How to validate environmental indicators

C. Bockstaller\textsuperscript{a,}\textit{*}, P. Girardin\textsuperscript{b}

\textsuperscript{a}Association pour la Relance Agronomique en Alsace (ARAA),
c/o INRA, BP 507, 68021 Colmar Cedex, France
\textsuperscript{b}INRA, Equipe Agriculture et Environnement, BP 507, 68021 Colmar Cedex, France

Accepted 1 May 2002

Abstract

Different types of environmental indicators have been developed to meet the increasing need of assessment instruments of environmental impacts. As any tool developed by research, indicators must be elaborated according to a scientific approach. One of the important steps of this elaboration is the validation. The overall objective of this paper is to present a methodological framework to validate indicators. According to the definition of an indicator, three kinds of validation are presented: the “design validation” to evaluate if the indicators are scientifically founded; the “output validation” to assess the soundness of the indicator outputs, and, the “end use validation” to be sure the indicator is useful and used as a decision aid tool. The output validation is inspired from validation of simulation models, which is shortly reviewed. Because indicators differ from models in many cases, validation procedures commonly used in modelling have to be adapted.

\copyright 2003 Elsevier Science Ltd. All rights reserved.

Keywords: Sustainability; Environment; Indicator; Validation; Nitrogen; Pesticide

1. Introduction

In the last decade, the development of indicators at the national, regional, local or field level, has become a commonly approach to meet the crucial need for assessment tools. Such tools are a prerequisite to the implementation of the concept of sustainability, and especially its environmental component (Hansen, 1996). The increasing use of indicators to face this challenge may be explained by the impossibility to carry out direct measurements in many monitoring programs, due to methodological problems, or practical reasons of cost and time, and to the lack of
feasibility of many simulation models which are developed as an alternative to direct measurements. The following definitions of an indicator refer to those considerations: An “Indicator is a variable which supplies information on other variables which are difficult to access (...) and can be used as benchmarker to take a decision” (Gras et al., 1989). “alternative measures (...) enable us to gain an understanding of a complex system (...) so that effective management decisions can be taken that lead towards initial objectives” (Mitchell et al., 1995). Both definitions imply: (1) an informative function, i.e. to supply simplified information about a complex system (e.g. an agrosystem), or an unmeasurable criteria (e.g. biodiversity, sustainability). It should be noticed that those definitions do not mention an accurate simulation of a system or a prediction of a variable as it is for models; (2) a decision aid function to help to achieve the initial objectives, e.g. the sustainability of a farming system.

Indicators may result from a set of measurements, from calculated indices, or they may be based on expert systems. At least two types of indicators may be distinguished (Girardin et al., 1999): simple indicators resulting from the measurement or the estimation (e.g. by a model) of an indicative variable and composite indicators that are obtained by aggregation of several variables or simple indicators. Mineral nitrogen content in soil before winter (Schröder et al., 1996; Machet et al., 1997) or bioindicators as the number of earthworms per surface unit (Smith et al., 2000) are example of the former type whereas the index of sustainable economic welfare (ISEW) developed by Stockhammer et al. (1997) belongs to the latter group. The nature as well as the purpose of an indicator, as mentioned before, differ from that of a simulation model. Many indicators are not aimed at being used to predict an actual impact but to supply information about a risk or a potential effect (Halberg, 1999). Indicators can also inform policy makers about the progress that is being made towards achieving a policy objective (Crabtree and Brouwer, 1999; Vos et al., 2000). Thus, the role of the indicator may be to signal positive movements as well as negative ones. Someone are also aimed at “raising the alarm”, meaning they should give information on negative impact before it actually occurs (Reus et al., 1999). However, in some cases indicators are derived from simulation models in order to make the output easier to understand and to “relay a complex message in a simplified manner” (Fisher, 1998).

In any case, the methodology underlying the elaboration and development of indicators should fit scientific standards, which implies a procedure of validation (Girardin et al., 1999). Such a procedure is currently applied in the development of simulation models and is a decisive step in the evaluation of model quality. However, few developers of indicators deal with this issue and propose a detailed procedure. Some authors mention the necessity for indicators to be scientifically valid (Mitchell et al., 1995; Crabtree and Brouwer, 1999; Smith et al., 2000; Vos et al., 2000) but those authors do not propose a procedure for validation. Neither van der Werf and Petit (2002) do find detailed information on this issue in a review describing and evaluating twelve assessment methods of environmental impacts. A qualitative approach based on the assessment of the potential to obtain undesirable outcomes for six methods can be found in Hertwich et al. (1997). Some authors like Sharpley (1995) go further. He gives an example of validation of a Phosphorus index, based on the comparison of the indicator outputs with field measurements, as it
is generally carried out in modelling whereas Appel et al. (1994) failed to validate in the same way a indicator based on the calculation of an nitrogen budget at field level.

Nevertheless, since the nature and the purpose of an indicator differ from those of a predictive model, it is questionable whether the same procedure as in modelling can be implemented to test indicators. The objective of this article is to propose a methodological framework to validate indicators. We will use the experience gained by modellers in validation and will focus on environmental indicators.

2. A general framework for the validation of indicators

From the definition given in the Oxford dictionary and discussed by Addiscott et al. (1995), something is validated if: “it is well founded and it achieves the overall objectives or it produces the intended effects”. Harrison (1990) and Kirchner et al. (1996) also refer to this last part of the definition and define validity as the adequacy for a specific purpose. The first point of the given definition requires the assessment of the scientific quality of the construction or design of a given tool. We will call this step a “design validation” (Gilmour, 1973). Concerning the second point of the definition, it comes out from their definition that indicators have a double function to supply information and help to take management decision. These two aspects of an indicator refer respectively to the soundness of the indicator output and to the usefulness of an indicator for potential users, which implies respectively an “output validation” (Gilmour, 1973) and an “end-use validation” (Girardin et al., 1999).

Thus, an indicator will be validated if it is scientifically designed, if the information it supplies is relevant, if it is useful and used by the end users. Three types of indicator validation correspond to these three conditions (Fig. 1).

3. Validation of the indicator design

In modelling, the design validation or conceptual validation (Mitchell and Sheehy, 1997) is carried by reviewers of scientific papers describing a model. However, this
first type of validation is often overshadowed by the validation of the model outputs or empirical validation (Mitchell and Sheehy, 1997), based on a comparison with a set of measured data, which is considered as the key step in model validation (Gilmour, 1973).

For indicators, the design validation may be very important when no other possibility of validation exists (see Section 6), as pointed out by Reus et al. (1999). It consists of submitting the design or construction of the indicator to a panel of experts. Van der Werf and Zimmer (1998) applied this procedure to their indicator of pesticide environmental impact, which leads them to improve their indicator before publishing it. Another example is given by Taylor et al. (1993). Expert consensus plays a major role in this kind of validation. The design validation may be also carried out, a priori, by expert judgements for the selection of the variables which should be measured as indicator. This is the case when the choice of indicators is made by a survey among a panel of experts, for instance by means of Delphi technique (Hess et al., 1999; Smith et al., 2000).

4. Validation of the indicator outputs

The purpose of this procedure focusing on the informative function of indicators is to evaluate their reliability. This is a very common approach in modelling. We will shortly review the validation procedures usually used for models, as this will help to define a specific approach for output validation of indicators.

4.1. What can be drawn from model validation?

First of all, it is important to know the limits of the objectivity and scientific value of a validation procedure. Based on philosophical and linguistic considerations, Konikow and Bredehoeft (1992) and Addiscott et al. (1995) show that validation never enables one to prove the veracity of a model but only to invalidate it. If the probability to refute a model is small, it can be considered as “validated”. Thus, validation is always linked to a certain element of subjectivity, because the model developer has to decide which level of probability is acceptable.

At least three types of procedure have been proposed by modellers for the implementation of validation (Mayer and Butler, 1993):

1. The visual procedure is commonly used by modellers and implemented by means of a plot, either comparing the time-trace of the model with the time series of the measured data, or comparing observed data with model predictions in relation to the line of perfect agreement. Kirchner et al. (1996) strongly criticise validation procedures based only on this kind of test. They demonstrate that an invalid model may satisfy the visual test and point out the lack of objectivity in the commonly used and vague statement “acceptable agreement with the data”. Actually, this visual step is only useful to put in evidence possible bias of a model.
2. The statistical procedure may give more objectivity to the test. Actually, the implementation of statistical tests raises some problems and is not recommended by several authors (Harrison, 1990; Mitchell, 1997). Harrison (1990) proposes to limit model validation to “descriptive (and not inferential) devices”. Several statistical variables assessing the deviance are proposed by Mayer and Butler (1993) and Wallach and Goffinet (1989). In any case, Mayer and Butler (1993) state the necessity to use several “validation measure” to “appreciate the whole picture”. Thus, Yang et al. (2000) analyse the correlations between some commonly used statistical variables in validation in order to avoid redundancy. An example of this approach based on several “validation measures” can be found in Mary et al. (1999) who used for the validation of their model several statistical variables. To face the lack of inferential method, Mitchell (1997), and, Mitchell and Sheehy (1997) developed an empirical test consisting of checking graphically whether 95% of deviations (prediction minus observation) lie within an acceptance envelope on the plot comparing deviations with observed data. The definition of this envelope depends on the purpose of the model and the precision of observed data (Fig. 2).

3. The third procedure is based on the judgement of experts. Such an approach as proposed by Mayer and Butler (1993) consists of selecting a panel of experts in the same field of interest and of submitting the output of the

![Fig. 2. A method of empirical validation of simulation models (after Mitchell, 1997 and, Mitchell and Sheehy, 1997) The test consists of checking graphically whether 95% of deviations (prediction minus observation) lie within an acceptance envelope on the plot comparing deviations with observed data. The acceptance envelope may have another shape, another range of variation (given by k). Both depend on the precision of measured data.](image-url)
models and real-world data to them. Some formalised tests exist like the Thuring-type tests for this kind of procedure. The comparison of the model predictions with data from the scientific literature may be also included in this type of approach (Chevalier-Gérard et al., 1994). Due to the subjectivity inherent to procedure based on expert judgement, Mayer and Butler (1993) authors consider this type of validation rather as a complement to more objective approaches mentioned previously. However, for specific types of models as expert systems for decision aid (Harrison, 1991) or multicriteria analysis models (Qureshi et al., 1999), such approaches based on expert judgements were also recommended and used for validation.

4.2. Application to indicators

A comparison of indicator output with measured data and the use of tests or descriptive variables as proposed in the previous paragraph for modelling would seem ideal. This is a first type of procedure for output validation: validation through comparison. However, two problems arise rapidly in such an attempt and concern the type of indicator outputs and the availability of measured data. In some cases, the only possibility is to get a global evaluation of the quality of the indicator value from expert judgements. This is can be called a global expert validation.

4.2.1. Validation through comparison

If the indicator results from the transformation of model outputs, its validity will logically depend on the validity of the model. This approach is applied by Pussemier (in Reus et al., 1999) for one part of his pesticide indicator. Some authors follow the approach currently used in modelling and restrict validation of their indicators to a visual procedure based on the comparison with measured data. This is the case of Brown et al. (1995) for the GUS index (indicator of pesticide leaching risk) or Sharpley (1995) for the Phosphorus index. The criticisms pointed out by Kirchner et al. (1996) to this procedure can also be made to those examples.

In many cases, indicators differ from a simulation model and may be based on an expert system, or result from mathematical equations or a system of qualitative ranks or scores. Furthermore, many indicators are not aimed at being used to predict an actual impact but to supply information about a potential impact, as mentioned in the introduction. Because of all those differences with simulation models, it is obvious that an linear relation between indicator output and field data cannot be expected. Taking into account those considerations, Girardin et al. (1999) propose a probability test which may be paralleled to the graphical test proposed by Mitchell (1997). As shown in Fig. 3, the test consists (1) of defining a “probability area” on a figure representing the indicator versus field measurements (2) testing whether this probability area includes at least 95% of the points, this acceptance level being recommended by Mitchell and Sheehy (1997). The definition of the probability area depends on the calculation method of the indicator and the precision of the measurements. The shape of a probability area given in Fig. 3a may be chosen when the
Fig. 3. A “probability test” to compare output of an indicator with observed data when no linear or non-linear relation is expected a priori (Bockstaller and Girardin, 2000). The test consists of checking whether 95% of the points are found in the probability area. (a) For a quantitative indicator; (b) for a qualitative ordinate indicator.
indicator assess a potential impact. The probability area can take another shape as shown in Fig. 3b for a qualitative procedure.

An example for a nitrogen indicator assessing nitrogen losses by volatilisation or leaching in spring and during winter (Bockstaller and Girardin, 2000) is given in Fig. 4. The risk is considered as non significant below 30 kg N/ha, and above this value, the probability area is delimited by a decreasing boundary line with a slope equal to 1/30, corresponding to the transformation of nitrogen losses into values of the indicator. The definition of the probability area means that high values of the nitrogen indicator can be only related to a low environmental risk, which, in this case, is a risk for groundwater quality. But a low value of the nitrogen indicator may be related to a high or low risk of winter leaching. This last possibility may be explained by nitrogen losses that have occurred before winter (through volatilisation or leaching in spring). The test carried out with 5 years measurements yields a successful rate of 97% of the points (74 points out of 76) in the probability area if the points close to the boundary line are included in the probability area and four points with a specific problem are discarded.

A second basic problem to implement a direct comparison with measured data is the lack of available observed data for different reasons: impossibility to measure

![probability_area.png](image)

Fig. 4. An example of a “probability test” for a nitrogen indicator (Bockstaller and Girardin, 2000). Measurements of soil mineral nitrogen were carried out for 5 years (1994–1998) before winter (after 15 November).
them (e.g. for sustainability or biodiversity indicators), problem of costs like in pesticides, etc. In this case, we propose to use other sources of data which do no result from measurements or observations. A solution may be to relate the indicator outputs with results obtained by simulation model that should be validated. Passemier (in Reus et al., 1999) shows for example, that there is a relation between the GUS index and one output of the PESLA model. If indicators cannot a priori be related to observed or measured environmental data, other procedures are possible. One is to compare the output of the indicator with those of other indicators having the same purpose but constructed in different way. It is assumed that if those indicators show the “same” results (e.g. the same ranking of risk for pesticides), their reliability may be reinforced. This type of work was carried out in the European Concerted Action CAPER (Reus et al., 1999). However, it can be argued that indicators may all be wrong for the same reason, although the probability is low when their construction is totally different, which was the case in CAPER. If indicators do not give the same answer, a analysis of the construction of each one is needed, to identify the reason of the discrepancy and which one may be “wrong”. In CAPER, one indicator was not linked to the other one, and it was clear that it comes from its simplified construction which was not scientifically founded. In case of lack of measured data for comparison with indicator outputs, another solution is found in Bockstaller and Girardin (1996) who compared the values for the effect of the preceding crop (one of the component of a crop sequence indicator) with values given by experts.

4.2.2. Global expert validation

If those methods are not possible, expert judgement is a last resort: outputs of an indicator may be submitted to a panel of experts who evaluate their relevance. For instance, this is, proposed by Lewis and Bardon (1998) although they do not give a formalised procedure to do it. As part of our ongoing work on agro-ecological indicators (Bockstaller et al., 1997), we developed a first version of an energy indicator assessing energy cost of cropping systems. We presented results of calculations to a group of farmers (having in this case the role of experts). They did not agree with the results on energy consumption for machines in the field according to their own experience of machine consumption. This put in evidence a weakness of the indicator for a given variable and lead to an improvement of the indicator (Pervanchon et al., 2002).

5. “End-use validation”

This part of validation deals with the usefulness of an indicator as benchmark for decision making. Several reasons may explain the lack of usefulness of an indicator: a target of great relevance for the user may be missed, some data needed to calculate them are not available or the outputs of the indicator are not understandable or legible. To better appreciate the quality of an indicator, Girardin et al. (1999) propose a “usefulness test” to get the end-users’ opinions. Through a survey, the users can point out the weaknesses of the indicator as diagnosis or decision aid tool. This
procedure should also help to ensure that end-users understand what is being indicated by the indicator or that the results are being interpreted correctly by end-users. This is particularly important for indicators of more complex issues like sustainability. An example of this type of validation is given by Douguet et al. (1999) who performed an end-use validation to evaluate the usefulness of the set of agro-ecological indicators developed by Bockstaller et al. (1997). Charollais and Schaub (1999) tested how easy an indicator assessing field erosion risk was to handle and the accuracy of the assessment made by the end-use. This last test showed some problems of reproducibility of the results between the participants to the test.

6. Discussion

Although several authors point out the difficulty of indicator validation in the sense of model validation (Reus et al., 1999; Lewis and Bardon, 1998) or do not mention this procedure at all, this does not mean that nothing can be done. This step of output validation is crucial to assure the credibility of an indicator approach among scientist involved in modelling and to the future end-users, governments or other decision-makers, NGO’s farmers, etc. of the tool. Nevertheless this last public are often less concerned by this point because they may trust in the outcomes of the research, or because they have no choice between several tools. With the increase of assessment systems, this type of procedure will take more importance beside other tests, for example, for the feasibility and the handiness of the tool, to help to the choice of the “best” method by a future user.

If many indicators cannot be validated in the same way as simulation models by comparison with measured data, validation based on expert judgement and expert consensus concerning the quality of the indicator design as well the quality of the indicator outputs is always possible. Such an approach should be considered as a minimum requirement for indicator validation. Thus, expert judgement plays a more important role than in modelling, which appears to be more subjective. However, this type of procedure should not be too much depreciated with regard to model validation which is generally considered as an objective procedure. Our short review of validation in modelling reveals that validation is not an absolute objective procedure and some validation procedures of model present a lack of rigour (Kirchner et al., 1996). Even more, we propose several possibilities to go further than this first level of expert judgement (Fig. 5). The availability of measured data on the one hand, and, the construction of the indicator on the other hand play a major role in deciding how to carry out the comparison between the indicator outputs and data. The different possibilities shown in Fig. 5 are complementary and do not exclude each other. Actually, a procedure of validation is never completely accomplished (Whisler et al., 1986). Thus, several alternatives shown in Fig. 5 should be used together in order to broaden the domain of validity of an indicator. In fact, measured data may be available for some situations (e.g. a given climate, soil, cropping systems, etc.) whereas for other situations, only expert data are available.
If no observed data are available, we propose to compare indicator outputs, with models outputs, with value of other indicators or with a data set based on expert judgements. In this case, the strength of validation appears lower than in validation through comparison with measured data. However, It should be noticed that this can also occur in model validation. In this case, Kirchner et al. (1996) recommend to use measurements obtained by laboratory experiments in simplified conditions. For specific models which differ from simulation models, even submission of models outputs to experts were recommended (Harrison, 1991; Qureshi et al., 1999). Actually, as it was pointed out by Addiscott et al. (1995), validation does not allow one to verify that a model or an indicator is true, but to assess its probability of refutation and to reject it if this probability is high. The proposed procedure based on comparison allows to reveal discrepancies between indicator outputs and corresponding data. If the data do not result from measurements, an assumption is made that if the indicator is not linked to this kind of data used as reference, it would not be linked to real-world data (Kirchner et al., 1996).

Concerning the construction of indicators, qualitative indicators or indicators aimed to put in evidence a potential impact seem to be more problematic to “validate” in the sense of modelling by a comparison procedure. For most of those indicators, climatic variable are not taken into account for reasons of simplification or to simplify the interpretation (Halberg, 1999). Thus, it is obvious that straightforward relation with measured or other data will not be observed. An example can be found in Appel et al. (1994) who fail to validate their nitrogen budget by looking for a linear relation with measured data. However, it is possible to go further in the

![Decision Tree](image)

Fig. 5. A decision tree summarising the possibilities of output validation for indicators (see text).
validation and we proposed a procedure of comparison which is based on a formalised graphic test, a probability test (Girardin et al., 1999) which can be compared with a test proposed in modelling (Mitchell, 1997). The definition of “probability area” or acceptance area is a key issue in this test and this should be done rigorously and a priori. Such a procedure may also be applied to qualitative indicators as well as observed qualitative data.

Another problem due to the construction of the indicator may arise when the indicator is a composite one taking into account instance different environmental component. For instance the pesticide indicator developed by van der Werf and Zimmer (1998) gives a value of global risk which results from the aggregation of three risks, for “air”, “surface water” and “groundwater”. If the data are available for each criteria, a $n$-dimension validation procedure should theoretically be carried out, which requests a high level of mathematical skill (Girardin et al., 1999). Another solution is at least to validate separately each one-dimensional module of the composite indicator. Another hindrance exists when the indicator is built with variables of different types. For example, environmental data and toxicological data. Examples can be given for some pesticides indicators, based on a risk ratio, i.e. the ratio between an estimated concentration of a pesticide in a given environmental compartment and its toxicity for relevant organisms (Reus et al., 1999). The direct comparison with measured environmental data becomes impossible. In this case it is only possible to validate the environmental part of the indicator, i.e. the estimated concentration in environment. In the example of the groundwater module of the pesticide indicator developed by van der Werf and Zimmer (1998), the validation can be restricted to the key environmental component of the module: the GUS index for which we have mentioned two validation procedures (Pussemier in Reus et al., 1999; Brown et al., 1995).

The last type of validation, the “end-use validation” may appear more original than the design and output validations, both being currently used in modelling. A similar procedure to the proposed “end-use validation” is not often mentioned in modelling which is regretted by Cox (1996) and is strongly recommended by Harrison (1991) for decision-aid models (Lundkvist, 1997). From the results of this work, Cox (1996) points out the necessity to include in a validation procedure a comparison of the model with other methods like currently used technologies or management systems which are not based on the given model. For simulation models, Kirchner et al. (1996) also propose to test them against other decision-making methods such as expert opinions in order to know the additional gain which can be expected by using a model rather than another method. An example is given in Meynard (1998). If the necessity of this last stage of validation is clearly stated, its implementation is not formalised. Thus, the surveys done by Cox (1996) or Douguet et al. (1999) are rather descriptive and are not based on a formalised test. In any case, such a procedure allows a tie between the developers of the indicators and the users. As mentioned by Fisher (1998), it is crucial to establish such interactive ties with the groups of users involved in conducting environmental risk assessment, in order to provide a user-feedback on how to improve the indicator.
7. Conclusions

Referring to models, validation of indicators is often requested. However, many indicator developers do not deal with this issue, probably because they think that the long term acceptance of their indicators by users is sufficient to assure their credibility. In this paper we propose for the validation of indicator a three-stage approach, based on a “design validation”, an “output validation” which can be achieved by comparison with measurements or other type of data, and the “end use validation”. This latter stage is particularly relevant for indicators which are built to be used as decision aid tools, to avoid that they remain confined to the office of the developer. Actually, the use of any tools should be thought about from its design. In any case, this procedure needs to be better formalised in a real test and not simply limited to a descriptive work.

Acknowledgements

This work is sponsored by the EU (Interreg 2 programme), the Land Baden-Wurttemberg (Germany) and the Alsace Region (France) as part of the ITADA programme. We are grateful to A. Blatz for his work of collecting data for the probability test and the farmers’ survey. Thanks also to J. M. Meynard, N. Munier-Jolin and H. van der Werf for their helpful comments on a draft of this paper and to the colleagues from INRA involved in modelling for their stimulating questions on the validation of indicators.

References


