limited criteria list of this project resembles the list from the US EPA report (section 4.3.4).

Table 17. (Sustainable transportation indicators for Canada (Gilbert et al 2002)
A qualifying variable should concern sustainable transportation , as elaborated in COST's definition, or provide a clear answer to one of the seven questions [Table 18]
A qualifying variable should be a time series , so that information would be provided on Changes in performance.
A qualifying variable, to the extent possible, should represent all of Canada .
A qualifying variable should come from what the project team considers to be a reputable and reliable source , usually a federal government source for Canada-wide data.

The policy relevance is to be ensured by making sure that indicators that address all the key policy questions are applied. Table 18 list the sustainable transport policy questions that are considered most salient by this study. These questions appears to be directly drawn from the European TERM indicator system.

Table 18. STPI topics and questions (Gilbert et al 2002)
1. Environmental and health consequences of transport Is the performance of the transport sector improving in respect of its adverse impacts on environment and health?
2. Transport activity Is transport activity changing in directions consistent with positive answers to the other questions?
3. Land use, urban form and accessibility Are land use, urban form, and transportation systems changing so as to reduce transportation effort?
4. Supply of transport infrastructure and services Are we increasing the efficiency of use of current infrastructure and changing the infrastructure supply in sustainable ways?
5. Transportation expenditures and pricing Are the patterns of expenditure by governments, businesses, and households, and the associated pricing systems, consistent with moving towards sustainability?
6. Technology adoption Is technology being used more in ways that make vehicle transport systems and their utilization more sustainable?
7. Implementation and monitoring How effectively are environmental management and monitoring tools being used to support policy- and decision-making towards sustainability?

3.3.7 Criteria for Sustainable Transportation assessment, USA

Zietsman and Rilett (2002) develop performance measures that are applied in a transport corridor assessment study in the US. The criteria in Table 19 are derived from a broad range of studies, and reflect most of the general criteria that are contained in the general literature. As such it is one of the most comprehensive of the transport criteria sets we have identified, and perhaps the most demanding one to apply in practice.

Their list has several criteria that aim to help discern the influence of transport on the environment ('Able to discriminate'; 'Appropriate level of detail'; 'Not influenced by exogenous factors'; 'Sensitive'). A specific item included in Table 19 but not directly in the previous lists is 'Acceptable'. It suggests that the general community who will be affected must assist in identifying and developing the performance measures. 'Multidimensional' could be relevant for sustainable transport de-

cision making generally. ‘Able to integrate’ may not necessarily be a requirement for each individual indicator.

Table 19. Zietsman & Rilett (2002)

Quality	Explanation
1. Able to discriminate	Must be able to differentiate between the individual components that are affecting the performance of the system.
2. Able to integrate	Must be able to integrate the sustainability aspects of environmental, social, and economic sustainability.
3. Acceptable	The general community must assist in identifying and developing the performance measures.
4. Accurate	Must be based on accurate information, of known quality and origin.
5. Affordable	Must be based on readily available data or data that can be obtained at a reasonable cost
6. Appropriate level of detail	Must be specified and used at the appropriate level of detail and level of aggregation for the questions it is intended to answer
7. Have a target	Must have a target level or benchmark against which to compare it.
8. Measurable	The data must be available, and the tools need to exist to perform the required calculations.
9. Multidimensional	Must be able to be used over time frames, at different geographic areas, with different scales of aggregation, and in the context of multimodal issues.
10. Not influenced	Must not be influenced by exogenous factors that are difficult control for, or that the planner is not even aware of.
11. Relevant	Must be compatible with overall goals and objectives
12. Sensitive	Must detect a certain level of change that occurs in the transportation system
13. Show trends	Must be able to show trends over time and provide early warnings about problems and irreversible trends.
14. Timely	Must be based on timely information that is capable of being updated at regular intervals.
15. Understandable	Must be understandable and easy to interpret, even by the community at large.

3.3.8 Local sustainable transport indicator criteria, UK

Marsden et al (2005) develop indicators for sustainable transport at the local level in the UK. In the process to identify indicators they draw on criteria from a wide range of sources inside and outside of transport research (see Table 20). Their review generally yields the same types of criteria as we have seen in the general and the previous transport specific references. The criterion ‘controllable/attributional’ again highlights the need to be able to separate out the specific (in this case transport) effects from general ones. Ones like ‘limited in number’ is a criterion for a whole set, not at the individual indicator level.

Two new types of criteria are added to the previous ones, namely ‘Avoids perverse incentives/corruption’, and ‘Allows innovation’. Both criteria aim to help avoid indicators that can mislead policy action or management. The criterion to ‘avoid perverse incentives’ is about ensuring that one or a few indicators would not get all the attention at the expense of other important aspects of the problem. The criterion to ‘allow innovation’ means that performance indicators should

not be so detailed that they prohibit any new solutions being considered. Both criteria relate to general management and policy performance aspects of indicator use, and are not specific to (sustainable) transport applications.

Refs: see Marsden et al (2005)	Audit Commission (2000)	Education Fitz-Gibbon (1996)	Sustainability indicators PASTILLE (2002)	Local Authority Carlin (2004)
Relevant to the organisation/ strategy	X	X	X	
Clearly defined/ easy to understand/ transparent	X		X	X
Based on available data/ measurable				X
Controllable/ Attributable	X	X		
Cost Effective	X		X	
Limited in Number		X		X
Timely	X	X	X	X
Avoids perverse incentives/ non corrupting/ not corruptible	X		X	
Statistically/ Scientifically Valid	X		X	X
Comparable/ consistent over time	X	X		X
Responsive	X			
Allows innovation	X			
Capable of aggregation				X

3.3.9 Summing up transport criteria

Summing up transport applications of indicator criteria we can see that many of the general criteria at level 1,2, and 3 reported in section 4.2 are reproduced in various references for the transport area and thereby obviously also relevant for this context.

The most significant additional element highlighted by the transport references vis a vis the general criteria is to consider if the chosen indicators adequately *reflect the responsibility of the transport sector in the considered impact* (Goger et al 2006), or, phrased differently, to what extent the indicator is *able to identify the transport part of the general impacts*, and thereby how to *separate the transport parts from other parts* of the problem (US EPA 1999; Farchi et al 2006; Zietsman & Rilett 2002).

One example of application of such a ‘*transport responsibility*’ criterion could be to assess if an pressure indicator can be split into a general part and a transport specific part (e.g. the transport share of overall emissions), or even better, if the indicator can be disaggregated further into contributions from different transport modes, vehicle types, travel purposes, etc. If it is possible to add such dimensions to an indicator, then this criterion would suggest to choose such indicators rather than ones where it is not possible to distinguish the transport parts of the problem from the general part. Hence it is relevant to add a ‘transport specificity’ criterion to the list of general criteria.

This leads to another issue that is brought forward by the transport references, namely the tension between indicators with a clear ‘transport’ focus versus ones with a clear ‘impact’ focus. In transport planning, transport- reflective indicators are typically chosen (e.g. Vehicle Kilometres Travelled) because they are responsive to transport policies or projects, while such indicators do often not reflect very clearly any specific environmental impact. Hence, in the example above, it is typically easier to identify a ‘transport’ part of the problem by using a ‘pressure’ type indicator (e.g. emissions) than by using ‘state’ indicators (e.g. concentration in air) let alone a direct ‘impact ‘indicator’ (e.g. number of people with health damage from transport pollutant X). In the latter case sophisticated calculation or modelling may be needed to identify the ‘transport share’, and data or models to ensure this may not always be available. Conversely, In the former case, transport specific data may be available (e.g. emission values per vehicle type or speed class), but such data do not indicate very accurately the actual health impacts. Indicators or both types may be needed.

There may be a few examples where it is clear that more or less all of an impact stems from transport because transport is the only source contributing to a particular impact. Examples could perhaps be noise disturbance indicators for people living near roads, or health effects of drinking water contaminated by Methyl Tertiary Butyl Ether (MTBE) leaked from underground gasoline storage tanks. But even in those cases there will also be other sources affecting the final impact endpoint (e.g. human health) .

There may also be examples where only alternative transport projects are compared, and the most important criterion for choosing indicators will in that case be if they are able to illustrate the *differences* in impact between the two cases, not if they can separate out a general ‘transport share’. Still, these tensions remains if the assessment is concerned with identifying differences in actual impact. If only relative performance is of interest, a ‘transport sensitivity’ criterion is less relevant, but if an absolute level is of importance (e.g. with regard to passing a threshold) then it might be critical to be able to identify the transport share or contribution. Therefore a ‘transport sensitivity’ criterion may sometimes, but not always be needed.

The point here is simply to note that such tension may exist, in terms of what an indicator represents; a trade-off may have to be made between choosing indicators that represents the transport cause of a problem versus ones that represent a final effect. Whatever is most important is likely to depend on the specific decision making context; what kind of plan or policy is considered, and what stage the decision making is in.

Finally, and following from the above, most of the transport examples above are not only concerned with accurate representation in a measurement sense, but also concerned with usefulness of indicators in a particular monitoring or management context (reflected by criteria such as policy relevance, links to targets, links to relevant legislation, avoidance of perverse management incentives, etc). Transport impacts are generally not measured for themselves, unless as is part of a monitoring or policy program or transport project. This emphasizes that user and policy criteria may should be considered alongside scientific and measurement aspects before finally selecting indicators for sustainable transport.

4 Developing indicator criteria for COST 356 and EST

In order to move forward this section compares the criteria described in the literature as reviewed in section 3 with the results of the initial work reported in Section 2 (Table 2) in order to identify a preliminary set criteria for EST indicator assessment to be potentially applied. The aim is to steer towards a set of criteria for EST indicators that are comprehensive, consistent, confounded and customized to COST 356 and EST needs.

The section progresses in the following steps:

Subsection 4.1 compares the literature criteria with the ones from the preliminary list derived in Table 2 and Table 3 in section 2: What is missing in the initial work, when looking at the literature?

4.2 compares the other way around: What was included in Table 2 and 3, that was not found in the literature? Does those elements need to remain, or can they better be skipped?

4.3 reviews the terminology and definitions tentatively set up in Table 2 and Table 3 in comparison with 'official' criteria and associated definitions in the literature. Several suggestions to revise or replace wordings of Table 2 are made.

4.4 provides a list of revised intermediate criteria with definitions.

4.5. reports conduct an internal exercise to test out the criteria and their usefulness

4.6 returns to discuss the categories, definitions and consolidation and reach a final suggested set of criteria with associated definitions and application examples for the subsequent work.

4.1 Correspondence of criteria: literature compared with Section 2 lists

First of all several criteria from the general literature were not 'discovered' or directly included in the initial work reported in task in Table 2 and Table 3. This goes for the following criteria.

4.1.1 Validity and validation

As also noted in section 2.2 most prominent among the apparently missing criteria in Table 2 was '**validity**', a concept that is widely used in the indicator criteria literature, and put forward in that literature as a fundamental requirement for indicator quality. WHO (2006) defines validity most simply: 'An indicator must actually measure the issue or factor it is supposed to measure'.

In Table 2 the concept of 'representativity' is formulated in away so it basically means the same as 'validity' defined in this way. ('To which degree the indicator is representing the phenomenon it is developed for'). However, validity contains several aspects beyond this simple notion of 'representativity'

NCHOD (2005) adds a bit more substance: "Will the indicator measure the phenomenon it purports to measure i.e. does it makes sense both logically and clinically?" Hence a distinction between logical (conceptual/theoretical) validity and some form of empirical, or practical (clinical) validity is introduced here.

Eyles & Furgal (2000) mention 'Coverage of important dimensions of concern' in their discussion of validity, and then goes further to introduce distinctions between various types of 'validities':

- 'face validity' (is it intuitively reasonable?)
- 'construct validity' (does it behave as expected?)
- 'predictive validity' (does it predict outcomes?)
- 'convergent validity' (different measures reacting in same way?)

Each of these notions have specific definitions and associated assessment methodologies in the technical measurement literature in e.g. psychology and the social sciences (e.g. Crocker 2001; Leviton 2001; Bollen 2004). 'Face validity' for example means an immediate (non-scholarly) assessment of plausibility. 'Construct' validity, on the other hand may apply a range of statistical techniques to assess whether the indicator is actually measuring variations on the phenomenon (construct) it is supposed to relate to - and not ones it is expected not to (Bollen 2004, p 7285).

Several other 'validities' exist or are proposed in the literature, for example 'internal' validity, which describe causal correspondence between an indicator and the phenomenon it directly measures, while 'external', validity refers to what other entities the indicator may be generalised to. (Leviton 2001, p 5195).

Innes (1990, p 215) suggest that validity is the most important criterion for an indicator, but unfortunately also an elusive concept to test for. Innes (1978) talks about 'operational', 'experimental', and 'theoretical' validity, as increasingly powerful notions, while Goertz (2001) invents the term 'concept-indicator' validity to denote the intrinsic correspondence between the theoretical structure of the indicator and the concept it is supposed to measure

Other authors again propose particular *strategies for 'validation' of indicators* (as opposed to validation of more 'normal' scientific products such as models or theories or observations). Bockstaller & Girardin (2002), for example, suggest three procedures, namely 'design validation', which evaluates if the indicators are founded in scientific theory; 'output validation', which assesses the (empirical) soundness of the indicator outputs, and, 'end use validation' evaluating the usefulness of the indicator as a decision support tool. Each type of validation will require different sets of methodologies (statistics, expert panels etc)

Similarly, Cloquell-Ballester et al (2006) talks about three types, 'self-validation' within an indicator working group; 'scientific validation', which involves external experts, and 'social validation', that includes public participation. "These three validation stages are complementary, so the indicators' credibility increases as they overcome the different validation stages (Cloquell-Ballester et al 2006, p 82)"

We will not go into further into the various validity and validation concepts here, but we must note the concept of validity as an important, but complex one that needs to be incorporated somehow. We will return to processes of indicator selection in Section 5.

4.1.2 Sensitivity and specificity

Another important criterion seeming to be missed in Table 2 is '**sensitivity**' (or 'responsiveness') – the ability to reveal important changes in the factor of interest (several sources have this).

This is particularly important combined with '**specificity**' – or '**attributability**' – the ability to reflect only changes in the issue or factor under consideration (WHO 2006; NCHOD 2005; Marsden et al 2005), or inverted '**Confounding**' - the risk that variations may be influenced by confounding factors (NCHOD 2005), or, stated otherwise, not being influenced by '**...exogenous factors** that are difficult control for, or that the planner is not even aware of' (Zietsman & Rilett 2002).

This relates again closely to the ability to '**Isolate transportation's share of the impact**' (US EPA 1999), which should be a key concern for COST Action 356. We propose to use the term '*transport sensitivity*'.

4.1.3 Other 'new' criteria

Additional potentially relevant criteria found in the literature review, that are not directly mentioned in the Table 2 headings include:

‘Anticipatory’, signify an impending change in key characteristics of the ecological system: Change in the indicator should be measurable before substantial change in ecological system integrity occurs. (Dale & Beyeler 2003).

Comment: This partly covered in Table 2 under ‘certainty’

‘Ethical’. An indicator must be seen to comply with basic human rights and must require only data that are consistent with morals, beliefs or values of the population (WHO 2006).

Comment: This may be less relevant in the case of indicators for sustainable transport (compared with health issues generally), but it could be included with management and policy issues.

‘Actionability’ - Can the factors which influence the phenomenon be positively influenced to induce a future health / cost benefit? (NCHOD 2006).

Comment: This is mostly related to the management and decision aspects, where it would be important to consider e.g. if indicators are measuring elements that may be influenced through policy or management actions or not.

‘Monitor progress’ toward quantified targets, management objectives, thresholds or reference values (OECD 2003; EEA 2004, Boule 1998)

Comment: This is related to management and decision aspects, where monitoring compared to a target or reference is typically important. Targets etc. can be built into an indicator itself (such adding a threshold line to a graph) or it can be part of the application framework. Targets or reference values should not be considered a universal indicator criterion, as this will depend on the problem to be measured or managed.

‘Perverse incentives / allow innovation’ - Will the measurement process encourage undesired behaviours by those under measurement? (NCHOD 2006 ; Marsden et al 2005).

Comment: Clearly a management and policy issue, potentially relevant for indicators that measure performance. We propose to consolidate them under the term ‘positive/negative incentivisation’

‘Integrative’: the full suite of indicators provides a measure of coverage of the key gradients across the ecological systems (Dale & Beyeler 2003)

Comment: This is mostly relevant for combining indicators to form a comprehensive monitoring or management program. It is not strictly an indicator criterion. We will not include it here.

‘Limitation in number’ - It is most effective to be selective and use smaller sets of well-chosen indicators (Segnestam 1999).

Comment: This is mostly relevant for combining indicators to form a comprehensive monitoring or management program. It is not strictly an indicator criterion. We will not include it here.

4.2 Correspondence of criteria: Section 2 lists compared with literature

More or less all of the previously proposed criteria in Table 2 were also found in some form in the indicator criteria literature.

One exceptions is ‘Discountability’, listed in Table 2, which does not seem to have a standard description or definition in the indicator criteria context, even if the notion of discounting is obviously in itself a well established term in economic theory. We would suggest not to include it as an *indicator* criterion, since all (quantitative) measures could in principle be discounted with a standard discount in the same way. It is more a question if this makes sense in terms of the *impact*

that is under review.. However, we retain it in the intermediate list in Table 21 and include it in the test in section 4.3.

Surprisingly, we also did not find a distinct definition of a criterion of 'Transparency' in the indicators literature. However, Hauge et al (2005) provides the following useful reflections:

"To judge the quality and relevance of an indicator, users need a transparent presentation of the scientific background and of the uncertainties involved (...) Knowing the underlying assumptions, simplifications, and other scientific judgements is useful, as is knowing how they affect the indicator and the objective to be agreed upon, and how well-founded is the underlying knowledge.(...) We regard four aspects as important for ensuring indicator transparency: a clear description in the context of associated knowledge, its scientific foundation, the robustness of its value, and its performance in a management context."

In this sense transparency may be understood more as a composite of a number of underlying aspects or criteria, rather than as a criterion in itself. Some of the same elements are thus covered by other references with terms such as, 'theoretical foundation', 'explicit methodology' or 'measurement' (NCHOD 2006; Rice & Rochet 2005). In contrast, the OECD defines transparency very generally as "access to information" (OECD 2008). The EEA (2004) defines transparent indicators also in a simplified way: "Indicator well documented and of known quality".

Because of the high importance attached to this concern among COST 356 members we will propose to retain transparency as a criterion with a definition revised according to the literature input (see Table 21).

Several of the other criteria in Table 2 appear to be overlapping, or the same concepts appear in the literature under different headlines, which mean that they could be rephrased (see the following section).

4.3 Reorganizing and rewording indicator criteria

Several criteria in Table 2 and Table 3 were defined and described intuitively and hence not in accordance with concepts or formal definitions in the literature. This can contribute to overlap and redundancy and sometimes confusion of such a list.

This seems to be partly the case for the following items in Table 2 (with the formulations given there repeated in 'marks') :

Preciseness - 'How precise the indicator can be measured (accuracy, reliability...) and/ or how precise the indicator is showing development of the phenomenon it is developed for'.

Comment: this seems to be a mix of various criteria. It is better subsumed under other more officially defined criteria such as representivity and validity.

Reliability - 'The ability of an indicator to perform its pre-defined functions in routine circumstances, as well as hostile or unexpected circumstances.... The IEEE [SIC] defines it as ". . . the ability of a system or component to perform its required functions under stated conditions for a specified period of time." Reliability of an indicator may also be 'the idea that something is fit for purpose with respect to time'.'

Comment: The tentative definition is not entirely clear. A number of different ways to conceive 'reliability' exist in the referenced literature. Eyles & Furgal (2000) mentions 'Repeatability across times and sources'. NCHOD (2005) more technically talks about 'Data reliability' defining it as 'agreement (kappa coefficient) between measured records and those collected by an independent source'. Farchi et (2005) and Goger et al (2006) a bit confusingly mixes it with validity. Nie-meijer & De Groot 2008 simply equates reliability with 'a proven track record'. Kusek & Rist (2004) have the following definition 'Reliability is the extent to which the data collection system is stable

and consistent across time and space. In other words, measurement of the indicators is conducted the same way every time'. See Table 26 for suggestion.

Measurability (included also forecast ability – will it change with time) 'Data required to figure out the indicators should be reliable. E.g. air quality measures should be taken through consistent procedures and using standard equipment. Likewise, population affected by different levels of pollution should be objectively calculated. When forecasting present values of magnitude, technically sound models should be preferable. Subjective assessment of significance is highly variable in time and space'

Comment: Should probably be split between measurability and forecastability. Possible too specific wording for a definition. for suggestions.

Data availability 'in terms of quality, quantity and timeliness (on time, how long does it take to produce an indicator from the data)- Indicators that can be accessible in time series and on a cross-geographical basis should be preferred. Decisions, mostly at the local level have to be based on local data and it is no use to recommend indicators that cannot be documented numerically or through generally agreed (undisputable) information. Detected lack/gaps of information at individual locations can, however, serve as basis for policy instructions to start data collection when particular indicators are generally used elsewhere. This criterion is related to one on space transferability, i.e. indicators that can be adequately measured and forecast in different locations should be preferable'-

Comment: relevant, but wording is too restrictive and detailed.

How frequently are data updated – 'See the previous description. The validity of extrapolated values is ruled by statistical significance. To establish trends a minimum number of values are required. Social surveys at regular intervals may be required to highlight changes in perceptions. Again, significance may vary locally in addition to temporally.'

Comment: More a discussion of method than actual criterion. Could be appendix/footnote

Certainty (monitoring and predictions) - 'Description can be derived from previous comments. Measures taken with reliable instruments and using internationally accepted procedures are less subject to challenge and can be used for comparisons. Robustness of forecasting models is essential. Beware of indicators based on subjective perceptions (e.g. value of scenery) and of allocated values of significance.'

Comment: More a discussion of method than criteria – could be appendix/footnote

Independence from each other – 'Indicators should be as much as possible independent from each other'.

Comment: This is important if the indicators are trying to measure the same thing. However, several indicators which are dependent can often be used effectively to illustrate different aspects of a problem, as long as this is clear, as e.g. in the DPSIR chain. More consideration is needed.

Simplicity (1) 'Condition, or quality of an indicator be simple or un-combined. This characteristic in some situations is better to turn easier the explanation of certain things than complicated ones'.

Comment: May not be required for all indicators. More of a general comment to whole set

Simplicity (2) How easy it is to understand the indicator: how it is constructed, how it is related to and varies with the main phenomenon etc, and/ or how easy it is to measure and calculate the indicator

Comment: May not always require simplicity. May be incorporated in' interpretability'.

4.4 Proposal for an intermediate revised set of criteria

In Table 21 we include the criteria from Table 2 and Table 3 that we could still consider appropriate, and combine them with definitions as well as revised and additional criteria drawn from the literature. We structure the criteria into more or less the same four groups as in Table 2, even if the literature review, as already mentioned, showed no clear consensus about how to categorize and group indicator criteria (another possible distinction would be the one between ‘measurement’, ‘monitoring’ and ‘management’ aspects proposed in section 3.1) The last group of criteria in table 2 is made broader and renamed ‘management and policy aspects’.

Table 21 represents an intermediate long, but consolidated list of criteria that are of potential relevance for assessing EST indicators, now mostly based on definitions established in the literature.

	Criterion	Proposed definition (Existing, adjusted or new)	Source
Conceptual and theoretical aspects	1. Representativity	Correlation between an indicator and the issue for which it is supposed to be a proxy	Hauge et al (2005)
	2. Conceptual validity	Is the indicator based on a well understood conceptual model? 1) The definition of the indicator and the concepts that comprise it up is suitable. 2) There is a correspondence between the indicator and the factor to be quantified. 3) The interpretation and meaning of the indicator are suitable	CGER (2000); Cloquell-Ballester et al (2006)
	3. Theoretical Foundation	Is the indicator explicitly defined by appropriate statistical units of measurement and standard international terminology? A clear theoretical definition of a concept to be indicated should, 1) identify the number of distinct aspects or dimensions of the concept. Each dimension requires a separate latent variable. 2) The theoretical definition should clarify whether the latent variable is continuous or not. 3) Each latent variable is ideally measured with several indicators	NCHOD (2005) OECD (2003) Bollen (2004)
	4. Predictive Validity	Does the measure correctly predict a situation which would be caused by the phenomenon being measured? The degree to which data values satisfy acceptance requirements of the validation criteria or fall within the respective domain of acceptable values. Data validity can be expressed in numerous ways. One common way is to indicate the percentage of data values that either pass or fail data validity checks.	Cole et al (1998) Batalle et al (2004)
	5. Sensitivity	An indicator must be able to reveal important changes in the factor of interest Do the measurement tools and timing of results allow changes to be observed over time	WHO (2006) NCHOD (2005)
	6. Specificity/ Transport specificity	An indicator must reflect only changes in the issue or factor under consideration The indicator should identify the effect of transportation rather than providing an estimate of environmental quality that may depend on numerous sources	WHO (2006) US EPA (1999)

Measurement aspects	7. Transparency	To which degree it is described in an understandable way how the indicator is constructed, how it varies with what it represent (the phenomenon in focus), and how it is influenced by uncertainties. This implies that input data, assumptions, methods, models and theories involved are described and justified.	COST 356
	8 Reliability	An indicator must give the same value if its measurement were repeated in the same way on the same population and at almost the same time; The ability of an indicator to perform its pre-defined functions in routine circumstances, as well as hostile or unexpected circumstances	WHO (2006) COST 356
	9. Measurability	Be easily measured: The indicator should be straight-forward and relatively inexpensive to measure Measurable indicators are based on data that should be readily available or made available at a reasonable cost/benefit ratio	Dale & Beyeler (2003) OECD (2003)
	10 .Data Availability	Data that are available and accessible, accurate, comparable over time, complete with historical information and covering sufficient geographic area	Boyle (1998)
	11. Timeliness	The degree to which data values or a set of values are provided at the time required or specified. Timeliness can be expressed in absolute or relative terms.	Batalle et al (2004)
	12. Threshold availability	Theory allows calculation of reference point associated with serious harm	Rice & Rochet (2005)
	Data analysis aspects	13. Aggregatability without loss of representativeness	How easy and to which degree indicators can be aggregated, to higher geographical levels, with other indicators etc.
14. Discountability		Discounting influences people's assessment and evaluation of impacts that will be perceived in different moments of time, as well as trade-offs with other effects characterized in other moments and through other indicators. Discounting factors are affected not only but subjective perceptions but, likewise, by changes in technology and by people becoming used to situations	No source found Own definition in COST 356

Management and policy aspects	15. Policy relevance	Relevant to the organisation/ strategy Policy relevant and useful indicators should: <ul style="list-style-type: none"> • provide a representative picture • be simple, easy to interpret and able to show trends over time • be responsive to changes • provide a basis for international comparisons • have a threshold or reference value against which to compare it 	Marsden et al (2005) OECD (2003)
	16. Linkability to targets	Be clearly relevant to articulated goals and objectives Monitor progress toward quantified targets	EEA (2004) Boyle (1998)
	17. Actionability	Can the factors which influence the phenomenon be positively influenced to induce improvements? At national level, an indicator must be able to act as a “marker of progress”... the data should also be useful locally, i.e. follow-on action should be immediately apparent	NCHOD (2005) WHO (2006)
	18. Transferability	This means the capability that an indicator has to be used in other similar contexts in order to compare different scenarios. This characteristic is useful in the case of cross-border issues.	No source found Own definition in COST 356
	19. Simplicity	Condition, or quality of an indicator be simple or uncombined. This characteristic in some situations is better to turn easier the explanation of certain things than complicated ones.	No source found Own definition in COST 356
	20. Positive/negative incentivisation	Will the measurement process encourage undesired behaviours by those under measurement? Will the measurement process encourage desired behaviours by those under measurement?	NCHOD (2005)
	21. Ethical	An indicator must be seen to comply with basic human rights and must require only data that are consistent with morals, beliefs or values of the population	WHO (2006)

The following immediate remarks to Table 21 should be made.

First of all the list contains an abundance of potential criteria, where several of them may continue to be overlapping or redundant. This is partly due to the fact that criteria from several sources were combined. This means that the list of criteria for actual application could be further reduced and structured compared with Table 21.

Secondly, there are significant differences among how the criteria could be used. Some can be more or less immediately applied to assess a candidate indicator, for example whether an indicator has a theoretical foundation or not (criterion 3). Others may allow a statistical test of the accuracy of a candidate indicators e.g. in terms of its ‘predictive validity’, (criterion 4). Others again will require some knowledge of existing data sources (19, ‘Data Availability’). The operability of the criteria would depend on who are using them and how detailed information is available.

Hence further work to consolidate the list criteria and to consider how they can be operationally applied to actual indicators is necessary before a ‘final’ set of criteria can be recommended. The following section 4.5 reports how the criteria were internally tested with respect to appropriateness, applicability and potential overlap.

4.5 An internal trial exercise with indicator criteria

In response to the needs noted in the previous section a 'trial exercise' among COST 356 members was conducted. The aim of the exercise was to test the applicability and usefulness of the intermediate long list of proposed indicator criteria for COST 356, shown in Table 21. How easy to understand and use are these criteria, and how stable are they when applied to assess sets of potential indicators of EST impacts?

The test applies to a group of researchers and academics who are well informed about transport and environment issues and indicators, without necessarily being specialists in environmental assessment or indicator selection. The context is a situation without access to actual data or measurement methodologies.

4.5.1 Trial exercise design

The trial exercise was designed as a small questionnaire sent out to all COST 356 members in advance of a meeting in Riga in May 2008. It was also rehearsed at the Riga meeting to allow additional contributions.

The questionnaire had three sections:

Section 1 listed in a table all the criteria for evaluating indicators that had been drawn out from the literature and internal discussions (Table 21). For each criterion the participants were asked about four issues (hereafter *Trial 1*):

- 1) Do you understand the criterion OK?
- 2) If yes, do you think the criterion could be relevant for selection of appropriate indicators for ST?
- 3) Do you find a need for more detail to assess the relevance of this criterion?
- 4) Does it overlap too much with other criteria?

In *section 2* participants were asked to apply the criteria to three hypothetical indicators for the causal chain "Air quality impact on human health", imagining a context where the participant was to advice her/his government about the choice of indicators for this impact chain (hereafter *Trial 2*).

For each of the three potential indicators the participants were asked to score the indicator using each criterion. The scores were:

"2" The indicator is very good according to this criterion

"1" The indicator may be OK for this criterion

"0" The indicator is not good at all according to this criterion

"?" I have no idea about the quality of the indicator for this criterion

In *section 3* the participants were asked to review the exercise and comment if a similar criteria assessment approach should be considered further in the action.

In total eight (8) members of COST 356 submitted the trial questionnaire. Some responses were partial by only answering Trial 1 or 2. There were then six responses for each of the two main sections (trials).

In section 3 there were only few comments, which were mostly positive about the use of a similar kind of exercise in the further work. We do not address these comments here.

Even if the number of respondents was extremely small, the exercise is reviewed here in some detail because of its instructive value for developing a criterion based approach to indicator assessment.

Overall	8
Trial 1	6
Trial 2	6
Questions	4/comments

4.5.2 Results of trial 1

There was a variety of responses to the questions. All responses are shown in Table 23

In the first column “Y” indicates that the criterion is understandable to the reviewer, “N” that it is not. “(N)” or stronger “?” means that the criterion may need some clarification.

In column 2 a similar notation is used for relevance of the criterion. In column 3 “N” means the criterion does not need more detail.

Column 4 shows possible overlaps between the criteria as suggested by respondents.

For each criterion a colour code is used in Table 23 to suggest the interpretation of the results for each criterion as a whole.

Clear green means that the criterion by all respondents is considered understandable, relevant, has no need for further detail and no overlaps, with only one respondent maybe having a minor reservation.

Weak green means that most respondents agree with the criterion more or less, with one or two having some reservations in some of the categories.

Weak yellow means that some respondents has some reservations in most categories.

Orange means that significant reservations or disagreements exist about the criterion as it is defined in the note.

Only five out of the 18 criteria receive almost unanimous full support. They are,

- Predictive validity
- Transparency
- Reliability
- Measurability
- Data Availability

Seven more criteria are considered fairly OK (light green). A few of those are seen to overlap with others including such as between 2 “conceptual validity” and 3 “theoretical foundation”.

Further seven are considered to be somewhat unclear or problematic (weak yellow), suggesting a need for clarification, or consolidation with other indicators or perhaps abandonment.

Two are most seriously challenged namely 14 “Discountability” and 17 “Actionability”.

Before we suggest any further conclusions from these results, we will look at trial 2 where respondents were asked to apply the criteria in the assessment of three potential indicators. This will throw further light on the usefulness of the criteria for selecting indicators in practice.

Criterion	Understand?	Relevant?	More detail?	Overlap with?
1. Representativity	YYYYYY	YYYYY(Y)	Y?N	2.2
2. Conceptual validity	YYNY(N)Y	YY?Y(Y)Y	?NN	3? 1(2.2)
3. Theoretical Foundation	YYYY(Y)Y	YYY(N)(Y)Y	NYN	2
4. Predictive validity	YYYY(Y)Y	YYYYY(Y)	NNN	(1)
5. Sensitivity	YYYYYY	?Y?YYY	N?N	9
6. Specificity/ Transport specificity	Y?YYYY	Y?NYYY	?NN	
7. Transparency	YYYYYY	YYYYYY	NNN	
8 Reliability	YYYYYY	YYYYYY	NNN	
9. Measurability	YYYYYY	YYYYYY	YNN	19 10
10.Data Availability	.YYYYYY	YYYYYY	NN	9.2
11. Timeliness	YYYYYY	N(Y)YY(Y)Y	YNN	10 DATA 11(?)
12. Threshold availability	YYY?YY	Y(Y)???Y	YYN	9 16
13. Aggregatability without loss of representativeness	Y?YYYY	Y(Y)NYYY	YNN	9/8
14. Discountability	??YNY	??NY?Y	NN?	
15. Policy relevance	YYYY(N)Y	Y(Y)YYY(Y)	NNN	2 (1,2,5,7,12)
16. Linkability to targets	YYY?YY	Y(Y)Y?Y(Y)	NYN	9/4 12 12
17. Actionability	NYNYYY	? (Y)NY(Y)(Y)	??N	15
18. Transferability	YYYYNY	? (Y)NY?(Y)	NNN	1
19. Simplicity	Y?YY(N)Y	Y(Y)NYYY	NN (N)	13
20. Positive/negative incentivisation	NYYYYY	N(Y)NYN(Y)	NNN	21
21. Ethical	N?YYYY	NNNYYY	NNN	20

4.5.3 Results of trial 2

In Table 24 the results from using the criteria in hypothetical cases are shown. Each cell contains all the respondents' assessments of one indicator with regard to one criterion, using the score 2,1,0, or ?.

For example for the indicator 'Emission of NO_x from motor vehicles in a country', five respondent think that this indicator is very good (2) according to the criterion 10 "data availability". For the same indicator the response is more mixed with regard to the criterion 5. "Sensitivity", some say "2" (very good), some say "0" ,not good at all.

Indicator	Emission of NO from motor vehicles in a country	Average concentration of PM2,5	Number of people with respiratory disease exposed to air quality above limit	Summary assessment of uniformity of criteria application
1. Representativity	10011(0-1)	20(0-1)111	222221	Middle
2. Conceptual validity	2?022(0-1)	2?(0-1)221	2?(1-2)122	Low
3. Theoretical Foundation	210222	21(0-1)222	22(1-2)112	Low
4. Predictive validity	122110	222111	212211	Middle
5. Sensitivity	120120	12(0-1)121	22-221	Low
6. Specificity/ Transport specificity	0?1222	(0-1)?1201	1?-10(0-1)	Low
7. Transparency	102121	212111	21-111	Middle
8 Reliability	21-12?	21-12?	21-22?	Middle
9. Measurability	202221	21222?	10-02?	Low
10.Data Availability	222221	212221	10-02?	Middle
11. Timeliness	22220?	20220?	10-00?	Low
12. Threshold availability	1221?0	2221?1	(0-1)2-0??	Low
13. Aggregatability without loss of representativeness	022110	01211?	20-?12	Very low
14. Discountability	??-?20	??-?10	??-?10	Middle
15. Policy relevance	11-121	21-121	21-222	Middle
16. Linkability to targets	2222?1	2222?1	20-0?1	Middle
17. Actionability	22-?21	22-?2(0-1)	12-?2(0-1)	Low
18. Transferability	?2-121	?2-121	?2-121	Middle
19. Simplicity	212211	21021?	10-21?	Low
20. Positive/negative incentivisation	002?0-	00-?0?	01-?0?	Middle
21. Ethical	222?22	22-?22	22-?22	High

Again a color code has been applied. In this case the colours suggest if there is strong agreement or not among the respondents. Clear green in a cell would illustrate that all respondents think in the same way about one indicator according to one criterion. However, there are no clear green cells in the table. Weak green means that all respondents agree, with one deviation. Weak yellow means variations over one degree (e.g. a mix of “2” and “1” answers, or “0” and “1”). Orange means full variation over all degrees, 2,1,0, indicating very little agreement over the indicator with regard to the criterion.

In the far right column an aggregated assessment is suggested. How much agreement is there in general when using a criterion for all three candidate indicators?

“Very low” agreement in right column (orange) is when there is full variation (0,1,2) for all three indicators. “Low” is when there is full variation for two of the indicators, “Middle” when there is full variation for only one indicator, “High” agreement is when there is maximum one degree of variation (0,1, or 1,2) for any of the indicators. “Very high” would be when there is almost full agreement (no or almost variation) across all the three indicators for a criterion.

It can be seen that there is only one criterion with “high” agreement, number 21 “Ethical”. There are no criteria here with “very high” agreement. Most criteria have ‘Middle’ or ‘low’ agreement.

4.5.4 Combining the two trial outputs

We can now compare the two tables, Table 23, which displays the general or ‘theoretical’ view of the criteria, with Table 24, which displays the results when criteria are applied to ‘actual’ (if hypothetical) indicators.

In general there is much more agreement about criteria at the general level (Table 23), than when using the same criteria for actual assessment of indicators (Table 24).

For example, even if all respondents are very confident and approving of the criteria 4 (‘Predictive validity’), 7 (‘Transparency’), 8 (‘Reliability’), 9 (‘Measurability’), and 10 (‘Data Availability’), there is only ‘middle’ agreement about how to assess candidate indicators when these criteria are actually applied to candidate indicators; one even has ‘low’ agreement (9, Measurability).

Generally, if we place the tables next to each other we see a rather random pattern, even if there is a weak tendency to better (or less poor) agreement in the indicator assessment (Table 24) for criteria which has higher rank in terms of understanding (Table 23). Hence, presumed good understanding and acceptance of a criterion does sometimes but not necessarily lead to a more uniform assessment of specific indicators with regard of the criterion.

All in all it seems to be easier to agree on criteria in principle, than to make sure that using the same criteria actually leads to a consistent assessment of a set of potential indicators.

4.5.5 Observations and interpretations

Obviously it was a very limited exercise based on simplistic criteria definitions and arbitrary, hypothetical indicators. Very few respondents were involved. The results should not be overstated.

The following observations could nevertheless be drawn:

First of all the exercise was generally welcomed by the participants, who found that it was a useful way to become well acquainted with indicator criteria and how they could be applied in indicator assessment. It was recommended to continue development of a criterion based approach to indicator assessment.

Secondly, the many available criteria from Table 21 makes the procedure somewhat burdensome but also allows a rich palette of options to draw the assessment from. More importantly, however, limited and partly overlapping *criteria definitions* makes the assessment difficult and vulnerable to misinterpretations. A need to improve some definitions and reduce overlaps was generally noted.

Thirdly, the assessment of potential indicators in trial is impaired by a lack of guidance about *how* to test the proposed indicators with regard to each criterion. For example, if a method was prescribed for how to test a particular indicator for ‘predictive validity’ this may lead to more uniform and robust indicator assessments. Also, if actual data sets was provided, more precise assessments for criteria such as for ‘data availability’, and ‘timeliness’ would be possible.

Fourth, the participants in the test were not all specialists in the impact chain “Air quality impact on human health”. It is possible that a group of experts in the field would be able to reach a higher degree of consensus about each indicator, e.g. for criteria such as ‘reliability’ and ‘sensitivity’.

Finally, the (hypothetical) context for the assessment is poorly defined. Especially criteria like ‘policy relevance’, and ‘positive and negative incentivisation’ seems to require that a more specific decision making context is considered. Not all criteria would be equally relevant for any ‘measurement’ or ‘decision making’ context.

4.6 Further refinement of the criteria list

Following the conclusions of the trial a discussion and further development of the criteria was conducted.

First it was agreed that before recommendations to the following work tasks and criteria use generally could be given, the intermediate long list of criteria would need to be reduced, criteria that were not fully understandable eliminated or revised, overlaps between criteria reduced, and some definitions reconsidered.

Second it was agreed that some kind of guidance to illustrate how a particular criterion can be applied to assess an indicator would be useful for making each criterion more operational.

An attempt to move this work significantly forward from the trial results was a written contribution from one Action participant (Joumard 2008). In this contribution all of the preliminary criteria with their associated preliminary definitions were critiqued, and re-evaluated for consistency, overlap/redundancy and logic position in an overall criteria set. This analysis led to a suggestion for an entirely revised, reorganized and much shorter tentative set of criteria, with revised definitions. Also, for some criteria examples of indicators fulfilling versus not fulfilling the criterion were provided (see Table 25).

This substantial contribution was welcomed by other WG2 participants. Especially the examples of agreement/disagreement over indicators were seen as a useful addition that should be completed. However, the discussion also revealed that no consensus about the new proposed list or structure of criteria could be reached, not even that this new condensed list necessarily represented an improvement compared to the previous longer list.

Counter-arguments to Joumard's proposal included that it disregards results of the work in task 2.2 so far including sections 2 and 4 and most of the literature review (section 3). Hence, the new contribution was seen as proposing a perhaps too drastic consolidation of several former criteria into much fewer ones, making the application of each criterion potentially more difficult, and harder to interpret. For example the concept of ‘validity’, by some considered as the most important indicator criterion of all (e.g. Bockstaller & Girardin 2003; Innes 1978, Bollen 2003) is apparently absorbed into ‘theoretical foundation’ together with several other criteria. It was felt that a long list with potential overlaps may be more operational for a screening exercise, especially because that list was developed by the COST action participants who are supposed to take part in the screening of indicators themselves. Also for example Joumard's critique and rejection of ‘policy relevance’ as a relevant criterion met some opposition in the discussion.

Category		Criteria	Ref. to criteria of Table 2	Proposed definition (in blue, could be deleted)	Examples of agreement ----- Counterexamples (disagreement) from Goger (2006) and Goger & Joumard (2007)
Indicator as measurement tool		1. Theoretical foundation	2.1, 7a, 3.1+3 .2, 3.3, 19	<p>An indicator must be based on a conceptual model which is well accepted by the scientific community of the concerned environmental impact..</p> <p>The indicator should be defined explicitly by a standard international terminology and should identify clearly its input parameters, continuous, discrete, quantitative or qualitative.</p> <p>The conceptual model of the indicator has to be transparent for the scientific community of the concerned environmental impact: To which degree it is described in an understandable way, how the indicator is constructed, how it varies with what it represent (the phenomenon in focus), and how it is influenced by uncertainties. This implies that input data, assumptions, methods, models and theories involved are described and justified.</p> <p>(As a consequence of its measurement tool characteristics, an indicator is simple or un-combined.)</p>	<p>A large number of scientists from a wide range of disciplines work on the greenhouse effect, aided by strong internal cooperation, particularly within IPCC. This organisation provides an indicator known as global warming potential (GWP), which is the subject of widespread international agreement (IPCC, 2001). This indicator establishes a simple relation between the emission of gases and the average increase of the Earth's temperature.</p> <p>-----</p> <p>Chemists have developed a global potential odour indicator (PO), built in the same way as the GWP, that establishes a relation between an intensity of odour and a quantity of pollutant emitted (Guinee et al., 2002). The global odour is given by the total emissions of pollutants weighted by a coefficient corresponding to an olfactory perception threshold. However, this indicator has not achieved consensus since many specialists underline the fact that sensitive pollution is characterised by annoyance, which is not directly related to the intensity of an odour, but far more to its variation through time.</p>
		2. Reliability	8.1	<p>An indicator must give the same value if its measurement were repeated in the same way on the same population and at almost the same time</p>	<p>An indicator based on a mathematic formulae using measured (or estimated) variable as input parameters is reliable and replicable.</p> <p>-----</p> <p>?</p>
		3. Representativity	1+6.2 (4.1, 5.1, 6.1)	<p>An indicator has to represent the environmental impact for which it is supposed to be a proxy, and not something else.</p> <p>The indicator should measure the effect of a transport project (in its broad meaning) on the environmental issue considered, and therefore must be able to reveal changes in the impact.</p>	<p>The global warming potential (GWP) establishes a simple relation (weighted total) between the emission of six greenhouse gases and the average increase of the Earth's temperature, which is the initial impact of the chain of impacts of the greenhouse effect. It permits evaluating the initial impact of any transport system or sub-system. It does not represent the final impact but an intermediate one.</p> <p>-----</p> <p>The acidification potential (Huijbregts, 2000) represents the quantity of H⁺ ions released by the pollutant emissions expressed in equivalent SO₂. It is more a maximal potential than a real impact, which depends a lot on the local conditions. Therefore this indicator does not represent the impact of acidification due to pollutant emissions.</p>

	Indicator as decision making tool	4. Data availability	10 (5.2, 9.1, 9.2, 11)	Indicators are based on (input) data that should be really available or made available at a reasonable cost. The data have to be accurate, comparable over time, complete with historical information and covering sufficient geographic area.	? ----- ?
		5. Interpretability	2.3, 7b, 15.3	An indicator must be easy to interpret by the users. The conceptual model of the indicator has to be transparent for the user: To which degree it is described in an understandable way how the indicator is constructed, how it varies with what it represent (the phenomenon in focus), and how it is influenced by uncertainties. This implies that input data, assumptions, methods, models and theories involved are well described.	The GWP being proportional to the initial impact of the chain of causalities of the greenhouse effect, it is easy to interpret. The methods is based on widely spread reports, with summaries for policymakers. ----- The Lyon conurbation developed some years ago an indicator of air pollution, based on pollutant concentrations (Rousseaux, 1994). As this indicator is a decreasing function of the concentrations, it is easy to misinterpret its outputs. In parallel the multi-criteria analysis tool Electre (Kunicina, 2008) is not understandable by most of the users, as it is a black box.
		6. Comparability to threshold	12, 15.6, 16.2	If the environmental impact concerned is quantifiable (quantitatively), an indicator should make possible a comparison with threshold or reference value (standard, political target...).	? ----- ?
		7. Ethical	21	An indicator must comply with fundamental human rights and must require only data that are consistent with morals, beliefs or values of the population.	? ----- ?

A further analysis revealed some more fundamental issues with regard to identify, delimit and define the final list of criteria. The following summarises the most important issues considered.

A major problem is the concept of 'representativity'. Representativity is of course fundamentally important, but it seems also very inoperational considered as an indicator criterion. In this respect it seems to encompass or overlap with several other criteria, such as 'validity' (does the indicator measure - *represent* - what it is supposed to?), 'reliability' (is it accurate - *representative* - through repeated measurements under different circumstances?) 'theoretical foundation' (has a cause-effect relation between the indicator and the phenomenon it indicates - *represents* - been theoretically established and accepted?), and 'sensitivity' (does the indicator reveal - *represent* - important changes in the factor of interest?). 'Representativity' can also refer to indication of a *wider* phenomenon than the variable being measured, which brings it close to the notion of 'external validity' which means generalisability of the indicator beyond the entity it directly measures (Leviton 2001). Moreover 'representativity' can be considered beyond the context of objective measurement to mean an indicator being perceived or accepted as appropriate - *representative* - of a problem by those involved in using the indicator. Hauge et al, for example, place 'representativity' as a criterion related to policy relevance (Hauge et al 2005 p 552). Joumard (2008) also place it as overlapping the 'measurement' and 'decision making' category. In sum it is not easy to operationalise 'representativity' as a criterion without risking considerable overlap with, or redundancy, of other important criteria. Neither Table 21 nor Table 25 have solved this problem.

Another problematic concept is ‘transparency’. As shown in the review in section 4.2 the concept is actually not clearly defined in the indicator criteria literature. Hauge et al (2005) suggest it to be a construct of a number of several underlying sub criteria. It is also argued by Joumard (2008) that this notion consists of two components that are better captured by other criteria, namely a measurement part (captured by ‘theoretical foundation’) and a user related management part (to be subsumed under a criterion of ‘interpretability’). In contrast, several participants in COST 356 have insisted that transparency be included as a fundamental criterion in the process, since lack of transparency is a persistent problem noted in the literature on indicator utilization (Hauge 2005; Innes 1998). The compromise could be to maintain the criterion name ‘transparency’ since it overlaps closely with the aspects Joumard describe under the less intuitive label of ‘interpretability’ .

Thirdly, ‘policy relevance’ is obviously important if indicators are to inform management or policy processes. However, as pointed out by Joumard (2008), the elements of policy relevance as e.g. proposed by the OECD definition (Table 21) involves a somewhat confusing mix of measurement and decision making aspects. Policy relevance can hardly be assessed in isolation from whatever objectives or tasks are assumed in a policy or organizational context. Rather than maintaining it as a separate criterion it could arguably be broken down to a number of components or subcriteria that are relevant in a policy context, such as ‘relation to objectives and targets’ and ‘actionability’. It would then still remain to be discussed how ‘policy relevance’ in general is to be determined – it might for example well be that an indicator is relevant for a policy even if this is not explicitly recognized with targets. One option to consider could be the use of ‘significance criteria’, together with indicators. Significance criteria refer to if some impact can be considered important or not with regard to several aspects such as potential irreversibility, potential for controversy, or potential inequality in the distribution of effects (Gibson 2000). According to Tomlinson (2004), stakeholders need to be involved in defining significance criteria. At this point there is no standard way to assess if some indicators would allow assessment of ‘significance’ in this wide respect.

More generally it is hard to maintain a consistent distinction between categories of criteria relating to either ‘measurement’ versus ‘decision making’ aspects, or a distinction into the four categories used in Table 21. Several criteria that have been formulated in the literature and discussed in the work so far (including the three examples above) bridge or challenge these distinctions. On the other hand it is possible to break at least some overall criteria down into separate components that refer more to one category than another. Categories would still be helpful when criteria are to be used across a range of different indicator types, assessment situations, and review teams.

To summarize, the discussion made it clear that a full consensus about a recommended list of indicator assessment criteria with complete definitions is hardly a feasible output from COST Action 356. This is due to a number of circumstances, including,

- conceptual inconsistencies in the literature with regard to criteria definitions and categorizations
- the notion that criteria shift in significance depending on the status of the indicators
- the notion that different situations may require emphasis on different criteria
- different professional opinions among members of the COST Action with regard to what is the preferred approach and output from the work in the Action.

A more realistic goal to complete the work in this area is to provide a tentative ‘best available’ set of criteria based on the above contributions and reflections and apply this set in the subsequent work tasks in COST 356 internally. Further conclusions about the use of criteria, which criteria to recommend, how to define and categorize them etc could follow from the results of such applications are made. Even this goal will require compromise.

The following approaches for consolidating a list of criteria were considered:

- a) Recommending to use either Table 21 or Table 25, depending on situational needs or preference. This seems not acceptable considering the critical comments made to both.
- b) Finding a ‘compromise’, adjusting and merging Table 21 and Table 25. This seems challenging considering the large differences.

c) Reverting to literature to select a list of criteria defined in one single literature reference. Overlaps may be reduced (if a ‘well proven’ reference is chosen). This may be possible, but the choice of reference could be arbitrary, and it would neglect the results reached in the work so far.

The way forward chosen was a variation of b) and a) taking Table 21 as the starting point while accommodating several of the points raised by Joumard’s contributions in Table 25 as well as the subsequent discussion. The aim was to produce a much smaller list than the 21 criteria of Table 21, with clear definitions, fewer overlaps, and extended explanations. This list was to guide and interpret the use of criteria in the subsequent work. This could be a ‘basic’ list, where more criteria from the longer list could be brought in if the situation requires or allows it (referring to a) above).

The result is shown in Table 26 a,b,c. All in all only ten criteria are included.

The indicator criteria are now grouped according to the *three categories levels* originally applied in section 3.1 ff, ‘Representation’, ‘Operation’ and ‘Application’. It is closely related to the original distinction of criteria found in the literature into ‘measurement’, ‘monitoring’, and ‘management’ criteria (see section 3), but it avoids the implied separation into different indicator sets for each purpose. This distinction was found to be more clear and less restrictive than the four categories used in Table 2 and Table 21, while more reflective of the results reach in the discussions than the two categories proposed by Joumard in Table 25.

Category 1 ‘Representation’ are considered as the criteria that are most fundamental for indicators from a ‘measurement tool’ point of view. ‘Representation’ should here be taken to refer to the analytic aspects discussed under this category (in the sense ‘analytically sound representation’ of a particular item).

Category 2 ‘Operation’ criteria refer to the operationalization of indicators for actual use in for example a continuous monitoring program. These criteria refer to what would be required for providing a stream of actual data (values) for the indicators defined, regardless of how well they represent the phenomenon being measured and reported in the ‘measurement’ sense.

Category 3 ‘Application’ criteria are considered as relevant if a policy realistic situation is assumed where indicators are to be used in some form of assessment, planning or decision making process, where users are involved and the indicators may have influence on policy processes or outcomes. These criteria could involve assessment conducted by policy dialogue or user review.

Hence *Category 1* can be considered basic in the context of COST Action 356 WG 2. Here the aim is to identify suitable indicators representative of transport impacts on the environment. The two other categories are possible add-ons, in the context of other parts of the work. However, this does not necessarily mean that ‘application’ aspects have to ‘come last’ in a process of indicator selection after the measurement aspects have been addressed. Policy relevance criteria could also be applied as a filter before assessment of representation or operation of potential indicators is performed (according to the procedure discussed by e.g. NCHOD 2005, see below section 5).

A *definition* for each criterion is added in Table 26 a,b,c,. The definitions are inspired by, and sometimes quoted directly from the literature as reviewed in Section 3. In addition a verbal commentary to assist the interpretation and application of each criterion is added. The commentary draws on literature and on the internal discussions in the present Action.

Also examples of agreement/disagreement to illustrate each criterion are added (to the extent that any examples or references are found). Some examples are cited from literature, others are more speculative.

Again, the table with criteria and associated definitions etc should *not* be considered to represent a recommendation from COST Action 356 about correct criteria and definitions. It is meant to serve as input for work in the subsequent tasks within the action. Later in the COST Action a review of the criteria and the approach should be made again, and recommendations possibly given.

Cat.	Criterion	Definition & commentary	Examples of agreement ----- Counterexamples (disagreement)
REPRESENTATION	Validity	<p>A valid indicator must actually measure the issue or factor it is supposed to measure. (WHO 2006)</p> <p>A valid indicator must be based on a conceptual model that justifies how the indicator and the issue is causally connected. The model should be well accepted by the scientific community involved in the particular field (conceptual validity). The indicator should be defined explicitly by a standard international terminology and should identify clearly its input parameters and causal mechanism. The validity of indicators can be reinforced by statistical tests of the agreement between a prediction obtained from the indicator and other, more direct or 'objective' measurements of the same phenomenon (predictive validity) . Predictive validity without conceptual validity can however be misleading and should not be considered a substitute (Innes 1990)</p>	<p>A large number of scientists from a range of disciplines work on the greenhouse effect, aided by strong internal cooperation, particularly within IPCC. This organisation provides an indicator known as global warming potential (GWP), which is the subject of widespread international agreement (IPCC, 2001). This indicator establishes a simple relation between the emission of gases and the average increase of the Earth's temperature.</p> <p>-----</p> <p><i>Chemists have developed a global potential odour indicator (PO), built in the same way as the GWP, that establishes a relation between an intensity of odour and a quantity of pollutant emitted (Guinee et al., 2002). The global odour is given by the total emissions of pollutants weighted by a coefficient corresponding to an olfactory perception threshold. However, this indicator has not achieved consensus since many specialists underline that sensitive pollution is characterised by annoyance, which is not directly related to the intensity of an odour, but to its variation through time.</i></p>
	Reliability	<p>A reliable indicator must give the same value if its measurement were repeated in the same way on the same population and at almost the same time (WHO 2006)</p> <p>If a scale is used 10 times to measure something that weighs 100 kg, and it reads "100" each time, then the measurement is reliable and valid. If the scale consistently reads "150", then it is not valid, but it is still reliable because the measurement is very consistent (after Wikipedia). Reliable indicators allow different people to obtain the same results when operating the indicator. Reliability is therefore often more difficult to obtain for qualitative indicators that involve interpretation as part of the measurement process. Reliability also refers to the consistency of the indicator results when it is applied across the domain (e.g. subgroups, time periods) of the phenomenon it is supposed to represent (representative reliability).</p>	<p>An indicator based on a mathematic formula using measured (or estimated) variable as input parameters is reliable and replicable if it produces the same results every time the same data are entered, with little influence of random error. The formula used to calibrate quicksilver thermometers allows to make a reliable prediction of the temperature because the expansion of the material does not vary randomly but only with temperature (and, to a negligible extent, air pressure).</p> <p>-----</p> <p><i>Eskler et al (2007, p 57) review a range of potential indicators to characterize accident protective measures, including the function of airbags. As they observe using a qualitative indicator such as the very presence of airbags in cars would not adequately reflect the great variety of airbags present on the market and within the vehicle fleet. It would hence not be a reliable indicator of the effectiveness of in-vehicle protective systems.</i></p>
	Sensitivity	<p>A sensitive indicator must be able to reveal important changes in the factor of interest (WHO 2006)</p> <p>Indicators should generally react clearly and promptly to significant changes in the phenomenon being indicated. The main concern here is <i>transport sensitivity</i> meaning how well the indicator shows the contribution of transport changes in the considered impact evaluated by the indicator (Goger et al 2006). A transport sensitive indicator should identify the effect of transportation rather than providing an estimate of environmental quality that may depend on numerous sources (US EPA 1999) Transport sensitive indicators would be ones that could be broken down to subcomponents of the transport system to allow detailed assessment of the cause of the change (e.g. measured by transport mode, vehicle type, speed level etc).</p>	<p>Drivers sometimes suffer from fatigue, which is a potential traffic hazard. Systems to detect fatigue must use indicators that are sensitive to be able to rapidly diagnose signs of fatigue. Fairclough (1997) found some measures of car driving such as measured variation in short term steering adjustments to be sensitive indicators of driver fatigue, while others (like standard deviation of speed) were sensitive to other factors</p> <p>-----</p> <p><i>Black (2002) found that variations in Vehicle Kilometres travelled (VKT) could to a very high degree explain variation in a set of nine other transport variables. However, Black also noted that VKT ignores differences in fuel efficiency. For example, if California would shift completely to zero emission vehicles, it would have (almost) no influence on VKT and we would misinterpret the state's transport sustainability using VKT only. In this regard VKT suffer from low sensitivity as an indicator of transport sustainability</i></p>

Cat	Criterion	Definition & commentary	Examples of agreement ----- Counterexamples (disagreement)
OPERATION	Measurability	<p>A measurable indicator should be straightforward and relatively inexpensive to measure (Dale & Beyeler 2003)</p> <p>Measurability is an operational concern. It is important that indicator can be measured or calculated using easy tools and using simple data that are easily achievable and at a raw level (non elaborated) (Goger et al 2006) Indicators can be measured in different ways using nominal, ordinal, interval or cardinal scales. Qualitative (nominal) indicators may be easier to observe than some quantitative measures but more difficult to measure in an accurate (reliable) way if it involves interpretations. Indicators on a cardinal quantitative scale is typically the most measurable, and able to provide the most information through measurement.. Simple indicators are easier to measure than aggregate ones combining several data streams.</p>	<p>The number of motor vehicles in a country is measurable rather exactly via the legally required vehicles licensing and registration. Other ways to measure the number of motor vehicles include manual or automated traffic counts, satellite and areal cameras, or surveys and interviews. Each method may allow different degrees of accuracy and different attributes of the vehicles to be measured together with the simple numbers.</p> <p>-----</p> <p><i>The average degree of satisfaction with the public transport service in European cities cannot be measured in studies where the satisfaction is expressed on an ordinal Likert scale (Ferrari & Salini 2008). The fate over the next 100 years of each molecule of CO2 emitted from all motor vehicles cannot be measured, and hence the exact contribution to global warming from each motor vehicle is not known</i></p>
	Data Availability	<p>Data available indicators are indicators based on (input) data that should be readily available or made available at reasonable cost and time (=OECD 2003)</p> <p>The data have to be accurate, comparable over time, complete with historical information and covering sufficient geographic area. (Boyle 1998) Time, cost, ownership or work required could be considered as parameters in the assessment of data availability for an indicator. Some data are readily available immediately (e.g. on www). Some are less available while Some could potentially become available with the use of new technology.</p> <p>Timeliness is a particular concern associated with data availability. Timeliness can be defined as the degree to which data values or a set of values are provided at the time required or specified. (Batalle et al 2004). An operational measure proposed by NCHOD (2005) is the average time (months) between measurement and results</p>	<p>Comparable data for urban traffic systems in Europe are often lacking Through the work in the so-called 'European Common Indicators' and the 'Urban Ecosystem Europe 2007' report (AMBI-ENTEITALIA 2007) comparable data on a number of indicators have become available, via a coordinated effort of data collection and reporting involving 32 cities. Hence it is now possible to compare e.g. the average length of dedicated cycle lane per inhabitant, as one indicators for 'Better Mobility'</p> <p>-----</p> <p><i>The TERM indicator set contains an indicator (TERM 39) 'Uptake of environmental management systems by transport companies'. The indicator has been defined conceptually but is has only been produced once (in EEA 2000). The indicator has been omitted from all subsequent annual TERM report since data have not been collected since 1999.</i></p>
	Ethical concerns	<p>An indicator must comply with fundamental human rights and must require only data that are consistent with morals, beliefs or values of the population (WHO 2006)</p> <p>The criterion has been introduced in the human health assessment context to ensure that health data collection does not violate privacy or other ethical concerns of people. Similar concerns might be appropriate with regard to other aspects of human and social activity (e.g. transport behaviour, criminal records, property exposure to environmental pressure, etc) An indicator should not be based on data that are offensive for people to report or could be used against them.</p>	<p>In travel surveys such as the Danish TU (TransportUndersøgelse, DTU Transport 2009) information is collected about travel activates including 'private' information about peoples choice of destinations, travel purposes, timing of trips, etc on a certain day. The use of the data is restricted by privacy safeguards. Users have to sign up to confidentiality agreements</p> <p>-----</p> <p><i>Collecting data to produce performance indicators on drunk driving as a cause of accidents is hampered by a number of factors, one of which are privacy concerns, which in some countries disallows for example police to collect blood alcohol data from test made at autopsies. It is an ethical question if privacy of deceased persons should be violated to improve data quality for accident reporting.</i></p>

Table 26 (c) Proposed criteria with regard to application			
Cat	Criterion	Definition & commentary	Examples of agreement ----- Counterexamples (disagreement)
APPLICATION	Transparency	<p>A transparent indicator must be feasible to understand and possible to reproduce for intended users</p> <p>The conceptual model must describe in an understandable way how the indicator is constructed. Input data, assumptions, methods, models and theories must be accessible. Transparency allows the user to check the calculation and therefore to trust in the figures. Transparency is associated with but not identical to simplicity. A simple indicator may be more attractive because it is easier to show how it is produced. However, complex indicators may also be transparent if the methodology is well justified, well defined and well explained.</p>	<p>Innes describes a process involving an environmental management plan being developed for The San Francisco Bay in California in the 1990s. A number of stakeholder organizations set out to establish a consensus about how to measure water quality in the Estuary. Transparency emerged because the information was discussed and validated within the broader consensus building process, rather than by using measures predefined by external experts (Innes 1998).</p> <p>-----</p> <p><i>Sager and Ravlum (2005) report a case where the cost-benefit ratio was used as an indicator to inform political decision about a rail freight terminal in Norway. The politicians had no way to control how the results were produced. It is not the method, but how it is applied that fails.</i></p>
	Interpretability	<p>An interpretable indicator allows an intuitive and unambiguous reading</p> <p>It must be possible to draw clear conclusions from reading the indicator. Interpretability depends on how well the indicator varies with what it represents (the phenomenon in focus), and how it is influenced by uncertainties. It should move in an analogue fashion to the phenomenon.</p>	<p>When the Global Warming Potential of an emissions source increases it means an increased forcing of the global average temperature: Higher GWP means warmer planet.</p> <p>-----</p> <p><i>The Lyon conurbation developed some years ago an indicator of air pollution, based on pollutant concentrations (Rousseaux, 1994). As this indicator is a decreasing function of the concentrations, it is easy to misinterpret its outputs.</i></p>
	Target relevance	<p>A target relevant indicator must measure performance with regard to articulated goals objectives, targets or thresholds</p> <p>If the environmental impact concerned is quantifiable (quantitatively), an indicator should make possible a comparison with any relevant threshold or reference value (standard, political target...). If there are no quantified targets or thresholds the indicator should be considered in terms of its relevance for non-quantified policy objectives or goals. Indicators that do not or can not measure performance with regard to any goals or targets are less supportive of management and decision making function of indicators.</p>	<p>The European Commission has established the European Road Safety Observatory. In the Basic Fact Sheet Main Figures (ERSO 2007) we find the number of road accident fatalities in Europe 1990-2006. This figure is comparable to the road safety target for Europe of a 50% reduction in the number of annual fatalities from 2001 to 2010. The report provides the indicators together with an assessment of target fulfilment.</p> <p>-----</p> <p><i>"The lack of targets for some of the indicators (e.g. all-cause mortality and childhood poverty) may be a deterrent to monitoring" (Zucconi & Carson 1994, p 1645).</i></p>
	Actionability	<p>An actionable indicator is one which measures factors that can be changed or influenced directly by management or policy action</p> <p>Actionability refers to the role of indicators as tools to support decisions and management. The indicator can be directly actionable by measuring a parameter that is also a policy variable (e.g. number of police controls to check vehicle emission control equipment), or indirectly by measuring something that can be influenced by policy (e.g. population exposure to air pollution above limit values). An indicator directly measuring the parameters of decisions (e.g. funding decision) are more actionable than indicators measuring the general environmental conditions (e.g. temperature rise of the atmosphere). The point of actionability is that follow-on action to the indicator should be immediately apparent (WHO 2006).</p>	<p>Road construction has significant negative impact on habitats. The US Federal Highways Administration has adopted a performance target measuring the number of so-called Exemplary Ecosystem Initiatives (EEI), which are actions or measures that will help sustain or restore natural systems and their functions and values. Each EEI is counted and the results compared with an annual target value of 50 projects, which was just reached for 2007 (US DOT 2007). The measure is actionable considering that the FHWA can control the number of initiatives initiated.</p> <p>-----</p> <p><i>"in the context of European road safety, variables describing differences in weather conditions in different countries might help an understanding of why accident rates differ across Europe. However, such variables are not "actionable" in the same sense that variables describing variations in infrastructure quality, for example, would be." (Rackliff 2008)</i></p>

5 Frameworks and methods for assessing indicators

5.1 Introduction

This section will consider how criteria sets can be applied to assess or develop indicators, beyond simply listing individual criteria as discussed in section 4. The consideration of methodologies and procedures for indicator assessment and validation represents an important aspect of indicator work, as noted in another context by the US National Commission on Science for Sustainable Forestry (NCSSF):

“The bottleneck in effective selection and use of indicators is not a lack of good indicators or good science, but rather the lack of [. . .] a clear process for electing indicators [. . .] The reliability of identified measures is frequently questioned, at least in part because selection of indicators often has lacked transparency, social inclusiveness, and/or a logical structured process of selecting indicators.” (NCSSF 2005, cited from Niemeijer & de Groot 2008)

Although there may be several ‘bottlenecks’ for the identification of appropriate EST indicators (including lack of both good candidate indicators and science), this section will follow this line of reasoning by seeking to review and establish procedures for indicator assessment and selection.

At the most general level three different pathways to the development of indicators have been described: so-called data-driven, policy-driven, and theory-driven approaches (Hanafin & Brooks 2005; Niemeijer 2002; Niemeijer & De Groot 2008). *Data-driven* approaches mean that indicators are mainly selected on the basis of the availability of data that are suitable as indicators. Existing data sets are exploited inductively to develop a range of potential indicators. In *policy-driven* approaches indicators are developed for issues that are currently on the political agenda and for which indicators are politically in demand, for example based on policy objectives and targets. A *theory-driven* approach is defined as one that focuses on selecting the best possible indicators of a particular system or problem from a theoretical or scientific point of view (Niemeijer 2002). Hanafin & Brooks (2005) suggest that all of the three approaches should be combined in order to arrive at appropriate sets of indicators that would be both measurable, representative, and useful.

The three approaches roughly corresponds to the three types of criteria for selecting indicators that have been identified in this report, namely as criteria related to representation/measurement (‘theory driven’), to operation/monitoring (‘data driven’) and to application/management (‘policy driven’). Each group of criteria could thus support primarily one part of a process towards the identification of broadly acceptable indicators. Ultimately the aim of COST 356 concurs with the idea of combining the approaches, as in the attempt to connect ‘measurement’ and ‘decision making’ aspects of indicator selection. The starting point has been taken in the measurement or ‘theory driven’ dimension, with the question of how well existing or possible new indicators describe individual impacts of transport activity or policy interventions on the environment. Monitoring and in particular management aspects have been considered as additional important concerns. The question here and now is how to make the approaches operational, and possibly combine them.

Meanwhile, authors like Hoppe (2005) and Turnout et al (2007) suggest that the different ‘approaches’ are not randomly chosen and that harmony between them is not a given opportunity. The acceptance of scientifically based indicators in policy and decision making may for example depend on the degree of consensus about the basic underlying knowledge, and also about the degree of shared values involved in decision making. In cases of conflict or uncertainty policy and theory driven approaches may never meet. Where to start the process, and which type of criteria to build on may well depend on the status of the knowledge in each particular area to be measured by proposed indicators. We will return to this problem later in this section.

The section will first consider a number of frameworks and procedures proposed in the literature and will then consider ways to apply and adapt those to the area of EST and more specifically to the types of indicators and situations considered in work of COST Action 356.

5.2 Validation frameworks and selection procedures

In the indicator literature can be found a number of more or less elaborate methodologies for how to perform the identification, evaluation, selection and application of indicators using criteria in various ways. The references identified all roughly follow the general logic proposed by Boyle (1998) involving three main steps,

- Generation of indicator selection criteria
- Generation of potential indicators
- Selection of indicators.

A number of contributions seek to establish logical frameworks and general procedures of indicator validation. Three examples are Innes (1978), Bockstaller & Girardin (2003), and Cloquell-Ballester (2006) who all refer to the need for indicator *validation*. By validation they generally mean procedures and criteria to ensure acceptance of indicators as appropriate by scientists, but also by indicator users. A few works reported in the literature define more specific practical step by step approaches for using criteria with associated guidelines or sub methods for each step. Examples include (again) Cloquell-Ballester (2006), Jackson et al (2000); Kurtz et al (2001); NCHOD (2005) and not least Rice & Rochet (2005), who in an accompanying paper (Rochet & Rice 2005) even reports a test of their methodology.

Below we walk through a number of these references moving from the more general to the more detailed, practical and reflective approaches.

5.2.1 Innes on validation of policy indicators

Innes (1978) focuses mainly on the scientific validity of indicators used in policy making. She identifies a hierarchy with three types of validation, each one providing a stronger degree of validity than the former. The first is operational validity, meaning simply that a (statistical) correlation can be found between the indicator and the target or concept it represents. Such a correspondence does not provide a strong degree of validity, as the correlation may change unpredictably when the circumstances shift. Hence it is not a firm basis for selecting an indicator. A second stage is experimental validity, when some experience is gained to confirm the indicators validity under different circumstances. It adds more weight to the indicator. However the third and strongest level is theoretical validation, when a correspondence between indicator and target is based on a theoretically validated model. Often only very simple models are applied, however. One example is when the volume of cars on a road is used as an indicator of the number of accidents (Innes 1978, p 175). It is important that such implicit models behind indicators are made explicit and to discuss their limitations so their validity or preconditions for validity can be made clearer.

A number of tests can be employed although full validity is not a likely to be achieved for many indicators serving in complex areas. Innes' point is that both scientists and users are anyhow likely to trust indicators better if they are backed by theoretical validation. Simple indicators are often easier to test and assess, but may oversimplify relations. Indicator designers could test candidate indicators against other potential ones for the same phenomenon. It may be better to have several 'competing' indicators, as users may refer to different frameworks that influence which indicators they see as most appropriate. Users should therefore be somehow involved in the validation processes. Overall messages are that candidate indicators should be assessed with regard to what kind of validation they are based on; procedures to establish theoretical validity should be sought if they are not in place, and users should be involved even in confronting evidence from theoretical validations, if indicators are to be used for policy.

5.2.2 Bockstaller & Girardin on validation of environmental indicators

A similar understanding is made more operational by Bockstaller & Girardin (2003), with their framework for validation of environmental indicators. The authors understand indicators as variables having dual functions; as information tools for complex systems, and a decision support function. Even if indicators are not exact models, their development and assessment should follow somewhat similar scientific standards. However procedures to ensure this are rarely specified.

Bockstaller & Girardin suggest three steps of indicator validation inspired from model validation, namely 'design validation', 'output validation', and 'end-use validation'. A 'decision tree' for parts of the process is proposed as shown in Figure 2.

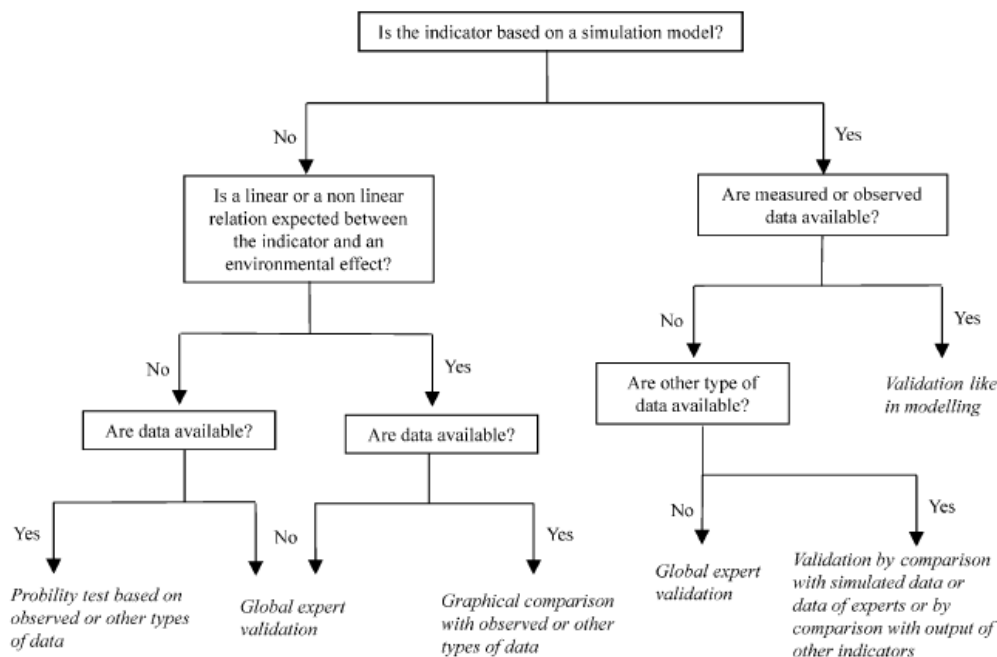


Figure 2 Decision tree as defined by Bockstaller & Girardin (2003)

Design validation is concerned with confirming the conceptual quality of the indicator, how well founded in theory the representation of the indicator is. This is typically done in experts reviews of proposed indicators. This is especially important in the indicator areas where there may be no other way to control the quality of the indicator (Bockstaller & Girardin 2002, p 642).

Output validation focuses on the information function of indicators, and if the indicator produces reliable results (values). This is where the parallel with model output validation may be most appropriate, especially if there are other data series, alternative indicators or model results to compare the indicator with. Extending the example of Innes, above (not given by Bockstaller & Girardin) one could assume volume of cars as indicator of accident risk. Data for car volumes could be plotted against accidents rates for the same road system, and compared with data for other possible indicators, such as average speeds on the network, or the character of the neighbourhoods. However, Bockstaller & Girardin recognize that indicators are often difficult to test like models, as sufficient studies may not be available. Moreover subjective judgments will also be needed to decide how much a trend is allowed to deviate from actual measurements to be considered an 'appropriate' indicator. The acceptable range should be defined before making any tests. Anyway 'expert validation' will often be the only method to assess the output of an indicator.

Finally the end-use validation concerns the usefulness of the indicator for decision making. According to Bockstaller & Girardin, such a validation requires the input from users, e.g. via a survey where users point out weaknesses of a potential indicator for making a useful diagnosis of problems or assist in decisions. One could also imagine (not mentioned by Bockstaller & Girardin) that experts well informed about decision situations may be involved in the end-use validation if no actual users are available. Summing up, largely the ideas are similar to Innes (1978), but a more systematic approach is suggested. Validation is divided into design, output and end user processes. The starting point is design validation, and the output validation should be done by thinking in parallel to models as far as possible. Which methods to use for output validation depend on whether there is (only) casual assumptions, or a simulation model behind the indicator, and what kind of data are available. Users are bought in as part of end-use validation.

5.2.3 3S Methodology for validation by Cloquell-Ballester et al

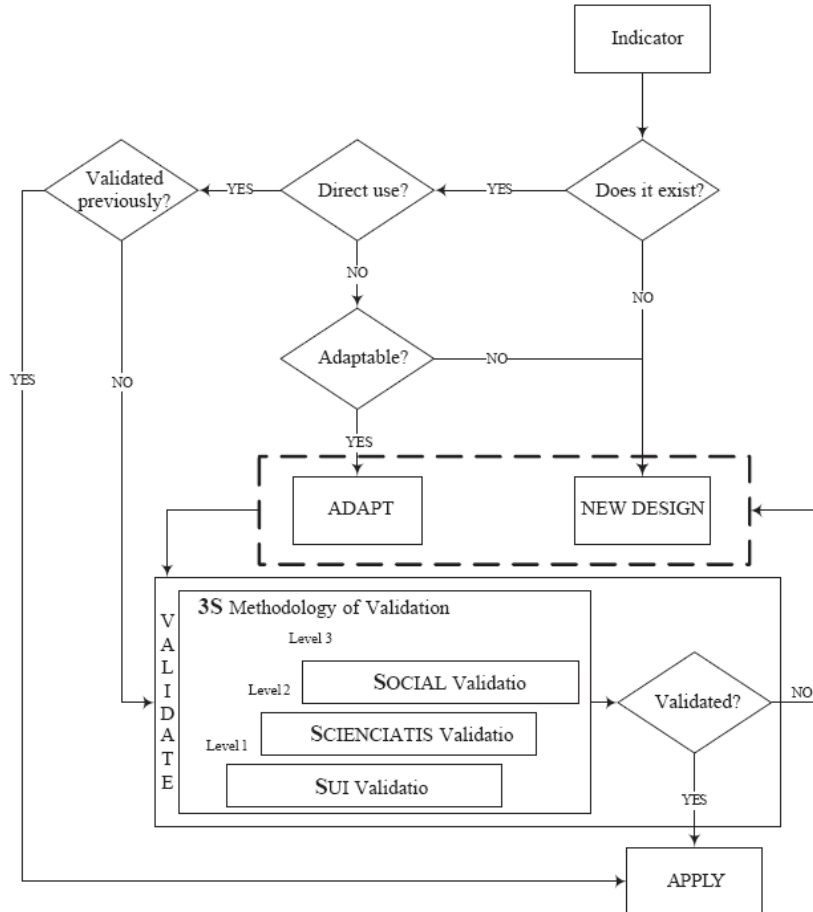
Cloquell-Ballester et al (2006) build on the ideas of Bockstaller & Girardin but develop the notion of validation further by breaking it up into three complementary procedures conducted by different groups involved in indicator development and use ('Self-validation', 'Scientific validation', and 'Social validation' = 3S). They also further systematize the process by using a multi-criteria methodology, and then demonstrate the procedure in a concrete case. The methodology applies to new indicators proposed in an area, in the sense that these indicators have not previously been validated and accepted. The aim is to make indicator assessment more transparent and comprehensive. The assumption is that '3S-validated' indicators will not only guarantee quality and reliability but will also support public participation and broader consensus in the use of indicators for assessment (Cloquell-Ballester et al 2006, p 81).

The starting point for the methodology is a new proposed indicator design. Then follows a series of steps to describe and evaluate the indicator(s). First a basic 'indicator report' is drawn up. The report describes the environmental aspect that the indicator addresses, defines the indicator itself, and informs about its various conceptual, measurement, data and relevance aspects. It also lists available documentation. Next a set of criteria for assessment are defined. Cloquell-Ballester et al propose 12 criteria, organised in three groups, 'conceptual coherence'; 'operational coherence' and 'utility'. (The categories closely corresponds to the ones proposed in this report (see Table 26); the individual criteria are also more or less the same as here, but the allocation of individual criteria to each category is quite different). The indicators are then assessed by three different groups representing the three 'S's'. The first S refers to 'self validation' which involves the working group undertaking the indicator development itself (similar to e.g. COST Acton 356 members). The second S is 'scientific validation' where a group of external experts undertake the same assessment in a Delphi setup. For the third S 'Societal validation' groups of stakeholders are invited to take part in a similar process. A 'process-controller' is engaged to assist and encourage the work. The assessments process use a similar methodology for each group involving a number of steps. First the indicators are scored according to the individual criteria on a five-point Likert-scale. Then the results are aggregated to the level of the three categories to reach an overall assessment for each category using weights suggested by evaluators and a multi-criteria methodology. This leads to a judgment of the indicators in four categories, from 'validated' (high scores and low deviation in all categories) to 'unacceptable (the opposite).

A case is described where four indicators relevant for assessing the location of industrial facilities are tested with the 3S method. The three teams give rather similar scores to the indicators. Their relative weights of the three categories differ strongly however, where 'scientists' place great emphasis on 'operational' criteria, while stakeholders not surprisingly emphasize 'user' criteria. They agree on the importance of 'conceptual' criteria. The aggregate scores are found to differ significantly depending on the multicriteria method used to reach a result on the level of the category. 'ELECTRE Tri' method is recommended. The case assessment is reported to have taken 36 days in total, a speed up compared to previous methods. The validation of indicators is partly achieved. The subsequent practical legitimacy of the assessment (to confirm if the method actually supports

consensus etc) is not addressed. A peculiar detail is that it was not possible to involve civil servants in the ‘societal validation’ process, presumably because of time constraints.

In summary the ‘3S’ method provides a rigorous framework and procedure for indicator assess-



ment. Its core methodology is a qualitative (expert and stakeholder based) assessment of pre-defined indicators using individual criteria combined with multicriteria methodology. The method as such is not dependent on the exact criteria or categories suggested by Cloquell-Ballester et al. It could be applied to any of the three (or other) ‘S’ groups. It is assumed but not verified if application to all three groups enhance the overall legitimacy. The procedure is illustrated in Figure 3.

Figure 3. ‘3S’ methodology for indicator validation as proposed by Cloquell-Ballester et al (2006)

5.2.4 Rice and Rochet’s framework for selection of a suite of indicators

Rice and Rochet (2005) provides one of the most detailed reports of approaches to the selection of indicators, as applied in the context of fisheries management. It involves a procedure with eight steps, as shown in Table 27. The potential indicators to be assessed in the case example measure either the conditions of the fish stocks or the environmental conditions for fishing. After deriving the method, two of the critical steps are tested (see Rochet and Rive, 2005) using trial groups of experts assessing candidate indicators for a range of marine ecosystems.

As the detailed account is very helpful for considering the practical application of indicator criteria each step of the process is briefly summarised here.

1. Determine user needs
2. Develop a list of candidate indicators
3. Determine screening criteria
4. Score indicators against criteria
5. Summarize scoring results
6. Decide how many indicators are needed
7. Make final selection
8. Report on the suite of indicators

1) In the first step the purpose of the indicator set is defined. This involves considering the functions the indicators should serve (e.g. management or information), the specific management objectives that may have been defined, and the identification of the relevant associated user groups (e.g. experts/advisors, decision makers/managers, general public).

2) Candidate indicators are identified and listed. This requires expertise about the impact area (in this case marine ecosystems), the source of the impact (in this case fisheries, and other influencing activities), and may also require understanding of the range of values and concerns generally present in the area, in case the final indicators are to be used for management or policy purposes. The available knowledge that the indicators are based on must be presented

3) Criteria are defined for assessing the indicators. General criteria are identified and summarized from the literature (as shown in Table 8). The relative importance of each criterion in the specific context must then be established. No categorical grouping of the criteria is undertaken, but an assessment is made of how three different user groups (Technical experts, Decision Makers, General audience) are likely attach importance (High, Moderate, Minor) to each criterion.

4) The scoring of the candidate indicators involves two elements, the scoring itself, and an assessment of the quality of the evidence available about the indicator. It is recommended to base scoring on simple ordinal rankings (e.g. 1-5), and avoid overly sophisticated scoring methods that may disguise the inherently subjective judgments made. The assessment of the evidence is helpful to retain as much objectivity as possible. A typology of evidence types is provided, and a simple hierarchy of evidence is proposed to support assessment for each criterion (see Table 28). For example 'High confidence' in an indicator for the criterion "theoretical basis" is assumed if it is not contested in the scientific literature, as confirmed in multiple publications. Table 27 shows the types of evidence proposed, organized in a tentative hierarchy. Each criterion is sustained by a different combination of potential evidence bases.

Conclusive published experimental research using Strong Inference
Multiple independent Publications providing consistent findings
Formal designed Surveys
Multiple independent Models producing consistent results
Interdisciplinary Consensus of weight of evidence
Research Team professional Judgement

5) The results of the scoring is summarized and presented. It is advised against ranking indicators based on simple multiplications of weights and scores by criterion. For example there may be some criteria where a low performance is unacceptable. Differences in quality of evidence should also be noted and kept in mind.

6) Deciding how large the indicator set should be. This refers not to the assessment of individual indicators but to a situation where joint consideration is required. Generally there is a conflict between having all system components represented with indicators versus a need for a manageable list for relevant functions and uses.

7) Final selection. It may be necessary to repeat the assessment process after some time, as evidence and experience may develop.

8) A report is drawn up. A number of ways to present and aggregate indicators are discussed with a view to associated strengths and weaknesses. Bias can be introduced in the presentation, even if the previous steps have been conducted rigorously.

The accompanying paper (Rochet and Rice 2005) reports a test of steps 3 and 4 of the methodology (weighting of criteria, and scoring of indicators) for a series of marine ecosystems by a group of 16 experts. A number of conclusions of general relevance are drawn. First of all the process of selecting indicators was found to be affected by subjectivity and value judgments, despite the relative rigour of the method. This is partly because most criteria consist of several sub-dimensions, that may not be formally resolved. Experts were found to assign different weights to criteria, and to score indicators differently. The scoring was considered by participants and the most difficult and partly arbitrary step. The variations were smaller among experts who were specialists of the same ecosystem. The requirement to write down justification of indicator scoring did not improve consistency, but contributed to improve transparency and communication in the process. Using a longer list of simple criteria gave less controversial results than using fewer and more complex criteria. According to the study the solution is not to introduce more formal quantitative approaches but to confront the differences of view explicitly.

In summary the work reported by Rice and Rochet (2005) provides a more detailed framework and guideline for assessment of indicators than the previous studies. It specifies the practical steps involved, and suggests specific approaches and methods for each step, from definition and weighting of criteria, to assessing available knowledge, to scoring, to reporting results for a suite of indicators. Like in the previous studies it acknowledges the different perspectives of various groups (experts, decision makers, etc) although here this is addressed by experts assigning presumed criteria weights for each group. The method is considered useful and applicable overall, but not as a way to eliminate subjectivity or personal perspectives from the selection of indicators, which is in fact ruled out. Parts of the guidance is specific to the area of fisheries management (e.g. the ranking of types of evidence, and conclusions regarding specific indicators).

5.2.5 Indicator assessment guidance from NCHOD, UK (2005),

Finally we consider the guidance for indicator assessment provided by the National Centre for Health Outcomes Development (NCHOD) in the UK. The guidance is based on assessment of a range of criteria and methods found in 18 different references. Criteria are divided into four groups: scientific criteria; policy criteria; methodological criteria; and statistical criteria.

For each criterion a guideline is provided in the form of a question to ask. For example for the criterion 'validity', the proposed question is "Will the indicator measure the phenomenon it purports to measure i.e. does it make sense both logically and clinically?"

In addition to the criteria a range of possible sources of evidence to help apply the criteria to potential indicators is mentioned. The source types are not ranked. They include,

- Expert opinion using rating scales

- Systematic literature review
- Audits/ survey of the measurement process
- Statistical analysis of output

The proposed scoring system is simple and ordinal, with 5 groups from 'very poor' to 'very good'

As a unique element of the studies considered here, the NCHOD report suggest a distinction among criteria according to which state of development the indicator is in, whether it is under development, whether it is in the measurement phase, or whether the results are to be interpreted. These stages are to be considered consecutive and exclusive, meaning that an indicator should not proceed to the next stage (e.g. 'measurement') if it does not score sufficiently in the previous one (e.g. 'development')

In the 'development' phase both scientific and policy should be applied, involving criteria such as 'policy relevance' and 'actionability' on the policy side, and 'validity' and 'explicit definition' on the scientific one

In the 'measurement' phase, methodological criteria, such as (obviously) 'measurability', 'sensitivity', 'timeliness' and 'cost-effectiveness' should be applied.

In the 'interpretation' stage statistical and criteria becomes relevant, including (obviously) 'interpretability', 'comparability', 'data quality' and 'reliability'

Summing up the NCHOD (2005) provides a very simple general guideline for the assessment of potential indicators for health and health policy. The criteria are similar to ones proposed in many other references, as is the use of simple ordinal scoring. Special consideration is given to which criteria are relevant in different stages from the development to the interpretation of the indicator.

5.3 Summary discussion of criteria methods and frameworks

The review of criteria based methods makes it clear that there is more to the selection of indicators than simply to assess them using a set of universal criteria. First of all a rich palette of criteria – more or less well-defined - is available to pick from the literature, but a universal list of criteria for assessing indicators does not exist. Although similar, each reference has its own set of criteria, and definitions do often deviate to some degree among references. The criteria as defined and reported in literature are rarely simple and uni-dimensional. Rather they are typically composed of several concepts or sub criteria, which may disallow a fully objective and transparent application of any one criterion. Hence, application of criteria is prone to inherent subjective bias and interpretation. An option is to break down criteria into sub-components, leading to more, and possibly more unique criteria. According to Rochet and Rice (2005) such longer criteria lists may be less controversial to apply than more condensed ones. However they could also increase the risk of overlap, excessive work loads, and missing ability to score some criteria .

Next, It is generally recognized that the relevance and applicability of criteria vary according to a range of aspects. Some criteria are useful for scientific assurance of accuracy, others concern the operation of monitoring systems, and others again cater to policy making and management. Hence the purpose of the indicator matters for how relevant each criterion is and how much it should then count when scoring indicators in a particular situation. However, there is not a generally agreed way to classify criteria according to situation or need. One way to approach this problem is to let indicator users apply weights to each criterion before the indicators are scored. In the framework of Rice and Rochet (2005) this is performed by experts assuming different usage positions. In other approaches like the ones proposed by Bockstaller & Girardin(2003) and Cloquell-Ballester et al (2006) the assessment of the indicators is undertaken in consecutive 'validation' rounds involving different groups e.g. of scientists, epistemic communities, managers or public users. Generally it is assumed that scientists and experts are the ones most concerned

with criteria for accurate representation, such as validity, reliability and sensitivity, but these criteria are also not irrelevant to or generally disregarded by other groups.

The application and weight of each criterion may depend on what kind of evidence is available. Different types of evidence may be used, depending partly on which type of phenomenon the indicator is supposed to measure (e.g. if it measures a physical dose-response relation, or if it measures the satisfaction with a given condition). In some research areas (like fisheries management), a hierarchy of methods may be established, allowing a transparent assessment of the strength of the evidence behind the indicator scoring, while this is not necessarily the case in all areas (or a hierarchy may have to be established). As pointed out by Innes (1975), the types of evidence is likely to affect the trust that policy makers and other users bestow on the indicators, where generally indicators based in theory as well as confirmed by statistical correlation is likely to be most easily accepted. However, in many cases evidence in the form of 'expert judgement' seems the only feasible approach. According to NCHOD (2005) the stage in the development of the indicator can also be a consideration in connection with choice or weight of criteria.

The actual assessment of indicators is typically done by individuals or groups, using simple scores with a limited number or ordinal levels, say 3 -5, and a similar or lower number of ordinal criteria weights. In some situations more refined numerical scales may be applied, and some methods applies mathematical tools to reach aggregate scores and ranks of indicators, as exemplified by the multi-criteria approach of Cloquell-Ballester et al (2006). In that example the outcomes was aggregate scores of indicators for categories of criteria like 'conceptual coherence'; 'operational coherence' and 'utility'. However Rice & Rochet (2005) strongly warns against too sophisticated or purely quantitative methods for assessment. This assumes homogeneity in the knowledge available for each indicator, and may mask subjective interpretations of criteria. In short overly sophisticated methods may belie the ambiguousness of the underlying knowledge.

A general observation is that explicit criteria are useful or even essential for the rational assessment and selection of indicators, but application of criteria is sensitive to purpose, type of problem addressed, users applying them, available knowledge, stage in the process, and other factors, and the processes should in no way presume to neutral or objective. As it has been formulated,

"The creation and use of standard procedures for the selection of ecological indicators allow repeatability, avoid bias, and impose discipline upon the selection process, ensuring that the selection of ecological indicators encompasses management concerns" (Dale & Beyeler 2001, p 6),

Systematic approaches may eliminate some of the randomness, and in any case help to increase transparency and dialogue,. However they are generally not yet developed to prescribe 'secure' methods to find the best possible indicators.

5.4 Discussion of criteria and 'joint consideration' of indicators

A main theme of COST Action 356 is aggregation and other forms of so-called joint consideration of indicators. This has been the topic of task 3.2 of the Action and the work reported in Chapter 6 of the Scientific Report.

The present report has not directly addressed criteria or methods for joint consideration, although some aspects have appeared. There are at least three – partly overlapping - ways in which indicator assessment and joint consideration are connected:

- Criteria for putting together series, sets, or suites of indicators to be considered jointly
- Criteria to assess indicators in the form of indices or composite.
- Criteria for methods to produce aggregate results

These ways will be discussed briefly one by one in the following, and a summary. made.

5.4.1 Criteria for putting together indicator series to be considered jointly

Indicators are usually not developed and applied one by one, but used in a context of a set or suite of indicators. The set of indicators may be defined by the aim to provide a comprehensive picture of a system, or by the need to cover a defined set of policy objectives. It may also reflect different management needs, such as early warning, system surveillance, or policy evaluation. As discussed in section 2 of the Scientific Report of COST Action 356 the structure of indicator sets are some times guided by conceptual frameworks such as DPSIR. Thus the extension of a suite of indicators is dependent on the purpose and framing of reporting the indicators. In the context of Environmentally Sustainable Transport around 50 indicators would be needed to ensure a description of all environmental impacts by at least one indicator (see chapter 2 in COST Action 356 Scientific Report) In practical applications fewer impacts may need to be addressed in many situations, while more than one kind of indicator may on the other hand be needed to indicate separate stages in the causal chain from cause to effect. Aggregation methods A (see below) may again allow consolidation across chains, and so on. The aim should be to have a complete and non-redundant representation of all dimensions of the phenomenon to be indicated, however defined and delimited.

According to e.g. Niemeijer and de Groot (2008), it is often not clear how the context of indicator use or the particular framework influences actual indicators chosen. Many studies do not report precisely how they compose their sets, and it can be impossible to trace why indicator reports with seemingly similar purpose and framework report different indicators for the same impacts. This is not only a flaw with some reports, but seem to reflect a more general methodological gap. In their review of indicator assessment criteria, Niemeijer and de Groot finds that only three criteria out of 34 reported in the literature (see Table 9) refer to criteria concerning the interlinkage of indicators. The authors sketch two different approaches to structure indicator sets. One corresponds to a traditional procedure where groups of experts seek the scientifically best indicators for individual ecosystems or the like. Another one is to apply a causal logic to derive an interconnected suite of indicators for a whole causal network.

Rice and Rochet (2005) make some observations concerning the composition of an indicator set for joint consideration even if their starting point is still a 'traditional' one. First of all there is the dilemma of balancing between few and many indicators. On the one hand there is a need to minimize the set of indicators to what is necessary to manage a system (e.g. only one key indicator per management objective); on the other it is desirable to have trustworthy indicators for all key components of the observed system, especially if outside factors could influence it independently of management interventions. Hence Rice and Rochet (2005) suggest that the number of indicators needs to be sufficient to distinguish changes from the source of interest (say fishing, or transport) from other influencing factors. If many similar causes are able to provoke the same change in the impact, more indicators may be needed to separate the causations. A simple example in the transport area could be taking into account other sources to, say, ambient noise, when readings of monitoring stations are used as indicators of traffic noise. Hence the appropriate number of indicators could partly be a function of the complexity of the system itself, partly of the management objectives, and partly of the degree of interactions with other systems. Another suggestion from Rice and Rochet is that the suite of indicators should be composed of items that perform differently with regard to the chosen criteria, in order to avoid 'blind spots'. It should be avoided too have many indicators in a set that perform well on one criteria dimension (say, reliability) but none that comply well with others (say, transparency) .

All in all there is no absolute guidance for composing indicator suites to fulfil the requirements of complete and non-redundant representation. Factors to take into account seem to include complexity of the observed system, management objectives/reporting purpose, interfering interactions with other systems, and a balance of the indicators themselves with regard to performance on the criteria mix.

5.4.2 Criteria to assess indicators in the form of an index or a composite

Many indicators for individual environmental impacts have the form of an index, say of air quality (Franceschini 2005), soil quality (Barbioli et al 2004) or biodiversity (Mace, G.; Baillie 2007). Other indices or composites cut across impacts or even across sustainability dimensions (economic, environmental etc). Indices and composites are specially derived measures, but once produced they are supposed to serve as indicators, in a principally parallel way to other, less elaborate ones. In this respect it must be appropriate to subject such aggregates to fulfil at least the same criteria as any other indicators. Hence, as a starting point, the criteria reviewed in sections 3-4 of this report, and the consolidated list of criteria presented in Table 26 could also be applied to indices and composites.

Some of the 10 criteria may seem more obvious to apply in this context than others. However, most of them would be applicable. In terms of scientific representation the 'validity' of a composite indicator would for example depend on the existence of a plausible conceptual assumption uniting the components (e.g. notions of health, wealth, entropy, or other). Validity would be low if the index is simply a combination of disconnected variables. The 'reliability' is also crucial. If a composition is based on a single clear algorithm any unreliability would come from 'noise' in the measurement of the subcomponents. Several underlying noise sources could however amplify unreliability at the composite level. Böhringer and Jochem (2007) found that 11 sustainability indices all suffered from various fundamental flaws in terms of scientific requirements.

Operational concerns such as 'data availability' can be critical for composites building on large data sets for a number of entities. An earlier version of the so-called Environmental Sustainability Index was for example criticised for introducing bias via the many averaging assumptions made to compensate for missing values in the data set (Niemeijer 2002).

Applicability in general communication or policy debate is often highlighted as a main service of composites (Saisana and Tarantola, 2002). Ideally they can allow interaction or decision processes to rid itself of unnecessary detail. While 'transparency' does not have to be sacrificed in composite indices, it can of course be a problem in some cases. 'Actionability' can also be low, as a meaningful response to 'alter levels' of a composite measure will require disaggregation to the critical subcomponents. One exception can be aggregate indicators with a predefined threshold set for accepting or rejecting cases, such as Net Present Value indicator used in economic assessment of transport projects (Lee 2000). This assumes sufficient agreement over the meaning of the threshold, which goes back again to validity.

All in all it seems clear that the proposed criteria could be applicable to indicators built for joint consideration. However it should also be clear that much is ignored as the criteria only scratch the surface of the underlying methodologies questions, as will be considered next.

5.4.3 Criteria for methods to produce aggregate results

DeMontis et al (2004) is an example of a text explicitly addressing criteria for aggregation methods, more specifically a set of seven multicriteria decision aid (MCDA) methods. The study finds that the methods differ in regard to which criteria they use, how they are assessed, how weights are assigned and so on. DeMontis et al define a set of criteria to characterise and evaluate the methods regarding their usefulness in a context of sustainable development assessment.

The criteria (see DeMontis et al 2004, p 2) are organized in three groups somewhat similar to what is used in this report:

- 1) operational components of MCDA methods,
- 2) applicability of MCDA methods in the user context, and
- 3) applicability of MCDA methods considering the problem structure.

While the first category refers to theoretical aspects and the second one is straightforward, the third one is particularly diverse and interesting as it aims to provide guidance on choosing methods with regard to different decision making situations (e.g. applicability at different geographical

scales, possibilities to combine methods, and possibility to include sustainability thresholds and sustainability. The criteria are given in Table 29.

Quality Criterion	Description
<p>Operational Components</p> <p>Interdependence Completeness Non-linear preferences Transparency of weighting Meaning of weights Solution procedure Results</p>	<p>Allowance for the interdependence of different criteria Need for the completeness of the criteria Possibility to express non-linear valuation patterns Type of the procedure of deriving values for the weights Interpretation and role of weights in the evaluation process Type of procedure used for the comparison of alternatives Interpretation of the results generated by the use of method</p>
<p>User Context</p> <p>Costs Time Stakeholder participation Problems structuring Tool for learning Transparency Actors communication</p>	<p>Implementation costs in the specific user situation Implementation time in the specific user situation Possibility to include more than one person as decision maker Existence of mechanisms supporting the structuring of the problem Support of learning processes Promotion of transparency in the decision making process Support of the communication between opposing parties</p>
<p>Problem Structure</p> <p>Geographical scale Micro-macro-link Societal/technical issues Methods combination Type of data Risk/Uncertainties Data processing amount Non-substitutability</p>	<p>Applicability of different geographical scales for one case Applicability of different institutional scales for one case Possibility for the consideration of both societal and technical issues Possibility of methods' combination Type of data supported as values for the indicators Possibilities for the consideration of risk and/or uncertainties Processing amount needed to compile data Possibility to consider sustainability and non-substitutability</p>

While there are some overlaps to the criteria for indicators discussed in this report (like a concern for transparency), the MCDM quality criteria are clearly quite different as they characterize aspects of methodologies rather than individual output variables (composite or not). It also seems that the DeMontis et al 'criteria' are to some extent descriptive rather than strictly evaluative. In the assessment of the methods a different way to 'score' the methods is applied for each criterion, only a few involving a clear ranking (namely 'transparency' which can be from 'high' to 'low'), meaning that an overall assessment is not straightforward to make.

The message is that a 'best' method cannot be defined independently of the case to which it is applied and a more detailed assessment of the methods. However four groups of criteria are highlighted as particularly relevant for situations where sustainability is to be assessed. According to this analysis, for example, only the methods Multi-Objective-Programming (MOP), Goal Programming (GP) and ELECTRE Tri, allows to set explicit sustainability thresholds. Again a clear ranking is not possible, as each method has its strengths and weaknesses (DeMontis et al 2004, p 21).

The final result is a kind of checklist where key features of sustainability assessment are highlighted to allow a tentative general allocation of MCDA methods according to which sustainability assessment feature the potential MCDA user may emphasize, including the following:

- Correspondence with social welfare theory (could be equated with weak sustainability)
- Possibility to Involve conflicting interest groups (a governance approach)
- Supporting a learning process among decision makers (a 'process of change' approach)
- Assessment with regard to thresholds, constraints or non-substitutability (a strong sustainability approach)

Further discussion of criteria for aggregation or joint consideration methods will not proceed here, as the subject clearly extends beyond indicator criteria and methods to apply them. It is dealt with in another part of COST Action 356 work.

5.4.4 Summary of criteria and 'joint consideration' of indicators

Joint consideration of indicators may take several forms from building a suite of indicators within or across impacts to constructing indices and composites aggregating across several or all impacts or dimensions (here called 'aggregates'). Criteria may be defined for suites as well as for aggregates, although for suites methodologies based on 'criteria' seem not to be so developed and would possibly also be insufficient. For aggregates, the same criteria as for individual indicators could and (probably) should be applied, but also here such criteria are far from sufficient to assess, score or rank the underlying methodologies.

The methodological aspects of joint consideration methods are discussed in Chapter 6 of the Scientific Report from COST Action 356.

6 Proposed approach and recommendations

This chapter will propose possible next steps based on the results obtained. First two general approaches to assess indicators for transport and environment will be briefly outlined indicating possible avenues for further work, and then a set of more specific next steps will be suggested for work within COST Action 356

6.1 General approaches for the assessment of EST indicators

It seems generally feasible that indicators for environmental impacts and sustainability of transport could be assessed, developed or selected using general quality criteria as discussed in this chapter, to help ensure that the indicators measure what they are supposed to, and able to serve the functions for which they are intended. Such assessments could well be based on systematic methods to identify and apply appropriate criteria, like for any other area, as described in the literature.

However, 'EST', is not a well defined unit 'in nature' that can be delimited, described, measured and indicated in a similar way as, say, an ecosystem or a unique endpoint. Indicators for individual environmental impact chains and endpoints will have to be derived using indicator assessment criteria and methods similar to the ones described in section 4 and 5, The role of transport as an element of a wider range of sources to that impact must be considered, and then combinations or aggregates of indicators for transport and environment as a whole will have to be derived and assessed taking into account additional aspects such as system delimitations, policy context, and balance of indicator performance with regard to criteria.

As demonstrated in Chapter 2 in the COST Action 356 Scientific Report, transport is a contributor to a wide range of environmental impacts. For some impacts many indicators likely exist, while for others there may be few or none. It is not clear in advance for which impact chains indicators are available to allow a criteria based selection process, or indeed if such a generic process can provide a valuable result. Moreover, as described in Chapter 3 in the COST Action 356 Scientific Report the use of the indicators in transport also involves a large number of different policy and decision making situations. These are likely to affect which types and combinations of criteria would be appropriate, and how much each criterion should weigh in each case. It is not clear in advance exactly what the situational context means for indicator selection, and therefore which (or how many different) kinds of guidance would be appropriate.

Two types of approach could therefore be considered,

a) *Generic assessment*

a general, or generic assessment process, where potential indicators of the environmental impact of transport are assessed, for all impacts, one by one. The aim should be to establish for each area, to what extent good 'measurement' indicators exist, or if new or better indicators are needed. A further possibility is to score and rank candidate indicators if possible. This work should probably be conducted mainly by researchers /advisors, where contributions from policy makers/managers or other external users/stakeholders could be valuable additional input.

b) *Situation dependent assessments*

to develop procedures and templates for the identification, assessment and selection of indicators for specific policy, planning or decision making situations, taking into account how various criteria may be combined and weighted in order to reflect specific needs or situations. The most typical policy, planning or decision making situations should be taken as a starting point. This procedures and templates should probably be worked out in a collaboration between researchers /advisors, policy makers/managers and external users/stakeholders.

6.2 Approach and guidelines for subsequent work within COST Action 356

The work in task 2.3 in the work programme of COST Action 356 (reported in chapter 5 of the main Scientific Report of the action) aims to construct or select 'indicators per environmental impact' using criteria and methods as identified in the present report.

The context of that work fits with approach a) above, although the time, capacity and expertise to assess indicators in depth for all environmental impact chains identified in Chapter 2 is not available in the COST Action. Hence this approach will be tried out in a more limited scale than reported in the literature. It will involve scanning a small, but varied range of impact chains in order to identify possible indicators, and to undertake a first brief assessment of them according to criteria as identified in this report (Table 26). Key objectives of this work are to identify and review indicators per impact, to try out and reflect on the 'generic assessment' approach, and to discuss how indicator availability and quality vary across selected different impact areas.

The process can resemble the first step, 'sui validation' (or 'self' validation by a working group), in the three stage methodology of Cloquell-Ballester et al (2006), or the 'research team' efforts of Rice & Rochet (2005), but it should be emphasized again that the effort here is more limited, where, for example only one or two experts have been involved in identification and assessment of indicators for each impact.

The following guidelines refer to the literature review and discussions in this report and in particular using the list of criteria presented in Table 26. The aim is support the assessment of indicators for a limited number of environmental impacts of transport selected among those identified in Chapter 2 of the Scientific Report. The approach is intended to be simple, manageable and comparable.

The work on 'joint consideration' of indicators to be reported in Chapter 6 of the Scientific report, may fit with and support approach b) above. This aspect is not considered further in the following.

6.2.1 Considering what is to be indicated

For the assessment of each selected impact its title and main contents should be given to clarify 'what is to be indicated'? This involves a reflection of whether the chain or impact is clearly defined or not in terms of causes and effects. If it is not clearly defined it is more challenging to suggest good indicators. If the role of transport in the impact is unclear it is also more difficult to suggest good indicators. If there are several dimensions involved in the impact itself (e.g. different endpoints for the impact) this may also challenge the identification of adequate indicators.

6.2.2 Considering situation(s) where the indicators are needed

Assumed need and purpose of the indicators can further help to specify what the indicators are supposed to describe and evaluate. The basic option is to imagine that the indicator is assessed as a generic descriptor of the impact chain without any particular purpose in mind (as we here consider generic types of assessment). Reviewing the overall appropriateness of the indicator could however be helped further by imagining different application situations. We propose two case examples which are likely to require different types of indicators (and emphasize different criteria), to serve as imagined background

- a) indicators are to be used for an environmental monitoring program that will help scientists perform a major environmental assessment of a national road network in 5 years time from now.
- b) indicators are to be used in decision to locate near road link in location A or location B; the project is delayed and the decision timing allows only 3 months to prepare the assessment.

6.2.3 Weights and aggregations of criteria according to categories or situations

In this limited approach we do not propose to rank the criteria. It would nevertheless be possible as an experiment to let the three situations described in the previous section each introduce a filter,.

- the 'generic case' could suggest to consider only the three 'representation' criteria of Table 26.
- situation a) could suggest to consider all criteria equally
- situation b) could suggest to give higher weight to criteria like measurability, data availability and policy relevance.

For the present exercise, it is proposed to avoid weighting and provide only qualitative remarks..

6.2.4 Describing the candidate indicators

Potential, or 'candidate' indicators are described. The indicator descriptions can not avoid reflecting the specific character of indicators for each unique impact type. However the descriptions should be to some extent harmonized. Some key elements to consider (if not necessarily copy) for each candidate indicator include:

- definition
- formula (if applicable)
- single or multiple dimension (ex index) indicator?
- location in DPSIR type chain
- amount of documentation available, e.g. 'multiple scientific sources'/'few scientific sources'/'
- example of use in practice e.g. for transport assessment, monitoring, evaluation

6.2.5 Scoring each potential indicator with all ten criteria

The candidate indicators are scored using the criteria in Table 26. It is proposed to use a simple four level ordinal ranking, 1) 'Poor', 2) 'Limited', 3) 'Good', 4) 'Excellent'. The assessor (author) will have to use his/her own best judgment, and possibly consult literature.

There will obviously be different ways to use and interpret the scoring. The meaning of the 'quality' of an indicator and the associated score has to be considered individually for each criterion. It should be noted to what extent an assessment score refers to the actual quality of the indicator, or to the degree of available knowledge about it. Scores 'good' or 'excellent' should only be given to indicators that are well established in research or otherwise well documented.

The scoring can be undertaken for one or more of the hypothetical situations described above, but the default should be the one where all criteria weigh equally, unless there are reasons to deviate from this, which should then be explained.

6.2.6 Summary assessment

A summary assessment should be made for the set of indicators considered per impact taking into account scores on all criteria. The summaries could – if applied - be made for each of the hypothetical situations (see above) and then compared, e.g. looking for indicators that would perform well in several situations, versus ones that would differ if criteria were differentiated or weighted.

Generally, a summary assessment could aim to,

- Optimize: rank the indicators according to performance on all criteria to choose the best indicator
- Satisfy: allow to identify one or more indicators that are passing some defined threshold and become 'recommended' (e.g. hypothetically: "at least 'good' or 'excellent' for 6 out of 9 criteria, and none with 'poor' ")
- Reject: allow to discard some indicators
- Fuzzy optimization: allow a qualitative distinction between 'better' and 'worse' ones to choose

The suggestion here is to seek only a fuzzy or qualitative type of summary assessment, pointing out

- If the candidate indicators score differently or are more or less the same level
- If there are indicators which appear to be good or excellent with regard to all or most criteria
- if the indicators score differently in the different hypothetical situations (if applied)
- If there appears to be a need for building new better indicators

7 Conclusion

A process to derive criteria and methods for the assessment and selection of EST indicators has been undertaken. The process has evolved through a combination of literature review and working group discussions. The review has included general indicator literature in areas like environmental assessment, health, resource management, sustainability, as well as literature more specifically on transport and sustainable transport indicators. The working group discussion have addressed particular indicator needs and criteria of relevance for measurement and assessment in the area of environmentally sustainable transport.

It was found that there are many similarities in the indicator assessment criteria applied throughout the literature, although far from a full consensus. Often basic terms are defined in different ways while the same terms are sometimes assumed relevant for opposite ends of the indicator selection process. The transport indicator literature is not always explicit about criteria but tend to import similar criteria as used in other fields, while stressing also a special concern for the transport sensitivity of environmental indicators. The general literature documents a number of methods and frameworks for how to apply the criteria when indicators are to be assessed in different situation with regard to problem or indicator function. No such examples were found in the specific transport literature, but may exist.

An important aspect of the methodologies is the relative sensitivity or importance of indicator criteria with regard to different contexts such as different indicator purposes and functions, different development of the knowledge, or different user groups.. Many attempts are made to categorize criteria into types to reflect such contexts with low agreement over the exact categories to use. In the present work a distinction of contributions from the literature initially grouped criteria into 'measurement', 'monitoring' and 'management' oriented ones, which was subsequently further adjusted into the three groups of criteria for 'representation', 'operation' and 'application'. 10 criteria were highlighted and equipped with interpretation and examples . However the partly arbitrary character of such a list must be recognized, and a need to draw in additional or other criteria if relevant must be retained. The notion of conflicts over facts and values introduced in chapter 3 may also be an underlying factor behind situations where criteria are interpreted, accepted or suppressed. Still, it is common for published methodologies to suggest differentiation or (simple) weighting of criteria according to various contexts, as a superior alternative to considering all criteria always equally important.

Several studies posit indicator development as a process that should involve connections between indicator validation contexts, which means, first, that most types of criteria would be prone to play some role at some point in the process, and second, that the final indicator selection is likely to depend on many other factors than formal criteria and associated weights.. The scoring of indicators themselves are often made with simple ordinal scales administered by experts or sometimes wider groups of stakeholders. Sophisticated multi-criteria methods to allow ranking of candidate indicators have been applied in some cases. while other scholars posit that this may mask underlying inherent ambiguities and subjectivity, and advice against letting criteria based indicator assessment assume a disguise of 'rocket science'.

Based on the review it was recommended to promote further work in the EST area on indicator criteria along two routes tentatively called, respectively, *generic assessment*, meaning a general assessment of potential indicators per environmental impact focusing on scientific measurement aspects, and *situation dependent assessments*, involving methods to define how transport policy, planning, decision making or governance contexts would effect selection, ranking or application of criteria in particular cases where multiple impacts must be considered.

A simplified methodological guideline was drawn up for an internal trial of the generic type assessment type to be undertaken for selected environmental impacts in subsequent part of the COST 356 work.

Criteria to assess individual indicators may also be applied to aggregates, but further analysis is required to address the use of criteria or other methods to appropriately assess methodologies for joint consideration cutting across environmental impacts of transport.

References

- AMBIENTEITALIA 2007. Urban Ecosystem Europe, Report 2007. Ambiente Italia, Rome: Available: http://ec.europa.eu/environment/europeangreencapital/about_submenus/green_thinking.html
- Babisch, W 2006. Transportation Noise and Cardiovascular Risk. Review and Synthesis of Epidemiological Studies- Dose-effect Curve and Risk Estimation. Federal Environmental Agency, Berlin, January 2006
- Barbiroli, C; Casalicchio, G. Raggi, A 2004. A New Approach to Elaborate a Multifunctional Soil Quality Index. *J Soils & Sediments* 4, (3) pp. 201 – 204
- Batalle, Cambridge Systematics, Texas Transportation Institute 2004. Traffic Data Quality Measurement Final Report. Federal Highway Administration. U.S. Department of Transportation Washington, D.C.
- Black, William R. 2002. Sustainable Transport and Potential Mobility. STELLA's Focus Group 4 Meeting 1, May 3, 2002
- Bockstaller, C. & Girardin, P. 2003. How to validate environmental indicators. *Agricultural Systems*, 76, 639–653
- Bollen 2004. Indicators: Methodology. pp 7282-7287 in: *International Encyclopaedia of the Social & Behavioral Sciences*, Elsevier
- Bossel, Hartmut 1996. Deriving indicators of sustainable development. *Environmental Modeling and Assessment* 1, 193-218
- Boyle, Michelle 1998. An Adaptive Ecosystem Approach to Monitoring: Developing policy performance indicators for Ontario Ministry of Natural Resources. Master in Environmental Studies degree from the University of Waterloo, Canada (1998). URL: http://www.nesh.ca/jameskay/ersserver.uwaterloo.ca/jjkay/grad/mboyle/th_pdf.html
- Breckenridge; R. E; Kepner, W.G.; Mouat, D. A. 1995. A Process for Selecting Indicators for Monitoring Conditions of Rangeland Health. *Environmental Monitoring and Assessment* 36. 45-60.
- Broughton, B; Hampshire J. 1997. Bridging the gap : a guide to monitoring and evaluating development projects. Australian Council For Overseas Aid, Canberra
- Cameron, K., M.H. Beare, R.P. McLaren, and H. Di. 1998. Selecting physical chemical, and biological indicators of soil quality for degraded or polluted soils. Proceedings of 16th World Congress of Soil Science. Scientific registration No. 2516. Symposium No. 37. Aug. 20-26, 1998. Montpellier, France
- CGER 2000. Ecological Indicators for the Nation. Commission on Geosciences, Environment and Resources (CGER). National Academy Press, Washington DC
- Cloquell-Ballester, V-A; Cloquell-Ballester, V-A; Monderde-Diaz, R; Santamarina-Siuranaet M-C; 2006. Indicators validation for the improvement of environmental and social impact quantitative assessment. *Environmental Impact Assessment Review* 26 pp 79 – 105
- Cole D.C.; Eyles J.; Gibson B.L. 1998. Indicators of human health in ecosystems: what do we measure? *The Science of the Total Environment* 224, 201-213
- Crocker, L 2001. Content validity. pp in: 2702-2705 *International Encyclopedia of the Social & Behavioral Sciences*, Elsevier

Dale, V.H ; Beyeler. S.C 2001. Challenges in the development and use of ecological indicators. *Ecological Indicators* 1, pp. 3–10

de Bruijn, Hans: Performance measurement in the public sector: strategies to cope with the risks of performance measurement. *The International Journal of Public Sector Management*, Vol. 15 No. 7, 2002, pp. 578-594.

De Montis, Andrea; De Toro, Pasquale; Droste, Bert; Omann, Ines; Stagl, Sigrid 2000. Criteria for Quality Assessment of MCDA Methods. Presented at 'Transitions Towards a Sustainable Europe, 3rd Biennial Conference of the European Society for Ecological Economics, Vienna, 3. - 6. May 2000

Dobranskyte-Niskota, A; Pejrujo, A; Pregl, M. 2007. Indicators to Assess Sustainability of Transport Activities. Part 1: Review of the Existing Transport Sustainability Indicators Initiatives and Development of an Indicator Set to Assess Transport Sustainability Performance. European Commission, Joint Research Centre, Institute for Environment and Sustainability, ISPRA

DTU Transport 2009. Transportvaneundersøgelsen. Available:
<http://www.dtu.dk/centre/modelcenter/TU.aspx>

EEA 2004. EEA Core Set of Indicators (CSI) 2004. Background document for. Joint Meeting 21-23 April 2004. European Environment Agency, Copenhagen.

EEA 2000. Are we moving in the right direction?. Indicators on transport and environment integration in the EU. TERM 2000. Environmental issues series No 12. European Environment Agency, Copenhagen

Eksler, V.; Allenbach, R.; Holló, P., and Schoon, C (2007). Protective Systems. In: Hakkert, A.S; Gitelman, V. and Vis, M.A. (Eds.). Road Safety Performance Indicators: Theory. Deliverable D3.6 of the EU FP6 project SafetyNet.
Available: http://euroris.swov.nl/safetynet/fixed/WP3/sn_wp3_d3p6_spi_theory.pdf

ERSO 2007. Traffic Safety Basic Facts 2007. European Road Safety Observatory. Available:
http://ec.europa.eu/transport/road_safety/observatory/observatory_en.htm

Eyles, J; Furgal, C 2000. Indicators in Environmental Health: Identifying and Selecting Common Sets. *Canadian Journal of Public Health*. Selected Papers from the Quebec City Consensus conference on Environmental Health Indicators, VII 03, October 2000, pp 62-67

Farchi, S.; Molino, N.; Rossi, P.G.; Borgia, P; Krzyzanowski, M.; Dalbokova, D; Kim, R.; 2006. Defining a common set of indicators to monitor road accidents in the European Union. *BMC Public Health* 2006, 6:183

Ferrari, Pier Alda; Salini, Silvia 2008 Measuring Service Quality: The Opinion of Europeans About Utilities. FEEM Working Paper No. 36. The Fondazione Eni Enrico Mattei, Milano April 2008

Franceschini, F.; Galetto, M; Maisano, D. 2005. A short survey on air quality indicators: properties, use, and (mis)use. *Management of Environmental Quality: An International Journal* Vol. 16 No. 5, 2005. 490-504

Gibson, R. B 2000. Specification of sustainability-based environmental assessment decision criteria and implications for determining "significance" in environmental assessment. *Research and Development Monograph Series*, Canadian Environmental Assessment Agency, Ottawa
Available:
<http://www.ceaa-acee.gc.ca/default.asp?lang=Frn&n=086E7767-1&toc=show&offset=1>

- Gilbert, R; Irwin, N; Hollingworth, B & Blais, Pamela 2002. Sustainable Transportation Performance Indicators (STPI). Project Report On Phase 3. The Centre for Sustainable Transportation, Toronto
- Girardin, P; Bockstaller, C.; van der Werf, H 1999. Indicators: Tools to Evaluate the Environmental Impacts of Farming Systems. *Journal of Sustainable Agriculture*: Volume: 13 Issue: 4, pp 5 – 21
- Goertz, Gary 2001. "Increasing Concept-Indicator Consistency: The Case of Democracy." Unpublished manuscript, Department of Political Science, University of Arizona, 2001.
- Goger T. & Joumard, R. 2007 A method of building an aggregated indicator of air-pollution impacts. 3rd int. conf. Sustainable development 2007, 25-27 April 2007, Algarve, Portugal
- Goger, T. Maranda K, Stein W, Karkalis A, Gerassimos, A 2006. Integrated Assessment of Environmental Impact of Traffic and Transport Infrastructure – A Strategic Approach. Part C. Chapter 4. WG 3 Environmental indicators. Available: <http://cost356.inrets.fr/>
- Guinee J.B., M. Gorree & R. Heijungs 2002: Handbook on Life Cycle Assessment. An operational guide to the ISO standard. Kluwer Academic, London, UK
- Hanafin, S; Brooks. A-M. 2005. Report on the Development of a National Set of Child Well-Being Indicators in Ireland. The National Children's Office, Dublin
- Hardi, P.; DeSouza-Huletey, J.A. 2000. Issues in analyzing data and indicators for sustainable Development. *Ecological Modelling* 130, pp 59–65
- Hauge, K. H.; Olsen, E.; Heldal, H. E.; Skjoldal, H. R. 2005. A framework for making qualities of indicators transparent. *ICES Journal of Marine Science*, 62, pp 552-557
- Hoppe, R 2005. Rethinking the science-policy nexus: from knowledge utilization and science technology studies to types of boundary arrangements. *Poiesis Prax* 3, pp 199–215
- Huijbregts M.A. 2000. Spatially explicit characterization of acidifying and eutrophying air pollution in life-cycle assessment. *J. Industrial Ecology*, vol 4, n°3, p. 125-142.
- Innes, J.E 1998. Information in communicative Planning. *APA Journal* vol nr 64 No 1, 52-63
- Innes, Judith Elanor 1990 . Knowledge and Public Policy. The search for meaningful indicators. Second expanded edition. Transaction publishers, New Brunswick, 171-188
- Innes, J. De Neufville, 1978 Validating Policy Indicators. *Policy Sciences* 10 (1978-79),
- IPCC 2001. Climate Change 2001, the Scientific Basis. Intergovernmental Panel on Climate Change, Cambridge University Press, 2001,
- Jackson, Laura E.; Kurtz, Janis C.; Fisher, William S. (eds.) 2000. Evaluation guidelines for ecological indicators. U.S. Environmental Protection Agency, Washington DC.
- Jeon CM and Amekudzi A 2005. Addressing Sustainability in Transportation Systems: Definitions, Indicators, and Metrics. *Journal of Infrastructure Systems*, March 2005. 31-50.
- Joumard, R 2008. Tentative to homogenize and simplify the list of criteria for indicator assessment. Lab. Transport and Environment, INRETS, Bron, Franc. COST Action 356 memo]
- Jørgensen, S.E; Costanza, R; Xu, F-L (eds.) 2005. Handbook of ecological indicators for assessment of ecosystem health. Taylor & Francis, Boca Raton.

Kunicina, N. 2008. Evaluation of decision making methods and its application for transport problems. STSM report, COST Action 356. Oct. 2008, 13 p.

Kurtz, J.C; Jackson. L E; Fisher, W.S. 2001 Strategies for evaluating indicators based on guidelines from the Environmental Protection Agency's Office of Research and Development. Ecological Indicators 1, 49–60

Kusek, Jody Zall and Rist, Ray, C 2004. Ten Steps to a Results-Based Monitoring and Evaluation System. A Handbook for Development Practitioners. The World Bank, Washington DC.

Lahti, P; Calderon, E; Jones, P; Rijsberman, M; Stuij, J (eds.) 2006. Towards sustainable urban infrastructure : Assessment, tools and good practice. Results of COST Action C8 "Best practice in Sustainable Urban Infrastructure". European Science Foundation, & COST, Brussels

Lee, D.B 2000. Methods for evaluation of transportation projects in the USA. Transport Policy, 7 pp 41- 50

Lenz R; Malkina-Pykh I.G.; Pykh, Y. Introduction and overview. Ecological Modelling 130 (2000) 1–11

Leviton, L 2001 External Validity, Content validity. pp 5195-2000 in: International Encyclopedia of the Social & Behavioral Sciences, Elsevier

Litman T 2008. Well Measured. Developing Indicators for Comprehensive and Sustainable Transport Planning. Victoria Transport Policy Institute, Victoria BC. Available: <http://www.vtpi.org/wellmeas.pdf>

Marsden,G; Kelly, C; Snell, C.; Forrester, J 2005. Sustainable Transport Indicators: Selection and Use. DISTILLATE. Improved Indicators for Sustainable Transport and Planning. Deliverable C1. University of Leeds; University of York

Mace, G.; Baillie, J.E.M 2007. The 2010 Biodiversity Indicators: Challenges for Science and Policy. Conservation Biology. 21, No. 6, 1406–1413

May, A. D; Grant-Muller, S; Marsden, G; Thanos, S. 2007. Improving the collection and monitoring of urban travel data: An international review. TRB 2008 Annual Meeting CD-ROM

McCoy, K. Lynn; Ngari, Patricia Njeri; Krumpel, Edwin E. 2005. Building Monitoring, Evaluation and Reporting Systems For Hiv/Aids Programs . Pact, Washington DC

Mitchell, G., May, A. and McDonald, A. 1995. PICABUE: A Methodological Framework for the Development of Indicators of Sustainable Development, International Journal of Sustainable Development and World Ecology, 2, 104-23.

NCHOD 2005. Compendium of Clinical and Health Indicators User Guide. National Centre for Health Outcomes Development (NCHOD), London site – London School of Hygiene and Tropical Medicine. URL: www.nchod.nhs.uk

NCSSF, 2005. Science, Biodiversity and Sustainable Forestry: A Findings Report of the National Commission on Science for Sustainable Forestry. National Commission on Science for Sustainable Forestry, Washington, DC.

Niemeijer, David; de Groot, Rudolf S. 2008. Conceptual framework for selecting environmental indicator sets. Ecological Indicators 8, pp 14 – 25.

Niemeijer, D. 2002. Developing indicators for environmental policy: data-driven and theory-driven approaches examined by example. Environmental Science & Policy 5 pp. 91–103

Niemi, Gerald J.; McDonald, Michael E. 2004. Application of ecological indicators. *Annu. Rev. Ecol. Evol. Syst.* 2004. 35:89–111

OECD Public Governance Committee 2008. Report on the Consultation with Stakeholders. Draft Checklist for Enhancing Integrity in Public Procurement. Organisation for Economic Co-operation and Development GOV/PGC(2008)6

OECD 2003. Environmental Indicators – Development, Measurement and Use. Reference paper, OECD, Paris 2003

Rackliff, L. 2008. Road Safety Performance Indicators: Database Development. Deliverable D3.12 of the EU FP6 project SafetyNet

Rice, J. C., Rochet, M.-J. 2005. A framework for selecting a suite of indicators for fisheries management. *ICES Journal of Marine Science*, 62: pp. 516-527

Rochet, M.-J., Rice, J. C. 2005. Do explicit criteria help in selecting indicators for ecosystem-based fisheries management? *ICES Journal of Marine Science*, 62: 528-39.

Rousseaux P. 1994. Proposition de descripteurs de suivi de l'état de l'environnement. Lyon (France) (Proposal of indicators to monitor the state of the environment in Lyon). Observatoire des changements écologiques du Grand Lyon. 39 p.

Saisana, M; Tarantola, S 2002. State-of-the-art Report on. Current Methodologies and Practices for Composite Indicator Development. Joint Research Centre for the European Commission. Institute for the Protection and Security of the Citizen, Technological and Economic Risk Management, Ispra

Segnestam, L 1999. Environmental Performance Indicators. Environmental Economics Series. Paper No. 71, World Bank, Washington D.C.

Tomlinson, Paul, 2004. The Role of Significance Criteria in SEA. Presentation at IAIA04 Workshop. 24-29 April, 2004, Vancouver.

Available: http://www.sea-info.net/files/general/The_Role_of_Significance_Criteria_in_SEA.PDF

Turnhout, Esther; Hisschemöller, Matthijs; Eijsackers, Herman 2007. Ecological indicators: Between the two fires of science and policy. *Ecological Indicators* 7 (2007) 215–228

US DOT 2007. Performance and Accountability Report FY 2007. United States Department of Transportation, Washington DC. Available: <http://www.dot.gov/perfacc2007/perfreportPST.htm>

US EPA 2006. Report of the 2005 peer review of proposed indicators for the U.S. Environmental Protection Agency's Report on the Environment 2007 (ROE07) technical document. FINAL REPORT January 4, 2006.

WHO 2006. Reproductive Health Indicators Reproductive Health and Research Guidelines for their generation, interpretation and analysis for global monitoring World Health Organization, Geneva.

Zietsman, Josias & Rilett, Laurence R. 2002 Sustainable Transportation: Conceptualization and Performance Measures. Report No. SWUTC/02/167403-1. Texas Transportation Institute, The Texas A&M University System, College Station, Texas

Zucconi, S.L; Carson, C.A. 1994. CDC's Consensus Set of Health Status. Indicators: Monitoring and Prioritization by State Health Departments. *American Journal of Public Health* 1645 October 1994, Vol. 84, No. 10.