11 Define Assessment Methods and Write Evaluation Plan

Figure 11.1: Define assessment methods and write evaluation plan within the STTRIDE Evaluation Process

Assessment methods

This step defines the assessment methods in more detail by determining which methods are to be used to meet each of the evaluation objectives defined earlier.

The assessment methods can be summarised in a table as shown Table 11.2. A flow chart is also recommended.

Once they are defined, the assessment methods are recorded in an evaluation plan. An outline of an evaluation plan is shown in Section 11.9.

Defining the assessment methods involves a series of stages as shown in Figure 11.2. They lead to the content of the evaluation plan as shown in Section 11.9 and in the template on the STTRIDE web site.

Figure 11.2: Stages in defining assessment methods

11.1 Define indicators

The indicators are the variables used to estimate the impacts and thus the extent to which the objectives have been met. They are relative measures, with a denominator that makes it possible to compare them per unit of distance, time, area, user etc. Indicators should be defined for each research question (and therefore linked to an evaluation objective), and in each case there should be a clear definition of how the indicators are to be measured or derived. Linking the indicators with evaluation objectives means that the data collection is focused on the main areas that have been identified for assessment, and less relevant data is not collected.

The indicators used should be able to meet two criteria:

- Clearly reflect the impact
- Can be assessed reliably using the methods available.

When defining the indicators, the availability of new types of data should be considered. It is often the case that new types of data are generated when new technologies are implemented; such data may be useful, if not central, to the evaluation. For example advanced fare management systems collect data on customers' use and habits, while cooperative ITS (C-ITS) services can collect data from vehicles to produce information on vehicle flows on routes at different times. In addition, technology advances are making new data sources available through 'Open Data' initiatives which can be useful for monitoring and evaluation.

The indicators will also vary depending on where the technologies are deployed; they may be in vehicles, carried or worn by travellers, or be part of a service delivery 'infrastructure' such as an operations room, a data 'hub' on or near roads, interchanges, etc.

With the focus on connected and multimodal journeys in STTRIDE, it will be important to include indicators reflecting use of the interfaces between modes; for example cycle parking at public transport stations, use of public transport interchanges and park and ride services.

Define indicators: HD Maps

The impact of and investment in HD Maps and road databases can be indicated at the output level in terms of the number of square miles mapped, or the frequency of updates related to roadworks, diversions etc.

At the outcome level, indicators might include the number of calls made by onboard technologies to these databases.

Impact indicators would relate to the rate of accidents and delays associated with roadworks, diversions etc. that have been added to HD Maps and databases.

Bearing in mind the STTRIDE focus on sustainable travel, it is worth noting that in some countries, specific indicators have been developed for assessing impacts associated with active travel (walking, cycling etc.) such as journey quality, physical activity, absenteeism, safety, environment, and time savings. Techniques have been developed to take account of the likely decay in impacts of some types of scheme over time and the variation in response to interventions between different types of user (such as commuters, leisure users and utility users).

It is recommended that a set of common indicators be defined for use in different National Road Authorities for assessing similar interventions. This will make it possible to make comparisons between areas and interventions, enabling road authorities to learn from each other.

Two types of indicator should be defined: those that are generic to any investigation of the impact of technologies on mode use, and those that are specific to certain technologies.

Suggested indicators that could be used to address the example research questions listed in the 'Frame the Research Questions' module on the STTRIDE web site are shown in Table 11.1 below.

Table 11.1: Example indicators

Example research questions	Example indicators
What is the scale of service delivery or amount of service provided, where, and when?	**Outputs** • Number of items of equipment/ vehicles sold/ provided/ equipped in different areas and times • Number of applications downloaded/ server calls • Number of km of network equipped with technology • Square km of communications coverage • Speed/ latency of communications • Number of leaflets distributed/ hits on information web site/ enquiries answered • Number of service providers engaged • Number of safety/ information messages delivered to different user groups
	Outcomes • Number of subscribers/ users • Number of new services established • Number of requests/ notifications during service delivery • Number of safety/ information messages received by different user groups • Number of incidences of users responding to information provided • Availability of service at planned time • Availability of service for on-demand use
Has there been a need to change other systems/ services as a result of introducing the intervention, and what is the impact of those changes?	• Number of systems/ services changed • Degree of impact of these changes on: operators, service providers, users

Example research questions	Example indicators
Has the new technology led to a new 'pain point' in the transport system?	• Features of pain point and number of users/ trips affected
How often does the service fail to operate as planned, and how long do these incidents last?	• Number of incidents in operating week when service fails to operate as planned • Mean, minimum and maximum duration of incidents in operating week when service fails to operate as planned
For what percentage of each day or week is the service fully operational/ what proportion of potential users have access to the service?	• Percentage of operating week when service is available for, and has capacity for, users • Percentage of all potential users who have subscribed to the service • Minimum time required between booking service and travelling
To what extent is the new system/ service interoperable with others/ existing services?	• Number/ proportion of existing/ other services that the new service is interoperable with
Are users aware of the service and informed about what it offers them?	• Number/ proportion of potential users who know about the service • Number/ proportion of potential users who have an accurate understanding of its key features
What proportion of attempted uses fail because the user does not have the equipment/ service needed to access it and how many potential users does this represent?	• Percentage of attempts to use the service which fail and why (because the user does not have smartphone, credit card, internet etc.) • Percentage of potential users who are unable to use for these reasons
Have users found the service easy to use, and if so to what extent/ how?	• Percentage of users rating the service as easy/ very easy to use • Shares/ likes/ ratings (where applicable)
How long does it take users to complete actions required to use the service?	• Time taken to complete actions to use the service
How frequently is the service used by individuals?	• Time elapsed between uses of the service
How is use of the service shared between different user groups?	• Percentage distribution of uses of service by different user groups

Example research questions	Example indicators
How do users think the new service compares with what they did before in terms of cost, quality, overall journey time etc.?	• Time before first use • Time between uses • Cost • Waiting time • Time spent using service • Number of links/ modes • Journey time • Number of referrals/ shares made to other potential users • Compromises made
How much were users willing to pay for the service, and how did this vary between types of user, types of journey etc.?	• Willingness to pay for service on commute journeys • Willingness to pay for service on business journeys (in course of work) • Willingness to pay for service on leisure journeys
Has there been a change in the number and/ or scale of delivery new products/ services that have been provided, and what has changed as a result?	• Number of products/ services available

Example research questions	Example indicators
Has the number of journeys made by each mode changed on the inter-urban network and or other roads, and if so in what way?	Data on indicators for all users and each user group: • Number of vehicles of each type on interurban and other roads • Number of trips by each mode on interurban and other roads • Number of connected/ multi-modal trips on interurban and other roads • Distance travelled by each mode on interurban and other roads • Percentage of trips/ distance by each mode on interurban and other roads • Percentage of trips on interurban and other roads that are connected/ multimodal • Correlation between distribution of messages and use of inter-urban network, such as percentage change in vehicles in response to messages about better routes, delays, alternative modes
Has the number of shared mode journeys changed on the inter-urban network and or on other roads, and if so in what way?	• Number of journeys by public transport or shared car on interurban and other roads/ routes
Has the number of multi-modal or connected journeys changed on the inter-urban network and or other roads, and if so in what way?	• Number of journeys involving two modes or more (excluding walk to access to public transport) • Number of journeys involving interchange within modes e.g. multi-stage public transport journeys • Number of journeys involving relevant combinations of modes • Number of journeys involving relevant interchanges e.g. use of cycle parking at public transport stations, park and ride, hire of bicycles/ e-bikes
Have existing journeys been replaced with shorter ones that serve the same purpose, and if so which ones, where, and how?	• Number of journeys shifted to a closer destination
Have journeys been avoided by home working, home delivery, teleconferencing etc., and if so which ones, where and how?	• Number of journeys avoided
Has the number of single car occupant journeys on the inter-urban network changed, and if so, at what times, and by how much?	• Number of single occupant car journeys on interurban and other roads

Example research questions	Example indicators
Have journey times changed on the interurban network and/ or other roads, for car and other modes, and if so, in what way?	• Mean journey time on interurban and other roads for each relevant mode
Have journey speeds changed on the interurban network and/ or other roads, for car and other modes, and if so, in what way?	• Mean journey speed on interurban and other roads for each relevant mode
Has the quality of journeys changed on the interurban network and/ or other roads, and if so, in what way?	• Standard deviation of journey time on interurban and other roads for each relevant mode • Standard deviation of journey speed on interurban and other roads for each relevant mode • Number of vehicles parked at the roadside
Has fuel consumption changed on the interurban network and/ or other roads, and if so, in what way?	• Estimated fuel consumption (petrol, diesel, electricity, other) by mode on interurban and other roads derived from number of vehicles, types, distance, speed
Has the number of casualties changed on the interurban network and/ or other roads, and if so, in what way?	• Number of fatalities by mode • Number of serious injuries by mode
Has the amount of active travel changed, and if so in what way?	• Number of trips by walk/ cycle/ access to public/ shared transport • Distance travelled on walk/ cycle/ access to public/ shared transport trips • Proportion of trips by walk/ cycle/ access to public/ shared transport • Proportion of distance by walk/ cycle/ access to public/ shared transport
Have vehicle emissions changed on the interurban network and/ or other roads, and if so, in what way?	• Estimated emissions of CO_2 derived from vehicle numbers, types, distance, speed
What is the monetary value of changes in safety, economy, health and the environment?	• Monetised valuations of changes in safety, journey time, levels of active travel, emissions
What were the overall costs of setting up and operating the scheme and over what timescale are these incurred?	• Investment costs for different organisations • Operational costs for different organisations • Maintenance costs for different organisations

Example research questions	Example indicators
What were the additional costs for service providers and operators of setting up and operating the scheme over what timescale?	• Difference between investment costs and usual costs for different organisations • Difference between operational costs and usual costs for different organisations • Difference between maintenance costs and usual costs for different organisations
What were the additional revenues accrued by service providers over what timescale?	• Revenues for transport operators/ service providers (may be derived from data on trips/ subscriptions etc.)
Has the National Road Authority made cost savings, and if so how and over what timescale?	• Costs incurred by the NRA to operate and maintain the network

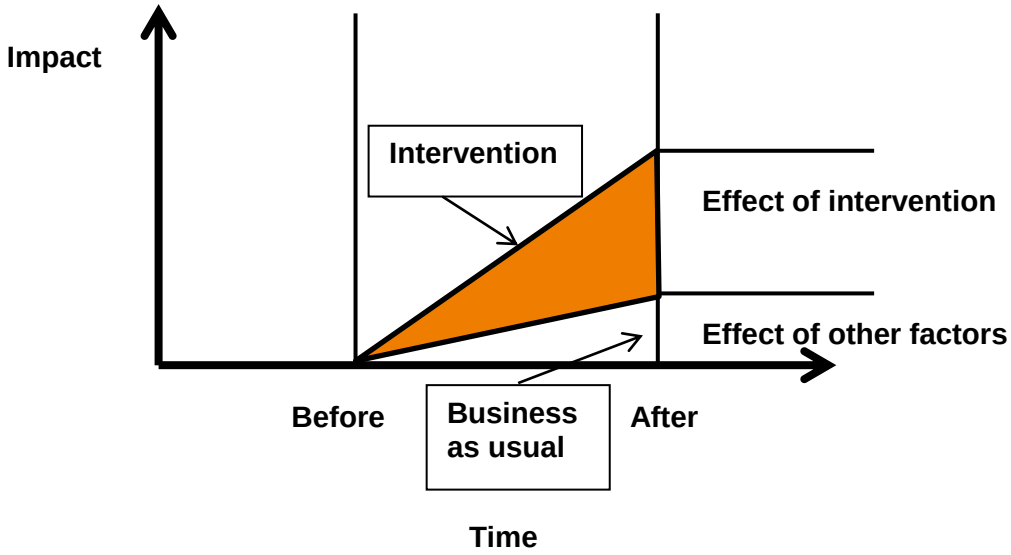
11.2 Define the reference case

The reference case, or baseline, defines the existing situation without the intervention; this is sometimes known as the 'do nothing' scenario, in which no other investment in the transport network takes place. Comparison of the situation as a result of the intervention with this reference case is what determines the impact of the intervention.

Given that the impacts of new technologies will often take some time to work through into behavioural changes affecting mode use, it is important to consider whether the reference case or baseline is likely to change during this time. If so, then the reference case is actually an estimate of the likely situation if existing policies continue without the intervention or other planned transport investments take place, rather than a reflection of the 'before' situation; this is often described as the 'business as usual' scenario or the 'do minimum' scenario. For example road collisions are tending to reduce across Europe as a result of sustained policies to reduce casualties; an evaluation of the safety impacts of new technology would take into account these forecast changes in casualties; a downturn in the economy could also have an influence on the demand for travel and levels of car use.

Figure 11.3 illustrates the baseline, business as usual and after scenarios.

Figure 11.3: Before (baseline), business as usual and after scenarios



When assessing the impacts of new technologies on mode use, the reference case that is relevant to most of the evaluation objectives is likely to be the 'before' or 'business as usual' situation in the area where the intervention is to be implemented. However in the case of an assessment of the performance of the technologies, the reference case may be defined in industry technical standards.

Another factor to consider in defining the reference case for interventions involving new technologies is that if the intervention generates data that can be used in the evaluation, it will be necessary to obtain equivalent data from other sources for the reference case, rather than relying on the intervention itself to provide all of the required data.

11.3 Define measurement methods

Investigations to assess the impact of new technologies on mode use will largely be based on measurements made during real life conditions. The measurement methods will vary depending on the intervention that is being evaluated; they may include:

- Questionnaire surveys of users and other stakeholders (such as household or 'destination based' surveys, roadside surveys, surveys on public transport vehicles or at public transport stops)

- Counts (automatic or manual) of levels of use at key points in the area influenced by the intervention, using continuous or periodic monitoring
- Collection of data from the technology or service delivered in the intervention
- Collection of operational and usage data from services
- Extraction of data from secondary and 'open data' sources such as mapping, population, census, local plans and routine monitoring data on traffic and travel; this can reduce the need to collect new data and reduce costs.

The limitations of these various sources of data will need to be taken into account when defining the measurement methods to be used. For example when considering the use of secondary and 'open' data, the timeliness, sample sizes, data quality, ownership and access rights may limit the suitability of the data for the evaluation. Collaboration with owners of such data may overcome these issues and by obtaining anonymised raw data through such collaboration, some of the limitations associated with using the processed data may be overcome.

A specific example of data quality considerations may arise when assessing impacts on cycling; it is important to bear in mind the limitations of using automatic traffic counting technologies to accurately detect the number of cyclists, while the location of automatic traffic counts for vehicles should be selected to avoid sites where queues are expected to form.

When planning real life data collection, it is important to consider legal, ethical issues and safety issues and other risks associated with collecting data from individuals and organisations, and to plan for how to deal with them. Legal and ethical issues will also need to be considered when extracting data from secondary sources.

One factor to be considered is the availability of data from commercial sources. For example transport operators are often unwilling to release data on levels of use and revenues for evaluation purposes so it may be necessary to use proxy measures, observations or surveys.

There are other reasons why measurements in real life conditions are not feasible, due to the scale or nature of the measurements involved. In such cases simulation or modelling may be used if there is sufficient data available on previous patterns of traffic and travel and other secondary data.

For example it may be necessary to use a modelling approach to assess the impacts of new technologies on travel behaviour across the network if it is not feasible to collect sufficient original data on traffic and travel behaviour for this purpose.

Another case when modelling is used is for measuring changes in vehicle emissions as a result of changes in the composition or speed of traffic; an evaluation will often estimate this change on the basis of known relations between vehicle types, flows and speed.

11.4 Define measurement conditions

The conditions under which data collection is carried out should be controlled and as stable as possible. It is well known that weather conditions, time of day, day of week and season of the year all influence patterns of travel, mode use and traffic levels, while traffic volumes and flows may influence the performance of some interventions, such as traffic management. Thus the difference between the reference case and the intervention may vary under different measurement conditions. A 'neutral' or representative data collection period should be defined, avoiding

public holidays, school holidays and the period immediately before and after these, and also large scale local events; in cases where education institutions are an important feature it is advisable to avoid exam periods as well as holiday periods.

Levels of cycling and to some extent walking can vary considerably with weather conditions, between winter and summer and between holiday and periods and normal working weeks. It is therefore recommended that where new technologies are designed to influence walking or cycling, data collection on use of these modes takes the form of continuous monitoring from automatic count sites at key points around the area expected to be affected by the intervention. If such count sites are not already established, it is advantageous to set them up as early as possible in the project – ideally at least a year before the intervention is implemented and preferably earlier. This will reduce uncertainty in interpreting the data on impacts on levels of walking and cycling after the intervention has been implemented. If it is not possible to set up automatic count sites to monitor walking and cycling, regular manual counts can be used but unless these are carried out frequently and at consistent times over an extended period, it will not be possible to ascertain with any certainty whether any changes can be attributed to the intervention rather than seasonal or weather-related factors. Established best practice for such manual counts is that they cover a 12 hour period on each day.

If a simulation model is being used, it may represent some situations more accurately than others, leading to differences in the apparent impacts of the intervention that reflect differences within the simulation model rather than differences in impacts.

It is also important to bear in mind that some of the indicators measured may have a strong correlation with parameters which describe the measurement conditions. For example travel time on a road network is closely associated with the traffic level. The measurements made will need to be carried out in a way that allows for this 'confounding factor' by taking account of variations in traffic levels in designing the data collection and analysis. This will have implications for the resources required for the evaluation.

There may also be confounding factors that need to be taken into account if a change in the transport network or services available in a neighbouring area has knock-on effects in the area where the intervention is planned.

11.5 Define statistical considerations and sampling

Where possible, a statistical approach should be used to define the number of measurements required to determine the impacts of the intervention in order to ensure that the appropriate quantity and quality of data is obtained in order to be able to attribute the impacts accurately. Depending on the nature of each indicator, the number of measurements required may be defined by one or more of the following: number of days, duration of measurements each day, number of units (vehicles, people), number of sites. It is recommended that a statistician is involved in this process.

For each indicator, the level of change that is expected to be brought about by the intervention should be estimated, using expert judgement and any available evaluation results from similar interventions. This is known as the 'overall definition of success'.

If a statistical approach is possible, the level of statistical confidence associated with this level of change that is acceptable or required should then be defined. For example it might be expected that a 5% reduction in single car use on motorways in the study area could be achieved, and that it is desirable to collect enough data to be able to state that there is 95% confidence that this level of change has taken place as a result of the intervention.

However even if a statistical approach is not possible, it is still important to ensure that there is a link between evaluation objectives and the overall definition of success through the definition of success for the relevant indicators.

For individual indicators, having determined the definition of success and level of statistical confidence required, the sample size should be defined – i.e. the number of measurements that are to be taken to represent the 'population' of all possible measurements. The following considerations should be borne in mind:

- Larger samples are usually needed for questionnaire surveys than simple counts because disaggregation of responses into sub-groups during the analysis means that fewer responses of any one type are available
- Larger samples are also needed as:
    - The expected level of change becomes smaller
    - The variation between individual measurements becomes larger
    - The level of statistical accuracy becomes greater
    - The number of sites that are to be compared increases (to more than one)
- Results that are based on objective or 'hard' measurements such as automatic vehicle counts can be treated as being more 'credible' than those based on subjective or 'soft' measurements such as questionnaire surveys of reported travel patterns.
- When planning counts of traffic, cycling and walking, the sample size is defined both by the times of day and number of days in the week when data are collected and by the number of sites where count data are collected. Automatic traffic counts would usually be designed to cover periods of two weeks or more in order to take account of variations from day to day. Particularly for cycling and walking, the number of sites needs to be calculated carefully. For example, experience of interventions to encourage cycling shows that in a medium-sized town, at least 15 automatic count sites are needed before a reasonable picture of changes in cycling can be obtained.

There will however be some situations when it is not possible to define the number of measurements statistically because the 'population' from which the sample could be drawn is too small. For example if assessing the impact of an intervention on the operators providing a service, there may only be a few operators available for inclusion in the assessment.

11.6 Define the measurement plan

Some data will be available from operational, monitoring and routine statistical sources and 'open data' – for example traffic flows and casualty numbers; in these cases the measurement plan defines the scale, scope and timing of the measurements that will be used in the evaluation. Other data such as user acceptance, number of walk trips and use of travel interchanges will usually need to be collected specifically for the evaluation.

One factor to consider when evaluating the impact of new technologies is who owns the data, which affects who has access to it for evaluating the impacts. It may be that only one party has access to the data, and that they may need to anonymise and pre-process the data before it can be used by others in the evaluation team.

Another factor which will be important in some cases is that data may migrate from legacy systems to new systems introducing the new technology. This may affect the availability of the data, its quality and scope, which could in turn affect the extent to which it is possible to compare between the baseline and the situation after the intervention has been implemented.

The duration of the measurements will need to take account of the fact that changes in travel patterns take time. The impacts of new technologies may also take some years to become evident. People are reluctant to change to new modes of travel or new ways of arranging their travel because this involves additional effort to gather the information they need to make the switch and potentially adapt other elements of their life to make such changes possible. It may be necessary to collect data at different time intervals to capture short term and longer term impacts.

The timing of the measurements for indicators of impacts of the intervention and for the reference case should be defined carefully so that this does not introduce any bias in the assessment. For example traffic and travel patterns should be measured at similar times (of day, week and year). It is helpful to produce a chart setting out time schedule for the various phases of the evaluation.

The area in which the measurements are to take place will need to be defined carefully to ensure that the evaluation objectives are met and the anticipated outcomes and impacts are captured. For at least some interventions designed to reduce single car use on the inter-urban network, this area is likely to include the surrounding road network and routes feeding into the inter-urban network. The location(s) will need to be clearly described when specifying the measurement plan, with maps, geographic coordinates and photographs as appropriate.

Once the measurement plan has been defined, the Resource Plan will need to be reviewed and it may be necessary to adjust either or both plans, to ensure that the measurement plan is appropriate for the resources.

11.7 Identify integrity of measurement

To ensure integrity of measurement, three types of factor should be considered:

- Completeness of the coverage to ensure that the measurements cover all of the significant impacts (not just those which are easiest or cheapest to measure)
- The scope of the measurements includes all of the factors which might influence the impact of the intervention or the characteristics of the reference case
- Accidental or intended bias to measurements may occur through factors such as respondent fatigue, policy response bias (respondents or participants who wish to influence the results), and justification bias (respondents give answers – knowingly or unknowingly - which they think the interviewer will find more acceptable); another form of bias may arise if the investigation assumes that users have received the necessary information to make rational decisions and informed comments relating to the intervention, when this is not the case.

11.8 Summary table

Table 11.2 provides a framework for summarising the details of the assessment method as described in this module. A template is available on the STTRIDE web site.

In a complex evaluation, it may also be helpful at this summary stage to draw a flow chart to show how all the sources of information fit together to inform the assessment objectives and research questions.

Table 11.2: Summary of assessment methods for each type of assessment, evaluation objective and research question

Type of assessment	Evaluat-ion objective	Research question	Indicators	Reference case	Methods of measurement	Measurement conditions	Statistical aspects			Measurement plan	Integrity of measurement
							Sampling	Statistical confidence	Definition of success		
Performance											
User acceptance											
Impact											
Socio-economic											
Financial											

11.9 Contents of an evaluation plan

An Evaluation Plan

An evaluation plan is a living document, built up by the evaluation team and agreed by the stakeholders involved. It provides a single source of reference for the evaluation.

A common structure is recommended for evaluation plans for interventions using the STTRIDE framework.

Recommended outline of an evaluation plan

1. Description of the intervention
   - 1.1. Objectives
   - 1.2. Technologies
   - 1.3. Area/ sites
   - 1.4. Timing of implementation
2. User needs for results
3. Description and mapping of intervention logic
   - 3.1. Expected impacts
   - 3.2. Types of change in mode use
   - 3.3. Intervention logic map
4. Definition of evaluation objectives
5. Definition of research questions
6. Scale and nature of expected outcomes and impacts
7. Selection of impacts to be evaluated
8. Assessment methods
   - 8.1. Definition of indicators
     - 8.1.1. Common core indicators
     - 8.1.2. Local indicators
   - 8.2. Definition of reference case
   - 8.3. Definition of measurement methods
   - 8.4. Definition of measurement conditions
   - 8.5. Statistical aspects
   - 8.6. Measurement plan and time schedule
   - 8.7. Integrity
9. Resources for evaluation and roles and responsibilities
10. Use of common approach to reporting results

A template for an evaluation plan is available on the STTRIDE web site. The templates for the tables and diagrams in the plan are available using the links in Figure 2.2 of STTRIDE D5.0 'Introduction to the Evaluation Process Guidelines' on the STTRIDE web site.