

Predicting User Satisfaction from Subject Satisfaction

Mark Notess

University Information Technology Services
Cook Music Library
Indiana University
Bloomington, IN 47405 USA
mnotess@indiana.edu

Margaret B. Swan

School of Library and Information Science
1320 E. 10th Street
Indiana University
Bloomington, IN 47405 USA
mbswan@indiana.edu

ABSTRACT

In this paper, we describe work-in-progress in comparing user satisfaction ratings after user tests with ratings obtained following actual use of a digital music library software. We identify some of the variables that hamper prediction, and we reflect on the utility of surveys for predicting user/subject gaps in satisfaction.

Keywords

User testing, satisfaction questionnaire, digital library.

INTRODUCTION

Questionnaires measuring user satisfaction are a frequent accompaniment to user testing. A recent article by Lewis [3] reports on five years' use of the PSSUQ instrument during user tests at IBM. This and other articles focus on the reliability and generalizability of the various instruments. Our own work has a different interest: Can these satisfaction data gathered during user testing predict satisfaction in real use? Measuring satisfaction during lab tests can help identify interface issues or guide selection between alternate designs, but if it cannot predict real-world usage satisfaction, one might well question the value of such measurements and of user testing generally. A secondary research interest is exploring the value of satisfaction questionnaire findings generally.

Our current work compares satisfaction of user test subjects (subject satisfaction) with ratings gathered from people using software for their normal tasks in a normal context (user satisfaction). The software in question is successive versions of digital music library software, Variations [2] and Variations2 [4], used by music students in the Indiana University Simon Music Library.

THE STUDIES

To date, we have completed three studies: a baseline of user satisfaction with the present Variations software, subject satisfaction during a usability test of Variations2, and user satisfaction with Variations2 during pilot usage in a real course assignment.

Variations User Satisfaction

Students in Simon Music Library were recruited as they entered the computer lab. 30 students who were planning to use Variations during their visit filled out QUIS-style [1] surveys after completing their planned work. They were asked to base their responses on the work they accomplished during that session. Most users reported using Variations 1-5 times each week, and half of the users had first used Variations over two years previous.

Variations2 Subject Satisfaction

Ten subjects were recruited from three music classes to participate in a Variations2 user test. Subjects were given two representative tasks to complete during the test, which lasted approximately one hour. None of the subjects had previously used Variations2, but all were frequent users of Variations. After completing the test, subjects filled out a satisfaction questionnaire.

Variations2 User Satisfaction

Students in a graduate vocal literature class were asked to complete a listening assignment using Variations2. Some improvements to Variations2 were made part-way through the study based on lab test results. Students were asked to fill out, upon completing the assignment, an on-line satisfaction survey. Twelve of the approximately 30 students in the class filled out the survey.

RESULTS

Table 1 shows the Likert scale data for the three studies. Two comparisons are of interest. First, we compare the usage data from Variations (V1 Use) and Variations2 (V2 Use). The mean rating is lower for the new software although only one of the item differences, straightforwardness of doing tasks, approaches significance; $t(21.5) = 2.048, p < .05$, two-tailed. The second comparison of interest is the difference between the subject and user satisfaction for Variations2. Test subjects (V2 Test) rated the software lower than did real-world users (V2 Use), but no differences approached significance at the .05 level.

Of primary interest is whether the V2 Test data correlate well with the V2 Use data—can test satisfaction data predict satisfaction in real use? A Pearson correlation revealed a low to moderate positive correlation ($r = .51$) between the two data sets, as shown in Figure 1.

COPYRIGHT IS HELD BY THE AUTHOR/OWNER(S).

CHI 2003, APRIL 5-10, 2003, FT. LAUDERDALE, FLORIDA, USA.

ACM 1-58113-630-7/03/0004.

Table 1 - Mean Item responses

Questionnaire Item (1-7 Likert scale)	V1 Use	V2 Test	V2 Use
1. overall: terrible...wonderful	5.57	5.05	5.50
2. overall: difficult...easy	5.73	4.60	5.17
3. overall: frustrating...satisfying	5.30	4.10	5.17
4. overall: dull...stimulating	5.17	5.30	5.67
5. overall: slow...fast	4.77	4.80	5.08
6. component navigation difficult...easy	5.90	4.30	5.17
7. tasks can be performed in a straightforward manner never...always	5.90	4.70	5.17
8. my location within system never/always apparent	5.86	5.80	5.81
9. characters hard/easy to read	6.13	6.50	5.73
10. organization of information confusing...very clear	5.53	5.20	5.55
11. number of screens/windows confusing...very clear	5.33	5.30	4.50
Mean	5.56	5.05	5.32

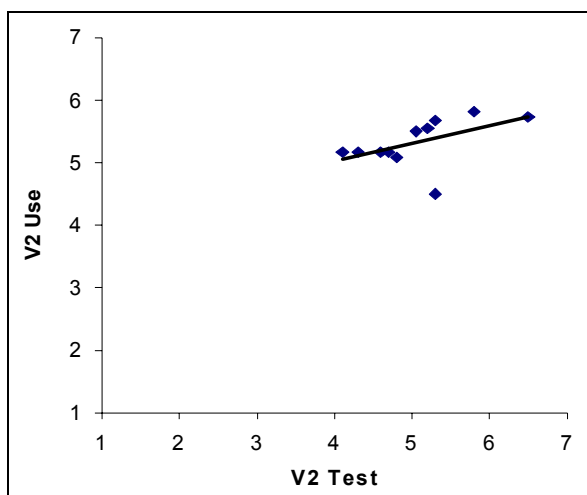


Figure 1 - Scatter plot comparing V2 Test and V2 Use

CONCLUSIONS AND FUTURE WORK

At this preliminary stage in our research, we cannot with confidence suggest that subject satisfaction is a well-understood, reliable predictor of user satisfaction. There are several confounding factors. First, even a strong correlation would not indicate the size of the gap between the two sets of means, merely that individual items tend to co-vary. If there is a standard gap factor (i.e., if subjects tend to rate software lower than users uniformly), only repeated studies can reveal its magnitude. A second issue is that the Variations2 development team made improvements to the software before most of the data was

in for the usage study. A third issue is our small sample size, which limits our ability to determine statistical significance. And finally, the tasks in the user test were not the same as the users' assignment. One of the user test tasks proved particularly difficult and likely skewed the satisfaction ratings. As yet, we do not know whether the somewhat higher ratings by users resulted from the difference in task, the contextual difference (e.g., being a paid test subject vs. a student with an assignment to complete), or other factors.

The difference in user satisfaction with the two versions of the software may result from the familiar being more satisfactory (but less stimulating) or may result from defects in the new software.

The confounding factors above are worth noting because, in real software development settings, most of these factors are likely to be present (e.g., changing software, small sample size, and task differences) and examination of them may be more likely to yield results useful to practitioners.

Our future work will focus on determining the extent to which the subject/user satisfaction gap can be minimized or measured. We hope to conduct further comparisons to understand the contextual differences between being a subject in a usability test and being a real user. What are the key differences? How do they affect satisfaction? If we can answer these questions, we may be in a position to better understand the usefulness of user test findings, especially in a digital library usage context.

ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Grant No. 9909068. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

1. Chin, J.P., Diehl, V.A., & Norman, K. Development of an instrument measuring user satisfaction of the human-computer interface, in *Proceedings of CHI '88* (Washington DC, May 1988), ACM Press, 213-218.
2. Dunn, J.W., and Mayer, C.A. VARIATIONS: A digital music library system at Indiana University, in *DL '99: Proceedings of the Fourth ACM Conference on Digital Libraries* (Berkeley CA, August 1999), ACM Press, 12-19.
3. Lewis, J.R. Psychometric evaluation of the PSSUQ using data from five years of usability studies. *International Journal of Human-Computer Interaction*, 14, 3 & 4 (September 2002), 463-488.
4. Variations2: the Indiana University digital music library project. Available at <http://variations2.indiana.edu/>.