JUS
JOURNAL OF USABILITY STUDIES

Vol. 4, Issue 2, February 2009, pp. 85-92

Usability Testing with *Real Data*

Alex Genov
Staff Experience Design Researcher
Intuit, Inc.
Alex_Genov@Intuit.com

Mark Keavney
Staff Experience Designer
Intuit, Inc.
Mark_Keavney@Intuit.com

Todd Zazelenchuk
Staff Experience Design Researcher
Intuit, Inc.
Todd_Zazelenchuk@Intuit.com

Abstract

Usability practitioners run the risk of misreading the results of usability evaluations, either identifying false positives when artificial user data interferes with a user's product experience or overlooking real problems when they use artificial user data. In this paper, we examine a strategy for incorporating users' real data in usability evaluations. We consider the value and the challenges of this strategy based on the experiences of product teams in a consumer software company.

Keywords

Usability method, teaching usability, real data, scenarios, tasks, privacy, security

Introduction

Usability testing is generally well equipped to help design teams identify issues with a wide range of products and systems, frequently identifying interaction problems that were initially overlooked in even the simplest of proposed design solutions. In certain instances, however, usability testing falls short of producing results that are both reliable and valid. Of particular concern are studies of products in which users heavily interact with their own personal data, for example a user test of an email or calendaring program. In these cases, the fact that data are often fabricated for the test (and are therefore unfamiliar and possibly unrealistic for any given user) can affect users' abilities to recognize, interpret, and interact with these data in an authentic manner. Thus, usability practitioners run the risk of identifying false positives (for example, if users are completely confused by a screen because they're unfamiliar with the data in it, when in the real world they would have been able to use their own familiar data to orient themselves) or overlooking real problems (for example, if users answer questions casually about their fake data, but would be much more concerned with the meaning of items when it applies to themselves). In this paper, we examine the strategy of incorporating users' *real data* into usability testing to avoid these issues and increase the validity of a study.

Real Tasks, Real Users … But What About Real Data?

In the field of user experience design and research, we routinely speak of the importance of having *real users* perform *real tasks* in order to effectively evaluate the usability of our products. The following are just a few examples:

- Nielsen (1993, p. 185): "The basic rule for test tasks is that they should be chosen to be as representative as possible of the uses to which the system will eventually be put in the field."
- Rubin (1994, p. 179): "Provide realistic scenarios…the participants will find it easier to 'stay in role'…if the scenarios reflect familiar situations, with realistic reasons for performing the tasks."
- Dumas & Redish (1999, p. 174): "The participants should feel as if the scenario matches what they would have to do and what they would know when they are doing that task in their actual jobs."

What is rarely discussed, however, are the pieces of information or data that participants actually encounter when they interact with our prototypes. In most cases, these data are fictitious, made to resemble the average users' data in the hopes of having participants successfully immerse themselves in the given scenarios. However, like the proverbial man with his head in the freezer and feet in the fire (but who's just fine, on average), having all the users interact with an average set of data does not mean that those data will be appropriate for any of them. A usability task with a 10-item list of stock sales is not a good test for anyone if nine in ten users typically have one stock sale in the list and one in ten has 100. And even for those users for whom the fake data is typical, the fake data will at the very least be unfamiliar, and will require the participant to assume the role of someone else in order to understand and relate to their experience during the research study.

For many products and systems, this may not be a huge problem. For example, participants involved in testing an online retail or travel Web site are likely to be familiar with the scenarios and able to adapt to the data presented. So a task to purchase a particular book or arrange a trip to Miami may be a good test of the interface even if a user has never bought that book before or if the user has never flown to Miami. Data are not critical in these cases, and the user is likely to be able to relate the test scenario to their own life experience, as long as they have some familiarity with these types of tasks.

But for some products, the participants' personal data represent an integral part of the product. For example, imagine testing an interface for setting photo sharing permissions or a system for online bill pay. In such cases, the particular photos that a user wants to share, and friends with whom they want to share them, or the number and timing of the bills to be paid will have a huge effect on how the user interacts with the interface. In these systems, testing with fake

data means testing a fake user experience, one that's potentially quite different from what users will have in the real world.

The way to get around this problem is to use the usability participant's real data in the study, rather than data artificially created to simulate the real thing. In other words, rather than making up reasonable approximations of users' banking transactions, annual income statements, or medical claim details, a researcher can find ways to incorporate users' actual information into the design being tested.

Numerous examples of research in psychology support the wisdom of this strategy. The first involves the "self-referencing effect" (Greenwald & Banaji, 1989) which suggests that information relevant to the self can be more easily encoded in memory and easily retrieved at a later date. Based on the self-referencing effect, we may hypothesize that by using artificial data rather than users' real data in a usability test we may impede a user's ability to perform as efficiently as he or she normally might.

A second line of research with similar implications for real data usability evaluations is related to the familiarity or novelty of content. This research has found that reactions to familiar vs. novel content are correlated with changes in different regions of the brain and may reflect different memory processes (Tulving, Markowitsch, Craik, Habib, & Houle, 1996), again potentially having a negative impact on a user's efficiency and effectiveness performance when completing desired tasks.

Finally, research has shown that a connection exists between personally relevant information and one's motivation to think about and elaborate on the topic at hand (Thomsen, Borgida, & Levine, 1995). In the context of usability studies, users' real data may motivate users to engage with a product in a more authentic, genuine manner than with artificial data. In the experience of Intuit product teams, this increased level of engagement has produced more accurate and richer findings than otherwise collected with fictitious user data.

The Benefits of Real Data Testing

While incorporating users' real data into your usability studies requires significant effort, there are some clear benefits that help make it worthwhile.

Ecological Validity

The primary benefit of incorporating users' real data in a usability study is increased ecological validity (Brewer, 2000)—that is, better approximating the real-life situation under study. And, by doing so, real data can also increase the study's external validity—in other words, the study results are more likely to generalize beyond the lab.

For example, consider a real data study done on a product called Intuit FinanceWorks. The product was used within a bank's Web site to allow the bank's customers to do certain personal financial management tasks that are more commonly done within a desktop software such as Quicken (for example, users could enter a check that had not yet cleared to account for that money in their balance). The team had run several fake data usability tests in which participants had no trouble with this concept, but in the real data test many were reluctant to even attempt the task. In this more realistic situation, participants assumed that there would be no way to do the task on their bank's Web site, that they "wouldn't do that here." This finding led the team to redesign to increase discoverability and to better educate first-time users about what the product could do. If the team hadn't tested with real data, this problem would have gone undiscovered until after launch.

Another example is in the user testing of Intuit's TurboTax tax preparation software. TurboTax has many screens that lead taxpayers through step-by-step explanations of how extremely complex and unfamiliar tax concepts may or may not apply to them. For example, based on the amount of income input by the user, the TurboTax software determines if the user is eligible or not for specific tax credits. Based on that determination, the program shows users a certain set of screens that is relevant to that specific tax situation. If, during a usability study, the participant is using made-up data including a specific level of income, the participant will see a set of screens that go with that level of income. These screens may include situations triggered by the pre-determined income level that are not familiar to the participant because that was not his or her actual income level. Such a scenario potentially introduces another source of

variability to the usability study, namely the difference between the usability of the software and the usability of the study task and non-real data. In the past, when TurboTax did non-real data usability testing (which has its own advantages in terms of standardizing measurements of some program-related tasks such as navigation), teams had difficulty testing how well users understood these kinds of questions in reference to their actual situations. When they tested with real data, they found that many of these questions were not clear to some users, and so the team rephrased and added explanations to make them easier for customers to understand.

Data Issues

The second benefit of real data testing is that by testing with a range of real-life data, we can uncover usability issues that would not be produced by a narrower set of representative fake data. For example, at Intuit a team had designed an interface for exporting data from an online payroll system to QuickBooks desktop accounting software. The export software was designed to cancel the export and display an error in certain (we thought) very rare exception cases, such as when an employee name in the online payroll system exactly matched something that was not an employee in QuickBooks, such as an item in the vendor list. When we usability tested this system with real data, we found that many users had added their employees to the *wrong* list in QuickBooks, and so the export frequently failed. When we discovered this, we made the matching rules more forgiving, allowing the export to go through in these cases. Without real data testing, we would not have discovered this issue until after launch.

In another very simple example, a team was designing a purchase process for a Web site. Instead of having participants enter artificial user data into a credit card field, the team asked users to enter in the actual name and expiration date from their company's credit card. In one instance, a participant entered in her very long company name (35 characters) that exceeded the field limitations of the design being tested. The most interesting thing was that this aspect of the design was already in production and was not even a priority of the test. By incorporating users' real data, the team discovered an unknown and previously unencountered design flaw that was easily and immediately fixed.

Reduced Cognitive Load

Another advantage of real data testing is that it doesn't require participants to bear the artificial cognitive load of remembering a fake scenario or recognizing data invented for the purposes of the usability study.

The TurboTax team noticed this benefit when they conducted real data testing. Previously, many test participants had asked clarifying questions about the artificial scenarios or data to the extent that the team was beginning to wonder if the test results were reflecting real usability problems or were just a result of the participants' confusion with the artificial data. When participants used their own data, they were completely immersed in preparing their taxes and did not ask any clarification questions about the scenarios.

The reduced cognitive load (and perhaps also the greater realism) led to an increased level of cognitive engagement in the interface than in previous studies. For example, when faced with a question about their income on one of the screens, participants made sure they understood the question and answered it correctly, fully aware that a wrong answer by them would lead to inaccuracies in their tax return. In contrast, in studies where participants used fictional data, they were more prone to make up answers and numbers and less motivated to be accurate.

So, rather than spending their time and mental energy focused on remembering artificial data, participants in a real data study are more focused on what they're doing, which is presumably their focus in the real world.

Greater Emotional Engagement

The final benefit to real data testing is that it makes participants more emotionally engaged in the situation. This often results in a better understanding of the product's strengths and weaknesses, and richer feedback to the design team. For example, when doing their own taxes, participants were very focused on exactly how much money they were going to get back or have to pay. TurboTax has a refund counter in the upper left of every page that shows the user's current refund or tax owed. Based on this feedback, the team decided to enhance the refund counter and make it a larger part of the interface.

The Quicken Health team, in their real data study of their product, found the same deep emotional engagement when participants encountered highly sensitive areas such as insurance claim denials and adjustments. The negative impact associated with these topics (e.g., unexpected fees) was heightened for participants because it was their own medical service that was being denied. This feedback was given to the design team, who considered the implications of this finding for the visual design and language around notification of claim denial.

The Challenges of Real Data Testing

Incorporating users' real data into a usability study also poses a number of challenges. Based on the authors' experiences across multiple product teams at Intuit, there are three primary challenges to incorporating users' real data into a usability study.

- Recruiting participants and getting their data
- Addressing security and privacy issues associated with users' real data
- Analyzing the results of the study

The remainder of this paper examines each challenge in turn, drawing on sample projects across product teams.

Recruiting Participants and Getting Their Data

In order to conduct a real data study, one must first determine how to obtain participants' real data. Depending on the nature of the product, this can be as straightforward as asking recruited participants to bring their data with them to the study to as complex as partnering with a third party to extract user data files from multiple backend systems prior to the study, followed by mapping and uploading those data into the prototype to be tested. The following are some methods that the authors have used for obtaining real data and the situations where researchers might use them:

- Have participants login to their own account (for online applications already in use) during the study.
- Have the participants bring their personal data with them to the session in paper or electronic format (for applications centered around data entry, like TurboTax).
- Have users send in their data in advance of the study so that the researchers can put it into the prototype (for applications with an easily transferable data file, like QuickBooks, or set of paper data, like Quicken Health).
- Work with a partner organization to collect the data on the backend, then map and upload these data into the prototype (for hosted applications in which the developer of the application is not the customer owner, e.g., Intuit FinanceWorks).

These methods vary widely in difficulty, but each has at least some issues. Having participants login to their own account or bring their own data to the study are logistically quite simple, but involve some recruiting and attrition challenges. People may be reluctant to share their details with others, especially when the product being tested involves sensitive information such as personal financial or medical information. Thus, recruiting becomes more difficult as recruiters have to approach a larger number of people before they acquire a sufficient pool of participants, and the recruitment costs for the study will typically increase.

Depending on the application, people may also have trouble finding or collecting the data to bring with them. In the case of studies done with Intuit TurboTax, this was not as much of a problem as it might have been, because even though the program requires user data from multiple sources, the users were already in the habit of pulling their tax data together to either complete their own taxes or to share them with an accountant. But other real data studies, such as some done for Intuit's payroll software, could not be done in the lab simply because too many participants failed to bring everything they needed and instead had to be done in participants' offices.

Recruiting participants to send their personal data to the research team in advance of the study can prove to be somewhat more challenging than having participants bring their personal data to each session. This approach takes participants' privacy and security concerns to a new level. With the growing fears of identity theft and phishing scams, people are increasingly wary of

sending their information to someone they do not personally know. As a result, research teams need to take extra care and precautions to ensure the security of data transmission and handling (see the Security and Privacy section) as well as reassure participants that their information is secure. Another challenge of this method is incorporating the data into the usability study prototype. In a real data study evaluating Quicken Health, participants were asked to mail in their data in advance of the study. This approach required extra time and significant manual entry into the prototype to be tested. Additional time was needed to allow for engagement from the Quality Assurance (QA) team who helped with ensuring the quality of the data entry prior to testing.

In the case of obtaining participant data ahead of the study via a third party partner, the logistical issues grow exponentially. Extra steps required by this strategy include convincing the third party of the cost and benefits associated with using real data, informing potential participants of the real data process and getting their permission to obtain their information, setting security and privacy procedures with the partnering group, and so on. In the case of the FinanceWorks project at Intuit, the application was intended to operate from within a financial institution's Web site, so the most useful method for obtaining customer's real data was through the customer's bank. Finding a bank willing to participate in this test, and then overcoming the legal and organizational hurdles to obtain permission both within Intuit and within the bank to run this study, took over two months.

However, the richness of using the customer's bank data in the tested prototype made the time spent well worthwhile (see the Benefits sections) and the success of the study established procedures and a relationship that opened the door for more real data testing, both with that financial institution and others.

Security and Privacy

With fictitious data, the concerns about security and privacy of information are usually quite minimal and generally related to permissions for videotaping and subsequent sharing of results. In contrast, when you incorporate users' real data, you may quickly find yourself dealing with institutional security policies and highly sensitive privacy issues that need to be considered with care.

In all the real data studies done at Intuit, extra efforts regarding security and privacy are taken to ensure that participants' data are not at risk. For the FinanceWorks and Quicken Health studies, these efforts included file encryption to ensure that users' data was protected at all times and the destruction of all data files within a four week period following the study. Explicitly communicating to participants that their personal data was going to be used in the study was highly effective in setting expectations with participants and ensuring them that the security and privacy of their data was a priority for the research team.

Efforts also have to be put in to place to handle any recordings of a real data study. In a real data study on the Intuit FinanceWorks product, the sessions were recorded using Morae. These recordings, in addition to the participants' banking data, had to be encrypted. The researchers kept records of anyone who had access to or viewed these recordings. In a few cases, the researchers asked for additional consent from the participants after the test was completed to show video clips to a wider audience. Only those video clips for which additional consent was obtained could be shown.

Scenarios, Tasks, and Data Analysis

With fake data, it's possible to define the successful path for each task as the same across all participants and to analyze the data accordingly. With real data, things aren't simple. Depending on the participant, the same task may involve different amounts and types of data; in some cases, the task may not apply at all or may need to be done multiple times. There is no simple solution to this problem. The best way to analyze the data will depend on your particular application and the questions you're trying to answer. However, there are two techniques that we've used in multiple real data studies that have been helpful.

One is to construct *task templates*, generic versions of a task that are then filled-in with the specific data for each participant. For example, a task template for a study on an electronic billpay application might be, "Pay your _____ bill that is due on _____." The blanks would then be filled in differently for each participant before their test session, based on the data that we

had received, and with an attempt to make the data in the blanks relatively similar across participants (e.g., selecting bills of similar amounts and due dates as much as possible). Thus there is some standardization of tasks, but each participant is actually paying a bill that is specific and familiar to them.

Another technique that is useful specifically for analyzing the data is to construct average scores for each participant in a given task before calculating success metrics across participants. So for example, if a task is to pay all bills that are due in the next week, in a real data study one participant might have five bills due and another might have only one bill. To avoid overweighting the participants with multiple bills, the best way to create an overall score for task success at paying a bill is usually to construct a composite score for each participant, and then average those composite scores.

Conclusion

In this paper, we have argued that using real data in usability testing is one way to reduce the artificial nature of lab usability testing and to increase the validity of the study results. Accompanying the benefits, however, the use of participants' real data introduces several methodological and logistic challenges. Our discussion of these challenges centered around recruiting participants and obtaining their real data, analyzing the data, and addressing security and privacy issues associated with users' real data.

Practitioner's Take away

- Whenever possible, obtain your participants' real data by having them bring their data with them to the test and enter it themselves.
- If it's not possible to have participants either bring or send in their own data in advance, make sure to allow sufficient time to develop and test the technical and logistical process of getting the participants' data into your test prototype.
- Involve your company's privacy and legal stakeholders in decisions of how to handle the participants' data. Consider encrypting data, obtaining additional consent, and restricting access to videos as possible ways to safeguard the privacy of participant data.
- Prepare a tailored script for each participant by creating task templates ahead of time and then filling them in with each participant's data as they are received.
- When analyzing your results, consider constructing average scores for each participant on a per task basis before calculating success metrics across participants. This will help to reveal overall performances when not all participants performed all tasks.

Acknowledgements

The authors would like to thank Kari Sortland, Sara Cole, and Matt Berent for contributing to the arguments and examples in this paper.

References

Brewer, M. (2000). Research Design and Issues of Validity. In Reis, H. and Judd, C. (Eds.) *Handbook of Research Methods in Social and Personality Psychology*. Cambridge: Cambridge University Press.

Dumas J.S. & Redish J.C. (1999). *A Practical Guide to Usability Testing*, Intellect Ltd., p. 174.

Greenwald A.G. & Banaji M.R. (1989). The self as a memory system: Powerful, but ordinary, *Journal of Personality and Social Psychology*, *57*, 41-54.

Microsoft HealthVault (http://healthvault.com/)

Nielsen, J. (1993). *Usability Engineering*. Academic Press, p. 185.

Quicken Health (http://quickenhealth.intuit.com/)

Revolution Health (http://www.revolutionhealth.com/)

Rubin, J. (1994). *Handbook of Usability Testing: How to Plan, Design, and Conduct Effective Tests.* John Wiley and Sons, Inc., p. 179.

Thomsen C.J., Borgida E., & Levine J. (1995). The causes and consequences of personal involvement. In Petty and Krosnick (Eds), *Attitude Strength: Antecedents and consequence*, (pp. 191-214) Mahwah, NJ: Lawrence Erlbaum Associates.

Tulving E., Markowitsch H.J., Craik F.I.M, Habib R., & Houle S. (1996). Novelty and familiarity activations in PET studies of memory encoding and retrieval. *Cerebral Cortex*, 6, 71-79.

About the Authors

Alex Genov
Alex is responsible for customer research and usability of TurboTax's products and services. He received his PhD in Experimental Social Psychology from Clark University. Areas of research: emotions, individual differences, non-verbal emotion measures, and usability. During his academic career, Dr. Genov developed and taught courses in Research Methods, Statistics, and Psychology.

Todd Zazelenchuk
Todd is a User Experience Researcher at Intuit in Menlo Park, CA. He earned his Ph.D. in Instructional Technology from Indiana University. Prior to the consumer software industry, Todd worked in academia (Indiana University) and consumer goods (Whirlpool Corp) helping to lead and refine their design and evaluation processes.

Mark Keavney
Mark is a User Experience Designer for Intuit's online payroll software. He earned a PhD in Psychology from Stanford University in 1993 and has since been working as a user researcher and UX designer for products as diverse as financial tools, mobile gaming, educational software, and interactive television.