- Furthermore, the conducting of an evaluation in itself can have a positive benefit for those participating in it. This is the case, for example, if the evaluation contributes to a change in the way the actors think about the evaluand itself, or if it stimulates changes in the programme organization or leads to further qualifications for those involved. This is referred to as the process use of an evaluation.

NOTES

1. This situation can be further complicated by the fact that a stakeholder may belong to more than one group. Self-evaluations are, for example, an extreme case, in which the members of the evaluation team at the same time also play functional roles in the programme to be evaluated, which often forces those involved into conflicting role which are very hard to resolve (see also the remarks in section 6.2).
2. The Standards for Evaluation of the German Evaluation Society (DeGEval) (2002), for example, provide a relevant guide here, covered in section 6.2.2.
3. The evaluation literature also refers in this context to the 'process use' of an evaluation (Patton 1997; see also section 5.4).
4. For further illustrations of a concrete participatory procedure see for example Patton (1997) or Weiss (1998a: 103–5).
5. In addition, or as an alternative of course, other sets of standards can be applied, such as the evaluation standards of the Joint Committee on Standards for Educational Evaluation (JC 2000) or the Guiding Principles of the American Evaluation Association (AEA 1995).
6. See Stevahn and King (2005) on conflict management in the context of evaluations from the point of view of the evaluators.

7 Measuring: indicators – scales – indices – interpretations

Wolfgang Meyer

Let us start right at the beginning: two cavemen run into each other and one asks the other how far it is to his cave. What do you think his answer was? Well, it certainly was not '200 metres', as it might be today in most countries in the world, because the metre as a unit of measurement has only existed for a little over 200 years. And he wouldn't have said 660 feet or 190 yards, or even 1000 dactylos[1] or 250 dhira,[2] in spite of the fact these units of length measurement came into use much earlier. Nevertheless, we may suppose that even the cave-dwellers – presumably even before they had any command of language – communicated with one another about distances. In very early times, for example, this would have been a matter of survival when they went hunting together.

Whatever that unit of measurement looked like, it is sure to have had a few things in common with the 'modern' units mentioned above:

- It will have been an *indicator* for the estimation of distances, which will have been based on easily available *standards for comparison* (for example, parts of the human body such as feet or hands, or objects found in nature such as willow rods or animal bones). The basics of comparing with the aid of indicators will be looked at in section 7.1.
- The indicator will presumably have been based on a common *scale* known to all those involved, which made it possible for them to *place* the distance they themselves had to cover in a predetermined *category*. It may even have been the case that rudimentary *scaling procedures* were used in the development of such scales. The latter will be covered in section 7.2.
- It is, on the other hand, rather unlikely that our ancestors used a complex *index* which comprised different *dimensions* as a basis for their strategic hunting decisions. Now we can no longer imagine our world without such indices, which is why section 7.3 is dedicated to them.
- It was not until the development of common indicators, standards for comparison and scales, applied by all those involved as *criteria*,

that a standardized or generally agreed *assessment* of distances became possible. The course such *interpretation processes* take is the subject of section 7.4.

Even if the *measurements* of our ancestors, based on those scales of theirs, defined as they were with insufficient exactitude, were of only modest *measurement quality* – that is, of low *validity* and *reliability* – they were obviously good enough to ensure the survival of the human race. It may even be assumed that the human race would by now have become extinct if, instead of following the *requirements of the situation* and its own modest *social, technical and economic possibilities*, it had committed itself to the development of perfect measurement techniques. This is a pointer towards the special qualities of *measurements in the context of evaluations*, which are committed rather to the pragmatic aim of their usefulness for outstanding decisions than to some lofty scientific ideal.

Having said that, some of the *discussion about indicators* nowadays certainly does leave room for doubt as to whether or not the human race really has continued to develop in a positive way since those primeval times. Communicating with others about indicators and the measurement of them is likely to lead to conflict, and it is rarely a purely technical issue. It is primarily attributable to divergent *objectives* and different *emphases* on the part of the groups of people involved. There is obviously also a *political side* to measuring, which is based on an assessment of the amount of *new insight* that can be gained through measurement and its *practical usability*. Measurement is a *social process*, which begins with talking about the features to be measured, continues with discussions about their operationalization to indicators and certain scale forms, and finally leads into debates about the interpretation of the results and the measurement quality attained. This applies especially to evaluations, in which assessments of circumstances are often (mis)understood as assessments of the actors' own positions.

In an evaluation it is the aim of each and every measurement *to provide the information required for outstanding decisions on measures, in adequate quality, at exactly the right moment*. It is not the absolute quality of the measurement that predominates, but its relative usefulness for the maximization of the effects being aimed at. Negative peripheral phenomena, like endless discussions about indicators, should be avoided for reasons of efficiency. The chapter which follows is intended to make a contribution to said avoidance.

7.1 INDICATORS

The term 'indicator' is among those most often used misleadingly in the conducting of evaluations. We only have to glance into the *Duden Dictionary of Foreign Words* (Duden Redaktion 1997) to find four different applications, one of which describes an indicator in very general terms as a *circumstance or feature which serves as a [conclusive] pointer to or evidence of something else*. Only the *indicating (that is, pointing) function* of indicators is emphasized in this definition. The other, more lavishly formulated explanations refer to applications which are technical, chemical and to do with librarianship, respectively, while neither the social scientific nor the economics usage is mentioned – in spite of the fact that the gross national product, for example, has become known far beyond the narrow circles of science.

Definitions of the term 'indicator' are of course also to be found in popular specialist social science lexicons and dictionaries. In the *Dictionary of Sociology* (Hartmann 2002: 223), for example, an indicator is understood as 'a factor which can be empirically directly ascertained (e.g. by means of an observation or a survey) and which provides information about something which is, itself, not able to be directly ascertained'.[3] Here, in contrast to the general Duden definition, it is the *empirical aspect* and the *non-measurability of the circumstance to be depicted by the indicator* which are emphasized.

Authors who occupy themselves with indicators in actual project work often relate them to the aims of the project only. The European Union (EU), for example, in its programme evaluations, describes an indicator as 'a characteristic or attribute which can be measured to assess a programme in terms of outputs or impacts' (Nagarajan and Vanheukelen 1997: 16). In development cooperation, indicators should 'offer a concrete description of that which is meant by the objectives of the project' (Werner 2000: 7) and are 'one of a variety of mechanisms that can answer the question of how much (or whether) progress is being made towards a certain objective' (USAID 1998: 16). Hence indicators are parameters which are to be recorded empirically (quantitatively or qualitatively), and which are intended to make it possible to compare targets and achievements with regard to the objectives of projects or programmes.

What is awkward about the definition of an indicator being restricted in this way to the objectives of projects or programmes is the fact that *unintended consequences of action or side effects* may be overlooked as a result. We are reminded again and again of just what catastrophic consequences failure to take such effects into account can have by cases from the pharmaceutical industry (such as the notorious case of the sedative

Thalidomide, made by the Grünenthal company). Restricting the definition of an indicator to the targets of projects or programmes would tend to encourage such misjudgements. In general, the most important feature of an indicator is not its *orientation towards certain objects to be assessed such as the objectives of a project or programme*, but its *function as an indicator of a theoretical construct which is not directly measurable*.

If indicators are to provide more than the *description of a circumstance*, they must be placed in relation to *comparative values*. This is of particular importance in evaluations, since these involve the assessment of objects, processes, statuses and so on, and an assessment as a general rule always necessitates a comparison. Comparative values can be *normatively based* (for example, by the stipulation of target values to be aimed at), *theoretically derived* (for example, by effect hypotheses about critical threshold values) or *empirically produced* (for example, by repeated measurements).

> Indicators are parameters intended to provide information about a specified circumstance which is not measurable or can only be measured with great difficulty. In an evaluation, as the basis of assessments, values empirically measured by an indicator are placed in relation to comparative values.

Thus the most important task of an indicator is *to compare*. 'Measure' always means 'compare': a person who, as at the beginning of this chapter, states the distance he has to cover in metres, is not really doing anything other than compare it with a mark on a bar made of platinum and iridium in the Paris National Archive.[4] Accordingly, saying that a thing 'cannot be compared' is tantamount to saying that it 'cannot be measured'. In discussions on the development of indicators this argumentation is especially popular, especially with people who want to avoid a particular measurement and use the funds required for the measurement for 'more important things'. Accordingly it makes very good sense to be aware of the actual possibilities and limitations of measurement – and thus of comparison.

An example from the language of everyday life fits in well here: someone who wants to emphasize the incomparability of two things might say 'You can't compare apples with pears'.[5] Next time you hear that sentence, you can quite happily place a wager with the person who says it. The section that follows shows how this can be done (and is done, every single day) by all of us.

When this experiment is actually carried out, it is not hard to see that the *measurement result* will be extremely good – independent of the people

HOW APPLES ARE COMPARED WITH PEARS

- Step 1: send someone to market with instructions to purchase some of the objects identified by the sales experts present there as 'apples' and 'pears', and put them in a big basket together. (These apples and pears are the 'set of objects being observed'.)
- Step 2: get three bowls and label them 'apples', 'pears' and 'unclassifiable'. (The bowls represent categories and the order in which they are arranged represents a scaling process.)
- Step 3: instruct someone to allocate the objects in the basket as appropriate to the three bowls according to their appearance and applying only a single criterion – that of shape. First, all round objects are to be placed in the bowl labelled 'apples', while all objects which are not round remain in the basket. (The appearance is the indicator for differentiating between the two types of fruit. By the operationalization of an allocation rule, the two shapes 'round' and 'pear-shaped', between which the sorter is to differentiate and each of which is theoretically clearly associated with one of the two types of fruit, a comparison of the objects is made possible, and with it a decision as to which object is an apple and which a pear.)
- Step 4: when all the round objects have been placed in the bowl labelled 'apples', ask the person to put all the pear-shaped objects in the bowl labelled 'pears'. Finally, the objects now remaining in the basket (for example, the aubergine which found its way in by mistake) are placed in the bowl labelled 'unclassifiable'. (This sorting of objects is the process of allocation.)
- Step 5: now have a random number of experts check the allocation on the basis of their expert knowledge of apples and pears, and ask them to confirm to you the success of the experiment. (The accuracy of the allocation is known as the validity of the measurement and its verification as validation.) You can repeat the experiment as often as you like – the allocation will be a perfect success time and time again (the 'sameness' of the allocation when the measurement is repeated being referred to as its reliability).

involved and even if someone is really found who has never seen an apple or a pear before. On account of the *dissimilarity* of the two objects, the indicator '*appearance*' is eminently suitable for the *comparison* of apples with pears, for the very reason that the possibilities for allocating them to the categories formed thanks to the *basic* dissimilarity in the criterion 'shape'. *In other words, comparison is particularly easy when the objects to be compared are quite different!*

Conversely, allocating the apples to different varieties (for example, with bowls labelled 'Golden Delicious', 'Boskop' and 'Granny Smith') applying the criterion of shape is hardly likely to be a success. Since the shape of all apples is 'round' and this aspect in particular is one in which the varieties do not differ, allocation under these conditions is bound to fail. When the objects to be compared are very similar, comparison is difficult and the measurement duly has little *validity* (that is, allocation errors occur), and little *reliability*. That is, when the allocation is repeated there will be errors which differ from those made in the first measurement. In other words, it is much more difficult to compare apples with apples than apples with pears!

> The task of indicators is to allocate objects to predetermined categories. The greater the difference between the objects in terms of this allocation criterion, the more successful the allocation will be.

It may already have become clear that this test only functions under one important *peripheral condition*: that the objects being observed only include round apples, pear-shaped pears and objects which are neither round nor pear-shaped. As soon as there are other round objects (for example, oranges) or pear-shaped objects (for example, avocados) in the basket, they will be incorrectly categorized as apples and pears when the allocation criterion (of shape) is applied. Using the indicator 'appearance', of course, it is not a comparison between 'apples and pears' that is being undertaken but of *different shapes with each other*, and the assumption is made that *all objects with a certain shape can be allocated perfectly to the two categories apples and pears*. If the objects being observed include others of the same shape, the indicator 'appearance' and its dimension 'shape' are no longer adequate to enable the sorter to put the objects in the desired categories without error.

In everyday life, however, it is child's play segregating these objects too from the apples and pears, by applying *other indicators and criteria* (for example, smell, colour, taste). Here, indicators are used which will ensure the best possible allocation on the basis of foreknowledge of the *differences*

between the objects. With a *set of indicators*, those objects are finally to be *selected* more or less consciously from a set of objects which correspond to the general concept of apples or pears (which can be as large as you like) by means of comparison. That is not all: allocating them into different varieties is, for experts at least, no great problem either, because there are defined similarities in all apples of the same variety, which distinguish them systematically from the defined similarities of apples of a different variety. Accordingly, by the application of further indicators such as colour, smell, taste, the apples can successfully be *sorted* into varieties.

In principle it is possible without any difficulty to determine new indicators and criteria for ever finer differentiations, so that at the end of a large number of comparisons one could come to the conclusion that each individual apple is *unique*. Indeed, it would not be possible to find two apples on Earth which had grown naturally and were absolutely alike in all detail, thus defying all attempts to tell them apart by means of comparisons using indicators. Hence the one extreme of all comparisons (and therefore also of measurement) is the specific uniqueness of an object, which *differentiates it from all other objects on account of a large number of comparable properties.*

Having said that, all the objects and life forms on Earth do have at least one thing in common – diverse though they may appear to be: they come from the planet Earth.[6] So it is also always possible, in spite of all their differences, to find *similarities* between any two objects in a determinable measurement dimension and using a selected indicator. Hence the other extreme of comparison (and therefore also of measurement) is the *reduction of the comparison to properties that all the objects observed possess.*

The *criteria for the allocation of objects to categories are determined arbitrarily* and can be re-determined in any situation. The sorted apples and pears can be put back in the basket at any time; as often as you like, new indicators and new criteria can then be thought up and the objects sorted accordingly. Depending on the number of selected categories and the sorting procedure, the result may be that at the end all the objects will be together in one bowl, or that a bowl will be required for each individual object. As a matter of basic principle, the spectrum of allocation criteria

> The allocation of objects to the categories is determined exclusively by the freely selectable allocation criterion. How many objects are allocated to a common category is decided by the definition of that criterion. The degree of differentiation is neither an immanent quality of indicators nor a quality feature of measurement instruments.

always ranges from the perfect segregation of all the individual objects to their all being allocated together into a single category; it depends on definition by the user and on nothing else.

The main task in the development of indicators is to find criteria which, when the comparison is made, will ensure the best possible allocation to the selected categories. This is what we expect from 'measuring' in *methodological* terms; in other words, the comparison should be able to be carried out as precisely as possible and it should be able to be decided upon in a clear, unambiguous way. (This is the background of the statement 'You can't compare apples with pears'. What is meant is that *different assessment criteria* are applied – for example, with regard to taste – and not that the objects are *different*!) Finding appropriate criteria and indicators is a matter of imagination, and since these are artefacts, that is human inventions, from a methodological point of view anything is possible: *nothing exists which is not measurable in some way using indicators!*

This only works, however, if we succeed in *theoretically* defining a *universally valid difference* which is then to be applied in practice in the comparison. This also applies to our example: the sorting of apples and pears using the indicator 'appearance' and the criterion of 'shape' only succeeds if the underlying assumption is correct that in general *all apples are round* and *all pears are pear-shaped*. Thus a *deterministic causality* is being assumed between the allocation criterion applied and the objects to be assessed. If this assumption is false, it will not matter how carefully the measurements are made (that is, how carefully the objects are compared applying the criteria used), there will still be errors in the allocation. This will obviously be a measurement error of a '*higher order*', as it will have arisen not in the comparison but in the formation of the criteria and definition of the indicators. If a thing is 'not measurable', it is not a technical or methodological problem but a question of the circumstance which cannot be depicted and its *operationalization*.

> Operationalization is the logical link between the non-measurable circumstance and the indicator to be measured by means of a rule of correspondence which provides information about the nature of this link. The allocation rules used in the measurement are derived from it.

Indicators are supposed to measure things which are thought to be immeasurable. Thus it is justifiably assumed that there is an essential and constant connection ('*rule of correspondence*') between the measured

indicator (the 'appearance of the object') and the immeasurable construct (the 'apple'). Thus the most important task in the design of an indicator is to *link* it *logically* with *a theoretical construct* ('all apples are round') and determine a derivable allocation rule ('an object observed as round will be sorted into the category "apples"').

The *number of categories to be formed* is also derived exclusively from theoretical considerations: if, for example, only the apples are to be taken out of the basket, the bowl for the pears will not be required – two categories ('apple' and 'non-apple') will suffice. The trader, on the other hand, will have to continue differentiating both the apples and the pears by varieties in order to be able to offer them for sale, duly sorted, on the market. Without theoretical assumptions for the purposes of measurement, the stipulation of categories remains arbitrary and ends in the well-known 'last-word' argument, that each and every object is unique!

So in the *process of operationalization* it must be established theoretically:

- what it really is that interests the researcher (the objective of the measurement, or 'what is to be operationalized and how?')
- what the relationship is between the measurable indicator and that objective (link between indicator and measurement objective, or 'why does the indicator used measure the construct?')
- how many different categories the objects are to be sorted into (precision of the measurement or 'how exactly must the indicator measure?')
- applying what criteria the decision is to be made on, as to whether the indicator developed is at all suitable for tackling the scheduled task (quality of the rule of correspondence, or 'when is the allocation between indicator and construct no longer adequate?').

THE PROCESS OF OPERATIONALIZATION

- Step 1: determination of the theoretical construct to be observed and of a potential indicator (*what is to be operationalized and how?*)
- Step 2: establishment of a rule of correspondence which is logically derived and not influenced by the measurement (*why does the indicator measure the construct?*)
- Step 3: determination of the required number of categories and the allocation rule (*how exactly does the indicator have to measure?*)

- Step 4: determination of falsifiers for said rule of correspondence (*when is this allocation between indicator and construct to be assessed as inadequate?*)

(See Corbetta 2003; De Vaus (2001); Laatz 1993: 30ff., Wottawa and Thierau 2003: 85ff.)

Altogether, indicators must satisfy four different requirements with regard to their quality, of which the above aspects of operationalization represent only one. The *quality of operationalization* decides from a *theoretical point of view* the quality of the indicators. As we have just said, statements need to be formulated about the connection between the circumstance measured by the indicator and the non-measurable construct which was actually being aimed at, and these statements must be unambiguously definable and able to be tested by means of suitable, identifiable test procedures. If this is not guaranteed, the indicator may be a good technical instrument, but it will not be suitable as an indicator for the circumstances in which we are interested in terms of their content.

From a *methodological point of view*, the *quality of the measurement* is the standard by which the quality of an indicator is to be measured. What is demanded of an indicator is that it should measure that which it was designed to measure (*validity*) and always achieve this in the same way when the measurements are repeated (*reliability*). An indicator, the content of which is meaningful, is hardly likely to be used if the measurement values it produces are extremely imprecise, or fluctuate so wildly that no decisions can be made on the basis of them.

For practical situations, the question also arises as to the extent to which a survey which meets the theoretical and methodological requirements for an indicator can be realized at all. The *practical perspective* thus refers us to the *quality of the implementation* as a yardstick for the assessment of the quality of the indicator. So it is that the limited resources of time, personnel and money available to us often prevent the deployment of the measuring instruments which are the best from a theoretical and methodological point of view, calling for a certain pragmatism in the implementation of those demands in the measuring process. It should be noted, however that the converse is also true: that which we consider to be just about still justifiable in practical terms is not necessarily still an option that makes sense from a theoretical and methodological perspective.

This begs the question as to the acceptability of an indicator, that is, the extent to which its measurement results are accepted as a basis for

decision-making by those involved, the extent to which the researchers are prepared to go about gathering the data in the correct way, and the extent to which the interpretation is recognized. This *political point of view* leads to an assessment of the quality of the indicator with regard to the *quality of recognition*. This 'psychological element of measuring' can even sometimes be the reason for the failure of theoretically and methodologically good, practically realizable indicators, for lack of acceptance on the part of those involved.

The quality of indicators is measured by the quality of operationalization (theoretical perspective), the quality of measurement (methodological perspective), the quality of feasibility (practical perspective) and the quality of recognition by those involved (political perspective).

Attention must be paid to these four requirements in the *process of development of indicators* and used in the assessment of the quality of various different suggestions for indicators. Researchers must start here by *clarifying the measurement objective*, that is, occupy themselves with the question of why a thing is to be measured and, indeed, what it actually is in the first place. That may sound a little disconcerting, but it is often the starting-point of endless '*indicator discussions*', which are in fact '*target discussions*'. If for example the success of a democratization project is to be measured with indicators, the problem relates not so much to the methodological shortcomings of the depiction of the construct 'democracy' as to the different attitudes of those involved regarding what is actually to be understood by the construct. Without an appropriate clarification of the contents and the various different dimensions of the construct, no suitable indicators can be developed to depict it.

In order for the best possible indicators to be found for the measurement of the circumstances in which the researchers are interested, it is advisable, having clarified the theoretical issues, to discuss several alternative suggestions for measurement instruments and to weigh up their advantages and disadvantages carefully.

Ideally, we can isolate *seven steps* in the *development of an indicator*. Apart from operationalization, these should include the development of suitable scales with sufficiently sophisticated categories and a test of the feasibility and the methodological qualities of the indicator. (See the box below; see also the similar descriptions in Hullinger 2002 and Shavelson et al. 1991.)

THE PROCESS OF DEVELOPING INDICATORS

- Step 1: design of the theoretical construct and development of rules of correspondence (*operationalization*)
- Step 2: derivation of observable circumstances and stipulation of allocation rules (*definition of scales and categories*)
- Step 3: design of realizable indicators which can measure those circumstances (*definition of indicators*)
- Step 4: development of instruments and determination of measurement procedure (*determination of measurement procedure*)
- Step 5: data collection and calculation of the indicators (pretest)
- Step 6: evaluation of measurement behaviour and difficulties in data collection (*assessment of the indicator's quality*)
- Step 7: decision as to whether to retain, further develop or reject each indicator (*selection of suitable indicators*).

As has already been made clear in the 'apples-and-pears' example, the formation of suitable *categories* is of vital importance in the comparison of objects. An optimum allocation of the objects observed to the respective categories is desirable. No more categories should be formed than are necessary in terms of content for the assessment process, so that the process of comparison can be carried out as economically as possible. So it is not necessary for a measurement to be carried out in the greatest possible detail. It is simply important for a *scale* to be formed which is *appropriate* in terms of the *requirements*, that is, a set of categories in a common dimension (like the bowls, for example, into which the fruit is to be sorted according to its shape). The next section provides more precise information about this process of scaling and the tasks connected with it.

SUMMARY

- Indicators depict circumstances which cannot be measured directly.
- The logical link between these circumstances and the indicator to be measured is known as operationalization.
- Indicators allocate objects to predetermined categories.

- A measurement is made by means of freely selectable allocation criteria according to which the objects are sorted into the selected categories.
- Measuring always means comparison.
- The quality of an indicator is determined by the quality of its operationalization, its ability to be allocated, its feasibility and its acceptability.
- In evaluations, values measured by indicators are placed in relation to comparative values and the evaluand assessed on this basis.

7.2 SCALES AND SCALING

Children are already confronted at school by the problem of assessing things using *scales*. It is their own marks which suggest the existence of an objective yardstick.

School marks, however, like all other scales, say nothing about the *quality of their measurement* or the *how they were arrived at*. This even applies to the results of measurement instruments which have been adjusted, calibrated and standardized better than the school marks. Here, too, is an example from our schooldays: the decision as to whether or not the pupils are going to be given the rest of a hot summer's day off on account of the excessive heat[7] is based on the diligently maintained school thermometer and whether it happens to be in the sun or the shade . . . and on the pupils' chances of giving the mercury a little nudge with a cigarette lighter.

Verbal, subjective assessments seem vague to most people and therefore not particularly worthy of being taken into account. Yet in the example above, the pupils' *subjective reaction to the heat* is certainly considerably more important in terms of the success of the learning process than the *objectively measurable temperature*. The thermometer may well be a more accurate measurement instrument than the pupils' own diagnosis, but it is doubtful whether or not it really *always provides better results in terms of protecting them* from being overtaxed and suffering negative effects to their health as a result of the heat.

Generally, we can differentiate between three different kinds of scale, each of which has its own specific properties.

- *Nominal scales*: the objects observed are allocated to categories on the scale exclusively according to the criterion of the correspondence