EVALUATING USER SATISFACTION OF AN INTERACTIVE COMPUTER PROGRAM

Thomas C. Hartrum, Thomas C. Mallary, and Jeffrey W. Foley
Department of Electrical and Computer Engineering
School of Engineering
Air Force Institute of Technology
Wright-Patterson AFB, Dayton, Ohio, 45433-6583

Abstract
A methodology is described for evaluating user satisfaction of interactive software tools. The methodology centers around the development of a survey questionnaire to be filled out by the users. This is an extension of the approach used by Bailey and Pearson [1]. The application of this methodology to develop a specific questionnaire to evaluate both commercial and internally developed CASE tools is described. The results are analyzed both to show an example of the information available from the questionnaire, and to validate the questionnaire as a relevant data gathering mechanism.

1 Introduction
There is considerable activity today in the development of easy to use, interactive computer tools to increase user productivity in areas ranging from business applications to engineering computer aided design (CAD) and Computer Aided Software Engineering (CASE). In order for such tools to successfully achieve their goals, they must be accepted and utilized by the user. Evaluation of user acceptance is important. It is useful to know if users are satisfied or if further investment is necessary. If alternate tools are available, some way is needed to evaluate them, especially with the high cost of many tools. Finally, when developing a tool, some method is needed to verify that new versions are more user acceptable than the older versions. This is especially important when employing rapid prototyping techniques to develop user interfaces.

Since one is dealing with human perceptions, there is no replacement for asking the user how he or she likes the software. However, informal user comments don’t provide sufficiently accurate information for practical decision making. A more quantitative methodology is needed. Although some approaches have been reported, there is a general lack of documented methodologies. Bailey and Pearson [1] and the Rushineks [2] dealt with the evaluation of overall satisfaction with a complete computer system. Coleman et al [3] developed a questionnaire tightly coupled to a specific editing tool.

This paper describes a methodology for the development and use of a questionnaire to measure user satisfaction of interactive software tools. The questionnaire is based on specific concerns of the tool designer while keeping the questions as tool-independent as possible. This methodology is then used to develop a questionnaire to evaluate CASE tools. Some results are presented both to illustrate the use of the questionnaire and to support the validity of this methodology.

The next section presents the methodology for development and application of the questionnaire. In Section 3 the methodology is used to develop a specific questionnaire. Some results of evaluating specific tools are presented in Section 4. Finally, Section 5 draws some conclusions from all this, while Section 6 discusses continuing research and recommendations for further studies.

2 Evaluation Methodology
We have developed a methodology for measuring interactive tool acceptance along three axes. The first is utilization, measured by instrumenting the tool to record when the tool is used and by whom. The second dimension, performance measurement, is accomplished by extending the instrumentation to record the start and stop times of various tool activities. The third measurement indicative of user acceptance is that of user satisfaction, determined through the use of a questionnaire filled in by the tool users. This paper discusses only the third dimension, the user questionnaire.

Our approach is based on that used by Bailey and Pearson [1]. They developed a general survey form with 39 questions or factors. Our goal was to develop a general questionnaire which could be used to support the analysis of user acceptance of any interactive tool. It might be noted that other investigators have taken the opposite approach, defining questions specific to the tool being developed [3].
Various general methodologies for data collection point out the need for a structured approach to experiment design [4] [5] [6]. From these ideas the following six steps were defined.

**Step 1. Establish the Goals of the Investigation:** This phase is intended to determine up front what information is desired out of the evaluation study, and what data is needed to provide that information.

**Step 2. Determine the Survey Questions:** In this step the specific survey questions are developed. In order to measure overall user satisfaction, those aspects of the tool which users feel are important should be measured. In addition, questions should be defined that will provide the data needed to answer the specific questions of the investigator that were defined in Step 1.

**Step 3. Design the Survey Form:** The objective of this step is to define a standard format for the questions defined in Step 2. In earlier experiments a simple qualitative survey was used to record user comments and feelings. Although of some use, this approach precluded any meaningful statistical analysis of the data. To solve this problem, a quantitative questionnaire was developed. The format of the actual survey form is based on the work of Bailey and Pearson [1], which uses the semantic differential technique developed by Osgood et al. [7] The questionnaire consists of \( N \) questions concerning different aspects of the tool interface being measured. Each factor is rated from -3 to +3 on each of four related adjective-pair scales, as shown by the sample question of Figure 1. A reaction score \( R_{ij} \) for user \( i \) on factor \( j \) can be calculated as the average of the four scores \( I_{ijk} \) for \( (k = 1 - 4) \):

\[
R_{ij} = (1/4) \sum_{k=1}^{4} I_{ijk}
\]

**System Feedback or Content of the Information Displayed.**
The extent to which the system kept you informed about what was going on in the program.

<table>
<thead>
<tr>
<th>insufficient</th>
<th>clear</th>
</tr>
</thead>
<tbody>
<tr>
<td>unclear</td>
<td>useful</td>
</tr>
<tr>
<td>useless</td>
<td>good</td>
</tr>
<tr>
<td>bad</td>
<td>satisfactory</td>
</tr>
<tr>
<td>unsatisfactory</td>
<td>unimportant</td>
</tr>
</tbody>
</table>

To me this factor is:

Comments:

\[
S_i = 1/(3N_i) \sum_{j=1}^{N} R_{ij}W_{ij}
\]

where \( N_i \), the number of meaningful questions for user \( i \), is given by:

\[
N_i = \sum_{j=1}^{N} \delta_{ij}
\]

where

\[
\delta_{ij} = 0 \text{ if } R_{ij} = 0
\]

\[
\delta_{ij} = 1 \text{ otherwise}
\]

This is then used to specify user satisfaction as defined in Table 1.

One addition to our form is a comment area for each factor. We have found that a comment area helps users accept the survey better by allowing them to express themselves beyond mere numbers, and in addition often provides more insight into specific problems than do the factor scores alone.

**Table 1: Bailey & Pearson Ratings [1]**

<table>
<thead>
<tr>
<th>NORMALIZED SCORE</th>
<th>TRANSLATION</th>
</tr>
</thead>
<tbody>
<tr>
<td>+1.00</td>
<td>Maximally satisfied</td>
</tr>
<tr>
<td>+0.67</td>
<td>Quite satisfied</td>
</tr>
<tr>
<td>+0.33</td>
<td>Slightly satisfied</td>
</tr>
<tr>
<td>+0.0</td>
<td>Neutral</td>
</tr>
<tr>
<td>-0.33</td>
<td>Slightly dissatisfied</td>
</tr>
<tr>
<td>-0.67</td>
<td>Quite dissatisfied</td>
</tr>
<tr>
<td>-1.00</td>
<td>Maximally dissatisfied</td>
</tr>
</tbody>
</table>

**Step 4. Collect the Data:** Data collection must also be carefully designed to support the investigator's questions that were raised in Step 1. This should include consideration of the test subject population, as well as the task to be performed with the tool. When performing comparison studies, care must be taken to insure that the two test groups differ only in those aspects of the comparison being considered.
Step 5. Analyze the Data: Since data analysis is involved with answering the questions raised during Step 1, it is somewhat dependent on the specific objectives of the study. In general, it involves performing a statistical analysis of the normalized user scores and individual factor scores, and running appropriate statistical tests to examine the differences for various control groups.

Step 6. Validate the Survey: The objective here is to determine whether the survey accurately measures the information of interest. Several indications of the validity of the questionnaire are available, as discussed by Bailey and Pearson. The simplest is an analysis of the user’s importance rating of each factor. A second validation step is to calculate the correlation of the reaction score $R_{ij}$ with the satisfaction score for that factor. Within the four adjective-pairs, each individual score can be correlated with the average of the other three. In addition, pairwise product moment correlations can be calculated between all pairs of the four adjective-pairs to find pairs that are not appropriate. Finally, it is possible to analyze the value of the comment field by assigning a comment score of +1, -1, or 0 to each factor, based respectively on the inclusion of a positive comment, a negative comment, or no comment at all.

3 Application of the Evaluation Methodology

3.1 Goals of the Investigation

The objective for this study was to evaluate the user satisfaction with an interactive CASE tool, both in absolute terms and in terms of improvements over an earlier version. Especially of interest was the ease of learning its use for a newcomer. In addition, we were interested in establishing the validity of the survey. Specific questions defined by the tool designer were:

- Did the user feel comfortable with knowing what was happening?
- How effective was the data display?
- How easy and natural was data input?
- How easy was it to learn to use the tool?
- How well did the program flow?
- Was error prevention and recovery adequate?

And as far as the survey itself:
- How relevant are the survey questions?
- When evaluating two versions, how much does experience with the one first used bias perception of the second one?

3.2 Survey Questions (Factors)

Eight factors were defined for this study. These were based on the above goals and on issues raised by users of the earlier qualitative survey. The resulting list of factors is shown in Table 2.

Table 2: Factors of Interest

1. System Feedback: The extent to which the system kept you informed about what was going on in the program.
2. Display of Information: The manner in which program control and relation information was displayed to the user. Consider the use of two display screens.
3. Program Pacing: The speed with which you were allowed to perform the normalization task.
4. Error Recovery: The extent and ease with which the system allowed you to recover from user induced errors.
5. Data Input: The ease with which you were able to communicate with the CAD tool in terms of inputting data into the system. Consider calibration and manipulation of the input device.
6. Ease of Learning: Ease with which you were able to learn how to perform the normalization task using the CAD tool. Please consider any on or off line help that was available to you.
7. Feeling of Control: Your ability to direct or control the activities performed by the normalization tool.
8. System Usefulness: Your perception of how useful the system is as an aid to a database administrator.

3.3 Data Collection

Data collection consisted of having two groups of graduate students use both versions of a database CASE tool to normalize a single relation. The first group (39 individuals) consisted of members of a database course who had performed normalizations manually and represented
the intended users of the tool. The second group (17 students) were members of an advanced course in software engineering. They were not familiar with normalization, but had been studying man-machine interface and user-friendly issues. They represented the first-time or casual user, and the interest was more in their evaluation of the user-friendliness of the tool. These two groups were further divided approximately in half, based on which version they tested first. Of the total of 56 subjects, 32 used the original version (NORM2) first and 24 used the “improved” version (NORM3) first. One student was unable to get NORM2 to operate, resulting in 55 surveys for NORM2 and 111 completed surveys in all.

All the students were first-time users. They were presented with the problem, told how to invoke the tool, performed the normalization of one relation, and then immediately filled out the survey.

4 Results

The objective of this section is provide an example of how the results can be used and the type of information that can be deduced. Three analyses are presented here, based on this study’s objectives. The first is an absolute evaluation of NORM3, the second is a comparison of NORM3 with NORM2, and the third compares the evaluation of two different subject groups, including both the effect of which tool was used first, and the effect of users with different backgrounds. Finally, the validity of the survey questionnaire is analyzed.

Descriptive statistics were calculated for each of the individual factors, for each tool. Differences between the two sets were examined to test the hypothesis that NORM3 was a “better” tool than NORM2. Scores on specific factors were examined to see where further improvements needed to be made. Then scores were grouped by background class and order of evaluation to see if there were any significant differences.

All data analysis was performed using the SAS data analysis tool [8] running under the VMS operating system [9] on a VAX-11/780 computer. (SAS is a registered trademark of SAS Institute, Inc., and VMS is a registered trademark of Digital Equipment Corporation).

4.1 Absolute Analysis of NORM3

In the absolute mode the survey is used to evaluate a tool by looking at the overall normalized score and at the raw scores of the individual questions. This is the way the original survey was used by Bailey and Pearson, and by Deese [10]. All data for NORM3 was combined, giving a sample size of 56. The normalized score of 0.459 falls about 38% of the distance between “slightly satisfied” and “quite satisfied” on Bailey and Pearson’s scale of Table 1. The

<table>
<thead>
<tr>
<th>Score</th>
<th>Description</th>
<th>Occurrences</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Neutral</td>
<td></td>
</tr>
<tr>
<td>0.33</td>
<td>Slightly dissatisfied</td>
<td></td>
</tr>
<tr>
<td>0.67</td>
<td>Quite dissatisfied</td>
<td></td>
</tr>
<tr>
<td>0.33</td>
<td>Slightly satisfied</td>
<td></td>
</tr>
<tr>
<td>0.67</td>
<td>Quite satisfied</td>
<td></td>
</tr>
<tr>
<td>1.00</td>
<td>Maximally satisfied</td>
<td></td>
</tr>
</tbody>
</table>

Figure 2: Histogram of Overall Scores for NORM3

histogram of the normalized score in Figure 2 shows that 15 of the 56 users (27%) scored NORM3 between “quite satisfied” and “maximally satisfied”. The raw scores on the individual questions appear relatively consistent, and are all on the positive side of the scale. These scores are shown in more detail in the histograms in Figure 3.

4.2 Relative Comparison of NORM2 and NORM3

The second approach is a relative mode, where the survey results are used for a comparison between the two tool versions, to see if the users perceived an improvement. Figure 4 shows the average raw scores for the individual questions for NORM2 and NORM3, indicating a noticeable improvement for all questions. The overall normalized score of 0.056 for NORM2 was rated just above “neither satisfied nor dissatisfied,” as compared to the “quite satisfied” score of 0.459 for NORM3. All such differences were statistically significant at the 0.0004 level.

4.3 The Effect of Which Tool Was Used First

Previous qualitative data had indicated that the tool used second tended to receive a higher score. In these tests, 32 used NORM2 first and 24 used NORM3 first. The average overall normalized NORM3 score was 0.563 for the group who used NORM2 first, and was 0.320 for the
group using NORM3 first, with a level of significance of 0.003. All of the individual question scores and overall scores consistently followed this pattern.

There are two possible explanations. Those who used NORM2 first had experience using some interactive normalization tool before using NORM3, while those using NORM3 first had no prior experience whatsoever. Or, those using NORM3 first had nothing to compare it to, and so performed an absolute rating based on the survey scales. The group who had used NORM2 first, however, had prior experience with an inferior tool, and rated NORM3 based on the perceived improvement over NORM2.

4.4 The Effect of User Background
Since one of the areas of concern with the tool was its user friendliness, especially in an infrequently-used environment, the test group included students in an advanced software engineering class. Figure 5 shows the comparison between these two groups for NORM3. The untrained user group consistently scored the tool higher than the database group. Based on written comments, the database users were comparing the tool to the manual way of doing things, and felt that using the tool was not much easier than doing it by hand. The software engineers, on the other hand, were rating the user friendly aspects of the tool on its own merits. This is also indicated in the results of question 8 (System Usefulness), where the database users rated the usefulness of the tool much lower than those who had no experience with the manual alternative.

4.5 Validation of the Survey Questionnaire
The other aspect of data analysis was to see whether the survey itself was valid. All questionnaire results were combined, providing a sample of size 111. Several variables were examined, based on the approach of Bailey and Pearson [1] and of Mallary [11].

4.5.1 User's Rating of Question Importance
The first area examined was the importance assigned to each question by the users themselves. Table 3 ranks orders the questions along with the corresponding average importance.

| #1 | System Feedback | 0.880 |
| #4 | Error Recovery  | 0.872 |
| #8 | System Usefulness| 0.862 |
| #2 | Display of Information | 0.855 |
| #6 | Ease of Learning | 0.838 |
| #5 | Data Input       | 0.815 |
| #7 | Feeling of Control | 0.773 |
| #3 | Program Pacing   | 0.762 |

4.5.2 Predictive Validity
For each question, the calculated average question score was correlated with the user-rated satisfaction score for that question. The results were excellent. The two parameters had a correlation greater than 0.93 for all except question 2 (Display of Information), where the correlation was 0.877.
4.5.3 Content Validity

The choice of appropriate adjective pairs was tested in two ways. First, each individual adjective-pair score was correlated with the average of the other three for that specific question. There were only three correlations less than 0.7: two adjective pairs on question 2 and one adjective-pair on question 6. The second test was a pair-wise correlation between the four adjective-pairs for each question. Correlations less than 0.7 were found for questions 2, 3, 4, and 6. For each question, one adjective-pair poorly correlated with the other three. In every case but one, the correlation was greater than 0.6.

4.5.4 Use of Comments

One final point of validation was whether the comment field was worth including. Of the total 888 comment fields, 70% contained no comment, 21% contained negative comments, and 9% contained positive comments. (The decision as to whether a comment was positive or negative was a judgement call by the researcher.) Thus the comment field was used to some extent by most of the respondents, and was felt by the researchers to contain sufficiently useful information to warrant its continued inclusion.

5 Conclusions

A methodology for evaluating user satisfaction with interactive software was developed that generated a questionnaire based on the goals of a tool developer. Two versions of a specific CASE tool were tested by two groups of users.

Overall we were pleased with the validation of the questionnaire. All of the questions were rated consistently by both groups of users as important. With a few exceptions, the adjective pairs chosen correlated well, and their ratings correlated with the users' general satisfaction scores well. The exceptions underscore the need to carefully select the adjective pairs, and to perform a validation to be sure that questions and adjective pair descriptors are perceived by the users in the same way they were perceived by the designers. It was shown that although most individual questions did not have the comment field filled in, the 30% with comments warranted keeping the comment field on the form. 70% of those comments were negative, which helped clarify just what aspects of each question caused dissatisfaction. Comments also gave some indication of areas where the survey question was unclear or was being interpreted in an unintended fashion.

There is some support in the data that a tool will be rated higher if the user has some experience with another tool first. However, the evidence is not totally conclusive. It is recommended, however, that in a test involving more than one tool that the order of usage be varied randomly across the test group.

6 Plans for Further Studies

We feel that the basic questionnaire format, as defined by Bailey and Pearson, works well. The problem is to determine the proper set of questions to be measured, as well as the proper set of adjective pairs for each question. We are currently evaluating a new version of the questionnaire with changes based on this initial test [12].

The three-pronged approach defined earlier is felt to be a good approach. More work is needed in instrumenting tools, collecting usage and performance data, and correlating it with the user survey data. Data should be collected over a relatively long period of time, during which the user has alternative tools (including manual techniques) that can be used. This will measure voluntary user selection of the tool he or she prefers as well as a more experienced evaluation of the tool on the questionnaire.

Given that user acceptance of computer aided tools is critical to their utilization, some form of tool evaluation is needed to aid tool designers in developing more acceptable tools. We encourage other investigators to report results of such studies in that literature usually available to the software engineering community. Quantitative measurement and analysis techniques are an important part of any engineering discipline.

References


