Improving the User Experience of Medical Devices with Comparative Usability Testing

Anneliis Tosine and Hala Al-Jaber

Abstract A comparative usability test is an evaluation that helps to determine how a particular product performs in relation to similar products by having end users attempt to complete the same set of tasks for each product. This type of usability test assesses if a product is better or worse than others from a usability perspective and reveals relative strengths and weaknesses. When conducting a comparative usability test, a number of variables make the execution more complicated than a standalone usability test. This paper identifies variables to consider, based on a recent comparative test involving three ultrasound systems. Some variables that need to be considered are defining and recruiting appropriate test participants, selecting a suitable test environment, preparing and executing training, creating consent forms, applying a proper test methodology, selecting usability metrics to capture, and analyzing data. This paper identifies what a comparative usability test can offer and the latest techniques of executing such a test.

Keywords Usability test · User experience · Comparative · Ultrasound

1 Introduction

A comparative usability test can help gauge the position of a company's product in comparison to indirect and direct competitors. It can identify each product's strengths and weaknesses from an end user's point of view. The comparison can be made through ranking products by overall usability metrics or can be quite focused on comparing features, functions, or content. A comparative usability test provides product management, research, and development teams with information about what works and what does not from an end user's perspective, by having a group of

A. Tosine (✉) · H. Al-Jaber
User Experience, Macadamian Technologies,
165 Rue Wellington, Gatineau, QC J8X 2J3, Canada
e-mail: atosine@macadamian.com

V.G. Duffy and N. Lightner (eds.), *Advances in Human Factors and Ergonomics in Healthcare*, Advances in Intelligent Systems and Computing 482, DOI 10.1007/978-3-319-41652-6_23

representative users perform the same set of tasks with each product. The results from these tests can help form baseline performance metrics and identify areas for improvement. Product teams can use this information to create and improve upon strategies for upcoming product release cycles.

When planning for and conducting a comparative usability test, a number of variables make the execution more complicated than a standalone, more traditional, usability test.

This paper suggests some variables to consider and will highlight a best practices approach to conducting a comparative usability test. By identifying variables to consider, this paper should help teams plan and execute a comparative usability test more successfully. The paper also provides specific examples based on a recent comparative usability test that was conducted involving three premium ultrasound systems.

2 Case Study

The subject vendor had made a concerted effort to improve the overall usability of its premium ultrasound system. While the vendor felt confident that it had achieved its goals through a user-centered design approach for the new system, it needed to be able to measure the outcomes of its efforts in an objective way.

The authors conducted a comprehensive comparative usability test of the subject vendor's beta system and two other similar premium ultrasound systems from two other vendors. The goal of the test was to compare the effectiveness, efficiency and satisfaction of the three different ultrasound systems using standard usability metrics and methodologies.

Eleven common tasks in abdominal sonography were utilized to assess ease of use, task completion, number of errors and deviations, and overall assessment of usability. Twenty practicing sonographers with a specialty in abdominal sonography were recruited to participate in the test.

3 Best Practices

3.1 Defining and Recruiting Participants

For a traditional usability test that focuses on evaluating one product, selecting participants is a primary challenge because the right participants need to be recruited in an efficient manner. Variables that need to be considered during recruitment may include age, gender, attitude, computer and web experience, and professional and academic backgrounds. For a comparative usability test that focuses on evaluating two or more products, selecting participants can be even more of a challenge because there are additional variables to consider. Three additional variables to consider are prior experience, brand and product attitude, and domain skills and frequency of using those skills.

One of the easiest and most intuitive approaches to handle these additional variables is to balance them across participants (i.e. ensuring that an equal percentage of participants with particular types of variables are involved). This approach is one of the easiest and most intuitive but it comes at a price, as it takes more time and could cost more to recruit these proportionally equal percentages of participants.

Paying close attention to the following performance-affecting variables is key when recruiting participants.

Prior Experience Prior experience with the product being tested has a direct impact on performance success in a usability test. Tasks may have higher completion rates and take less time to complete for participants with more experience. Prior experience also affects the participants' attitudes towards the product and experienced participants typically have more favorable attitudinal data [1].

In the subject comparative usability test, one new ultrasound system was compared alongside two other existing systems from different vendors. Test participants were to be current users of one of the two existing systems. Therefore, during recruitment, prior experience with certain brands of ultrasound systems was key to selecting test participants. This identification meant that each participant would be a first-time user of the two ultrasound systems s/he would evaluate during his/her usability session (e.g. current users of one system evaluated the other existing system and the new beta system).

Brand and Product Attitude Brand and product attitude affect usability test data so measures of favorability towards the products under evaluation are important to capture during the recruitment process [1]. To avoid existing bias towards a brand or product, screening potential participants at either ends of the favorability spectrum regarding a particular brand or product is a selection consideration.

It may also be interesting to manage and examine the differences of favorability. Ideally, the favorability responses should be relatively equal across participants.

For the subject comparative usability test, a Likert scale question was included during the recruitment process so that potential participants could rate their favorability towards each of the three brands. Participants were asked to rate their overall opinion of each brand on a scale from 1 to 5, where 1 represented 'very unfavorable' and 5 represented 'very favorable'. Participants that rated any of the three brands with a '1' were not included in the test. This approached helped to screen out individuals who really disliked a particular brand, as this feeling or attitude may have prejudiced the reliability of their data. This test included an equal number of participants for whom each brand was 'very favorable'. 'Very favorable' reflects a positive direction and strong intensity of feelings toward a brand. For this test's twenty participants, the favorability ratings for the three brands were within a point difference, based on the 5-point scale.

Domain Knowledge and Skills and Frequency of Use Domain knowledge and skills and frequency of use affect performance in a usability test. Specific domain knowledge or specialized skills usually have more impact on performance than differences in interface or design elements [1]. Domain knowledge refers to a set of concepts and terminology understood by practitioners in that domain and domain skills are specific skills useful only for that certain area of expertise [2].

During the recruitment process for the subject comparative usability test, professional demographics and sonography expertise were collected in order to determine the selection of participants. Information such as number of years worked as a sonographer, specialty, work environment, and academic background were collected. Effort was made to have an equal amount of sonography experience in the whole group of test participants.

During the recruitment process, another selection criterion was the participant's frequency of using certain skills, knowledge and brands of systems. For example, only full-time sonographers who regularly scanned patients were considered; students or part-time sonographers were not considered.

Participants with the same skill set were recruited to ensure that their skills would support them during the usability test, as these skills were related to what was being evaluated in the usability test (i.e. only participants with a specialty in abdominal sonography were recruited in order to complete tasks for a typical abdominal exam).

Defining soft and hard metrics for recruitment can allow for adjustments to be made along the way, as the success of the process is examined. For example, demographic information such as participants' sex and age were soft metrics for the subject usability test because they were deemed less impactful to the data as opposed to sonography expertise.

3.2 Training Researchers

In some instances, the user experience (UX) researchers who are conducting the comparative usability test also need product training before planning and executing the test.

For the subject comparative test, experienced end users of the three ultrasound systems were contacted to train the researchers and help them better understand typical workflows, features, terminology, etc. even for those aspects that were not considered a focus area of the test. This helped the researchers in a variety of ways such as documenting when test participants went off the ideal path or did an irreversible error while trying to complete a task. Documenting deviations from an ideal path is often used to measure product efficiency and errors to measure effective product design. The ideal path is considered the intended navigation route to complete a particular task. In some instances, there may be multiple ideal paths.

3.3 Usability Test Location and Space

For a comparative usability test, it is advisable to try to use a lab space that is in a neutral, third party location. This helps encourage participants to provide their honest opinions and feedback during the evaluation.

Comparative usability test sessions can be long because more than one product is being evaluated. Therefore, researchers need to incorporate breaks for the participants between evaluating one product and the next and to consider providing a lounge space and refreshments for participants so as to mitigate fatigue.

Client or stakeholder participation and collaboration during all user research phases has many benefits. The more researchers share, listen, accept and learn from clients or stakeholders, the higher the chances are that they will act upon test findings when they become available [3]. One way to involve a client or stakeholders is to have them observe the usability test sessions. However, to ensure that client or stakeholder participation does not affect the usability test data, ensure that test participants remain unaware of any direct connection that observers may have to any one of the products under evaluation.

When evaluating multiple brands of products, the sponsoring vendor's name must not be evident on any materials in the lab space (e.g. test protocols), as this may impact data as well.

During comparative usability tests, it is best to remove from view the product(s) that is/are not being used in the current usability test session. In the case of the subject test, participants could have become distracted by the third system in the lab space, as that was the brand of system they currently use for their job, so it was removed from the space. The participants' attention needs to be directed to the tasks at hand so that the test session can be kept on schedule.

3.4 Training Test Participants

In some instances, participants need to be trained before they begin a usability test (e.g. providing participants with training on the primary functions, interaction mechanisms, and associated domain knowledge of the device) [4]. Providing effective training should not be taken casually. Training participants should be formal, structured, and given to each participant so as to remain consistent. A standard means of training participants will ensure consistency from one session to the next. Then every test participant begins their usability session with the same skills and exposure to the system(s) or product(s) being tested.

For training to be effective, identify:

- The purpose of the training,
- What skills and knowledge participants are to learn, and
- How the training will be conducted [5].

For the subject comparison test of three ultrasound systems, each participant had to know how to perform some basic tasks with the systems s/he would be evaluating.

Creating training videos for each system, covering the same types of tasks and features in the same detail, was essential to ensuring that the test was not biased towards any of the systems. It addition, it freed the test facilitator from having to train each participant in the exact same manner.

Besides this standardized training, a set time was given for each participant to further familiarize themselves with the systems and for the researcher to observe first impressions. No additional assistance was given at this time in order to avoid swaying results.

The same overall time for training was allowed for each participant and each was encouraged to use the full time available to him or her. Access to user manuals, if available, is an additional construct to consider.

3.5 Waiver and Consent Forms

Typically, test participants are required to sign a waiver or consent form to participate in a test, especially if any parts of the usability test are being recording with notes or video/audio. They may also need to sign a non-disclosure agreement (NDA) form.

If a NDA is needed for a test, consider combining it with the consent form so that participants can sign one form at the beginning of the test instead of two. Since summative usability tests, such as the subject comparative usability test, require a large number of participants and may take a long time to complete, ensure that the consent form explicitly includes a statement requesting that participants not discuss the details about the test with others.

In the subject comparative usability test, the sponsoring ultrasound vendor wanted participant approval to share favorable findings and video recordings for promoting and marketing the new system. This approach required a consent and release form for marketing purposes.

The UX researchers advised the vendor that the most important part of the subject test was to conduct it in an efficient manner with the correct participants; the ‘nice to have’ addition was acquiring participants’ permission for subsequent marketing. Prior agreement on this principle allowed the researchers to keep the two different consent forms separate and have the test participants review them at different times during their usability test session.

Each participant was required to sign the consent form detailing his or her participation and acknowledging the reason for recording the test. Participants did not have to sign the consent and release form for subsequent marketing.

The participants reviewed the consent and release form for marketing at the end of their test session so as to better differentiate what would and would not be shared from their session, if they chose to sign it. Signing the marketing release form at the end of the session helped to keep participants focused on providing honest feedback during the test instead of worrying or becoming distracted by the matter of releasing their views for marketing purposes. Copies of the forms were provided to participants, if they wanted to keep any for their own records.

3.6 Methodology for Executing Test Sessions

For a traditional usability test that focuses on evaluating one product, the approach is rather straightforward and well documented. For a comparative usability test that focuses on evaluating two or more products, conducting the test is more of a challenge because there are more variables to consider.

The following six aspects are ones to pay closest attention to when conducting a comparative usability test.

Product Order If the same group of participants evaluates two or more products in the usability test sessions (called a "within-subjects" test), then it is important to alternate the order in which the participants evaluate the products. This technique is used to avoid the introduction of confounding variables and ensures that the same percentage of participants is exposed to each product first. A confounding variable is a variable, other than the independent variable, that may affect the dependent variable [6]. Variables that can be affected by the order could include practice or learning effect as users get "warmed-up" and/or participants becoming fatigued. By counterbalancing the order of exposure, one can ensure that these unwanted effects are uniformly spread among all the products being evaluated [1]. This approach should be used for any training sessions as well.

Types of Tasks Selecting the test tasks to be included in a usability test is based on the key goals that end users of the product under evaluation are trying to accomplish with it. For a comparative usability test, it may not be that straightforward, as competing products may not help users achieve all of the same goals. Therefore, test tasks for a comparative usability test need to be selected based on the participants that are involved (i.e. abdominal sonographers evaluating abdominal exam scanning tasks) and tasks that can be done on each of the systems or products. Since there will be a learning effect in comparative tests, consideration should be given to multiple test tasks that evaluate the same functions or features with each system or product. In order to compare usability metrics, keep tasks specific instead of open (e.g. "explore the homepage" would be considered an open task) so that the same task can be repeated in the same way on each of the other systems or products and can then be more easily compared.

Test Task Phrasing In most situations, the UX researchers conducting the test are not themselves expert users of the products or systems under evaluation so it is important to get some external help for phrasing the test tasks. For the subject comparative usability test, prior assistance was sought from a clinical team and practicing sonographers in phrasing the test tasks. The improved task phrasing helped to convey the same message and a clear goal for participants to accomplish so they could have the same baseline understanding for each task. Different products or systems often use different words to mean the same things (e.g. 'Erase', 'Clear' and 'Delete'); therefore, tasks need to be worded similarly for each system or product. Be aware of your target audience and pose tasks to usability test participants in a manner that naturally resonates with them. Participants can take tasks very literally so it could be helpful to use plain English and no slang or product-specific language. Every participant needs to interpret the tasks the same way.

Realistic Testing Environment As in traditional usability tests, try and simulate the lab space as closely as possible to a typical work environment. For comparative usability tests it is also important to keep the environment as consistent as possible from test session-to-test session and product-to-product so that the environment does not impact the test's data.

Usability Metrics A research goal and purpose of a usability test determine the usability metrics and data to be collected during test sessions. For a comparative test, the collection of quantitative data is highly recommended as it allows for direct product comparisons to be made and statistical significance calculated.

If well-presented, quantitative results can be very meaningful, easy for stakeholders to understand, and straightforward to market and promote. For example, completion rates provide a simple metric of success and system effectiveness and the rates are easy to collect. Qualitative data is also important to collect, as it provides details about human behavior, emotion, and personality characteristics. The tradeoff for a test could be to collect think-aloud comments over task times. For example, qualitative data provides an understanding of participant attitudes by observing them directly and helps answer questions about 'why' or 'how to fix' a problem. Task times provide researchers with a glimpse into understanding system or product efficiency.

It is also very useful to debrief after each participant's test session and flag 'gold moments' to revisit at the end of the test. This helps keep the initial feedback and findings for the different products clear and organized.

Usability Test Wrap-Up Traditionally, a wrap-up session follows the completion of the test tasks and it is a great chance to ask follow-up questions based on what occurred during the usability test. It also gives a chance for participants to complete standardized questions or questionnaires such as the System Usability Scale questionnaire (SUS) [7]. For a comparative usability test, try going one step further during the test wrap up. For example, follow-up on the SUS responses by calculating the SUS score for each system or product immediately after the test and further probe about the reasons why participants provided their ratings. The wrap up is also a great chance to ask a product preference question to get a 'bottom-line' response from all participants. Also, consider at this point breaking down the product preference question further to address finer variables that lead to the choice.

3.7 Analysis

The type of analysis performed on comparative usability test data depends on the data collected and who was involved in the test.

For most usability tests, quantitative data is a combination of completion rates, errors, deviations, task times, task-level satisfaction, help access, and lists of usability problems (typically including frequency and severity) [8]. As mentioned in the previous section, qualitative data provides details about human behavior, emotion, and personality characteristics in the form of think-aloud comments and responses to test questions.

When calculating test results, it is helpful for readers to understand how precise the estimates are, as compared to the unknown population value. Try to report results with confidence intervals around any mean to provide readers with the most likely range of the unknown population mean or proportion.

For comparative usability tests, test participants can attempt similar tasks on all products (within-subjects design) or different sets of users can evaluate each product (between-subjects design).

For a comparative test, it is necessary to compare results to a specific benchmark or goal in order to determine whether the difference between products, designs, versions, etc. is greater than what would be expected from chance [8]. From calculating confidence intervals, the boundaries of the interval can be used to determine whether a product or system has met or exceeded a goal. Keep in mind that the test design, within-subjects or between-subjects, impacts the calculations that determine if the difference is statistically significant or not.

For the subject comparative usability test, there was great value in conducting a within-subjects design. Besides benefits such as conducting the usability test in the same period of time and with the same recruitment effort and lab space, a major source of variation between sets of data could be removed because of the involvement of the same participants in each test group. Another fundamental advantage of a within-subjects design is statistical power because in effect, the number of subjects has been increased relative to a between-subjects design.

In order to keep the analysis as straightforward as possible, a carefully planned master spreadsheet of all the data to be collected often works best. One spreadsheet tab per product works well with each participant's data de-identified. Using one program to document the data and complete calculations saves time as opposed to copying data from one program to another. Spreadsheet software applications often have many built-in functions that will help with analysis such as t-tests calculations to acquire precise *p*-values.

4 Summary

Comparative usability testing provides product management, research, design, and development teams with a wealth of data and a glimpse into how a product sizes up to its competitors. Since results from these tests provide baseline performance metrics and comparative data to use for claims of product successes, it is important to collect and analyze the data accurately.

In general, planning and executing a comparative usability test is more challenging than a traditional usability test because of a number of variables that make it more difficult.

Test Planning Stage During the planning stage, it is important to define and recruit test participants by focusing on prior experience, brand and product attitude, and domain skills and knowledge and frequency of use. When training the UX researchers who will facilitate the test, experienced end users of the products under evaluation can help the researchers better understand typical workflows, features, terminology, etc. For an in-person usability test evaluating products from multiple vendors, a test location outside of one of vendor's locations helps to make participants feel more comfortable so that they can better provide honest opinions and feedback during the test. At the test location, a break area with refreshments for participants is a consideration. Observations by clients or stakeholders may be permitted but their connection to the product(s) or system(s) must not be obvious to participants.

Test Execution Stage During the execution stage, decisions about participant training that need to be made are: the purpose of the training, what skills and knowledge the participants should learn, and how the training will be conducted. When using a waiver and consent forms, a combined consent and NDA form leaves just one form for participants to sign at the beginning of a test session. Also, if a separate release form for any marketing purposes is needed, consideration should be given to include it at the end of each test session. In order to execute a successful test, careful selection of the product order, types of test tasks, task phrasing, and usability metrics is essential. Additionally, the usability test space should provide a realistic testing environment and there should be ample time to conduct a meaningful test wrap-up session.

Test Analysis Stage When calculating test results, include confidence intervals around any means, compare results to specific benchmarks or goals, and compare results to the other products to determine if a significant difference exists.

References

1. Measuring, U.: http://www.measuringu.com/blog/comparative-variables.php
2. Software Engineering Institute.: http://www.sei.cmu.edu/productlines/frame_report/rel_domains.htm
3. Tomer, S.: It’s Our Research. Morgan Kaufmann, Waltham (2012)
4. Wiklund, M., Kendler, J., Strochlic.: Usability Testing of Medical Devices. CRC Press, Boca Raton (2016)
5. Dumas, J.S., Redish, J.C.: A Practical Guide to Usability Testing. Intellect Books, Exeter (1999)
6. McDonald, J.H.: Handbook of Biological Statistics, 3rd edn. Sparky House Publishing, Baltimore (2014)
7. Brooke, J.: SUS: a “quick and dirty” usability scale. In: Jordan, P., Thomas, B., Weerdmeester, B. (eds.) Usability Evaluation in Industry. Taylor & Franci, London (1996)
8. Sauro, J., Lewis, J.R.: Quantifying User Research. Morgan Kaufmann, Waltham (2012)