Usefulness of Nonverbal Cues from Participants in Usability Testing Sessions

Karen Long, Lara Styles, Terence Andre, and William Malcom

Department of Behavioral Sciences and Leadership
United States Air Force Academy, CO 80840
c05karen.long@usafa.af.mil; c05lara.styles@usafa.af.mil; terence.andre@usafa.af.mil;
c05william.malcom@usafa.af.mil

Abstract

We investigated the effect nonverbal cues have on the ability of usability experts to detect usability problems within a stimulus set. The pilot study included nine voluntary participants who completed five tasks using the Internet Movie Database website. These sessions were recorded using TechSmith Morae software, where there were two types of recordings produced: one included video of the screen, user audio, and picture-in-picture (PIP) of user video (PIP group) and the other included just video of the screen and user audio (no-PIP group). Usability experts identified usability problems by watching video clips of users performing five tasks with a web-based program. In a formal experimental study, 24 human factors students were trained in usability evaluation and shown one of the video clips from the pilot study. These students identified usability problems based on a standard training presentation they all received. Results for the experimental study showed slightly higher agreement on identifying usability problems for the no-PIP group as compared to the PIP group. Agreement for both groups was rated as “fair” using the kappa statistic, indicating a low amount of agreement in identifying usability problems for both groups.

1 Introduction

The awareness level of usability testing has increased due to the importance of “doing usability” in the design of computer software products (Andre, Hartson, Belz, & McCready, 2001). Each year companies are expending more resources and training more usability testing experts in this focus on the perfect usability relationship between user and computer interface. High-tech usability testing facilities or laboratories are common in most software application organizations. Today usability problem reporting often involves the use of screen capture and audio/video recordings of a subject as a way to identify usability problems and errors.

In recent years, usability testing methods have become more main-stream as a result of the growing complexity of the human-computer interface. Usability is the relationship between the user or operator and the system with which they interact. Despite the growth of testing methods, there continues to be a need for more focused usability inspection methods so that usability problems can be accurately described (Hartson, Andre, Willinges, & Van Rens, 1999).

Usability inspection is quite popular in industry application and is used to evaluate software user interfaces using rules guidelines, or heuristics, rather than acquiring behavioral end-user feedback (Mack & Montainz, 1994). Mack and Nielsen (1994) listed eight different inspection methods: heuristic evaluation, guideline reviews, pluralistic walkthroughs, consistency inspections, standard inspections, cognitive walkthroughs, formal usability inspections, and feature inspections. Though the list of inspection methods are numerous in length, they share much in common. The similarity lies in the use of objectives to identify usability problems and then compare them to the respective objectives.

This study focuses on determining the usefulness of nonverbal cues from video captured during usability evaluation. That is, does video data of nonverbal cues help usability experts more accurately detect usability problems than data sets with audio alone? According to Patterson (1983), nonverbal cues such as observed behavior are more representative of the true characteristics, feelings, and attitudes of a person. Nonverbal behavior is often unconscious and sincere while the verbal output of an individual is more conscious and easily manipulated to sound as the user believes necessary (Patterson, 1983). There are many nonverbal characteristics to consider when observing an individual’s behavior. These characteristics include posture, body position, facial expressions, gestures, hand
movements, and self- and object-manipulations. Previous research has indicated that nonverbal cues can enhance verbal communication which is used through a participant’s introspection of his/her performance on a designated task with the program (Argyle, 1972; Argyle & Dean, 1965; Argyle, Lalljee, & Cook, 1968; Kendon, 1967).

Furthermore, one of the most basic functions of nonverbal cues is providing information that is otherwise nonexistent without the use of video imagery. For these reasons further study on the importance of nonverbal cues for usability problem identification is needed to determine value in usability recording sessions.

Our examination of this issue involved a pilot study and a formal experimental study. Our goal in the pilot study was to learn how experts identified problems on several user tasks given a particular stimulus set of video recordings. In the formal experimental study, we narrowed our focus to one user task and standardized the way in which the evaluators examined the video recordings and documented their results.

2 Pilot Study

2.1 Method

We first conducted a pilot study to determine how usability experts identify problems and what effect the type of video recording stimulus set had on these experts.

2.1.1 Participants

To create a stimulus set of usability problems, we recorded 9 participants (4 female and 5 male) interacting with a commercial website. These participants were selected on a voluntary basis from an introductory psychology course at the United States Air Force Academy. Six usability experts from industry and academia were used to interpret the stimulus sets and identify usability problems in the web-based program. The experts had experience in their specialty areas ranging from two years to thirty years. The mean experience of the experts was 14.5 years. All experts had higher education in the area of system engineering and evaluation; two had a Ph.D. and four had Master degrees. All of the usability experts had experience in the specialty area of testing/evaluation. Some experts also had experience in research, design, management, and consulting. These experts voluntarily analyzed the data sets on their free time and received no compensation for their participation.

2.1.2 Materials

Equipment used to create the usability problem stimulus set included a computer, usability recording software, and a website. The software used for recording was TechSmith Morae version 1.0.1; composed of Morae Manager, Morae Recorder and Morae Remote Viewer. The participants used the Internet Movie Database (www.imdb.com) to complete the tasks.

2.1.3 Design and Procedure

This experiment used the Morae software to create two highlight films of novice users performing tasks on the Internet Movie Database (IMBD). IMBD was chosen as our web-based program due to its multitude of functions and its unknown reputation to the public. Each participant was brought in individually to perform 5 different tasks, as determined by the experimenters, using IMDB. The tasks ranged from easy to very difficult in nature.

The participants were asked to think aloud while completing the tasks. The participants were told to act as if the experimenters were behind a wall and that they could not see what the participants were doing on the computer. The participants needed to tell the experimenters exactly what task they were doing and how they were going to complete the task, articulating every action they performed and every thought they had when interacting with the website.

Each participant’s recording was split into 5 segments matching the 5 tasks that were performed. For each task, anywhere from 2-4 different clips of participants performing that task was used in the final highlight clip. For example, in order to create the first task highlight clip, 3 participant clips were spliced together. The video of these 3 participants provided the best sample and represented the different usability problems found in that particular task. Each highlight clip was subsequently created using the same process.
The highlight video clips were produced in two formats. One format contained video of the screen, user audio, and picture-in-picture (PIP) of user video (PIP group) and the other included just video of the screen and user audio (no-PIP group). Figure 1 shows an example of the PIP stimulus set while Figure 2 shows an example of the no-PIP stimulus set. The two highlight video clips were then sent to 6 different usability experts on compact disks with 3 analyzing the PIP version and 3 analyzing the no-PIP version.

Figure 1: Example of PIP Version of Stimulus Set

Figure 2: Example of No-PIP Version of Stimulus Set

The experts received compact disks comprising of instructions to complete the analysis, a critical incident reporting sheet, and video clips of the tasks. The critical incident reporting sheet gave the experts an operational definition of a critical incident and fields to report the time the incident occurred, the comments or expressions made by the participants in the clip, and a description of the usability incident that the user encountered. A critical incident was
defined as something that happens while a participant is working that has a significant effect, either positive or negative, on task performance or user satisfaction, and thus on usability of the interface. A bad, or negative, critical incident is typically a problem a participant encounters that causes an error, blocks progress in task performance, or results in a pejorative remark by the participant. We instructed the usability experts to report only negative critical incidents experienced by the participants. The experts in turn each sent back the critical incident worksheet identifying each usability error they uncovered along with a description of the error and time when the error occurred.

2.1.4 Data Analysis
We analyzed the amount of usability problems each individual expert found in the videos and the level of agreement for each group of usability experts. To do these analyses, we first went through each individual critical incident reporting sheet and counted how many individual problems they found. We combined some problems that the experts reported as the same incident because they occurred at approximately the same time and had approximately the same description. To determine agreement we started with the expert who found the most problems and listed the problems chronologically. Next, we took each of the other experts in the groups and matched the problems they found in chronological order and matched their incidents reported with the incidents reported by the first expert. This process enabled us to mark every incident that the experts found in common.

To determine the level of agreement among expert evaluators, we used Cohen’s Kappa (Cohen, 1960). Kappa is commonly used to examine observer agreement for categorical lists or taxonomies, especially when chance agreement is a consideration. Kappa is a measure of the proportion of agreement beyond what would be expected on the basis of chance. Kappa is scaled between -1 and +1. Positive values of Kappa correspond to greater than chance agreement, zero represents only chance agreement, and negative values correspond to less than chance agreement.

2.2 Results
We found that the small sample of experts used in our pilot study differed greatly in the level of detail when identifying problems. For example, the mean number of usability problems found by the PIP group was 63.67 problems while the mean number of identified problems for the no-PIP group was 18 problems. Because of the small sample size, one expert greatly influenced the mean number of total problems reported. Because of the large individual difference in the number of problems, the level of agreement among the experts was very low. Using the Kappa statistic, the level of agreement for the PIP group was -0.08 \((p > .10)\) while the agreement for the no-PIP group was -0.12 \((p > .10)\). These results clearly show that expert agreement was less than what would be expected by chance. Such large differences led us to conduct a more controlled study to further understand the differences between watching PIP versus no-PIP video of user interaction.

3 Formal Experimental Study

3.1 Method
Based on our findings in the pilot study, we conducted a formal experimental study to focus in on any differences between the PIP and the no-PIP video stimulus sets. The formal experimental study focused on just one task and involved more control over how the usability stimulus sets were presented.

3.1.1 Participants
Twenty-four students in an advanced human factors design course at the Air Force Academy were trained in usability evaluation. They were given the same operational definition of a critical incident and a portion of the video clip that the experts utilized in the pilot study. Each participant was considered a novice at usability evaluation with all having the same level of experience and knowledge. Half of the participants received video clips containing PIP and half received video clips with no-PIP.

3.1.2 Materials
The equipment used in the experiment was the same as in the pilot study. The only difference in the formal experiment was that the video the individuals used to identify the critical incidents in the website was shorter in
length and only showed one of the original tasks. The video clip they received provided a representation of common usability problems made in a single task rather than all five of the tasks.

3.1.3 Design and Procedure
The excerpted highlight video clip was produced in two formats similar to the pilot study; one format as the PIP group and one as the no-PIP group. The two highlight video clips were then shown to the 24 human factors students in a classroom setting. Twelve students analyzed the PIP version and twelve analyzed the no-PIP version.

The students received the same instructions as the experts received along with the same critical incident reporting sheet and the same definition of a critical incident. They also received a question to answer at the end of their reporting sheet. The question asked the PIP group whether or not they thought having the PIP enhanced their ability to detect critical incidents and asked the no-PIP group if they thought having PIP would have helped them more when identifying critical incidents. Each group watched the video clip in a controlled group setting where a facilitator started and stopped the clip. Each group watched the video clip followed by five minutes to record usability problems they observed. This was repeated two more times so that each group watched their respective video clip a total of three times. After viewing the video clip a third time, the participants had 5 minutes to record usability problems which ended the session.

3.1.4 Data Analysis
We analyzed the number of usability problems each student found in the videos and the level of agreement for each group of students in the same manner as the pilot study. We combined similar critical incidents when they occurred at approximately the same time and had similar descriptions. The incidents were then listed in chronological order and the usability problems were matched between the individual students.

3.2 Results
Figure 3 shows the mean number of usability problems found by the students in the PIP group \((M = 3.75)\) versus the no-PIP group \((M = 3.67)\). This difference was tested using an independent groups t test, and was shown to be nonsignificant, \(t(22) = .192, p = .850\). Figure 3 clearly shows that the two groups found approximately equal number of problems on average. More variability appears to exist in the no-PIP group as noted by the higher standard deviation \((SD = 1.23)\) as compared to the PIP group \((SD = 0.87)\). The total number of unique problems found in the PIP group was 7 compared to 11 in the no-PIP group.

We also looked at the level of agreement among student evaluators using Cohen’s Kappa (Cohen, 1960). Using the Kappa statistic, the level of agreement for the PIP group was 0.25 \((p < .001)\) while the agreement for the no-PIP group was 0.35 \((p < .001)\). These results show that there is a “fair” amount of consistency in determining errors that would be expected by chance. The “fair” rating is based on recommended values by Landis and Koch (1977) where the level of agreement is between .2 and .4.

Figure 4 shows the results from the question asked to each group at the end of the experiment. The question asked the PIP group if they agreed that having a picture-in-picture video of the user helped them identify usability problems. A similar question was given to the no-PIP group, asking them if it would have been helpful to have picture-in-picture video of the user included in the video clip. The students answered their respective question using a 5-point Likert scale, with 1 being “Strongly Disagree” and 5 being “Strongly Agree.” Students who received the PIP recordings reported moderate agreement \((M = 3.75)\) in terms of the user video helping them identify usability problems. The no-PIP reported slightly stronger agreement \((M = 4.08)\) that having user video available to them would have been useful in identifying usability problems.
Figure 3: Mean # of Problems Found for each Group

Figure 4: Rating of Whether Picture-in-Picture Video of User is Perceived to Help or Not
4 Discussion and Conclusions

We discovered a large amount of variability in the data analyzed in the pilot study. Agreement was extremely low among the raters. This is possibly due to the various levels of training the experts had in usability evaluation, and the amount of time and effort they put into their analysis. The pilot study led us to do a focused experimental study to determine the differences in PIP versus no-PIP video of the user. We were able to control for the amount of time each student had in finding usability problems as well as the process for recording these problems. All students in the experimental study also had the same level of experience and training. Because we worked with a focused stimulus set (i.e., one task instead of all five), we were able to more easily test the level of agreement. The level of agreement was “fair” and is limited to this one task with the Internet Movie Database. Other experiments are needed to evaluate other tasks and user interface applications.

Although the mean number of problems reported for each group in the experiment was approximately equal, the no-PIP had higher variability as indicated by a higher standard deviation and a higher total number of problems reported. This may suggest that those in the PIP group concentrated more carefully than the no-PIP group because they must attend to three things (i.e., the PIP video of the user, video of the screen, and user audio). The no-PIP group only has to divide their attention between two things (i.e., the video of the screen and user audio), possibly freeing them to find more usability problems. We did not evaluate the quality or severity of problems identified by each of the groups. Identifying more problems may not necessarily mean that more “real” problems are found.

This study indicates that including user video can have an impact on the number of usability problems identified. Additional factors may have had an impact, such as how much the user “talks” about their interaction experience and the experience of the evaluator. This study did not show agreement level to be a conclusive metric for determining the benefits of including video of the user’s nonverbal interaction cues. There appears to be moderate agreement among evaluators that PIP is perceived as beneficial to identifying usability problems. Future research should develop a known stimulus set of “real” usability problems based on data from the field, and then use this set to compare the number and types of problems that are found with PIP or no-PIP video clips.

References


**Acknowledgements**

The work reported here was supported by research funding from the Institute for Information Technology Applications (IITA) at the United States Air Force Academy and the Air Force Research Laboratory in Mesa, Arizona.