COMPARISON OF LOCAL AND REMOTE MODERATED
USABILITY TESTING METHODS

A Project Report

Presented to

The Faculty of the Graduate Program in Human Factors and Ergonomics

San Jose State University

In Partial Fulfillment

of the Requirements for the Degree of

Master of Science

By

Scott Wheelwright

May 2014
APPROVED FOR THE GRADUATE PROGRAM IN
HUMAN FACTORS AND ERGONOMICS

Prof. Dan Rosenberg

Date
ABSTRACT

When competing products have similar technology, the usability of the products becomes a key differentiator. Measurements of product usability are commonly drawn from usability studies which involve a moderator prompting users with tasks and questions while observing their behavior and interactions with the product. Remote usability testing offers an alternative to the traditional local (in-person) usability testing method. In this study, the same product was evaluated in both local and remote testing settings to determine whether there were differences in the findings depending on the testing method used. The results were similar for both satisfaction-based and performance-based metrics from each testing method. However, there were different trends in overall ratings between the two groups. Participants in the local testing group tended to have higher expected usability ratings, while participants in the remote testing group tended to have higher actual usability ratings. The results show that similar findings can be obtained from local and remote usability testing methods, but that there may be important differences in the subjective experiences of participants.
ACKNOWLEDGEMENTS

There are many people who contributed to the successful completion of this project. Thank you to Mick McGee who provided access to the EchoUser software and guidance with implementing the magnitude estimation technique. Thank you also to Professor Dan Rosenberg for his support and advice throughout the entire project. And of course, thank you to all of my family and friends who supported me through this journey.
# TABLE OF CONTENTS

Abstract ........................................................................................................................................................ iv
Acknowledgements ....................................................................................................................................... v

## INTRODUCTION

Problem Statement ........................................................................................................................................ 1
Hypothesis .................................................................................................................................................. 2
Limitations ................................................................................................................................................ 2
Delimitations ............................................................................................................................................. 3

## REVIEW OF LITERATURE

Usability Methods ..................................................................................................................................... 4
User Testing Methods ............................................................................................................................... 5
Local vs. Remote User Testing .................................................................................................................. 6
Synchronous vs. Asynchronous Remote User Testing ............................................................................. 7
Magnitude Estimation ............................................................................................................................... 8
Summary ................................................................................................................................................... 9

## METHOD

Problem Statement .................................................................................................................................. 10
Participants .............................................................................................................................................. 10
Apparatus or Instruments ........................................................................................................................ 12
Procedures ............................................................................................................................................... 15
Design ..................................................................................................................................................... 18

## RESULTS

Expected Usability Ratings ...................................................................................................................... 20
Actual Usability Ratings .......................................................................................................................... 22
Actual Ratings – Perceptual Tasks ........................................................................................................... 24
Expected Ratings vs. Actual Ratings ...................................................................................................... 26
Errors ...................................................................................................................................................... 28
Time on Task .......................................................................................................................................... 29

## DISCUSSION

Limitations .............................................................................................................................................. 33
Future Research ...................................................................................................................................... 33

## APPENDIX A. Informed Consent Form

## APPENDIX B. Test Protocol
INTRODUCTION

When a new technology is invented, the functional capacity of the product alone is often enough to set it apart from the competition. However, as the technology becomes more widely available, the usability of a product becomes the main differentiator between various brands. The usability of a product is described in the ISO standard as parameters of ease of use, efficiency, and satisfaction (ISO, 1998). Measurements of these parameters are commonly drawn from usability studies which involve a moderator prompting users with tasks and questions while observing their behavior and interactions with the product.

Usability studies have typically been carried out in labs, with one-way mirrors installed for observations and video recording. This setup incurs significant costs, including the construction of the lab and transportation of users to and from the lab. Now that computing power has increased to the point that video conferencing and screen-sharing software are readily available to the average consumer, the use of remotely moderated usability testing is a viable alternative to local, in-person studies. While there is potential to save time and money by conducting studies remotely, there may be differences in the results obtained through this method compared to local testing. Determining whether there are in fact differences in the results of these two user testing methods is of great interest from a business perspective. If remote user testing provides sufficiently similar results at significantly reduced costs, it represents a strategic advantage over traditional in-person testing in the product development process.

Problem Statement

The purpose of this study was to compare the accuracy and variability of the results of local versus remote moderated usability testing methods. In order to examine the specific differences between the two methods, it was important to use the same moderator and protocol.
throughout the study. Differences in the results of the usability testing methods could be either performance-based or satisfaction-based. While performance-based parameters (i.e. time on task or number of errors) are straightforward to compare against each other, subjective ratings of satisfaction can be more difficult to compare. Since satisfaction ratings are abstract and will vary by interpretation of each user, a method called magnitude estimation was used to compare subjective usability ratings. The magnitude estimation method allowed items to be compared across a continuum, using ratio-based number assignments. Usability ratings were compared between the local and remote study methods using analysis of variance in order to make an assessment of any differences in the results.

**Hypothesis**

The hypothesis for this study was that there would be no difference in either performance-based or satisfaction-based results between the local and remote studies given an identical usability test protocol, similar subject profiles, and the same test moderator. The participants would, in theory, have the same experience with the product in either scenario.

**Limitations**

Participants in this study consisted of students enrolled in the Psychology 1 course at San Jose State University (10), graduate students at San Jose State University (10), and working professionals in the San Francisco Bay Area (10). San Jose State University is located in Silicon Valley, a region in Northern California known for being on the cutting edge of technology. The typical student in this part of the country may not be representative of the typical student in the general population in terms of technical knowledge and experience. Similarly, the working professionals in the bay area may be more technically savvy than the typical professional in the general population.
Delimitations

The study was meant to evaluate the differences in local versus remote usability testing. Differences between various remote testing software options, test protocols, or moderators were not included or evaluated in this study. The study was not conducted in a formal usability lab, but rather was conducted on the San Jose State University campus and in a local office conference room. This may limit the generalizability of the results to usability practitioners who exclusively conduct their studies in formal usability lab settings.
REVIEW OF LITERATURE

Usability is defined by the International Organization for Standards (ISO) as a measurement of the effectiveness, efficiency, and satisfaction of a product (ISO, 1998). Usability as a concept, however, is largely subjective. There has been debate over the definition of usability and over which metrics are considered best practices to use in usability evaluations (McGee, Rich, & Dumas, 2004). To address this issue, McGee et al. (2004) created a usability concept survey to evaluate the user perspective on what usability means. From the results of this survey, McGee et al. (2004) created a taxonomy of usability consisting of three levels of usability characteristics: core, secondary, and tertiary; and two categories of non-usability characteristics: satisfaction and style. Ultimately, McGee et al. (2004) defined usability as “your perception of how consistent, efficient, productive, organized, easy to use, intuitive, and straightforward it is to accomplish tasks within a system” (p. 909). This definition of usability will be utilized in the present study.

Usability Methods

There are several methods by which the usability of a product can be evaluated (Gulati & Dubey, 2012). These include both inspection methods as well as empirical methods. Inspection methods are typically conducted by usability and subject matter experts and include methods such as the cognitive walkthrough, heuristic evaluation, feature inspection, and standards inspection. Empirical methods involve observation of users and include methods such as the think-aloud protocol, focus groups, remote and local user testing, interviews, and field observations. Many studies have been conducted to compare the advantages and disadvantages of various usability evaluation methods (Jeffries & Desurvire, 1992; Karat, Campbell, & Fiegel, 1992; Molich & Dumas, 2008; Tan, Liu, & Bishu, 2009). In particular, these researchers were
interested in whether there were differences in the findings of inspection methods (i.e. heuristic evaluation and cognitive walkthrough) and empirical methods (i.e. user testing).

Results from these studies have been mixed. Molich and Dumas (2008) found that there was no difference in the results of the evaluations based on the method used. Contrary to this finding, Karat et al. (1992) found that the empirical usability test discovered more issues in the evaluated interface than the individual and group walkthroughs. Jeffries and Desurvire (1992) provided a counterpoint to the Karat et al. (1992) study and others similar to it, citing that the findings were leaving out important limitations of each method. For example, the heuristic evaluation method is highly dependent on the knowledge and skill level of the evaluator. On the other hand, user testing is typically expensive and can only be conducted late in the development process. Jeffries and Desurvire (1992) contend that the two methods are complimentary; when user testing is impractical or unavailable, inspection methods are a suitable substitute. Tan et al. (2009) similarly found that the heuristic evaluation and user testing methods are complimentary, with neither holding an absolute advantage over the other. They recommend that heuristic evaluation be used early in the development process and user testing be used later in the development process.

**User Testing Methods**

When conducting user testing, there are a variety of options for the practitioner: local user testing, synchronous remote user testing, and asynchronous remote user testing. Local user testing is typically performed in a usability laboratory, although it can also be performed less formally with the moderator and test participant sitting together in a room. Remote user testing emerged as an alternative to laboratory testing in the mid 1990’s (McFadden, Hager, Elie, & Blackwell, 2002). Synchronous remote user testing essentially mimics the in-person user testing
experience, though the moderator and test participant are in separate locations. The moderator can communicate with the test participant in real time using web conferencing software and screen sharing features. Asynchronous remote user testing removes the moderator from the scenario, leaving the test participant to complete the tasks of the study at their own pace and without interaction with the moderator. The tasks of the study are typically displayed in a dialog box on the screen and the test participant manually advances through the task prompts as they complete them. Data from asynchronous user testing may be collected in a variety of ways, such as screen recording, mouse click tracking, task completion time, and survey responses.

**Local vs. Remote User Testing**

There are advantages and disadvantages to both local and remote user testing methods (Alghamdi, Al-Badi, Alroobaea, & Mayhew, 2013; McFadden et al., 2002; Thompson, Rozanski, & Haake, 2004). In-person testing provides the moderator the ability to directly observe body language and facial expressions during the testing procedure (Thompson et al., 2004). This can provide valuable insight into the experience that the user is having, particularly when they are not verbally expressing it. A remote testing scenario with only a camera facing the product would not capture this information. Laboratory testing also affords the experimenter more control over the testing environment and data collection than remote testing (McFadden et al., 2002). On the other hand, remote testing that allows the user to participate from their home or office may be more representative of a real use case (Thompson et al., 2004). Additionally, the user may feel more comfortable in their home or office environment compared to the laboratory. Andrzejczak and Liu (2010) conducted an experiment to compare both the objective results of local and remote testing methods and the subjective experience of the testing process itself on the test participants. They found that not only was there no significant difference in the
number of critical incidents identified between the two methods, but that there was no significant difference in the levels of stress experienced by the test subjects in each condition. Another issue to be considered when comparing local and remote testing is the time and money required to conduct the study. Laboratory testing is typically expensive and requires test participants to physically travel to the testing site. Remote testing removes the travel requirements entirely and typically offers a significant reduction in cost.

**Synchronous vs. Asynchronous Remote User Testing**

While remote user testing removes the requirement that the moderator and test participant be in the same location, asynchronous remote testing removes the moderator entirely. This allows test participants to participate in studies regardless of time differences and reduces scheduling constraints. Test participants are typically prompted by a software interface to complete the various tasks of the study that the test facilitator has created. This added convenience has its drawbacks, most notably that the moderator and test participant can no longer communicate with each other in real time. Without the ability to ask questions to clarify task objectives, Alghamdi et al. (2013) found that the asynchronous test participants in their study were less successful at completing the tasks compared to the synchronous test participants. While the absence of communication may have caused some confusion for the asynchronous testers, they averaged faster task completion times than the synchronous testers (Alghamdi et al., 2013). When it came to satisfaction ratings, the synchronous testers rated the product slightly higher; however the asynchronous testers were more satisfied with the testing experience itself. In the end, the results showed only minor differences, and no statistically significant differences in the number of issues found or user satisfaction ratings (Alghamdi et al., 2013). These findings suggest that the tasks themselves should be given significant consideration when deciding which
remote testing method to use. If the tasks are considered easy enough for participants to follow
and complete on their own, the test administrator can expect similar results from either
synchronous or asynchronous remote testing. However, when evaluating complex tasks or
interfaces, real-time communication may be necessary to guide participants through the study.

**Magnitude Estimation**

When conducting usability testing, there are a number of metrics that are often used to
assess a product (McGee, 2003). These can include objective metrics such as time-on-task,
number of errors, and completion rates. Additionally, subjective measures such as preference
and satisfaction are also frequently gathered. One common method for quantifying these
subjective metrics is the Likert scale. McGee (2003) notes several limitations to the Likert scale,
particularly their limited, pre-defined ranges and ordinal nature. The limited ranges can reduce
the variance among participant responses, a metric which might otherwise have been a source of
insight for the experimenter. Additionally, the test participant is only able to make ratings at
one-unit intervals along the scale, which may not accurately reflect their opinions. To address
these shortcomings, McGee (2003) suggests using a psychophysical measurement method called
magnitude estimation when assessing usability. Magnitude estimation allows test participants to
evaluate the psychological sensation of physical stimuli and has been shown to be successful in
scaling vague and complex perceptions such as guilt and stress (McGee, 2003).

When implementing magnitude estimation, test participants are given a definition of
usability with which to evaluate all of the products included in the study. Participants are then
given a series of example products to evaluate. Each successive product is evaluated in relation
to the preceding products, which establishes a ratio-scale continuum of perceived usability. The
usability study is then conducted in the same manner as it would be with any other scaling
method, but with test participants assessing usability based on the ratio scale they have just established themselves. In order to analyze the data from several test participants using this method, the data must be normalized using geometric averaging (McGee, 2003). The data can then be analyzed in a variety of ways, including parametric statistics. Results from usability studies using magnitude estimation have shown to possess high construct validity and lead to data-driven descriptions of the subjective experience of usability that are easy to communicate to external audiences (McGee, 2003).

Summary

In this literature review, product usability has been established as an important differentiator among brands. A variety of usability testing methods have been discussed, including heuristic evaluation, cognitive walkthrough, local user testing, synchronous remote user testing, and asynchronous remote user testing. In addition to the various methods for conducting usability evaluations, different methods for assessing and defining usability have also been discussed. From a business perspective, the time and cost savings potential of remote user testing represents a potential advantage, provided that the results are sufficiently similar to other testing methods. Additionally, the data-driven results afforded by the magnitude estimation assessment method can be useful in presenting business cases related to usability issues. A comparison of remote and local user testing methods using magnitude estimation may contribute new insights to this discussion of usability evaluation methodology.
METHOD

The methods that were used in this study are described in this section. This includes a description of the purpose, participants, apparatus, procedures, and design of the experiment. The methods of analysis of data obtained from the study are also described.

Problem Statement

The purpose of this study was to compare the accuracy and variability of the results of local versus remotely moderated usability testing methods. In order to examine the specific differences between the two methods, it was important to use the same moderator and protocol throughout the study. Differences in the results of the usability testing methods could be performance-based or satisfaction-based. While performance-based parameters (i.e. time on task or number of errors) are straightforward to compare against each other, subjective satisfaction ratings can be more difficult to compare. Since satisfaction ratings are abstract and will vary by interpretation of each user, a method called magnitude estimation was used to compare subjective usability ratings. The magnitude estimation method allowed items to be compared across a continuum, using ratio-based number assignments. Usability ratings were compared between the local and remote study methods using analysis of variance in order to make an assessment of any differences in the results.

Participants

Participants in this study consisted of students enrolled in the Psychology 1 course at San Jose State University (10), graduate students at San Jose State University (10), and working professionals in the San Francisco Bay Area (10). The students received credit in their classes for participating in the experiment. Neither the students nor the professionals received financial compensation for their participation. The participants who were recruited ranged in age from 18-
Both males (14) and females (16) were included in the study. Prior to running the
study, an estimated sample size was calculated using the G*Power software (Version 3.1.7) with
a power factor of 0.80, a confidence interval of 0.05 and an effect size of 0.4. In order to achieve
this power level, 52 participants would need to be included in the study.

Figure 1 – A Priori Sample Size Calculation

However, the study only included 30 participants. A post hoc calculation using the G*Power
software gives the actual power achieved in the study of 0.56, shown in Figure 2.

Figure 2 - Post Hoc Power Calculation

All participants in the study were kept anonymous. Data associated with the participants
was linked to an identification number that is only known to the experimenter. Recordings of the
test sessions only included screen recording and audio recording. The participants themselves
were not video recorded during the sessions. All recordings were kept anonymous and were not
distributed. After all of the test sessions had been completed and the data had been analyzed, the
key which links participants to their identification number was permanently deleted from the experimenter’s computer. The video recordings were also permanently deleted after the data analysis has been completed.

**Apparatus or Instruments**

The study was conducted in-person for half of the participants and remotely for the other half. The participants who were included in the in-person portion of the study used a MacBook Air running Mac OS X Version 10.8.5 to complete the study. The participants viewed the website using the Chrome browser. Quicktime screen recording software (Version 10.2) was used to record both the on-screen actions of the participants and the audio of their comments and responses to questions. The moderator ran the EchoUser software on an ASUS A53S laptop with a 15-inch widescreen monitor running on the Windows 7 operating system.

*Figure 3 - Local Testing Setup*
Participants who were included in the remote portion of the study used their own laptop to complete the experiment. The laptops had to be powerful enough to run video conferencing software while browsing a website. Participants were required to use either Firefox or Chrome browsers during the study. The operating system used on participant laptops needed to be either Microsoft Windows (Version 7 or newer) or Mac OS X (Version 10.6 or newer). Remote participants communicated with the moderator via the video conferencing software GoToMeeting (Version 5.4).

![Figure 4 - Participant's Shared Screen in GoToMeeting](image)

This software ran on both the moderator’s and the participant’s computers. The participant shared their screen with the moderator while they were performing the tasks. The moderator used the GoToMeeting screen recording software to record the participant’s shared screen actions and audio comments. The moderator separately ran the EchoUser software on the ASUS A53S laptop.
For all participants (local and remote), the moderator used the EchoUser software to follow the protocol script, track responses to questions, and record task performance. The EchoUser software, shown in Figure 6, is web-based and is built on the Salesforce SAAS platform. The moderator created the test protocol within the EchoUser software prior to running the study with participants. The software was also used in the data analysis after all participants had completed their test sessions. Specifically, the software was used to normalize the data for the magnitude estimation technique and to graphically compare the participants’ expected usability ratings with their actual perceived usability ratings for each task.
Procedures

The study was conducted in two different ways: in-person (local) and remotely. Since the purpose of the study was to compare these two methods, all participants interacted with and evaluated the same website: www.songkick.com. The local study was conducted in a library room on the campus of San Jose State University for five participants and in a local office conference room for ten participants. In both locations, the participant and the test moderator were the only people in the room and the door was closed to avoid distractions. The moderator greeted the participant at the test room entrance and showed them where to sit. The moderator sat next to the participant in order to view what they are doing on their screen while prompting them with tasks. The moderator explained to the participant what the purpose of the study was and gave a brief overview of the process. The participant was then given an informed consent form to read and sign. Once the participant had signed the informed consent form, the moderator began recording the screen and audio on the participant’s laptop (MacBook Air).
The remote participants were located in their own homes. Remote participants were required to be alone in the room in which they performed the study to avoid distractions and to create a testing environment similar to the local testing scenario. The moderator greeted the remote participants over the GoToMeeting video conferencing software. The remote participants were emailed the same informed consent form that the local participants were given prior to their scheduled testing session. They printed and signed the informed consent form and then scanned it and emailed it back to the test moderator. The moderator verbally verified that the participant had read and signed the informed consent form prior to starting the test session. Once the participant had confirmed that they read and signed the informed consent form, the moderator asked the participant to share their screen via the GoToMeeting software. The moderator then began recording the shared screen with the GoToMeeting software on the moderator’s computer.

For both local and remote participants, the moderator then began the test protocol. The tasks of the test protocol were built into the EchoUser software. In addition to the protocol tasks of the study, the EchoUser software had some built-in tasks that were included in the experiment. The participants completed these built-in tasks prior to starting the protocol tasks. The built-in tasks gave an overview of the magnitude estimation concept, a definition of usability that was used in the study, and established baseline values for the subjective usability ratings. Completing these tasks created the personalized scale that each participant used throughout their test session.

The moderator then began prompting the participant to complete the tasks of the study. Prior to beginning each task, the participant gave a rating of their expected usability for that task based on the task description and the rating scale they had created. They then attempted to complete the task without interruption from the moderator. If the participant was confused or
became stuck, they could ask the moderator for help. The moderator used discretion in answering any questions and either marked the task as unsuccessful or helped the participant to complete the task if completion was necessary to continue with the rest of the session. The tasks to be completed were:

1. Navigate and Explore: Open a web browser (Chrome or Firefox) and navigate to www.songkick.com. Take a few minutes to explore the site.
2. Browse Local Listings: Browse the local concert listings (SF Bay Area).
3. Search Near SJSU: Search for concert listings near SJSU.
4. Sign Up: Sign up for an account with Songkick.
5. Search for an Artist: Think of a band or an artist that you would like to see. Search for that artist on the site, and then track that artist’s concert listings.
6. Track More Artists: Find two more artists that you are interested in and track their concert listings.
7. View Tracked Artists: View a list of all the artists that you are tracking.
8. View Upcoming Shows: Now view all of the upcoming concerts for the bands that you are tracking.
9. Mark Attending: Mark three events as if you will be attending them. These can be for any event, not just for the artists that you are tracking.
10. Track Local Events: Track three local events that you’d like to know about.
11. View Attending Events: Now view all of the events that you plan to attend.
12. Import Bands: Import bands from iTunes, Facebook, Pandora, or last.fm

The moderator tracked time-on-task within the EchoUser software, beginning to track the time as soon as the participant was told to proceed with the task and ending when they had either
successfully completed the task or had abandoned the task due to an inability to complete it. The number of errors and the usability ratings (both expected and perceived) were manually recorded in the EchoUser software by the moderator. Errors were counted equally whether they were slips (correct intention but incorrect action) or mistakes (wrong intention). Expected usability ratings were recorded in the EchoUser software by the moderator after each task had been described but before the participant began to attempt the task. The actual (perceived) usability ratings were recorded after each task had been either completed or abandoned. At the end of each session, the participant was asked to rate the usability of the testing process itself, unrelated to the website that has just been evaluated. The participants were then asked to rate three familiar products based on their memory of recent usage: Google Search, Amazon, and Google Maps. This gave additional perspective to the subjective ratings of each participant.

The moderator thanked the participants for their time and participation in the study. The moderator then either walked the local participants to the door or ended the meeting in GoToMeeting with the remote participants. The screen recordings were saved through either Quicktime (local) or GoToMeeting (remote) to the local hard drive. Data collected through the EchoUser software was saved on remote servers (cloud storage) in real time during the study sessions.

**Design**

The experiment was a between-subjects, two-group design. The independent variable was the testing method used: local or remote. The dependent variables were the expected usability ratings, actual usability ratings, number of errors, and time on task.
### Testing Method

<table>
<thead>
<tr>
<th>Protocol Task (12)</th>
<th>Local</th>
<th>Remote</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>• Expected Rating</td>
<td>• Expected Rating</td>
</tr>
<tr>
<td></td>
<td>• Actual Rating</td>
<td>• Actual Rating</td>
</tr>
<tr>
<td></td>
<td>• Errors</td>
<td>• Errors</td>
</tr>
<tr>
<td></td>
<td>• Time on Task</td>
<td>• Time on Task</td>
</tr>
</tbody>
</table>

**Figure 7 - Experimental Design**

The dependent ratings were measured for each of the 12 tasks in the protocol. Actual usability ratings were also taken for the testing process itself and the three familiar products (Google Search, Amazon, and Google Maps). Half of the participants were randomly placed in the local testing group, and the other half were randomly placed in the remote testing group. All participants performed the same set of tasks.
RESULTS

The experimental data was analyzed using the analysis of variance (ANOVA) method to determine whether there were statistically significant differences in the quantitative metrics between the treatments of local and remote usability testing. These metrics consisted of expected and actual usability ratings, error rates, and time-on-task. The results were compared across the different testing methods for each task. ANOVA calculations were computed using IBM SPSS Statistics software, version 20. Comparisons of expected usability ratings and actual usability ratings were also evaluated in conjunction with the magnitude estimation technique.

Since each participant created their own scale for expected and actual usability ratings, the data was normalized so that they could be compared across all participants. This data normalization was computed with the EchoUser software. Plots of the expected and actual usability ratings for the various tasks among all participants were also created in the EchoUser software. In addition to rating each task, the participants also rated their perceived usability of the testing methods themselves. These ratings are included to give additional insight into the subjective experience of each testing method for the participants. To obtain further perspective on the ratings scales, participants were asked to rate three familiar products based on recent usage: Google Search, Amazon, and Google Maps.

Expected Usability Ratings

After each task was described, but before beginning the task, the participants gave a rating of their expected usability for that task. Expected usability ratings for both local and remote participants are shown in Figure 8.
The expected usability ratings for each task were similar for both treatment conditions. There were no statistically significant differences in the expected ratings between the local and remote participants. These results support the hypothesis of the experiment, that there should not be a difference in the experience between local and remote participants. Each participant was rating their expectation of usability based on the task description, which should not be affected by their location. Two of the tasks did approach statistical significance: Search Near SJSU \([F(1,28) = 3.07, p = 0.09]\) and View Upcoming Shows \([F(1,28) = 3.15, p = 0.09]\). It is not clear why the expectations were somewhat different between local and remote participants for these two tasks. In both cases, the participants in the local group rated their expectations higher than the participants in the remote group. Participants in the local group gave higher expected ratings.
than participants in the remote group for 8 out of the 12 tasks. However, in many cases these were by very slim margins and in all cases they were not statistically significant.

<table>
<thead>
<tr>
<th>Task</th>
<th>Local Avg. Rating</th>
<th>Remote Avg. Rating</th>
<th>F</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Navigate and Explore</td>
<td>22.50</td>
<td>21.16</td>
<td>0.23</td>
<td>0.63</td>
</tr>
<tr>
<td>Browse Local Listings</td>
<td>21.96</td>
<td>23.01</td>
<td>0.15</td>
<td>0.70</td>
</tr>
<tr>
<td>Search Near SJSU</td>
<td>23.01</td>
<td>18.47</td>
<td>3.08</td>
<td>0.09</td>
</tr>
<tr>
<td>Sign Up</td>
<td>27.06</td>
<td>23.84</td>
<td>1.23</td>
<td>0.28</td>
</tr>
<tr>
<td>Search for an Artist</td>
<td>26.29</td>
<td>25.60</td>
<td>0.08</td>
<td>0.77</td>
</tr>
<tr>
<td>Track More Artists</td>
<td>29.03</td>
<td>30.39</td>
<td>0.16</td>
<td>0.69</td>
</tr>
<tr>
<td>View Tracked Artists</td>
<td>27.52</td>
<td>25.03</td>
<td>0.58</td>
<td>0.45</td>
</tr>
<tr>
<td>View Upcoming Shows</td>
<td>25.38</td>
<td>21.03</td>
<td>3.15</td>
<td>0.09</td>
</tr>
<tr>
<td>Mark Attending</td>
<td>24.16</td>
<td>21.96</td>
<td>0.72</td>
<td>0.40</td>
</tr>
<tr>
<td>Track Local Events</td>
<td>25.98</td>
<td>26.36</td>
<td>0.02</td>
<td>0.90</td>
</tr>
<tr>
<td>View Attending Events</td>
<td>26.15</td>
<td>25.93</td>
<td>0.01</td>
<td>0.93</td>
</tr>
<tr>
<td>Import Bands</td>
<td>20.80</td>
<td>23.74</td>
<td>1.08</td>
<td>0.31</td>
</tr>
</tbody>
</table>

Table 1 - Expected Usability Ratings Statistics

Actual Usability Ratings

After each task was completed, the participants gave their actual usability rating for that task based on their experience. Actual usability ratings for both local and remote participants are shown in Figure 9.
The actual usability ratings for each task were similar for both treatment conditions. There were no statistically significant differences in the actual ratings between the local and remote participants. Again, these results support the hypothesis of the experiment, that there should not be a difference in the experience between local and remote participants. Two of the tasks did approach statistical significance: Track More Artists \([F(1,28) = 3.41, p = 0.08]\) and Track Local Events \([F(1,28) = 3.21, p = 0.09]\). The remote participants rated the actual usability higher than the local participants for both of these tasks. This was a trend overall, with remote participants giving higher actual usability ratings for 10 out of 12 tasks. While the differences in the ratings were not statistically significant, it is an interesting finding that may reflect some underlying differences in the testing conditions. It is possible that the remote participants were more comfortable in their environment at home and therefore more likely to give favorable ratings. Additionally, the local participants may have felt an obligation to be more critical in
their evaluations given the more formal testing environment they were in. Having the moderator physically present may have affected the local participants’ ratings as well. It is also interesting to note that local participants were somewhat more likely to give higher expected ratings and lower actual ratings compared to the remote participants.

<table>
<thead>
<tr>
<th>Task</th>
<th>Local Avg. Rating</th>
<th>Remote Avg. Rating</th>
<th>F</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Navigate and Explore</td>
<td>20.32</td>
<td>22.69</td>
<td>0.82</td>
<td>0.37</td>
</tr>
<tr>
<td>Browse Local Listings</td>
<td>18.90</td>
<td>22.33</td>
<td>0.87</td>
<td>0.36</td>
</tr>
<tr>
<td>Search Near SJSU</td>
<td>10.43</td>
<td>9.52</td>
<td>0.10</td>
<td>0.75</td>
</tr>
<tr>
<td>Sign Up</td>
<td>25.67</td>
<td>27.50</td>
<td>0.33</td>
<td>0.57</td>
</tr>
<tr>
<td>Search for an Artist</td>
<td>26.92</td>
<td>28.28</td>
<td>0.19</td>
<td>0.67</td>
</tr>
<tr>
<td>Track More Artists</td>
<td>28.33</td>
<td>36.61</td>
<td>3.41</td>
<td>0.08</td>
</tr>
<tr>
<td>View Tracked Artists</td>
<td>30.86</td>
<td>33.67</td>
<td>0.34</td>
<td>0.57</td>
</tr>
<tr>
<td>View Upcoming Shows</td>
<td>21.97</td>
<td>20.77</td>
<td>0.10</td>
<td>0.76</td>
</tr>
<tr>
<td>Mark Attending</td>
<td>23.27</td>
<td>26.19</td>
<td>0.80</td>
<td>0.38</td>
</tr>
<tr>
<td>Track Local Events</td>
<td>25.16</td>
<td>34.50</td>
<td>3.21</td>
<td>0.08</td>
</tr>
<tr>
<td>View Attending Events</td>
<td>27.89</td>
<td>32.37</td>
<td>1.00</td>
<td>0.33</td>
</tr>
<tr>
<td>Import Bands</td>
<td>25.89</td>
<td>28.93</td>
<td>0.54</td>
<td>0.47</td>
</tr>
</tbody>
</table>

Table 2 - Actual Usability Ratings Statistics

Actual Ratings – Perceptual Tasks

In addition to the tasks that were performed, participants were also asked to rate the overall usability of the testing process itself. Participants were instructed to use the same rating scale that they had been using for the tasks they performed. All aspects of the testing process were to be taken into account, including their comfort level, any technical difficulties, getting to the testing location, and the magnitude estimation technique. Participants were then asked to rate three familiar products, again using the same ratings scale. This was to help give perspective to the ratings scales that each participant had created for themselves. The three products rated were Google Search, Amazon, and Google Maps. While the Google Maps application had recently
been updated, participants were told to rate whichever version they were more familiar with.

The ratings for these perceptual tasks are shown in Figure 10.

![Usability Ratings - Perceptual](image)

**Figure 10 - Usability Ratings - Perceptual**

The results from the perceptual tasks were consistent with the results from the performed tasks. The perceptual task ratings were similar for both local and remote participants, with no statistically significant differences. Again, the remote participants tended to give higher usability ratings for each perceptual task compared to the local participants. Including both the performed and perceptual tasks, the remote participants’ average actual usability ratings were higher in 14 out of 16 tasks compared to the local participants’ ratings.
<table>
<thead>
<tr>
<th>Task</th>
<th>Local Avg. Rating</th>
<th>Remote Avg. Rating</th>
<th>F</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test Method</td>
<td>26.81</td>
<td>34.95</td>
<td>2.16</td>
<td>0.15</td>
</tr>
<tr>
<td>Google Search</td>
<td>40.56</td>
<td>41.34</td>
<td>0.01</td>
<td>0.92</td>
</tr>
<tr>
<td>Amazon</td>
<td>29.09</td>
<td>35.13</td>
<td>1.09</td>
<td>0.31</td>
</tr>
<tr>
<td>Google Maps</td>
<td>28.66</td>
<td>34.76</td>
<td>1.25</td>
<td>0.27</td>
</tr>
</tbody>
</table>

Table 3 - Perceptual Task Ratings Statistics

Expected Ratings vs. Actual Ratings

Comparing the expected and actual usability ratings provides insight into the subjective experience of the participants. It can be inferred that when the actual results did not meet or exceed expectations, the participant may have felt frustrated or disappointed by their experience. When evaluating which tasks to address or improve, the tasks that were the farthest below expectations should be prioritized first in order to minimize the frustration of the users.

Figure 11 - Expected vs Actual Usability Ratings - Local Participants
The charts in Figure 11 and Figure 12 show the same trends that have been discussed previously. The ratings for the local participants are clustered below the line, indicating that the actual experience was below expectations. The ratings for the remote participants are clustered above the line, indicating that the actual experience was better than expected. The Search Near SJSU task stands out in both charts as being well below expectations. This supports the hypothesis that there should not be a difference in the results of the local and remote testing methods. In either treatment condition, the usability practitioner would conclude that this task should be a top priority for redesign and improvement. However, when looking only at the general trends, the local and remote testing conditions give somewhat different results. Looking at Figure 11, it seems that participants’ actual experiences were typically below expectations. Looking at Figure 12, it seems that participants’ actual experiences were typically above
expectations. This would lead to differing conclusions from the usability practitioner depending on which testing method they had used when evaluating the product.

**Errors**

Errors were tracked while participants performed the various tasks. This included both slips (correct intention but wrong execution) and mistakes (wrong intention). Errors were tracked in real time by the moderator using the EchoUser software. The total numbers of mistakes for each task are shown in Figure 13.

![Total Number of Errors](image)

*Figure 13 - Total Number of Errors*

Again, the results from the local and remote participant groups are very similar, with no statistically significant differences between them. Remote participants tended to commit more errors, with a higher total number of errors in 8 out of the 12 tasks.
<table>
<thead>
<tr>
<th>Task</th>
<th>Local Avg. Rating</th>
<th>Remote Avg. Rating</th>
<th>F</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Navigate and Explore</td>
<td>0.00</td>
<td>0.00</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Browse Local Listings</td>
<td>0.13</td>
<td>0.27</td>
<td>0.80</td>
<td>0.38</td>
</tr>
<tr>
<td>Search Near SJSU</td>
<td>0.93</td>
<td>0.93</td>
<td>0.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Sign Up</td>
<td>0.27</td>
<td>0.40</td>
<td>0.35</td>
<td>0.56</td>
</tr>
<tr>
<td>Search for an Artist</td>
<td>0.33</td>
<td>0.40</td>
<td>0.07</td>
<td>0.79</td>
</tr>
<tr>
<td>Track More Artists</td>
<td>0.00</td>
<td>0.07</td>
<td>1.00</td>
<td>0.33</td>
</tr>
<tr>
<td>View Tracked Artists</td>
<td>0.20</td>
<td>0.27</td>
<td>0.17</td>
<td>0.68</td>
</tr>
<tr>
<td>View Upcoming Shows</td>
<td>1.00</td>
<td>1.20</td>
<td>0.21</td>
<td>0.65</td>
</tr>
<tr>
<td>Mark Attending</td>
<td>0.40</td>
<td>0.53</td>
<td>0.33</td>
<td>0.57</td>
</tr>
<tr>
<td>Track Local Events</td>
<td>0.07</td>
<td>0.20</td>
<td>1.12</td>
<td>0.30</td>
</tr>
<tr>
<td>View Attending Events</td>
<td>0.07</td>
<td>0.00</td>
<td>1.00</td>
<td>0.33</td>
</tr>
<tr>
<td>Import Bands</td>
<td>0.53</td>
<td>0.20</td>
<td>1.45</td>
<td>0.24</td>
</tr>
</tbody>
</table>

*Table 4 - Total Errors Statistics*

**Time on Task**

The time to complete each task was recorded using the EchoUser software. The time started once the participant was told to proceed with the task and ended when they had either completed the task successfully or given up on the task.

*Figure 14 - Average Time on Task*
The time on task results were similar for both local and remote participants. There was not a clear trend of one test condition leading to longer task completion times. However, there were two tasks which did have statistically significant differences between the local and remote conditions: Track Local Events \([F(1,28) = 4.01, p = 0.05]\) and Import Bands \([F(1,28) = 8.60, p < 0.05]\).

The Track Local Events task was likely different due to varying interpretations of the task by the participants. Some participants took their time searching for events they were actually interested in; while others simply marked the first three events they saw. These two different approaches were essentially two different tasks, where one approach required searching and browsing while the other approach simply required identifying three events. These different interpretations of the task lead to significantly different times for task completion.

The Import Bands task was somewhat dependent on the software of the laptop being used. If the participant chose to import bands from iTunes, they were taken to a page that required a Java plugin. Some of the remote participants’ laptops already had this installed and some did not. Some of the participants did not notice that this plugin was needed, and waited for a while before realizing that nothing was happening with the import task. Once it was clear that a plugin was missing, some participants took the time to install the plugin, while others did not feel comfortable installing additional software on their personal computer. Local participants had less incentive to choose iTunes as their import source since they did not have their personal music collection on the laptop that was provided for them. For these reasons, the remote participants had a different experience than the local participants on this task. While the remote testing condition limited the ability to control environmental factors such as this, it also represented a realistic use case that participants would encounter outside of the lab. Requiring
participants to install the plugin prior to participating in the study would also have mitigated this issue.

<table>
<thead>
<tr>
<th>Task</th>
<th>Local Avg. Rating</th>
<th>Remote Avg. Rating</th>
<th>F</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Navigate and Explore</td>
<td>0.00</td>
<td>0.00</td>
<td>1.43</td>
<td>0.24</td>
</tr>
<tr>
<td>Browse Local Listings</td>
<td>0.00</td>
<td>0.00</td>
<td>1.40</td>
<td>0.25</td>
</tr>
<tr>
<td>Search Near SJSU</td>
<td>0.00</td>
<td>0.00</td>
<td>0.09</td>
<td>0.76</td>
</tr>
<tr>
<td>Sign Up</td>
<td>0.00</td>
<td>0.00</td>
<td>0.53</td>
<td>0.47</td>
</tr>
<tr>
<td>Search for an Artist</td>
<td>0.00</td>
<td>0.00</td>
<td>1.09</td>
<td>0.31</td>
</tr>
<tr>
<td>Track More Artists</td>
<td>0.00</td>
<td>0.00</td>
<td>1.22</td>
<td>0.28</td>
</tr>
<tr>
<td>View Tracked Artists</td>
<td>0.00</td>
<td>0.00</td>
<td>1.74</td>
<td>0.20</td>
</tr>
<tr>
<td>View Upcoming Shows</td>
<td>0.00</td>
<td>0.00</td>
<td>0.35</td>
<td>0.56</td>
</tr>
<tr>
<td>Mark Attending</td>
<td>0.00</td>
<td>0.00</td>
<td>1.02</td>
<td>0.32</td>
</tr>
<tr>
<td>Track Local Events</td>
<td>0.00</td>
<td>0.00</td>
<td>4.01</td>
<td>0.05</td>
</tr>
<tr>
<td>View Attending Events</td>
<td>0.00</td>
<td>0.00</td>
<td>1.85</td>
<td>0.18</td>
</tr>
<tr>
<td>Import Bands</td>
<td>0.00</td>
<td>0.00</td>
<td>8.60</td>
<td>0.01</td>
</tr>
</tbody>
</table>

*Table 5 - Time on Task Statistics*
DISCUSSION

The results from the local and remote usability studies conducted in this experiment suggest that each method yields similar results, both in terms of performance-based and satisfaction-based metrics. There were no statistically significant differences in the usability ratings or the number of errors between the two experiment groups. The Search Near SJSU task stood out as the task with the lowest usability ratings, falling well below expectations for both local and remote participants. While there were two tasks with significant differences in time on task, they could be explained by differences in task interpretation and equipment.

However, there were differences in the overall rating trends between the testing methods. Local participants tended to have higher expected usability ratings than the remote participants. Remote participants tended to have higher actual usability ratings than the local participants. When looking at the comparisons of expected and actual ratings rather than each set of ratings on their own, different results are seen depending on which test method is being evaluated. As shown in Figure 11 and Figure 12, the local participants generally evaluated the tasks as below expectations, while the remote participants generally evaluated the tasks as above expectations. It can be inferred from these results that the local participants were not as happy with their experience with the product as the remote participants were. It is interesting to note that the remote participants also rated their overall testing experience higher than the local participants.

There are many factors that may be affecting the difference in expected versus actual ratings between the local and remote participants. It is possible that remote participants are more comfortable due to participating from their own home on their own computer. This may reduce their stress levels and lead to higher actual ratings. Their familiarity with their computer hardware and software may also lead to the perception of the tasks being relatively easy to
accomplish. The local participants may be struggling to get familiar with their test equipment, increasing their cognitive load, and transferring some of that struggle into their actual usability ratings. Local participants may feel an obligation to be more critical in their evaluation given the more formal setting of the in-person setup. The presence of the moderator in the local testing method may also be influencing the participants’ ratings.

**Limitations**

It should be noted that there are some limitations to the results from this experiment. Due to the number of participants (30), there was a relatively low power factor of 0.56. There were also some inconsistencies in the local testing setup, as the sessions were conducted in two different locations. While the conditions were very similar in both locations, it is possible that this could have affected the results. The participants were all located in the San Francisco Bay Area and may not be representative of the general population due to the region’s highly technical culture. The product evaluated in this experiment was a website. It is not clear whether the same trends would be seen when evaluating other types of products in both local and remote testing settings.

**Future Research**

The results from this experiment suggest that local and remote usability testing methods produce similar findings. Further studies should be conducted to determine whether the pattern of higher expected ratings from local participants and higher actual ratings from remote participants remains consistent. If this pattern is found in additional studies, research into why this is occurring would be of great interest. Comparing local and remote usability testing methods on a variety of products would provide additional insight into the effects of the testing methods. In this experiment, it was easy to have remote participants share their screen in order
to observe their actions. It is more difficult to set up remote observations with mobile devices and physical products. A more complicated remote setup may negatively affect the remote participants’ ratings. Additional research comparing testing methods as well as different product types would be helpful in determining the impact of each factor.
REFERENCES


doi:10.1145/259963.260531


APPENDIX A. Informed Consent Form

Agreement to Participate in Research

Responsible Investigator: Scott Wheelwright, San Jose State University Graduate Student

Title of Study: Comparison of Local and Remote Moderated Usability Testing Methods

1. You have been asked to participate in a research study investigating the difference in outcomes between two usability testing methods.
2. You will be asked to perform tasks on a laptop computer. Your voice and your laptop screen will be recorded. The research session will take approximately 1 hour.
3. No risks are anticipated in the research.
4. This research will contribute to the general understanding of effectiveness of local and remote usability testing methods.
5. Although the results of this study may be published, no information that could identify you will be included. All individualized data collected during the research study will be stored in a password protected laptop accessible only to the primary investigator. All video images will deleted upon completion of the data analysis. No video images of you will be published nor disseminated.
6. Compensation for participating in this research study will be provided by the SJSU Psychology department in the form of research participation credit.
7. Questions about this research may be addressed to Scott Wheelwright at scott_wheelwright@yahoo.com. Complaints about this research may be presented to Dr. Louis Freund, Director, Graduate Program in Human Factors and Ergonomics, Department of Industrial and Systems Engineering, San Jose State University, at (408) 924-3890. Questions about a research subject’s rights or research-related injury may be presented to Pamela Stacks, Ph.D., Associate Vice President, Graduate Studies and Research, at (408) 924-2427.
8. No service of any kind, to which you are otherwise entitled, will be lost or jeopardized if you choose not to participate in the study.
9. Your consent is being given voluntarily. You may refuse to participate in the entire study or in any part of the study. You have the right to not answer questions you do not wish to answer. If you decide to participate in the study, you are free to withdraw at any time without any negative effect on your relations with San Jose State University.
10. At the time that you sign this consent form, you will receive a copy of it for your records, signed and dated by the investigator.

☐ The signature of a subject on this document indicates agreement to participate in the study.
☐ The signature of a researcher on this document indicates agreement to include the above named subject in the research and attestation that the subject has been fully informed of his or her rights.

______________________________  ______________________________
Participant’s Signature          Date

______________________________  ______________________________
Investigator’s Signature         Date
APPENDIX B. Test Protocol

Instructions:

Today you will be rating usability/user experience proportionally. This means that every rating you give must be related to those that came before it.

For example, if you give a rating of 10 to one task, and 20 to the next, the second task is about twice as usable as the first. Similarly, a task rating of 5 would be half as usable as the first task.

Remember, usability is your perception of how easy to use, well designed, and productive an interface is to complete tasks.

Tasks:

1. Navigate and Explore
Open a web browser (Chrome or Firefox) and navigate to www.songkick.com. Take a few minutes to explore the site as you normally would.

2. Browse Local Listings
Browse the local concert listings.

3. Search near SJSU
Now search for concerts near SJSU.

4. Sign Up
Sign up for an account with Songkick.

5. Search for an artist
Think of a band or artist that you would like to see. Search for that artist on the site, then track that artist’s concert listings.

6. Track more artists
Find at least two more artists that you are interested in and track their concert listings.
7. View tracked bands
View a list of all the artists that you are tracking.

8. View upcoming shows
Now view all of the upcoming concerts for the bands that you are tracking.

9. Mark attending
Mark at least three events as if you will be attending them. These can be for any event, not just for the artists that you are tracking.

10. Track local event
Track at least three local events that you’d like to know about.

11. View attending event
Now view all of the events that you plan to attend.

12. Import bands
Import bands and artists from iTunes, Facebook, Pandora, or last.fm.

Overall
How would you rate the usability of the testing experience you have just completed? This is not referring to the Songkick website, but rather the entire process that we have just gone through together (completing tasks, answering questions, etc).

Google
Using the same rating scale, how would you rate the usability of Google?

Amazon
Using the same rating scale, how would you rate the usability of Amazon?

Google Maps
Using the same rating scale, how would you rate the usability of Google Maps?