INTERNATIONAL DESIGN CONFERENCE - DESIGN 2004
Dubrovnik, May 18 - 21, 2004.

DESIGN 2004

EVALUATION OF A LIFE CYCLE ASSESSMENT TOOL - ADJUSTING THE SOFTWARE DEVELOPERS' VIEW TO THE EXPECTATION OF THE USER

T. Felsing, M. Dick, H. Birkhofer and B. Rüttinger

Keywords: design for environment (DfE), life cycle assessment (LCA), heuristic evaluation, software evaluation, IsoMetrics^{L}

1. Introduction

The *Collaborative Research Center (CRC) 392* at *Darmstadt University of Technology* develops methods and instruments that enable product developers to assess the environmental impact of the product they design. The research is particularly focused on a computer-based design environment, which is capable of a prospective and holistic assessment [Anderl, Weißmantel, Daum, Pütter & Wolf 1999]. The *Product Development Environment (PDE)* consists of three components: (1) a 3D CAD system, (2) the *Life Cycle Modeller (LCM)* and (3) an assessment tool *(LCAD – Life Cycle Assessment for Computer Aided Design) (see figure 1)*. This paper describes the evaluation of the PDE.

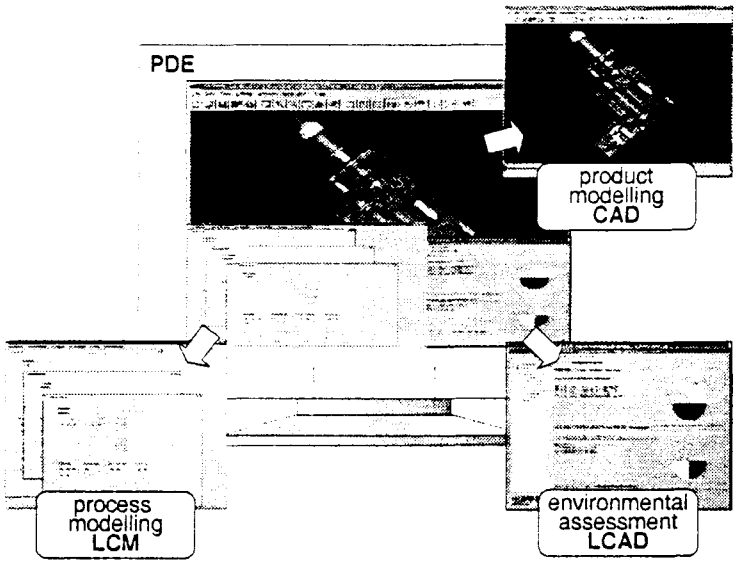

Figure 1. Components of the Product Development Environment (PDE)

The evaluation is based on a four-phase concept: In the first phase, potential users evaluated the usability of the implemented *PDE* prototype. In the second phase, after revision of the system, its usability was tested again. This was done to get some empirical evidence about the success of the revision process. In the third phase, the PDE will be compared with other ecological assessment tools

and evaluated with respect to a set of subjective and objective criteria. Finally, in the last phase, the PDE will be evaluated during and after the implementation in a company.
Within the second phase, which is of central interest in the present paper, the usability of the *PDE* was measured by a mix of qualitative and quantitative methods, concerning the following dependent variables: (1) well-being of users before and after working with the *PDE*, (2) the subjective assessment of the system by users and (3) usability problems noted by users after working with the *PDE*. The results were compared with the results of phase one, which are documented in [Wiese & Rüttinger 2001]. Furthermore, the participants of the second evaluation phase were classified with regard to their knowledge and expertise in the field of product development and especially with reference to the *PDE*. After this, the different groups of participants were compared with respect to the dependent variables (see above).

2. Method

2.1 Participants

Fourteen participants took part in the study. They were recruited from the staff of the CRC 392 and were mainly engineers (one participant was a psychologist).

2.2 Procedure

The participants had to work about ninety minutes with the *PDE* in the Design for Environment Laboratory (DfE-Lab) of the CRC 392. Before and after doing this, they had to fill in questionnaires, which are described in detail in the following sections.

2.3 Design for Environment Laboratory (DfE-Lab)

The study was carried out in the *Design for Environment Laboratory (DfE-Lab)* at *Darmstadt University of Technology*. The DfE-Lab is equipped with several workstations with the PDE installed on them. A video observing system recorded the participants, working with the PDE, as well as the video signals of the workstations. The audio and video signals were converted into digital files, making it possible to record the whole process while simultaneously marking the video file with a time and code scheme. This enables an ex-post evaluation of critical situations.

2.4 Data collection methods: Quantitative Data

The subjective assessment of the PDE was done with a slightly modified version of the IsoMetrics^{L}. IsoMetrics is an instrument for the formative and summative evaluation of software according to the international norms of the ISO 9241 Part 10 [Gediga, Hamborg & Düntsch 1999]. With respect to these norms, the IsoMetrics includes the following subscales: suitability for the task, self-descriptiveness, controllability, conformity with user expectations, error tolerance, suitability for individualisation and suitability for learning. The IsoMetrics is described as a "reliable and valid tool, supporting formative and summative evaluation of software systems" [Gediga, Hamborg & Düntsch 1999, p. 162]. Two versions of the IsoMetrics were developed: One for summative – IsoMetrics^{S} – and one for formative evaluation – IsoMetrics^{L} [Gediga, Hamborg & Düntsch 1999]. In the present study, a slightly modified version of the IsoMetrics^{L} was used. With this instrument, formative evaluation is done in two steps: First, the software is assessed with items according to the ISO 9241 Part 10 ("rating score"). In a second step, the importance of every item for the general impression of the evaluated software is gathered ("weighting index"). The distance between rating score and weighting index is seen as a measurement for usability problems: High distances are interpreted as indications for usability problems whereas little distances are seen as indications for good usability. The IsoMetrics^{L} is described in detail in [Willumeit, Gediga & Hamborg 1996].
The well-being of the test persons was tested with a shortened version of the Multidimensional Mood Questionnaire of [Steyer, Schwenkmezger, Notz & Eid 1994]. With this psychometric instrument, three mood dimensions can be assessed: "pleasant-unpleasant", "awake-sleepy" and "calm-restless".

For these scales, reliabilities between .85 and .97 are reported [Steyer, Schwenkmezger, Notz & Eid 1994].
Finally, the expertise of the test persons was measured by their self-reported knowledge about the PDE and their self-reported experience in product development.

2.5 Data collection methods: Qualitative Data

Qualitative data were collected with a so-called *Heuristic Evaluation* [Nielsen & Molich 1990]. The Heuristic Evaluation requires putting down in writing subjective impressions on the object of evaluation, after having worked with it. In the present study, the participants had to write down the usability problems they encountered during working with the *PDE*. To make this process easier, the test participants got an evaluation sheet, which was structured according to the single working steps, they had done before. As a result, people were able to assign their problems to these single steps. To stimulate answers, they got a list of nine usability heuristics that should gave them an idea about what features a good computer based working environment should have.

3. Results

3.1 Quantitative Data

The results of the IsoMetrics^{L} show for all subscales clear differences between ratings and weightings, with the ratings always lower than the weightings (see figure 2). With regard to the subscales suitability for the task/"st", self-descriptiveness/"sd", error tolerance/"et" and suitability for learning/"sl" this differences became significant ($p < .05$).
Five of the seven ratings are of middle size (suitability for the task, controllability/"ct", suitability for learning, error tolerance and conformity with user expectations/"ce"), the others are a little weaker (suitability for individualisation/"si" and self-descriptiveness). Overall, the results are very similar to the results of the first evaluation phase, which are described in [Wiese & Rüttinger 2001].

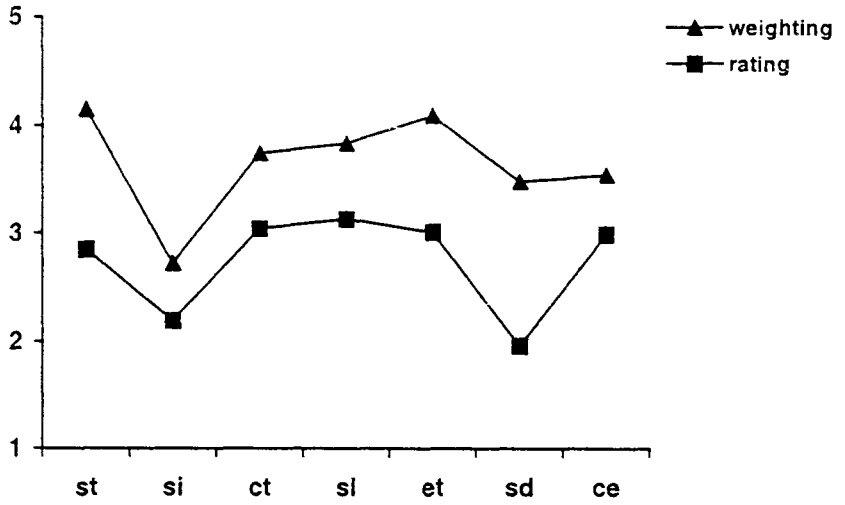

Figure 2. Results of the IsoMetrics^{L}: Whole sample

To analyse the IsoMetrics^{L}-results with reference to the expertise of the participants, the expertise of the participants was measured by their self-reported knowledge about the PDE. Along this criterion, the sample was splitted in two subgroups of novices ($n = 8$) and experts ($n = 6$). In addition to their significant higher knowledge about the PDE ($F = 40.18$; $df = 1$; $p < .01$), experts had also significant more experience in product development ($F = 4.9$; $df = 1$; $p < .05$).
As figure 3 shows, experts ratings could be found in all subscales lower than novices ratings. In the case of the subscale "suitability for individualisation" ("si") the difference became significant

(F = 5.08; df = 1; p < .05). With respect to the individual weightings of the subscales, results of experts and novices are fairly similar.

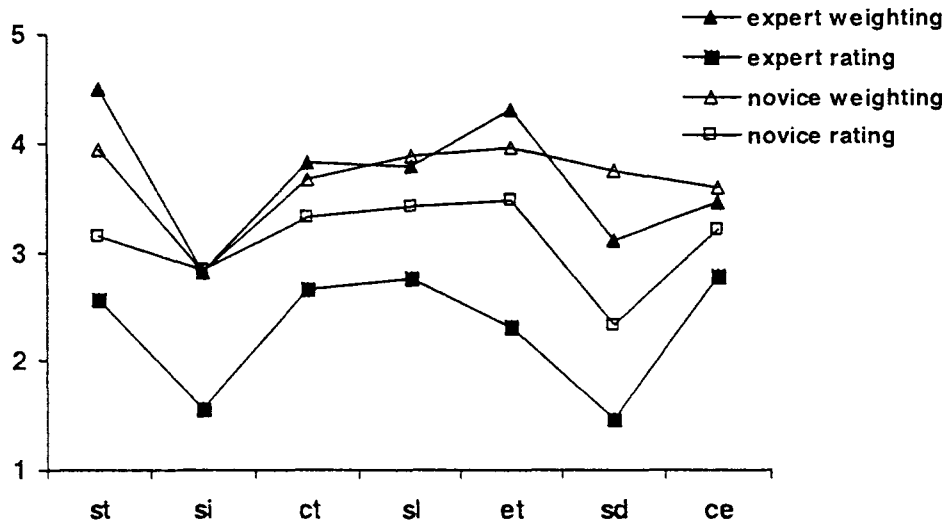

Figure 3. Results of the IsoMetrics^L: Differences between experts and novices

The well-being of the participants was after working with the PDE significant lower than before (F = 4.61; df = 1; p = .05). This effect results from significant differences within the subscale "pleasant-unpleasant" (F = 8.61; df = 1; p < .05). Within the other two subscales, no significant differences occurred. The comparison between experts and novices shows that only in the subgroup of the experts, the well-being was significantly detracted from working with the PDE (F = 20,00; df = 1; p < .01) whereas in the group of the novices such an effect could not be observed. The results with respect to the well-being are illustrated in figure 4.

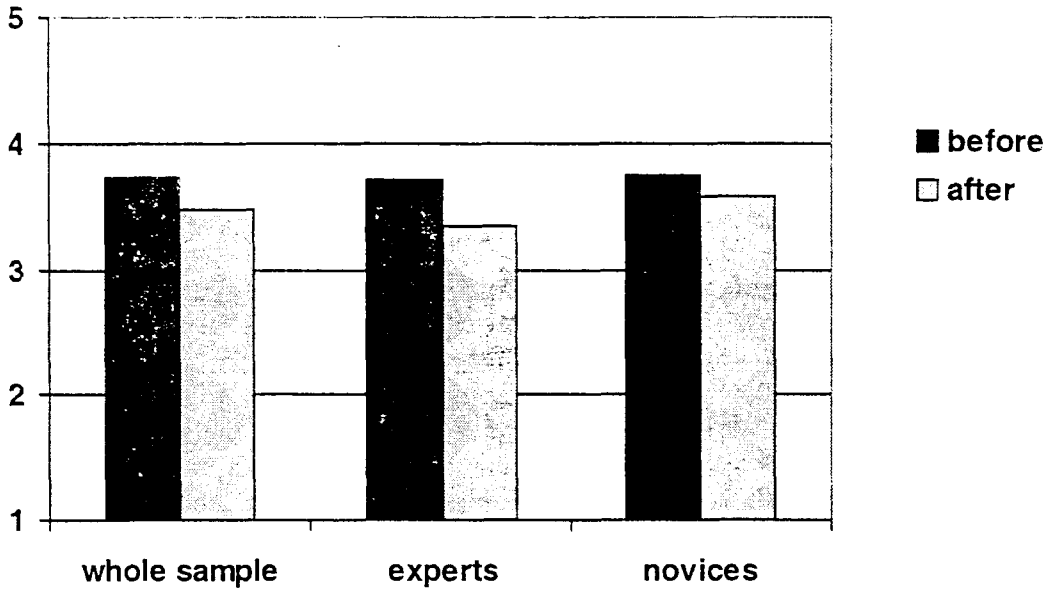

Figure 4. Well-being before and after working with the PDE

With regard to the quality of the used scales, satisfying reliabilities between .58 and .84 for the IsoMetrics^L and .66 and .93 for the Multidimensional Mood Questionnaire could be obtained. Items with a corrected item-total correlation < .30 were excluded from the analysis.

3.2 Qualitative Data

In addition to the quantitative data, the qualitative answers serve to detect usability problems in a more detailed and more concrete way. Especially those problems were of interest that persons wearing "professional blinkers" do not encounter any more.

The filled in questionnaires have been analysed as follows: (1) The questionnaires have been transcribed as a basis for further information processing. (2) In order to understand the - often colloquially expressed - comments in a right way a semantical interpretation followed. (3) For structuring the information synonymous answers have been clustered. (4) Finally, the importance of the reported problems has been classified. The steps (3) and (4) have been carried out in interdisciplinary workshops involving psychologists as well as engineers. The main findings were "mirrored back" to the developers of the PDE and are summed up in Figure 5 and in the following lines.

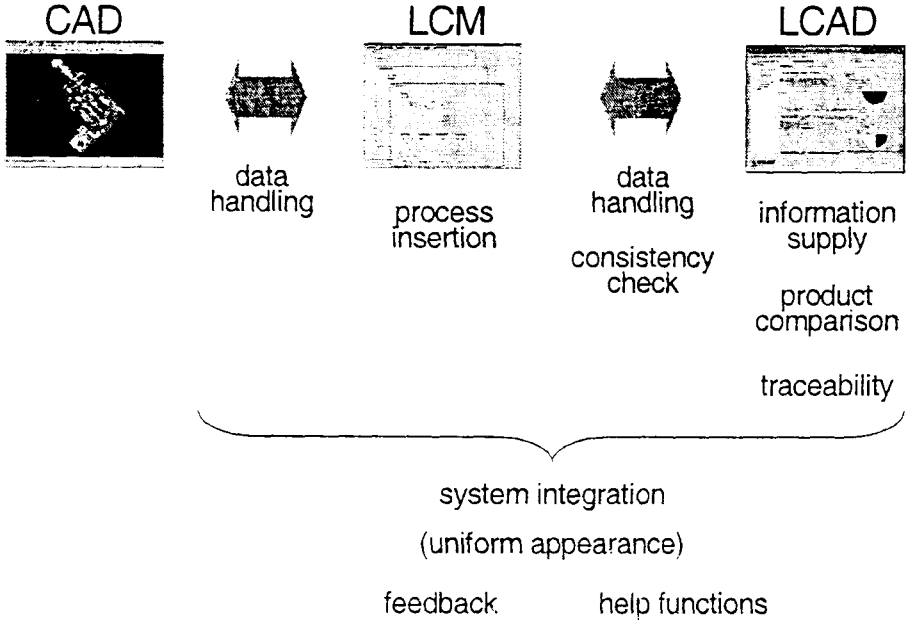

Figure 5. Identified problem fields

Above all, the data handling caused offence. Most of the users reacted with incomprehension as they were "forced" to roll product models and process plans in a database and to retrieve them again awkwardly when switching between different PDE components. Moreover, the three constituents of the PDE lacked an uniform appearance. A corrective measure will be the integration of the PDE system in a so-called "Eco Design Workbench" within the next revision phase.

While modelling the product life cycle with the LCM the participants had some problems with the insertion of processes. The processes belonging to patterned features had to be inserted as often as they appeared in the product model. The users remarked that the system could not automatically assign processes to special features such as drillings, chamfers, etc. Finding the "right" process out of an unstructured list of processes was judged to cause problems when the limited number of processes will be raised in future. Apart from that, the total number of processes implemented in the software prototype has been seen as still too small for the environmental assessment of complex industrial products.

Users had problems to cope with the information supplied by LCAD. It turned out that the front-end of this evaluation system was conceived by environmental experts that did not take the product developers' need of information into account. The user felt overloaded with detailed information about impact categories while a simple comparison between the cumulated environmental impact of two products in one window was not supported. Also tracing back the environmental impacts to the processes, which cause them, was difficult and in some cases not possible. These findings have led to a complete revision of the LCAD front-end that guides the user from aggregated values down to detailed information.

Concerning the user interface, most test persons missed feedback and help functions that have not been implemented yet in the examined prototype.

4. Key Conclusions

All in all the study gives some worthy indications about the usability of the PDE. As the data show, a further revision of the system has to be done. Within the first revision, the focus laid obviously too much in technical improvements, whereas the usability of the system was not as much taken into account, as it seems to be necessary. This may be a typical fault of people, which are focussed in their daily work more on the "hard" technical characteristics of technical systems than on their more "soft" aspects.

The application of a combination of quantitative and qualitative methods in the present study can be seen as very successful. As intended, the different types of data complement each other very well: Whereas the quantitative data give a more general overview about the "status quo" of the system, the qualitative results give a lot of individual and more concrete indications, which aspects of the system should be improved and which not.

The dividing of the participants into subgroups of different expertise was also a successful step: The data show, that the expertise of a software evaluator is connected with his assessment of the software he is working with. With respect to typical results in the field of expertise research, it seems to be probably that the assessments of experts are more realistic than the assessments of novices [Anderson, 2000]. In spite of this, we argue that the inclusion of novices in the evaluation process can be helpful, too. Novices may bring another point of view into account and may for example recognize aspects, which experts do not remark, because they may be postulated from them. Furthermore, the inclusion of novices makes sense if someone wants to estimate how difficult and strenuous it is to work with a special system without basic knowledge about it.

The data about the well-being can be interpreted as an indication for a slight frustration or anger within the group of experts because the PDE may not met their requirements. In sum the well-being of the participants was not affected very strongly by working with the PDE, so working with the PDE seems to be not too strenuous – neither for experts nor for novices.

References

Anderson, J. R., "Cognitive Psychology and its Implications", New York, 200.0

Anderl, R., Weißmantel, H., Daum, B., Pütter, C., Wolf, B., "Life Cycle Modelling. A Cooperative Method Supports Experts in the Entire Product Life Cycle", Proceedings of ICED 99, Munich, Germany, 1999.

Gediga, G., Hamborg, K.-C. & Düntsch, I., "The IsoMetrics usability inventory: an operationalization of ISO 9241-10 supporting summative and formative evaluation of software systems", Behaviour & Information Technology, Vol. 18, No. 3, 1999, pp. 151-164.

Nielsen, J. & Molich, R., "Heuristic Evaluation of User interfaces", CHI Proceedings, April 1990.

Steyer, R., Schwenkmezger, P., Notz, P. & Eid, M., "Testtheoretische Analysen des Mehrdimensionalen Befindlichkeitsfragebogens (MDBF2) ("Testtheoretical analyses of the Multidimensional Mood Questionnaire"), Diagnostica, 40, 4, 1994, pp 320-328.

Wiese, B. S. & Rüttinger, B., "Akzeptanz IT-gestützter Methoden der umweltgerechten Produktentwicklung: Vorschläge für eine theoriegeleitete Evaluation" („User acceptance of IT-based methods for Design for Environment"), Darmstadt: Institutsbericht 2/2001.

Willumeit, H., Gediga, G. & Hamborg, K.-C., "IsoMetricsL: Ein Verfahren zur formativen Evaluation von Software nach ISO 9241/10"("IsoMetricsL: An instrument for the formative evaluation of software with respect to ISO 9241/10"), Ergonomie und Informatik, 27, 1996, pp 5-12.

Tobias Felsing, Dipl.-Psych.
Darmstadt University of Technology, Institute for Psychology
Hochschulstraße 1, 64289 Darmstadt, Germany
Telephone: ++49-(0)6151-16-2097, Telefax: ++49-(0)6151-16-4196
E-mail: felsing@psychologie.tu-darmstadt.de