Methods and Practice of Software Evaluation.
The Case of the European Academic Software Award

Peter Baumgartner and Sabine Payr

Baumgartner, P. & S. Payr. 1997. Methods and practice of software evaluation: The case of the European Academic Software Award (EASA). In: ***Proceedings of ED-MEDIA 97 - World Conference on Educational Multimedia and Hypermedia,*** **Edited by . Charlottesville: AACE. 44-50.**

Abstract: The paper discusses theoretical and methodological problems of software evaluation in Higher Education. It criticizes the commonly used approach in product evaluation using numerically weighted lists of criteria. On this basis, an alternative method with qualitative weighting is presented. The European Academic Software Award held in Klagenfurt/Austria in 1996 (EASA '96) serves as an example of the practical implementation of this method, whose outcome is reported at the end of the paper.

1. Motivation

The evaluation of interactive media becomes more and more important: on the one hand because of the growing offer of software from which good products have to be chosen, on the other hand in order to promote software quality and to set quality standards. In the domain of academic software, the EASA was created with these considerations in mind.

The aim of EASA is to promote the development and use of software in research and higher education. There are two main aspects to this aim, reflected by the definition of "academic" software as either software *for use in* higher education and research, that is, educational software, but also tools for teaching, learning, and research, or as software developed *in* higher education and research. The award is intended both to make software producers and distributors aware of innovative ideas and programs developed in the academic world and to point out to the academic audience good examples of technology-enhanced teaching and learning.

The original focus of EASA being the promotion of new developments excluded the use of certain evaluation approaches, like, for example, a quasi-experimental design (comparison of reference groups by pre- and post-testing). At the first EASA (Heidelberg, 1994), the evaluation was done instead by expert groups using lists of criteria.

2. Evaluation with checklists

In checklists, the features that are considered necessary and/or desirable in the product that is to be evaluated (the evaluand) are listed, and the product is rated for each feature. This method has a number of advantages: It is

- *cheap*: one expert, one licence of the program and the suitable hardware configuration are in principle sufficient to evaluate the software and fill in the checklist.
- *easy to organize*: the software need not be assessed in a situation of real use (classroom), evaluation can be assigned to a specialized (and centralized) group or institution removed from the context of use.
- *at first view, methodically "clean":* by going through the same, often voluminous checklists in each examination, the method appears to be objective.

However, this method is insufficient for different reasons:

- The practice of establishing lists of criteria and of working through them in the evaluation procedure does not avoid the problem of weighting the different features to come to a final conclusion: "No software is perfect. Every item has strengths and weaknesses. It is the evaluator's job to identify those virtues and defects and then decide which outweigh the others. - It is a judgement call." ([Doll 1987], p. 58). But it is exactly this "judgement call" that has to be methodically clean, and this is the methodological problem the present paper focuses on.
- Another problem lies in the assumption that the usefulness of educational software can be assessed in a context-free judgement. Often, however, it is not the product (piece of software) itself but the innovative use that is made of it that guarantees the positive learning effect. For this reason, future EASA evaluations will include case studies and shift the focus of the assessment from the product to the learning situation.

3. The Logic of Evaluation

Basically, "evaluation is the determination of a thing's value" ([Worthen and Sanders 1987], p. 22), or, as Scriven ([Scriven 1991a], p. 1) puts it: "Evaluation is the process of determining the merit, worth and value of things, and evaluations are the products of that process."

This definition implies that in the center of every evaluation is the formulation and assignment of a value judgement. In this sense the logic of evaluation are different from other disciplines (e.g. social research).

- Formulation of value criteria: First, those criteria that the evaluand has to satisfy in order to count as valuable or good are selected and defined.
- Formulation of standards: For each criterion, a standard or norm has to be established. Only if the evaluand comes up to this standard, it is considered as satisfying the criterion.
- Measurement and comparison (analysis): The evaluand is analysed and measured on each criterion and compared to the predefined standards.
- Value judgement (synthesis): In this last and most difficult stage of evaluation, the individual results of measurement and comparison have to be integrated into a single value judgement.

Each of these steps or stages gives rise to certain problems:

- *Formulation of criteria*: Which criteria should be chosen? What importance (weight) should be given to each criterion for an overall assessment? Note that this is mostly a theoretical question whose answer needs a thorough analysis of the evaluand. The attempt to arrive at a complete list of criteria by putting together all existing lists (as done, for educational software, by [Thomé 1989], who collected 324 individual criteria from which she compiled a "Long Checklist for Educational Software" with 221 criteria) does not solve the problem, because it does not guarantee neither that the criteria are independent of each other nor that they are really of equal importance.

- *Formulation of standards*: This implies that criteria have to be operationalized, which has not yet been done for the majority of the criteria usually applied to educational software. For example, how should "interactivity" be measured? Based on the analysis of the evaluand (in our case: of theoretical assumptions about academic software) we can distinguish three main levels of standards:

  - Level 1 = Requirements or needs (necessitata): These must be satisfied in any case if the evaluand should be included in the further evaluation procedure (k.o. criteria). They are essential criteria that every evaluand must have or fulfill, for example: a certain educational software cannot be used if it does not run with the existing hardware.

- Level 2 = wishes or desiderata: These are functions and features that are not mandatory but useful. They increase the value of the evaluand and are important for the final value judgement (e.g ranking). They have to be defined and weighted in a way that mirrors the evaluation goal. (For examples cf. the list of criteria used in the EASA Competition in [Tab. 1])

  Two commonly made mistakes are: (a) Too many and too detailed criteria, whose number and collective weight inundate the evaluation process and complicate the synthesis needed for the final value judgement. (b) A bias towards certain features results from criteria that ”overlap” (= are not independent of each other).
- Level 3 = ideals: Although they can hardly be reached, they provide an important perspective by pointing to ways in which the evaluand could be further improved. This is especially important for formative evaluation, where the evaluand is still under development. Most of the EASA criteria can also be used for program improvement, and actually this is one of the motives of this award: to raise the quality of academic software.

- *Measurement and comparison*: This stage depends on the previous operationalization of criteria. Where this is - as in the case of interactive media - still under way, no generally approved, tested and normed scales are in use. Instead of objective measurement, we therefore had to rely on dialogue and intensive discussion between jurors and authors and among the jurors themselves. Contrary to an objectivist point of view, we do not consider this to be a shortcoming. In expert groups a certain element of subjectivism helps to integrate different opinions. Often the qualitative assessment is a preliminary stage for the establishment of an accepted standard that finally can lead to a quantitative evaluation procedure. We doubt, however, that this will ever be the case with educational software.

- *Value judgement*: The integration of individually assessed criteria into one value judgement implies a prior definition of the role or weight of each criterion. Such a definition has to be based on the theoretical assumptions that have to guide the evaluation procedure. At last year‘s Ed-Media Conference, we proposed a detailed theoretical framework that can be used as a heuristic model for such an assessment ([Baumgartner and Payr 1996a]). In this paper, we will therefore concentrate on the methodological difficulties. Let us first compare two methods commonly used for the weighting of criteria in product evaluations - numerical vs. qualitative weight and sum.

4. Methods of Weighting

4.1 Numerical Weight and Sum (NWS)

This method is frequently applied and can take different forms. The general form is called multi-attribute utility analysis ([Scriven 1991b], p. 380f.): 1) The relevance (weight) of each criterion is set using a scale from e.g. 1-3, 1-5 or 1-10. 2) The evaluand is rated for each criterion. 3) Rating multiplied by the weight gives the result for each criterion, results are added up for each evaluand. 4) The final result is a single number for each evaluand. The evaluands can be ordered by this number (ranking), the one with the highest score being the "winner".

This method has a number of intrinsic problems:

a) By attributing and adding numerical values, this method assumes a linear scale of utility for all criteria. But this assumption is clearly wrong! At the moment there is no normed, tested, standardized and agreed linear scale for the quality of educational (or academic) software and it is doubtful if there will ever be one. As it stands, the different components or dimensions must not be added up to a single final number.

b) The multiplication used for weighting also presupposes a metrical scale where a zero value makes empirical sense, which is not the case with many criteria. It could maybe make sense-, for example, with the criterion of "documentation" (from "no documentation at all" to "all features are documented")~~.~~, but any other question like "is the documentation adequate, clear and useful?" can only be answered on an ordinal scale (like: "sufficient - good - excellent") that allows ranking, but not calculation.

4.2 Qualitative Weight and Sum (QWS)

This alternative method overcomes the methodological difficulties of the numerical approach and consists of three main steps:

1st step: Constructing the list of criteria
As in the NWS method a list of criteria is established and weighted. The crucial difference is that QWS is not based on the assumption of an interval or ratio scale. In order to prevent the possible confusion with numeric operations that are only legitimate for linear scales, Scriven ([Scriven1991b], p. 294) recommends the use of symbols for the weights. Frequently used symbols are: E = essential, * = very valuable, # = valuable, + = marginally valuable, and 0 = zero.

2nd step: Weighting of criteria
The weight of a criterion determines the range of values that can be used to measure an evaluand‘s performance. For a criterion weighted #, for example, the evaluand can only be judged #, +, or 0, but not *.

Three rules help to execute complex product evaluations, e.g. evaluations with many evaluands and many different components or dimensions:

1) *Elimination of evaluands:* An evaluand that does not satisfy a criterion considered essential (E) is eliminated. In our case, such a criterion could be, for example, if a software crashes. Such a criterion can already reduce considerably the number of evaluands. But the process of elimination is not always as easy as the example suggests. Essential criteria often are not "all or none" criteria, but establish a certain minimum standard of performance. For example, response time of a software must not be longer than a certain limit (set by psychological factors), but inside this limit the variation of response time does not make a substantial, qualitative difference.
2) *Elimination of 0-criteria*: Every criterion that gets the weight 0 in the weighting process can be eliminated - it is judged irrelevant.
   The remaining evaluands (after applying rule 1) are now evaluated by attributing a value for each remaining criterion (after applying rule 2) up to the maximum weight. If a juror does not feel sure or has any doubts he/she can put a value in brackets, so that this criterion for this evaluand is marked for further analysis or comparison.
3) *Elimination of criteria with uniform results*: Criteria where all evaluands have reached the same level can be eliminated from further consideration, based on the assumption that all evaluands are more or less equal in this regard and that the criterion therefore does not contribute any distinctions.

3rd step: Ranking
Counting the different symbols given to each evaluand results in three numbers for each evaluand - the number of *, of # and of +. The evaluands can now be ordered (ranked) according to these numbers.

This ordered list can either be used for grading, e.g. for cutting off a certain proportion of evaluands either on top or on bottom of the list, or it can be the basis for determining a "winner" and a ranking of results. For a definitive ranking, however, results have to be analysed more closely. There is no doubt that an evaluand with 3*, 4# and 2+ is better than one with only 2*, 5# and 2+, but it is not clear whether it is better than one with 2* and 7#. It may be necessary to analyse and compare these two candidates more closely.

It may also be necessary, in case of unclear results, to proceed to a second round of weighting. The method of QWS is not only complex, but also has this disadvantage of not offering a clear decision algorithm. Sometimes it has to be applied several times and the evaluations have to be redone in the light of previous results.

5. QWS Applied to EASA

The method applied in the final evaluation round of the EASA was a variety of QWS ([Baumgartner and Payr, 1996b]). It departed from the general method as outlined above mainly by not using the E symbol or essential criteria. The reason for this was that the programs admitted to the final stage of the competition had already been tested and pre-selected, so that they could safely be assumed to satisfy a certain minimum standard for the criteria that played a role here (they were, of course, functional and virus-free, but the pre-selection had also looked into aspects of educational use, language, design and innovation). The shortlisted projects (35 out of 157 submissions) were invited to the final event, together with 20 jurors drawn from the jurors' body established in the pre-selection round and 5 additional student jurors. The authors set up their software in an exhibition-style setting. The evaluation process took two days, starting with an introductory session where the "rules of the game" were established, and ending with a plenary session where the final decisions on the award winners were endorsed.

5.1. Weighting of Criteria

A list of 12 criteria ([see Tab. 1]) was proposed to the jurors. The jurors attributed the value of *(very important), # (important, relevant) or + (additional, less important) to each criterion. This specific expert group considered all 12 criteria at least "important", but most of them "very important" (criteria 3, 8 and 11 had #, the rest *). So the jurors gave themselves, for most criteria, the broadest possible range of values to attribute in the following evaluation round.

5.2. Evaluation

For the evaluation, jurors worked in five sub-groups. Each of these sub-groups had to evaluate seven programs. They had a whole day for examining the programs and the documentations and for requesting additional information from the software authors. The jurors filled in their evaluation form for each software individually and then discussed it in their sub-group to come to a single group evaluation. Only this group result was taken into account for the subsequent ranking.

5.3. Ranking and Discussion of Results

The next day, the evaluation results were presented to the plenary session of jurors in the form of tables where not only the integrated results (the number of * etc. for each software) were made visible, but also the result for each criterion, including bracketed symbols and cases where jurors considered a criterion not to be applicable. Moreover, these tables were arranged and presented under different aspects, like submission category (department projects, student projects, and commercial projects), country of origin, and discipline. This documentation was necessary as a basis for the discussion of the results. Certain aspects had to be analysed before coming to a decision:

- Are the results consistent, i.e. have the groups applied comparable standards when attributing values for each criterion, or are one group's results constantly lower or higher than the rest?
- Is an overall good result, but without any highlights, i.e. twelve #, but no * - preferable to a software with some outstanding features, but also with some weaknesses - i.e. six * and six +?
- How many awards should be given away, and in which categories? For example, if the best program of a certain award category comes very late in the overall ranking, is it still good enough to get an award, or should this award be left out?
- Does the resulting list of award winners represent the different disciplines, countries, submission categories etc. well enough?

EASA '96 Evaluation Criteria

1. Correctness: Is the subject material accurate and up-to-date? (For tools:) Is the program functional?

2. Relevance: Does the software correspond to real user needs? Is the contents relevant for teaching and learning in the subject area?

3. Coverage: Is the subject material sufficiently covered? Does the software cover an important part of the subject area? (For tools:) Is the range of functionality appropriate?

4. Interaction: Is the software highly interactive? Does the software encourage active/exploratory learning? Does the software create and maintain learner motivation and interest? (For tools: not applicable)

5. Learning: Is the material well structured and organised in order to support the learning process? Are learning objectives defined and can they be attained? (For tools: not applicable)

6. Usability: Is the software appropriate for the target group it addresses? Can the software actually and easily be used in research, teaching and learning? Does the software run on current students'/universities' computers?

7. Navigation: Can users always see clearly where they are in the program and what actions/functions are available? Does the software always show its current staturs, mode? Are reactions of the program to user actions clear and appropriate?

8. Documentation: Is online help available? Are manuals, tutorials etc. available? Is the documentation clear and useful for the target group?

9. Interface: Are contents and functions well organised on the screen, easy to learn and used, and well presented? Does the software follow the known standards of interface design? Does the software satisfy ergonomic requirements?

10. Use of computer: Does the software support activities, forms of teaching and learning that are not or not easily feasible otherwise? Does the software make adequate use of the medium?

11. Adaptability: Can the software be easily updated and adapted to new contents and teaching/learning requirements? Is the software portable to other European curricula and languages?

12. Innovation: Does the software contribute new and interesting aspects to educational computing and multimedia? (For tools: ... to computing?)

Table 1: Evaluation Criteria used for EASA '96

In the case of EASA, no single "winner" had to be found, so that the method was used for a process of apportioning: a group of award winners had to be cut off from the rest of the finallists (who, in their turn, were already considered "winners" of some sort, having reached the final stage of the competition).

The plenary discussion had, in this case, to take the role of the re-analysis and comparison of evaluands, simply because of the limited time frame.

The results showed a rather clear "top group" that was defined as the award winning group without much further discussion. It only had to be completed with some programs that were promoted from "lower down" in the list, the submission category turning out to be the decisive factor: these additional award winners were student projects, which jurors did not expect to offer the same level of performance and quality as department projects and commercial software.

6. Conclusions

We started out criticizing checklists as a method of evaluation for interactive media, went on to present alternatives and reported the practical experience with the qualitative weight and sum (QWS) method as used by an expert group in the EASA '96.

Applying this method of product evaluation to the EASA can serve as a good illustration of the necessity to come to terms with both theory and practice of evaluation in any specific setting. While theory can clarify the respective merits and shortcomings of different evaluation methods and formulates the critical questions that guide evaluation design, practice sets up inevitable limits of time and cost to any evaluation. What we hope to have shown is that by

- clearly defining the evaluation task at hand
- following the logic of evaluation and
- taking into account the practical limitations

it is possible to design evaluation settings that are theoretically grounded, practically feasible and adequate for their specific purpose and goal.

As there exist plans to broaden the scope of the European Academic Software Award to include not only software products, but also case studies and experiences in the innovative use of software in higher education, EASA will be progressively leaving the field of product evaluation. With these new pedagogic goals in the award, other methods of evaluation will have to be found and tested. So we can look forward to the next EASA which will take place in the U.K. in 1998 (see [http://www.york.ac.uk/inst/ctipsych/easa/]), both for interesting submissions and for new insights into their evaluation.

7. References

[Baumgartner and Payr 1996a] Baumgartner, P., & Payr, S. (1996a). Learning as action: A social science approach to the evaluation of interactive media. Proceedings of ED-MEDIA 96 - World Conference on Educational Multimedia and Hypermedia, Charlottesville. 31-37.

[Baumgartner and Payr 1996b] Baumgartner, P., & Payr, S. (1996b). European Academic Software Award 1996 (EASA '96) Internal Report, University of Klagenfurt: ASI.

[Doll 1987] Doll, C. A. (1987). Evaluating Educational Software. Chicago/London: American Library Association.

[Scriven 1991a] Scriven, M. (1991a). Introduction: The Nature of Evaluation. In: Evaluation Thesaurus, ed. M. Scriven, 4th ed. Newbury Park: SAGE. 1-43.

[Scriven 1991b] Scriven, M. (1991b). Evaluation Thesaurus. (4 ed.). Newbury Park: SAGE.

[Thomé 1989] Thomé, D. (1989). Kriterien zur Bewertung von Lernsoftware. Heidelberg: Hüthig.

[Worthen and Sanders 1987] Worthen, B. R., & Sanders, J. R. (1987). Educational evaluation: Alternative approaches and practical guidelines. White Plains: Longman.