

Evaluation of Educational Projects: Basic Conceptual Thoughts and Reasons for an Approach to Continuous Collections of Inventories

Why Evaluation?

When this term entered the vocabulary of Educational Science in Germany in the sixties/seventies, this meant the claim to do an empirically controlled practical test with the different approaches of the reform of education which happened on the level of content and on an organisational-institutional level. At that time the didactic development research mainly concentrated on the elaboration of curricula and evaluation dealt with the examination of the given learning objectives and possible side effects: they wanted, for example, to find out if pupils, doing a course in set theory, were learning "better" because of this and if they possibly enjoyed it more than other pupils in different curricular conditions. Already then, the question for criteria what could be determined as "better" was so multifarious that a consensus as to which instruments to collect which data could hardly be reached.

Central approaches and methods of didactic evaluation also developed in the context of these questions. One decisive aspect of evaluation measures was for example if the results of an evaluation could still be included in the intrinsic development process (e. g. of a curriculum) (=formative evaluation) or if they are "only" available afterwards (=summative evaluation). Different basic patterns for evaluation measures developed, which mainly considered the questions that could be deducted from the respective reform and innovation approaches. In order to understand and to be able to judge the background of these developments, it seems advisable to have a short look at the prehistory of this modern evaluation research.

First of all to use an evaluation in the context of education is nothing new. We are told that about 4000 years ago people in China, who fulfilled a public function, were subject to an inspection and appraisal of their capabilities. And, of course, pedagogics of former centuries already knew ways and methods for the inspection of capabilities in Europe and elsewhere. But these were only methods, which regarded the receiver in an educational system or the vocational aspirant as the subject of evaluation measures. He had to prove himself and, as the case may be, had to bear the consequences of an unfavourable result. Thus DOHSE reports, that the roots of what we know as school reports in our schools were some kind of benefit report at the beginning of modern times, which reported to the sponsors of pupils about the school performance of those pupils who – as we would say today – received private or public grants

But in the course of the European Enlightenment a new thought came up, namely the appraisal of the system context, in the framework of which teaching and learning processes were organised and happened.

We know for example a sentence of Kant, that one should set up trial schools before introducing new methods/procedures in normal schools. The point was not necessarily the appraisal of the learners but rather the question of how to best deal with learners. Today these two basic functions of evaluation and performance measurement are identified as receiver and system evaluation.

Which Means and Methods?

At the end of the 19th century there were attempts in the USA and Europe to develop and test new ways of a performance measurement in education, which would particularly resist all experiential-scientific demands, and at first it was this aspect of a system evaluation that was the decisive factor for the procedures chosen. Thus Joseph Meyer Rice, who developed the first proficiency test in the USA at the end of the 19th century, wanted to find out, which bad effects lessons can have, in order to prove his conviction that the time at school is inefficient.

During the following decades approaches and practices regarding the performance measurement in education and formation developed in the USA but also in Europe, which we know as modern test procedures (by the way very early in Germany too regarding the selection of vocational aspirants).

This does not only mean the classical test but also more open ways of a performance measurement like essay-tasks and practical situations. Particularly in the USA, but also with some pedagogues and psychologists in Germany and other European countries, the vision of such performance measurement in education and formation, which is based on such test procedures, came up.

But what else did developed from these approaches in the USA until the middle of this century?

- Evaluation and measurement procedures were equated, i. e. they thought that by developing and using measurement procedures they were evaluating already, without considering the need for additional decision-making processes and other measures.
- Basic pattern of evaluation and measurement procedures was a scientifically-oriented paradigm, i. e. the particularities regarding the inquiry of social processes were not sufficiently considered.
- Evaluation and measurement procedures are geared to inter-individual differences, the predominant measurement procedures were, for example, more suitable to order the learners in ranks instead of identifying if a given learning target was achieved or not.

In 1949 Ralph Tyler set a new milestone with his book "Basic Principles of Curriculum and Instruction", in which he judges the question of the objective as the most fundamental problem of any evaluation research. But he only considered the question if a given objective was achieved, not the evaluation of the objectives themselves, which he regarded as predetermined.

In 1957 after the Sputnik-shock quite a promotion of innovation in education started in the USA and with different inducements and to a lesser degree this was also the case in the Federal Republic of Germany 5-10 years later. On the one hand this opened up a new dimension regarding the function of evaluation because the sponsors and initiators demanded an account about the use of their provided funds (e. g. the Ford Foundation or the Foundation of Volkswagen). The thought of "accountability" (how much education do we get for our money?) arouse in this context, but in contrast to the benefit reports of the 16th century they were targeted to measures in education (e. g. a new organisational form of schools or a new special curriculum) and not targeted to persons like pupils or students. On the other hand it became clear, that previous methodological instruments were not sufficient to meet this aim.

In his paper "Course Improvement Through Evaluation" written in 1963 Cronbach formulated the problem as such:

1. If evaluation is supposed to be useful for – in this case – the development of new courses it has to be related to the decisions developers have to take in the course of this development process. Therefore, evaluators shouldn't ask for the objectives someone has in mind and how these were reached, but rather:
 - Who are the decision-makers?
 - What kind of decisions do they make?
 - Which criteria are used?
2. Evaluation has to be targeted to the refinement and improvement of the course developed, and in fact during the development stage.
3. If evaluation has to serve the amelioration of the course or the system, the comparison of the courses or the systems is of little use.

The extension of innovative programs called for evaluation activities in view of the increasing confusion about the right evaluation concept.

During the following years numerous article were published dealing with evaluation research, but to present them in detail would go beyond the scope of this paper.

In 1967 "The Methodology of Evaluation" of Michael Scriven was published, who partly reconciled the discussing parties with a "both...and": Both the formative evaluation (as the continuation of the thought of Cronbach regarding the improvement of the course) and the summative evaluation (in a way the identification of the sum of the outcome innovation was offering). Thus evaluation for different ends; professional, but also by an amateur; both intrinsic or process evaluation and pay-off or results evaluation. But Scriven, in contrast to Cronbach, clearly opted for the comparison of courses and systems. He even went so far as to speak of a hybrid evaluation, i. e. a blend in the practical line of action.

A further extension of the claim for evaluation was the outcome of an article by Elliot Eisner about "Instructional and Expressive Objectives" in 1969. According to Eisner, learning targets are not only not impartial, but the way they are formulated, developed and codified expresses a specific metaphorical thinking; there are 3 dominant metaphors:

- The industrial metaphor according to the pattern of scientific management;
- The behaviouristic metaphor (following the positivistic behavioural psychology);
- The biological metaphor (following the biological development theory).

According to Eisner, teachers, in contrast to researchers and evaluators, are used to think and act within the dimensions of the third metaphor. Evaluation following the patterns of the first two metaphors misses the thinking and acting of these experts from the practice. In view of this each evaluation has to bear in mind their way of thinking by "considering the uniqueness and significance of what has been produced".

We are able to profit by the concept of Eisner in two ways: On the one hand we realise that in each evaluation measure we do not only have to consider the ponderability, that we do not only have to include the commensurable into the aspects of evaluation. The opposite approach

is not wrong per se, but in view of its claim of absoluteness, if it tends to consider only quantities and quantifiable things.

The Research-Practice Dilemma

On the other hand it becomes clear that there are different ways of thinking in research and in practice, which cannot really be reduced to a common denominator. Even more: Research and practice are affected by different functions and conditions in view of evaluation measures and as a consequence also in view of evaluation concepts and requirements. This train of thought plays a main role for the following remarks:

In the sense of research evaluation is a series of complex and often very laborious procedures in order to check effects or side-effects of measures in education and to prepare and make decisions. There is no doubt any more, that in the context of a didactic development research the system evaluation should be favoured over the evaluation of the receiver.

But for practical requirements, as they are given in training institutions that are exposed to the permanent claims for a performance evaluation, there arose procedures and measurement practices, which could not meet the scientific demands of a standardised test construction: this is where the home-made small test for the practice came up:

In fact tests are control samples of performances, i. e. of apparent behaviours of the statements of persons, who are induced for the evaluation purpose: If for example the fitness to drive of a passenger car driver was to be checked, we would have to define typical requirements regarding a fit car driver, like parking the car in a parking space, the knowledge of the colours of a traffic light, the ability to coordinate the clutch, accelerator and gear change etc. Doing the test the person concerned would be exposed to such typical requirements situations or performance control tests; but it is not always necessary to go back to real behaviour, as some things could be simulated or found out by simple interrogations. Not all persons involved would have to be lead into the streets and to be put in front of traffic lights, they could be given a picture or a film and be asked for the meaning of the different colours. Their performances, i. e. what they do and say or express in any way as a reaction to these control tests, would be a hint regarding their respective competencies and readiness to act. It would be necessary to elaborate an evaluation key in order to determine the number of performances which limits the criterion if someone is able or unable to drive. In the theory of tests two ways developed how to set this criterion: Either its content is determined in advance (certainly not without having consulted an expert or for example based on the experience from accident reports) or the decision is due to the standards of comparison which is based on the data of a given number of persons who did this test. In the fields of education, formation and profession there are quite a number of tests of this kind, which developed, this way or another, but in general after very laborious procedures. To give an example: The elaboration and testing of a school performance test, covering about the pensus of a subject for one year, needs a time frame of one to two years and a team of about three people until this test is available and can be generally used.

The problem is even bigger in cases where expectations, contentment etc. of learners are supposed to be measured in order, for example, to provide a basis for the identification of attitudes towards specific learning procedures, environments or other. But these are fundamental variables of evaluation measures. Similar to what has been presented about the development regarding the performance measurement, self-made measurement instruments were normally developed and used in the framework of evaluation measures.

It is only in the context of more extensive projects, which either explicitly concentrate on the comparing evaluation (like the PISA-studies) or which work with such a high number of probands that there are sufficient internal possibilities of comparison within the population (like the studies by Geert Hofstede about the cultural dimensions), that this problem does not necessarily exist.

Elaboration of a Valuation Standard in Evaluation Measures

This does not lead to the possibility to classify the results according to their value: If, for example, 30% of the interviewees are very content with an eLearning course and 35% content, but 25% not content and 10% not content at all, it is unclear which is the scale for these results. Is this a good or bad result for the appraisal of the course or even a insignificant result?

As far as opinion researches in the field of psephology or market research are concerned the polling institutions therefore have long since gained their own accumulation of experience values; based on previous data they are able to judge the significance of changes, e. g. regarding the appreciation of parties or politicians and to refer this to social groups.

Such an accumulation of data which is aiming at a longer period is also needed in view of evaluation measures of educational projects. This shouldn't come from some "secret knowledge" of one institution, but some kind of "open source" model should be developed and offered. This is closely linked to the idea of inventories. This term has been used in psychological and social-psychological research for a long time in order to underline the completing character (comprehensive measuring of certain characteristics or traits, e. g. fear, preferred ways of learning) of measurement instruments, which allow sums or percentiles as grading value.

Inventories in the Framework of the eL3 Project

In the context of the eL3 project such inventories are to be prepared regarding the collection of the following features in or according to course offers with blended learning:

Contentment of the Learners:

What expectations and estimations do learners have at the beginning of a course, how do they judge the process and the outcome?

Characteristics of the Learners:

What knowledge or assumptions do the authors and producers of blended learning courses have about "their" learners and which information should they possibly gather.

Evaluation of the Quality of the Didactic Design:

How do the authors and producers of blended learning courses, but also other groups of persons, judge the quality of the didactic design and of the material used, in the sense of an inspection evaluation.

Evaluation of the Effect of Blended Learning Courses:

How do the authors and producers of blended learning courses, but also other groups of persons, judge the effect of the didactic design, in the sense of an inspection evaluation.

Requirements and Evaluation by Decision-makers:

Which requirements do people express, who can be called decision-makers in connection with the elaboration of blended learning courses, and which criteria are the basis for their decisions, e. g. for the acquisition of courses?

The collected items of these first inventories are analysed and commented regarding their content in a first draft by the project partners, i. e. the suppliers or developers of courses with blended learning; they are then translated into the national languages and are presented to persons of the respective project environment in order to be completed, so that first data-bases can be made available, to do a revision and new version of the inventories. This should be finished by autumn 2005, so that a revised version can be presented via the blinc-platform and can be offered to other interested parties; here the further idea is important that the inventories are regarded as "item-pool", i. e. that they are the basis for the composition of concrete questionnaires especially designed for the particular needs of the relevant course suppliers.

We still have to settle the question of the modus of a further use, which either goes back to text files, which the interested parties adapt, administrate and evaluate themselves as the basis for the elaboration of questionnaires, or which happens on the platform as the interested parties compose their version according to their needs and then make it available as online-questionnaire to their relevant circles of recipients. In any case the willingness to make these data available is a vital precondition for the own utilisation, so that the presented collection of information, which is the original aim of the procedure, is done via the items and the fundamental statistical parameters can continuously (each new usage offers additional information regarding variables like mean value, variation, internal correlations, findings about sub-groups etc., therefore it improves the informational value of the desired comparisons) be collected and used.

Bibliography

Cronbach, L.J.: Course improvement through evaluation. In: Teachers College Record (64) 1963) No: 8, p. 672-83.

Dohse, W.: Das Schulzeugnis – Sein Wesen und seine Problematik. Weinheim 1963.

Eisner, E.: Instructional and Expressive Educational Objectives: Their Formulation and Use in Curriculum. In: AERA Monograph Series on curriculum evaluation. Nr. 3. Chicago 1969, p. 1-18.

Haller, H,-D.: Evaluation von Lehre- ein Weg zu einer effektiveren Wissenschaft? In: D. Hoffmann/K. Neumann (Hrg.): Ökonomisierung der Wissenschaft. Beltz, Weinheim etc. 2003, S. 177-1191.

Hofstede, G.: Interkulturelle Zusammenarbeit. Kultur - Organisation - Management. Wiesbaden 1993.

Scriven, M.: The methodology of evaluation. In: R.W.Tyler et al.: Perspectives of curriculum evaluation. AERA Monograph Series on curriculum evaluation. Nr. 1. Chicago 1967.

Stufflebeam, D.L.: Evaluation as Enlightenment for Decision-Making, Improving, Educational Assessment and an Inventory of Measures of Affective Behavior. In: W.H. Beatty (ed.): Association for Supervision and Curriculum Development (ASCD): Washington 1969, p. 41-73.

Tyler, R.W.: Basic Principles of Curriculum and Instruction, University of Chicago Press. Chicago 1950.