



Published in final edited form as:

J Biomed Inform. 2015 August ; 56: 284–291. doi:10.1016/j.jbi.2015.06.001.

The Use of Think-Aloud and Instant Data Analysis in Evaluation Research: Exemplar and Lessons Learned

Jonathan Joe^{a,*}, Shomir Chaudhuri^a, Thai Le^a, Hilaire Thompson^b, and George Demiris^{a,b}

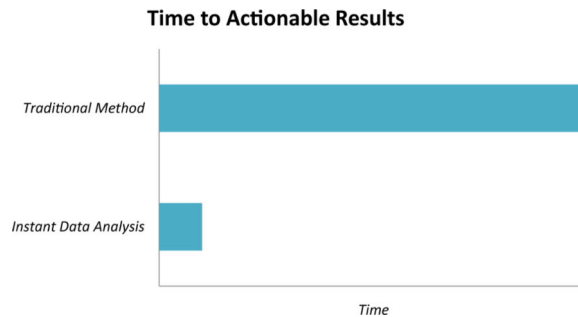
^aBiomedical and Health Informatics, University of Washington, Box SLU-BIME 358047, 850 Republican St, Building C, Seattle, WA 98109, USA

^bSchool of Nursing, University of Washington, Box 357266, 1959 NE Pacific Street, Seattle, WA 98195-7240, USA

Abstract

While health information technologies have become increasingly popular, many have not been formally tested to ascertain their usability. Traditional rigorous methods take significant amounts of time and manpower to evaluate the usability of a system. In this paper, we evaluate the use of instant data analysis (IDA) as developed by Kjeldskov et al. to perform usability testing on a tool designed for older adults and caregivers. The IDA method is attractive because it takes significantly less time and manpower than the traditional usability testing methods. In this paper we demonstrate how IDA was used to evaluate usability of a multifunctional wellness tool, discuss study results and lessons learned while using this method. We also present findings from an extension of the method which allows the grouping of similar usability problems in an efficient manner. We found that the IDA method is a quick, relatively easy approach to identifying and ranking usability issues among health information technologies.

Graphical Abstract



*Corresponding author: Jonathan Joe, jjoe@u.washington.edu, 206-569-8403.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Conflicts of interest

The authors declare that there are no conflicts of interest.

Keywords

Instant Data Analysis; Usability Testing; Usability Research; Older Adults; Wellness Tools

1 Introduction

Usability testing is an important component in design, as it aims to assess ease of use and identify learnability issues within a tool. Performing usability testing typically involves users of the target user group, as these “real” users may think or act differently than expected by the designers or developers[1]. Often, these issues are only identified when testing with real users, reinforcing the importance of doing “real world” usability testing. Furthermore, this testing can be done during early stages of development, leading to easier and cheaper fixes compared to finding issues after the product has been built and released. However, the usability testing process can be time consuming and labor intensive, which may lead designers to omit testing, as the upfront cost is perceived to be too high even though the process could be useful. Instant data analysis may be one solution to address this challenge providing real world testing while reducing the time and labor involved.

1.1 Usability Testing

Traditional usability testing involves a think-aloud protocol combined with a video recording of a user from the target group while they interact directly with the device or tool in question to complete specified tasks[2–4]. This recorded video allows for observation of the user to identify points of frustration, confusion or other issues. The video is transcribed and often analyzed qualitatively or referenced for issues. These issues are then reconciled between researchers and scored by severity, depending on the frequency of the issue and how much it delayed or frustrated the user on completing the tasks. While such observational analysis identifies what causes the user to be frustrated or delayed, the reason or *why* this causes frustration is not evident. In order to better understand the users’ thought process, this observational method is often combined with a think-aloud protocol. A think-aloud protocol asks the user to verbalize their thoughts as they perform the tasks required in a usability test giving insight into their mental model, and has its roots in Ericsson and Simon’s work[1,5]. With these data, researchers can then examine the differences between the participants’ mental model and the system’s interaction model to identify errors and changes that need to be made. These thoughts can address what users like, what they dislike or how to improve the interface and tool from their perspective. Combining these two techniques with qualitative analysis of a transcript comprises the traditional method for usability testing. At the end of the analysis, researchers or designers are able to generate a list of usability issues and a related a score/severity ranking for each issue. Such usability tests have been used successfully to assess the usability of home-based telemedicine systems[6], medical diagnostic and research tools[7], and online self management tools[8], among others[9–12].

Traditional usability testing however, is not without its own challenges. While such an approach is very thorough, it can require significant amounts of manpower and time. Transcription of user comments and verbalizations, along with specifying user actions in

relation to the interface can require significant amounts of manpower, which is then followed up by qualitative coding and analysis. Thus, the time between when the actual usability tests occur and when the final results are generated can span several weeks. For example, Jeffries et al.'s empirical usability study, with 6 users each participating in a 2 hour usability session took 199 man-hours to analyze[13]. This may delay or discourage system improvements.

Other methods, such as heuristic evaluation, rely on usability experts to compare a system against usability principles, in order to hopefully avoid major usability issues[14–17]. Once a device or application has been through a heuristic evaluation, various aspects of the tool will have been judged to be either in or out of compliance with recognized usability heuristics[18]. From this analysis, changes can be made to bring the device or application into compliance, and avoid user frustration. While this can save time compared to conducting the usability tests and can form an important component of the design lifecycle for tools, it lacks interaction between the system and real users. Additionally it is based on the expert's assumptions about user needs and preferences, rather than the users' perspective. Users may interact differently with the system than expected by the usability expert, with the result being many unidentified usability problems. Furthermore, the fact that multiple expert evaluators are needed to do a heuristic evaluation can be challenging within a single organization [19]. Heuristic evaluation can therefore be a useful complement to traditional usability testing, but is not a direct replacement.

1.2 Instant Data Analysis

Instant data analysis (IDA) aims to reduce the labor and time commitment required to perform and analyze a usability test[20]. In IDA, multiple individual sessions are held on a single day. After sessions are completed, those participating in the evaluation meet to discuss the usability issues that were identified. Meeting directly after the sessions allows a better recall of the events, and allows thoughts and ideas that may not be at the forefront of one's memory to be prompted by the other person involved. The idea behind this initial brainstorming session is to list as many usability issues remembered or seen down on paper. After these issues are exhausted, they are ranked based on severity and frequency with which the issue arose. This method is designed to make usability testing more accessible while retaining the advantages of "real" user testing by cutting down on the amount of time needed for analysis [20]. The majority of time involved in usability testing goes into understanding what issues were identified during the tests. Instant data analysis reduces the amount of time needed for analysis significantly, potentially allowing results to be seen the same day as the usability testing sessions. Previous studies have shown that using IDA can reduce the amount of time needed for analysis by 90%, while achieving 85% overlap in critical usability issues compared to the traditional standard video analysis, while a second study found 76% overlap between the two methods [20,21]. However, this method is relatively novel. To date, it has been used successfully to improve the design of medication lists to reduce adverse drug events, personal health applications, and electronic meeting support systems[9,22–25].

This paper details our experiences using the novel IDA method together with analysis mapping methods. We use an exemplar of this method in the evaluation of a multifunctional wellness tool designed for older adults. We provide insight into the feasibility of the IDA method and discuss our experiences and insights of this method to inform future researchers, designers and other stakeholders who evaluate the usability of technology tools.

2 Case Exemplar

The number of adults aged 65 or older in the United States is projected to grow quickly over the next few decades, climbing from 40 million in 2010 to 72 million by 2030[26]. As people age, they are more likely to have health issues and multiple comorbidities, leading to an increased need for health interventions[27] while the healthcare workforce is not increasing at a similar rate. Information technology is emerging for the delivery of health related interventions targeting both health maintenance and disease management. While the use of technologies has generally grown, the usability of these technologies have lagged for older adults, who have their own unique needs[28,29]. Usability concerns will play a larger role, potentially leading to greater user dissatisfaction and reduced effectiveness.

This paper is based on a pilot study for testing the usability of a multifunctional, commercially available wellness tool for older adults, hereafter referred to as “device A” using IDA as the usability testing approach. The purpose of the pilot was to evaluate and assess usability issues with the device in an older adult population. Older adult participants (N=5) were recruited at an independent retirement community via information sessions. Participants could not have had prior exposure to the device to be evaluated. All participants conducted usability sessions individually, and were given 3 tasks to complete using the device.

2.1 Design

Usability testing was accomplished with a think-aloud protocol that asks users to verbalize their thoughts as they complete various tasks, allowing investigators to gain insight on participants’ thought processes in relation to the interface and task[1]. Sessions included a single participant and a facilitator and designated note-taker, who observed and took notes as the participant worked through the various tasks. Testing involved a short questionnaire which asked about demographics, eHealth literacy (eHEALS) [30] and other technology use questions, followed by 3 tasks for the participants to work through. A brief post-session interview was then conducted to solicit further feedback regarding their overall impressions of the system, suggestions for improvement, and any particular frustrations they wanted to emphasize. The University of Washington institutional review board approved all procedures in this study.

2.2 Device

This study focuses on usability testing a commercially available multifunctional wellness tool, Device A. Device A is a multifunctional, touchscreen wellness tool installed in over a thousand communities across the US. It has features that were selected to address many different dimensions of wellness, including social wellness (email, video chat, reminiscence

features), cognitive wellness (brain exercises, puzzles), spiritual wellness (videos, relaxation), and physical wellness (exercise videos, aerobics), among many others.

Physically, the device consists of a touchscreen computer, with a keyboard, mouse and speakers on a movable stand. The entire device is mounted to allow user-adjustable height. The main navigation consists of a 3×3 grid, where each point is a button that specifies a category or folder, with a hierarchy that is several levels deep. Generally, there is also persistent navigation along the top to allow users to go back to the previous page and change the volume. The device was developed for senior communities with the activities targeted towards older adults. This particular device was selected for this study due to the popularity of the device; however there was sparse published information regarding usability available.

2.3 Procedures

Since we were interested in first-time use and learnability[31], participants were not to have used or seen the device before, as assessed verbally by the researcher. Following informed consent, participants were asked to complete a questionnaire which asked general demographic questions such as age and education, questions about eHealth literacy (via the eHEALS instrument), and other technology usage questions such as how often they used a mobile phone or computer[32].

Participants were introduced to the system and walked through some brief example tasks to understand how the system worked and familiarize themselves with the think-aloud protocol. Participants were guided through the evaluation by a facilitator, who was responsible for prompting thoughts from the participants if they stopped thinking aloud. The facilitator was also responsible for keeping the sessions on track, and intervening when needed if the participant was excessively frustrated[33]. The next task was presented when participants indicated that they thought they had completed the task or if they did not feel that they could complete the task. A second researcher served as the note-taker, recording issues, frustrations, and comments made by the participants during the session. The note-taker observed the participant and participants' actions and thoughts without directly interacting with them.

Participants were given 3 representative tasks depicting a range of difficulties and applications within the interface and were to: 1) play music, 2) read their home newspaper, and 3) play tic-tac-toe and then watch a relaxing waterfall video and aquarium application. These tasks were selected to be a spectrum of difficulties, from easy to hard to complete. Since leisure activities have been associated with slower cognitive decline, these activities fit well within the context of a wellness tool[34,35]. Participants were asked to complete these tasks navigating through the device's interface while thinking aloud to give insight into their thought processes and their thoughts on what they liked, what was confusing, and where they thought they needed to go within the interface next to accomplish the task at hand along with other feedback. Throughout the process, participants gave their thoughts on the difficulties they were experiencing, where things did not match their mental model and suggestions for improvement. To encourage honest feedback and thoughts on the system as it was being used, researchers assured the participants that the device was being tested, not them. The session concluded with an exit-interview asking for additional comments from

participants that they did not already verbalize during the sessions. This included what aspects of the system they found particularly frustrating, the utility of the system from their perspective as well as any suggestions for improvement.

2.4 Instant Data Analysis

Sessions were analyzed via IDA. To complete the IDA, initial brainstorming occurred at the end of each day to identify usability issues observed. Each issue was ranked as critical (unable to complete task), severe (significant delay or frustration in task completion), or cosmetic (minor issues). Each of these issues was then annotated with specific, clear references to the interface and other notes giving additional detail on the problem and participants' reactions.

2.5 Affinity Mapping

While the ranked list generated by IDA serves the purpose of identifying individual issues, we sought to gain a broader understanding on the major types of issues that were causing problems. In order to do this, we separated out all the issues and aggregated them into larger themes using affinity mapping once all sessions were completed[36]. The inductive process looks at all the issues as a whole, by aggregating like issues together until all of the issues have been sorted into groups. By keeping all of the issues on separate pieces of paper, it is feasible to re-categorize and regroup issues as needed as themes emerge. Once all the groups had been sorted, they were then labeled to create larger themes or categories. Thus, at the end of this process, we had identified major themes of usability issues as well as the specific issues associated with each one. This process is a bottom-up, inductive exercise, with categories emerging from the data at large. Using the process with 5 older adult participants, we identified 48 usability issues, which aggregated into 8 major themes. The process worked well for our population, and did not need a significant revision of our protocol to effectively use IDA.

3 Lessons Learned Using IDA

3.1 Think Aloud/Usability Session

The think-aloud process asked participants to verbalize their thoughts, feelings, and frustrations with the facilitator as they worked their way through the tasks. A good facilitator must make sure not to cut off or intervene the participant too early, as this may cause them to give up earlier in the subsequent tasks, or encourage them to look to the facilitator for help in completing the tasks early[1]. Ideally, the participant should act as if they were encountering the device in question within context in real life, where there would be no expert user nearby to offer immediate aid. Thus, it is important for the facilitator to be able to resist helping the user immediately after running into a problem so as to more accurately portray how a new user may act. The facilitator must also decide how much deviation from the task is acceptable, as participants may take a non-direct path to reach their goal in line with their mental model. In order to better standardize the process, we would recommend using the same facilitator through the sessions, if possible. As for the think-aloud process, some participants found it easier than others, and initially explaining the think-aloud process with an example seemed to help. Reminding participants when they became silent also

seemed to help keep them on track, and some participants made suggestions for improvement as they made their way through the tasks. The resulting insights that users provided via think aloud were useful to identify frustrations that may have not been obvious to designers when building the tool. Finally, emphasizing that the device was being tested and not the participant seemed to relax the participants before the tasks were given to them.

3.2 Brainstorming & Scoring

The goal behind brainstorming was to allow the facilitator and notetaker to elicit as many usability issues as could be remembered. We found great utility in conducting this together as it enabled prompting of remembered issues. The issues were generated quickly in the beginning but tended to slow down as time went on, so prompting each other was useful in identifying more issues. Furthermore, brainstorming by both team members who were present at the usability tests enabled researchers to add in detail or fill in gaps in the issues identified by the other member

The actual brainstorming took no more than an hour, followed up by another hour of writing out all the details and references of the interface per day. Compared to the standard of transcription and coding (or video annotation), the IDA method of analysis was much less labor intensive and time-intensive. Our findings were in line with those reported by Kjeldskov's 10-fold reduction in analysis time when comparing the traditional method to instant data analysis methods[20]. Consequently, while the instant data analysis method was likely not as exhaustive as the traditional method, we were satisfied with the number and quality of issues, as well as the immediacy to which we could see all the issues considering the cost-benefit ratio of the extra time the traditional method needs. This same day analysis is easier and more immediate, allowing for quick identification and potential system changes to be generated in the same day. By ranking severity in the same process, it was also easier to identify which issues should be tackled first when deploying limited resources to make changes in the device in question.

However, there were some tradeoffs to IDA that should be kept in mind when selecting a usability testing technique. First, the resulting usability issues list is generally not as exhaustive as the traditional method. Thus, if searching for the maximum number of issues out of a given number of usability sessions, researchers may want to consider using the traditional method instead, although the tradeoff would be greater analysis time. Furthermore, since IDA relies on several sessions to occur in a single day followed immediately by initial analysis, the logistics of scheduling everyone involved, both participants and researchers can be problematic. Depending on the length of each session, the session and analysis could take up a whole or several day(s) contiguously, which may be difficult to accomplish given competing priorities. In some contexts, the availability of representative users may be limited. In these cases, the benefits gained from using representative users should be weighed against the difficulty of scheduling users to test the technology system.

Overall, the use of the IDA method may be a good place to start for organizations wanting to do user testing, but do not have or want to commit the time and resources to traditional user testing. It can be used as a component in an organization's implementation of the human-

centered design framework, and can complement other design techniques and processes to more fully understand the user[37].

3.3 Affinity Mapping

The standard IDA method was extended by using affinity mapping/diagramming to generate larger themes to categorize the issues identified in an inductive manner, rather than have a collection of disparate issues. The process of laying out all the issues to group them together allowed an overview of the bigger issues that could suggest what needs to be done as a focus for future development. It also highlights the fact that similar issues came up multiple times, indicating that problems were not isolated issues. The affinity mapping process added to the results by allowing us able to easily see the number and severity of issues for each theme. Creating an uncategorized section for yet to be sorted issues or issues that did not fit into other larger categories was also effective, so as to not force an issue to be placed in a poorly matching category or with only a tenuous similarity to other issues in that category.

It was useful to have the researchers involved in data collection to be together in the same room to carry out the exercise, as the researchers could discuss the reasons of aggregating issues together, and easily make changes when another researcher brought up a better congregation or match. This allowed agreement across all the researchers involved, and was able to be completed relatively quickly. Furthermore, initially sorting the ideas in silence helped the process along so that each researcher was not influenced. Once the initial sort was mostly complete, discussion occurred to identify the shared meanings of each group, and if any changes should be made to create a better similarity match between issues. This process of silence followed by discussion allowed natural sorting without undue influence from the other researchers, while at the same time allowing consensus to be reached at the end. Moreover, it is important to not allow a single individual to dominate the affinity mapping process, which would not lead to a satisfactory consensual grouping as others involved may feel left out and their opinions not being taken into account. The silence in mapping can dissuade this somewhat, but care must be taken into not having a single person dominate the process. Finally, the affinity mapping process may not be necessary if only a small number of issues are identified (i.e., less than 10 or 15). In this case, it may be possible to skip this step, as well as reflect on if more issues could be identified from sessions.

4 Case Exemplar Results & Discussion

4.1 Demographics

A convenience sample of 5 older adults was recruited for this usability testing study, which is in line with the recommended number of users for usability sessions[38]. The mean age of the participants was 72.4 years (Range: 64–86), with the majority being women (60%). Four participants (80%) identified as white, and one participant identified as mixed racial background (White/American Indian). For their highest completed degree, 2 participants had completed 4 years of college or higher (40%), 2 participants had completed 2 years of college (40%), and 1 participant had completed a high school degree (20%).

4.2 eHealth Literacy Scale (eHEALS) and Technology Use

The mean eHEALS score (Possible range: 8–40, higher score indicates higher electronic self-sufficiency) for participants was 32.8 (range: 21–37). Three participants had high electronic self-sufficiency (60%, Score: 30–40) and the other two participants had moderate electronic self-sufficiency in eHealth (40%: Score: 19–29). The majority of participants (3, 60%) indicated they used a mobile phone, while two participants reported they did not own a mobile phone. Of the three mobile phone users, one each reported use across each of the following categories: use rarely, use moderately, and use frequently. All participants indicated they owned a computer, and most of these participants (n=4, 80%) used their computer “Frequently” while one participant (20%) indicated he used it “rarely.”

4.3 Usability Issues

Among the participants, 48 usability issues were identified, and 19 were ranked as critical (40%), 21 as severe (44%) and 8 as cosmetic (16%). “Critical issues” were defined as those issues that prevented task completion, “severe issues” were defined as those issues that caused significant slowdown or frustration, and “cosmetic issues” were the ones that were left and caused minimal issues. These issues were then grouped into themes using affinity mapping into 8 major categories.

4.3.1 Unintuitive Categorization & Nomenclature—The layout of the device homepage consists of large icons and labels that act as folders for the content lower in the hierarchy. Many participants found the categorization of various applications within the system to be unintuitive. The system uses a multi-level categorization method to separate out the various applications. For example, participants had issues with categories or applications that appeared in multiple places, asking “have I been here before?” They may have recognized some of the icons, but not others. This led to confusion and frustration of users and was categorized as a critical error. Furthermore, the categories were perceived as having an unclear naming scheme. The various category names did not intuitively describe the applications and objects stored inside the folder. Frequently, the applications within a category were not consistent with user expectations. The categorization was challenging for participants to deal with and it “takes [them] awhile to find what [they are] looking for.”

Moreover, participants did not think that the pictures attached to the categorical label matched the content, leaving some participants to wonder which one more accurately reflected the content hidden underneath. Some participants were unaware of what “Skype” was, while Device A seemed to assume that users would know what it is, and left the Skype name as an option without any further explanation. Adding further to the confusion of the participants was the feeling that the categories were not mutually exclusive, and thus were confused about where to move forward to find what they were looking for. Terminology was also unclear, with one participant typing “E-S-C” in response to a prompt to push “ESC” to escape.

4.3.2 Unclear Iconography—An issue that was seen frequently was unclear or confusing iconography. Critical issues included those where the icons could be interpreted as symbols for something else. For example, an icon with a globe intended to represent that “internet

needed for this feature” was taken by participants to be various things, including that the application with the globe icon meant it had to do with the environment or something global. Other icons intended to help distinguish between various features, such as a film icon to indicate it was a playable movie had unintended consequences, since some participants thought that every video had that icon, so non-movie videos were not seen as playable. Other icons were reused for different purposes, causing some confusion on whether the icon represented an actual application or a folder. In addition, zoom icons using the plus or minus icons were unclear to a participant on whether a plus or minus made the text bigger or smaller. Finally, participants found the use of images for categories useful, but often found the images didn’t match what they would expect, causing confusion and the inability to complete tasks.

4.3.3 Unclear Place in Navigation Tree—Many critical issues were related to participants being unclear where they were in the navigation tree. The device is set up so that it has several folders a user can click on and subfolders to organize the applications. Several participants were unclear where they were in the hierarchy, and were unable to successfully navigate between folders to complete the tasks. For example, a participant was stuck in the “entertainment” folder and did not realize he could move up another directory to the home screen, to reach the correct folder. This is also related to the confusion of category labels, in that once participants didn’t remember which category they selected, and were liable to select the same category again when backing out and trying to complete a task, which caused participants significant frustration. Finally, multiple participants were unclear on the concept of a “homepage”, and when directed to do so were unclear on the meaning and why they needed to be there. The confusion led to participants being unable to complete tasks, which is classified as a critical error.

4.3.4 Misuse of Conventions and Misleading Perceived Affordances—The term “affordance” was originally coined by Gibson, to refer to what an object offers to an organism to perform an action[39]. Norman later coined “perceived affordance” which refers to the perception of properties of an object that suggest what actions can be done to the object or how it could be used[40]. Conventions are a learned way to understand or interact with an object established by usage [40]. Breaking conventions or having misleading perceived affordances could greatly increase the frustration and make tasks more difficult to complete. Severe issues included graying out the back button even when it could still be pressed, leading a participant to assume they were on the home page since they associated greyed out with unable to be pushed. This can lead to great delay and frustration in task completion. Another severe issue observed was the lack of clarity when a keyboard input was required versus when a touchscreen input was allowed. A participant switched to a mode of input once from the touchscreen interface to the keyboard, and didn’t switch back to the touchscreen even when the new input method was not allowing her to be successful in her intended task. Moreover, a participant was unable to complete a task since the box to input text was not selected, and it was not clear to the participant that it needed to be. This could have potentially been avoided if the device had stuck to using standard conventions from the web or computing areas. Other conventions were misleading, such as a participant thinking that a white box that looked like a text entry box was actually a progress bar,

leading to her trying to click and type in a search entry unsuccessfully, and thus leading to task failure.

4.3.5 Accessibility Issues—Participants indicated having a touchscreen mounted vertically, such as a computer monitor or television, could lead to fatigue. The constant upward moving motion and placing one’s hand back down could become tiring, and those adults with shoulder issues would not be able to use the touch interface comfortably at all. Furthermore, while Device A was able to raise and lower to different heights, one participant wished the device could move lower. Since the device was unable to accommodate her request, she had to strain to look at the screen and had trouble reading the screen, causing some frustration. The same participant was also left handed, and while Device A allowed the mouse to be moved to the other side of the computer for left-handed access, the participant didn’t see an easy way to switch the orientation of the mouse buttons, so that the primary mouse click was on the right side of the mouse. She said that she wished it would switch automatically, and pressed the incorrect button for her intended action multiple times.

4.3.6 Physical Responsiveness of the Touchscreen Problematic—The touchscreen on Device A allowed easier menu selection, since users could directly touch what they wanted to select. However, the physical responsiveness of the touchscreen was problematic, creating severe issues such as how the delay of touching the screen to selecting on the device was enough that participants ended up pushing multiple times to achieve their intended action. Delays in system feedback due to this caused consternation on the part of the users. The participants also sometimes used a tap and hold gesture rather than a single tap on the screen, which caused the system to not act as the participants intended. Other times, the system failed to pick up the touch at all, causing confusion and increased retouching to make sure the selections were picked up by the system. Participants also had critical issues with the touchscreen. For example, one participant used three fingers to touch the screen, which caused a failure to select the intended object properly. Since she would always touch with 3 fingers, the system often forced her to touch again to select the intended action.

4.3.7 Inability to Exit Consistently—Participants had a difficult time consistently exiting the page or application they were in. For example, even when there was a labeled “close” button, a participant was unable to close the window because the finger’s touch target was off by a few pixels, which led to selection of the wrong portion of the screen. This participant had to try closing many times before clicking the “close” button successfully after a significant delay. Another participant was unable to close the same window successfully, and missed the button labeled “close.” Only after a delay did the participant recognize the button, even saying that it “wasn’t there before.” Other severe issues included the lack of clarity on what the “exit” button did on the always present navigation bar, such as whether the button exited the current application, the entire system, or something else. Participants were also unable to exit full-screen applications, leading to both severe (only exited after frustration or long delay) and critical (unable to exit) classification of the issues, since the navigation bar disappeared in the full-screen scenario.

4.3.8 Volume Issues—Participants indicated that they wished to change the volume but were unable to do so without significant frustration or delay. Participants did not discover the volume option initially, even when they complained that they couldn't hear anything because the system was too quiet. The navigation bar had a volume button to increase and decrease the volume, but it did not seem to be obvious to participants unless prompted by the facilitator. Furthermore, Device A uses hardware speakers with a hardware dial for volume control, but this was not discovered or acted upon by any of the participants.

4.4 Discussion of Case Exemplar

The usability testing study was performed to better understand issues that exist within a multifunctional wellness tool. These findings can inform the development of future wellness tools. Even popular, commercially available multifunctional wellness tools have many usability issues that should be addressed. This study has identified many usability issues that were categorized into 8 major categories. This suggests that designers should carefully consider how the content and organization of their multifunctional wellness tools are presented to older adult users. These themes should be used to inform future development of tools that cause less frustration and potentially happier users. This study also highlights the importance of testing devices with representative users, as even this popular commercial device has many issues that could be remedied to create a better experience.

Cognitive changes related to aging, such as a decline in working memory should be fully considered when designing wellness tools for older adults[41]. Taxing working memory should be minimized when possible. Many participants found issues with navigation, and especially with keeping track of where they were within the navigation tree. Thus, it was unclear how often they could move up a level, and difficult to remember what category they had clicked on to reach the page they were currently on. Difficulties in navigation are in line with previous studies that suggested that reduced working memory made it more difficult for older adults to navigate and use the web[42,43]. Designers could alleviate these issues by creating obvious titles on each page, and creating breadcrumbs to show what level of the hierarchy they are on, as well as how they reached that page. Previous studies have suggested the use of breadcrumbs for positive effects in performance and user satisfaction[44,45]. Alternatively, a designer could think deeply about what needs to be included in the system, and remove unnecessary features to simplify the number of levels and options and reduce the load of navigation on working memory, which is in line with previous research that suggests the use of shallow hierarchies[46]. Designers may also want to consider the external cognition framework, which describes how cognition does not solely occur in the individual, but also relies on external representations in the environment[47,48]. Increasing computational offloading could help a user experiencing cognitive decline continue to successfully use the system, so they will not have trouble using it.

Related to navigation were categorization and nomenclature issues. Many participants found that the categories in which applications were sorted were not memorable and did not match their mental model of how applications should be sorted. When using icons or pictures, designers should validate that the icons are intuitive and represent to users what they represent to designers and wellness tool to minimize confusion. This is underscored by

previous studies that suggest that older adult reluctance to use some technological systems was related to the issue of being able to understand the terminology and symbols used within the system[49,50]. While certain terminology and symbols may be understood by more technologically adept or younger people, if the intended user group is older adults, the use of these terms would not be satisfactory. Our experiences within the study reinforce the importance of user testing with the intended user group of older adults. Furthermore, seeking broad older adult input on categorization and labels to match their mental models would alleviate much of the frustration of the users and could make them more inclined to learn and use a wellness tool.

Another area that spurred frustration among participants pertained to conventions that led them astray or perceived affordances that misled them. If designers are going to use conventions, such as greying out a button to indicate that it can no longer be pressed, it should be consistent with norms that exist today. This would also mean that designers should be aware of what the conventions are, such as a white box usually representing a text box or a greyed button usually meaning it can no longer be pressed. Unless they have a very good reason to do so, interface designers should avoid breaking these conventions to avoid frustrating users.

Accessibility issues, such as physical changes related to aging including chronic conditions such as shoulder pain, should be considered when designing wellness tools for older adults. Participants worried about shoulder pain and fatigue when dealing with touchscreens set vertically, such as in computer monitors, especially for those adults who have chronic shoulder pain. Work should be done to observe users using the device to get a better idea of what range of motion is needed to satisfy them, as well as the positioning and type of input to reduce physical stress and fatigue on users. Participants appreciated the touchscreen, as it was more intuitive as they could touch what they wanted to select rather than a more abstract use of mouse or keyboard. However, the delay in processing their touch or issues caused them to press again causing actions to be taken that were unintended, and the touch points of where participants intended to touch compared to where the system registered touch were not always in sync. Future designers should test the touch interface with real users and make changes as necessary to reduce the burden on the user. Other alternatives could be to use other input methods, such as voice recognition to act as input. The use of voice input as an alternative to mitigate some of the issues older adults have with using technologies has been suggested by previous studies, although voice input comes with its own issues[51–53].

Other issues that arose, such as the inability to exit consistently or change volume as needed suggest that designers should test these features extensively with users early in the design process. Repeated, iterative testing could identify interface problems and facilitate the creation of potential solutions.

Our results are consistent with existing user interface and information architecture design guidelines[14–16]. These guidelines recommend allowing users to recognize what they're looking for rather than recall from memory with regard to systems. This aligns with our suggestion of reducing memory load to help navigation. Furthermore, systems should speak

the user's language to match the user's nomenclature. The match between user language and system language did not occur in our tested device but should be done in future designs. Finally, these guidelines suggest that systems should work in a way that is consistent with user expectations. This also reinforces the need to test with actual users in order to see their mental model of understanding with regard to navigation, nomenclature, categorization, and object function. While ideally all future wellness tools should employ a designer with extensive usability and information architecture experience, we hope that this study's guidelines will be useful for designers without this experience.

5 Conclusion

Minimizing usability issues of tools before releasing to a wide audience can increase user satisfaction and reduce user frustration with these tools. This paper presents the feasibility of and lessons learned while using the IDA method as a quick and less labor-intensive way to do usability testing.

IDA would be most useful to those organizations and designers who would benefit from usability testing from real users, but do not want to commit the time or resources to perform the traditional usability testing methods. The speed at which results are generated and issues identified and lower commitment can be useful for relatively quick feedback on the design of tools as they currently stand. Organizations that currently perform traditional usability testing may want to consider the cost-benefit ratio of the additional and more thorough results from the traditional testing compared to instant data analysis and determine whether the additional commitment of resources and time is worth it. Additionally, it is important to recruit individuals who are similar in nature to the intended users of a tool to get the best results. Finally, when considering the use of instant data analysis, researchers and designers should remember that IDA aims to identify the most critical issues, not all of them.

This method is quicker and less labor-intensive than traditional usability testing methods and leads to results that has many of the benefits of traditional usability testing with end users. Furthermore, the addition of affinity mapping was highly beneficial in the identification of themes and areas for further investigation. When combined with the severity rankings, these methods can lead future designers to correct and/or avoid previous mistakes. Future studies may want to compare the use of instant data analysis with other usability testing and inspection methods to more clearly understand the cost-benefit of each method relative to the other. Our study shows the feasibility of the IDA method for usability testing in analysis in a pilot study with older adults, and the use of the addition of affinity mapping to identify themes as feasible and pragmatic. This study has detailed the use of IDA in a clearly defined methodology with affinity mapping that could be beneficial for researchers who are interested in identifying usability issues with users and wish to attenuate these issues before the next iteration of the application.

Acknowledgements

This study was supported, in part, by the National Library of Medicine Biomedical and Health Informatics Training Grant T15 LM007442.

Bibliography

1. Rubin, J. Handbook of usability testing: how to plan, design, and conduct effective tests. John Wiley & Sons; 1994.
2. Nayak N, Mrazek D, Smith D. Analyzing and communicating usability data: now that you have the data what do you do? a CHI'94 workshop. ACM SIGCHI Bull. 1995; 27
3. Kushniruk AW, Patel VL. Cognitive computer-based video analysis: its application in assessing the usability of medical systems. Medinfo. 1995; 8(Pt 2):1566–1569. [PubMed: 8591502]
4. Mack RL, Lewis CH, Carroll JM. Learning to use word processors: problems and prospects. ACM Trans Inf Syst. 1983; 1:254–271.
5. Ericsson KA, Simon HA. Verbal reports as data. Psychol Rev. 1980; 87:215–251.
6. Kaufman DR, Patel VL, Hilliman C, Morin PC, Pevzner J, Weinstock RS, et al. Usability in the real world: assessing medical information technologies in patients' homes. J Biomed Inform. 2003; 36:45–60. [PubMed: 14552846]
7. Bond RR, Finlay DD, Nugent CD, Moore G, Guldenring D. A usability evaluation of medical software at an expert conference setting. Comput Methods Programs Biomed. 2014; 113:383–395. [PubMed: 24210868]
8. Simon ACR, Holleman F, Gude WT, Hoekstra JBL, Peute LW, Jaspers MWM, et al. Safety and usability evaluation of a web-based insulin self-titration system for patients with type 2 diabetes mellitus. Artif Intell Med. 2013; 59:23–31. [PubMed: 23735522]
9. Siek, Ka; Khan, DU.; Ross, SE.; Haverhals, LM.; Meyers, J.; Cali, SR. Designing a personal health application for older adults to manage medications: a comprehensive case study. J Med Syst. 2011; 35:1099–1121. [PubMed: 21562730]
10. Harris LT, Tufano J, Le T, Rees C, Lewis Ga, Evert AB, et al. Designing mobile support for glycemic control in patients with diabetes. J Biomed Inform. 2010; 43:S37–S40. [PubMed: 20937484]
11. Luckmann R, Vidal A. Design of a handheld electronic pain, treatment and activity diary. J Biomed Inform. 2010; 43:S32–S36. [PubMed: 20937483]
12. Neri PM, Pollard SE, Volk La, Newmark LP, Varugheese M, Baxter S, et al. Usability of a novel clinician interface for genetic results. J Biomed Inform. 2012; 45:950–957. [PubMed: 22521718]
13. Jeffries, R.; Miller, JR.; Wharton, C.; Uyeda, K. User interface evaluation in the real world: a comparison of four techniques; Proc. SIGCHI Conf. Hum. factors Comput. Syst. Reach. through Technol. - CHI '91; 1991. p. 119-124.
14. Nielsen J. 10 Usability Heuristics for User Interface Design 1995. <http://www.nngroup.com/articles/ten-usability-heuristics/>.
15. Tognazzini B. First Principles of Interaction Design (Revised & Expanded). AskTog. 2014:1–38. asktog.com/atc/principles-of-interaction-design/.
16. Brown D. Eight principles of information architecture. Bull Am Soc Inf. 2010; 36:30–34.
17. Nielsen, J. Finding usability problems through heuristic evaluation; Proc SIGCHI Conf Hum Factors Comput Syst - CHI '92; 1992. p. 373-380.
18. Nielsen, J.; Molich, R. Proc. SIGCHI Conf. Hum. factors Comput. Syst. Empower. people - CHI '90. New York, New York, USA: ACM Press; 1990. Heuristic evaluation of user interfaces; p. 249-56.
19. Jeffries R, Desurvire H. Usability testing vs. heuristic evaluation. ACM SIGCHI Bull. 1992; 24:39–41.
20. Kjeldskov, J.; Skov, MB.; Stage, J. Proc. third Nord. Conf. Human-computer Interact. - Nord. '04. New York, New York, USA: ACM Press; 2004. Instant data analysis; p. 233-240.
21. Best, MA De; Deurzen, K Van; Geels, M.; van Oostrom, F.; Pries, J.; Scheele, B. Quick and clean, fast and efficient usability testing for software redesign. Proc. CHI Ned. 2009; 2009:149–152.
22. Khan, DU.; Siek, Ka; Meyers, J.; Haverhals, LM.; Cali, S.; Ross, SE. Designing a personal health application for older adults to manage medications; Proc ACM Int Conf Heal Informatics - IHI '10; 2010. p. 849

23. Lawton K, Skjoet P. Assessment of three systems to empower the patient and decrease the risk of adverse drug events. *Stud Heal Technol Inf.* 2011
24. Zurita, G.; Baloian, N.; Baytelman, F.; Morales, M. A Gestures and Freehand Writing Interaction Based Electronic Meeting Support System with Handhelds. In: Meersman, R.; Tari, Z., editors. *Move to Meaningful Internet Syst.* Vol. 2006. CoopIS, DOA, GADA, ODBASE, Berlin: Springer; 2006. p. 679-696.
25. Siek, Ka; Ross, SE.; Khan, DU.; Haverhals, LM.; Cali, SR.; Meyers, J. Colorado Care Tablet: the design of an interoperable Personal Health Application to help older adults with multimorbidity manage their medications. *J Biomed Inform.* 2010; 43:S22–S26. [PubMed: 20937480]
26. US Census Bureau. *The Next Four Decades: The Older Population in the United States: 2010 to 2050.* 2010
27. Anderson G, Horvath J. The growing burden of chronic disease in America. *Public Health Rep.* 2004; 119:263–270. [PubMed: 15158105]
28. Nahm E-S, Preece J, Resnick B, Mills ME. Usability of Health Web Sites for Older Adults. *CIN Comput Informatics, Nurs.* 2004; 22:326–334.
29. Hawthorn D. Possible implications of aging for interface designers. *Interact Comput.* 2000; 12:507–528.
30. Norman CD, Skinner HA. eHEALS: The eHealth Literacy Scale. *J Med Internet Res.* 2006; 8:e27. [PubMed: 17213046]
31. Grossman, T.; Fitzmaurice, G.; Attar, R. Proc. 27th Int. Conf. Hum. factors Comput. Syst. - CHI 09. New York, New York, USA: ACM Press; 2009. A survey of software learnability: Metrics Methodologies, and Guidelines; p. 649
32. Norman CD, Skinner Ha. eHEALS: The eHealth Literacy Scale. *J Med Internet Res.* 2006; 8:e27. [PubMed: 17213046]
33. Barnum, C. *Usability Testing Essentials: Ready, Set . . Test!.* Burlington, MA: Elsevier, Inc; 2011.
34. Scarmeas N, Stern Y. Cognitive reserve and lifestyle. *J Clin Exp Neuropsychol.* 2003; 25:625–633. [PubMed: 12815500]
35. Scarmeas N, Levy G, Tang MX, Manly J, Stern Y. Influence of leisure activity on the incidence of Alzheimer's disease. *Neurology.* 2001; 57:2236–2242. [PubMed: 11756603]
36. Martin, B.; Hanington, B. *Universal Methods of Design.* 2nd ed.. Beverly: Rockport Publishers; 2012.
37. *Ergonomics of human system interaction - Part 210: Human-centered design for interactive systems (formerly known as 13407).* 2010 ISO FDIS 9241-210.
38. Nielson J. Why You Only Need to Test with 5 Users. Neilson Norman Gr Alertbox. 2000 <http://www.nngroup.com/articles/why-you-only-need-to-test-with-5-users/>.
39. Gibson JJ. The Theory of Affordances. *Perceiving, Acting, and Knowing*, vol. *Perceiving.* 1977; (332):127–142. doi:citeulike-article-id:3508530.
40. Norman, D. *The Design of Everyday Things.* New York: Doubleday; 1990.
41. Dobbs AR, Rule BG. Adult age differences in working memory. *Psychol Aging.* 1989; 4:500–503. [PubMed: 2619956]
42. Mead SE, Spaulding VA, Sit RA, Meyer B, Walker N. Effects of Age and Training on World Wide Web Navigation Strategies. *Proc Hum Factors Ergon Soc Annu Meet.* 1997; 41:152–156.
43. Becker SA. A study of web usability for older adults seeking online health resources. *ACM Trans Comput Interact.* 2004; 11:387–406.
44. Kurniawan SH, King a, Evans DG, Blenkhorn PL. Personalising web page presentation for older people. *Interact Comput.* 2006; 18:457–477.
45. Maldonado, Ca; Jlesnick, ML. Do Common User Interface Design Patterns Improve Navigation? *Proc Hum Factors Ergon Soc Annu Meet.* 2002; 46:1315–1319.
46. Gappa H, Nordbrock G. Applying Web accessibility to Internet portals. *Univers Access Inf Soc.* 2004; 3:80–87.
47. Scaife M, Rogers Y. External cognition: how do graphical representations work? *Int J Hum Comput Stud.* 1996; 45:185–213.
48. Rogers Y. New theoretical approaches for HCI. *Annu Rev Inf Sci Technol.* 2004; 38:87–143.

49. Holzinger, A.; Searle, G.; Auinger, A.; Ziefle, M. Informatics as semiotics engineering: Lessons learned from design, development and evaluation of ambient assisted living applications for elderly people. In: Stephanidis, C., editor. Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics). Vol. 6767. Berlin: Springer Berlin Heidelberg; 2011. p. 183-192.LNCS
50. Aula A. User study on older adults' use of the Web and search engines. *Univers Access Inf Soc.* 2005; 4:67-81.
51. Dias MS, Pires CG, Pinto FM, Teixeira VD, Freitas J. Multimodal user interfaces to improve social integration of elderly and mobility impaired. *Stud. Health Technol. Inform.* 2012; 177:14-25. [PubMed: 22942026]
52. Smith MW, Sharit J, Czaja SJ. Aging, motor control, and the performance of computer mouse tasks. *Hum Factors.* 1999; 41:389-396. [PubMed: 10665207]
53. Maguire, M.; Osman, Z. Designing for Older Inexperienced Mobile Phone Users. In: Stephanidis, C., editor. *Proc. HCI Int.* 2003. Mahwah, New Jersey: Lawrence Erlbaum Associates; 2003. p. 22-27.

Highlights

1. We present the use of a quick method for usability testing with a case exemplar.
2. The method presented in this study gives good insight into usability issues.
3. Others can benefit from this method as it's faster and requires less commitment.