HANDLING AWKWARD USABILITY TESTING SITUATIONS

Mona Patel and Beth Loring
American Institutes for Research
Concord, MA 01742 USA
mpatel@air.org; bloring@air.org

You are half way through a usability test session, and the participant s cell phone rings. She answers it and says it s her child calling and she has to leave immediately, but she ll be back in 45 minutes to finish the test. What should you do? In the course of running hundreds of product evaluations, we have encountered some awkward and unusual situations. These situations involve issues with participants, with the software or product being tested, and with the test protocol. Among the factors that we consider when resolving these issues include the participant s rights and psychological well-being, the cost to replace the participant s test session, the importance of adhering strictly to the recruiting criteria, and how eliminating the test session might affect the integrity of the protocol or data. In this paper, we describe some real-life sticky testing situations and discuss the factors we considered in resolving each issue.

INTRODUCTION

In the course of conducting hundreds of product evaluations each year at the American Institutes for Research, we have encountered a number of awkward and unusual testing situations which have require quick decision-making to resolve. These situations have involved issues with participants, with the software or product being tested, and with the test protocol. Although no two situations are exactly alike, we have identified a common set of factors to consider each time we seek to resolve these issues.

In this paper, we describe three real-life awkward testing situations, explain the list of factors we have developed, and discuss which factors we considered when addressing the three situations.

THREE SAMPLE CASES

Case #1

You are a consultant working with the developer of a software product to conduct a usability test. To reduce costs, the developers have decided to create the recruiting screener and schedule the test participants themselves. You have just started your first test session when you discover that the participant does not have the required profile; her job category is different from that which is representative of most users of the product. However, the participant does seem to be able to complete the first task successfully. She also seems good at thinking aloud and expressing her opinions. Should you continue testing or stop the session?

Case #2

You are conducting a the first round of usability tests of an e-commerce web site. The company developing the site has scheduled their launch date to occur two days after the usability test, so they are working rapidly to update the code even as you are conducting the test. While running the seventh participant (out of eight,) you discover that the developers have incorporated some changes based on the first six test sessions. Should you continue testing with the updated site or revert back to the previous version?

Case #3

You are testing the user interface of a television-related product, and the current participant is a 10-year-old boy. His parents have left the building and will return in an hour to pick him up. Suddenly, a fire alarm goes off and everyone must evacuate the building. It turns out to be a false alarm, but the child appears somewhat unnerved. However, he says he finds the product interesting and wants to continue with the evaluation so that he can receive his honorarium. Do you continue testing even though he appears rattled?

FACTORS TO CONSIDER

Although no two situations are alike, and problems vary in severity, over time we have developed the following list of factors to consider. With the exception of the participant s rights, these factors are not in order of priority because each situation is different.

Participant rights and psychological well-being. Our first priority is to ensure that people feel comfortable during the usability test session and we protect their rights as participants. As test administrators we need to be aware of a participant s emotional state and terminate the test session if the participant is at all uncomfortable. AIR has instituted an Internal Review Board (IRB) to review our proposals and procedures, ensuring that we meet the ethical standards for human behavioral research and that we do not expose participants to risk or harm.

As an example, participants should be able to leave without any pressure and without feeling they have failed (NIH, OnlineEthics.org, 2001). In fact, forcing, or even urging people to continue with a test when they don not want to or can not proceed is a violation of their rights as participants in behavioral research. (Dumas & Redish, 1994.)

When unexpected situations arise involving a participant s rights or well-being, we turn to the IRB s guidelines for help. We also use our best judgement and take into account the participant s wishes when deciding whether to continue or terminate a test session.

Extra costs when replacing participants. When there is an issue involving the suitability of a participant, we typically consider the extra costs (in terms of time and money) associated with replacing him or her. Not only does the replacement participant require an added honorarium, but there are additional labor costs associated with recruiting the new participant, running the extra test session and incorporating the participant s data into the findings.

To avoid delaying the project in the event that a participant is unsuitable, we recruit additional back-up or floater participants whom we can call at the last minute to make up for any no-shows, cancellations, or situations where the participant is unsuitable. Because we pay them only if we need to call them in, we incur minimal cost and disruption to the test schedule.

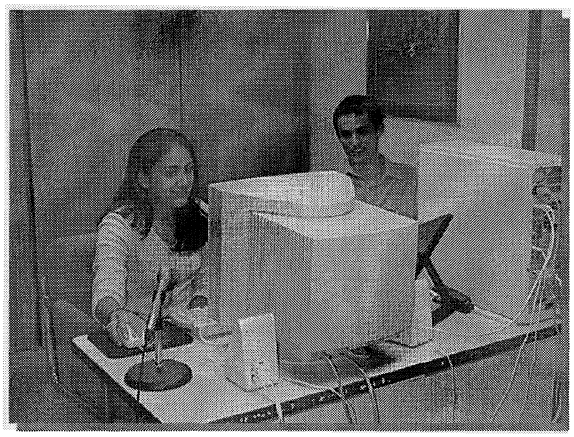

Photo of a usability test session

Size of user group. Deciding to cancel a test session can affect the test results in many ways. Before designing the study, we work with the client to determine what type of test is necessary, then how much flexibility we have in recruiting and running participants. In comparison tests or tests involving inferential statistics, for example, we must ensure that the sample size is adequate to be confident in our findings. If we cancel a test session, we must replace the participant with

another to maintain a set sample size. In qualitative tests, on the other hand, usability issues become apparent after six to eight test sessions per user group (Dumas & Redish, 1994). If we have scheduled eight participants and one is not suitable for whatever reason, we may not need to replace him or her based on how consistent the findings are.

Strictness of user profiles. One of the cardinal rules of usability testing is that the people who work with the product in the usability test must be like the people who will actually use the product. (Dumas & Redish, 1993). In cases where a participant does not strictly meet the recruiting criteria, we judge how important the criteria are, and whether that criteria is critical in deciding how users actually interact with a product. In some cases, participants must fit a narrowly-defined user profile, while in other cases incorporating slightly different points of view may add valuable data.

Sometimes when we encounter a participant who does not strictly fit the demographic requirements and the test includes a relatively small number of participants, we may relax the criteria since we cannot draw conclusions about the general population from one person anyway.

Consistency of data. Sometimes clients alter the product while we are conducting a usability test. For example, if we uncover a major usability problem with the first few participants, the client may not think it is valuable to watch the rest of the participants struggle with the same problem and decide to try a different design. Changing the product mid-way can lead to discrepancies in data because participants will react to different versions of a product. On the other hand, testing a refined version of a product may produce valuable feedback on an alternative design option.

Participant s value. If a participant leaves the session early, arrives late, or takes an unexpectedly long break, we must determine how much data we have gathered (and will be able to gather) from that participant during the shortened test session. We must also decide whether or not to use this participant s data when tabulating and analyzing the findings. If we will be calculating statistical significance, then we typically do not use the data and instead, call one of the back-up participants.

Goals of the usability test. If the developers are testing a product for the first time and are looking for a preliminary assessment, then we have the flexibility to modify the test script as testing progresses. We can also probe more about certain details of the product to gain new insights, if necessary. On the other hand, if the purpose of the study is to benchmark a product or service against previous versions, its competitors, or established heuristics, then we follow the test protocol stringently.

Conflicting Factors

In some cases multiple factors come into play, leading to conflicting solutions. For example, in a case where some scheduled participants do not fit the user profile, it may be best to replace them, but the client strict budget limitations may prevent you from adding participants.

The decision about what to do will depend on the weight of each factor, which in turn depends on the situation. When such conflicts arise, we discuss with our clients the potential benefits and drawbacks of various solutions and work together to determine the best way to proceed.

POSSIBLE SOLUTIONS TO THE THREE SAMPLE CASES

Below we describe how we handled the three sample cases mentioned earlier. Note that these are possible solutions and there may be alternatives to resolving these scenarios.

Case #1

In the case where the participant s demographic background did not fit target user profile, we discussed how her background might impact how she used and perceived the product. We chose to

continue the session, and conducted a debriefing with the client afterward to determine whether or not to use her data.

In this case, the participant s experience level was lower than we had wanted, but she was able to complete many of the tasks. We acknowledged that although the participant was not in the target population, we benefited from her insights. Many times in a usability test with six to eight participants, the criteria can be relaxed somewhat to accommodate a slightly different user profile.

In Case #1 we also considered the costs we would incur and the time we would lose if we terminated this session prematurely and recruited a new participant. As we did not participate in the recruiting process because the client had recruited the participants, we were unsure whether the participant misrepresented herself (either intentionally or unintentionally) or whether the recruiting screener was flawed. Although some companies do not reimburse participants if they intentionally misrepresent themselves, our IRB requires that we reimburse participants in every case. Therefore, we would paid an additional incentive.

Factors considered: strictness of user profiles, consistency of data, participant s value, extra costs when replacing participants

Case #2

In Case #2 where the developers changed the web site partway through the test, we considered the fact that the client was in the early design phase as opposed to the validation phase. In this case, the clients were experimenting with various designs to find the best way to layout their interface, so our solution was to pause testing, explain to the client the ramifications of continuing to test with a different design. They decided to continue testing the revised site, and we discussed the particular aspects of the new design they wanted to investigate and added additional questions to our test script.

When analyzing the findings and writing the test report, we documented the changes to the web site and took into account how the changes in the site affected the data we collected from participants, such as averages and comments.

The benefit of running the test with a different interface was to gain feedback about whether that particular design solution helped users complete their tasks. Our solution provided the designers with instant feedback about their new design, just in time for their launch date.

Factors considered: consistency of data,, goals of the usability test

Case #3

In this case we were concerned about the boy s psychological well being after the fire alarm. Making participants comfortable during the test and being acutely aware of their rights is a major part of running an effective usability test session, especially when testing children. In this situation, we made sure that the participant understood that he could leave the test at any time without losing his honorarium.

Because participant was interrupted during testing and visibly upset, we took a break from the session. We told the participant that he did a great job helping us evaluate the product and that his comments would be extremely useful in helping improve it. We took him out of the testing room and offered him a snack. After fifteen minutes, he was more comfortable and ready to continue. We explained again that he had the right to stop the testing at anytime, but he chose to continue evaluating the product.

If we had cancelled the session, we would have called in one of our back up participants.

Factors considered: participant rights and psychological well-being, extra expenses, consistency of data

CONCLUSION

When conducting usability tests, it is likely that testers will occasionally encounter awkward

situations. In this paper, we have described some real-life examples and a list of the factors to consider when handling these situations. Every case is different and requires quick and smart thinking. We hope the information presented here will help other practitioners when they encounter similar situations.

REFERENCES

Dumas, J.S., & Redish, J.C. (1993). A Practical Guide to Usability Testing. New Jersey: Ablex Pubishing Corporation.

Nielsen, Jakob. (2001). Homepage. *Usable Information Technology*. http://www.useit.com

OnlineEthics.org. (2001). The Online Ethics for Engineering and Science. http://www.onlineethics.org

National Institutes of Health (2001) http://www.nih.gov.

Schrier, J.R. (1992) Reducing Stress Associated with Participating in a Usability Test. Proceedings of Human Factors Society 36th Annual Meeting, pp. 1210-1214.