Chapter 14
Why and How Think-Alouds with Older Adults Fail: Recommendations from a Study and Expert Interviews

Rachel L. Franz, Barbara Barbosa Neves, Carrie Demmans Epp, Ronald Baecker and Jacob O. Wobbrock

14.1 Introduction

Older adults (aged 65+) constitute the fastest growing age group in most nations, projected to double by 2050 (World Population Ageing 2017 2017). Concurrently, older adults in developed countries are adopting new technologies (Anderson and Perrin 2017). Yet, older adults are still less likely to adopt new technologies compared to other age groups, and they are more likely to discontinue use with age (Berkowsky et al. 2015). One of the factors influencing adoption and use is the usability of technologies (Neves et al. 2015).

Understanding how to adapt usability testing methods to the needs of older adults is essential to developing innovative technologies that serve this population. Contrary to popular belief, the challenges of using technology do not appear to be a current generational challenge. Rather, they are a life course one: as we age, our skills, needs, and aspirations change and we are more likely to face cognitive and

R. L. Franz (✉) · J. O. Wobbrock
University of Washington, Seattle, WA, USA
e-mail: franzrac@uw.edu

J. O. Wobbrock
e-mail: wobbrock@uw.edu

B. B. Neves
University of Melbourne, Melbourne, Australia
e-mail: barbara.barbosa@unimelb.edu.au

C. D. Epp
University of Alberta, Edmonton, AB, Canada
e-mail: cdemmansepp@ualberta.ca

R. Baecker
University of Toronto, Toronto, ON, Canada
e-mail: ron@taglab.ca

S. Sayago (ed.), *Perspectives on Human-Computer Interaction Research with Older People*, Human–Computer Interaction Series,
https://doi.org/10.1007/978-3-030-06076-3_14

functional decline (Neves et al. 2018a). Cognitive and functional decline can lead to frailty, which is defined as having three or more of the following: low physical activity, muscle weakness, slowed performance, fatigue, and involuntary weight loss (Torpy et al. 2006). Frail older adults are likely to stop using technology; authors suggest that abandonment is related to poor usability of technology (Gell et al. 2015). Poor usability also influences older adults' self-efficacy and anxiety surrounding technology (Vroman et al. 2015).

Although the suitability of design methods for older adults has been investigated (e.g., participatory design) (Vines et al. 2012), little attention has been paid to the suitability of usability testing methods for frail older adults. If usability testing methods do not capture the usability issues that frail older adults experience, it is likely these technologies will be abandoned. Despite the employment of usability testing methods with older adults, these methods have not been directly compared. Undoubtedly, individual researchers have learned some of the challenges of these methods, but to the best of our knowledge, there are no systematic accounts that compare usability testing methods with frail older adults.

Given the lack of research on how to effectively employ usability testing methods with frail older adults, we conducted a study to compare three variants of a commonly employed usability testing method, the think-aloud (Ericsson and Simon 1980). We compared Concurrent Think-Aloud, Retrospective Think-Aloud, and Co-discovery with frail older adults (study 1). Additionally, we interviewed Human-Computer Interaction (HCI) experts working with older adults to gain insight into their usability testing practices (study 2).

The think-aloud, based on the verbal protocol by Ericsson and Simon (1980), is the most established usability evaluation. The most common variant is the Concurrent Think-Aloud (CTA), in which participants verbalize their thought process while performing tasks to give insight into their mental model of the system. In another variant, the Retrospective Think-Aloud (RTA), participants perform tasks silently and then think aloud while recalling how they performed those tasks. Co-discovery (CD) involves two participants working as a team to complete tasks: as they interact with each other they also verbalize their thought processes.

Our study with frail older adults showed that Co-discovery is the most suitable think-aloud variant because with it (and only it), older adults verbalized their thought process throughout the entire usability test. We also found that older adults perform impression management during usability tests (i.e., try to present themselves in a favorable light). Our findings from interviews with HCI experts suggest that usability tests can be leveraged to enhance/maintain participants' motivation to engage with technology in their daily lives. Experts also warned that participants' impression management efforts and low self-efficacy can impact the usability test by making the participant appear less competent.

Taken together, our results support a contextual approach—i.e., a perspective that includes context as both a core variable and framework—to usability studies. This approach considers the complexity of working with a heterogeneous but vulnerable older population and includes the need to understand users and empower them

in usability studies through adjusted and user-sensitive multi-methods. Thus, this contextual perspective contributes to methodological innovation in HCI, advancing methodological eclecticism (multi-methods) and flexibility when studying sensitive populations (Neves et al. 2018b).

14.2 Deployment with Frail Older Adults

To compare the three usability testing methods with frail older adults, we conducted a usability test with 12 participants in the context of a three-month technology deployment study (Neves et al. 2017). Participants experienced one of three usability testing methods: Concurrent Think-Aloud (CTA), Retrospective Think-Aloud (RTA), or Co-discovery (CD).

14.2.1 Participants

We recruited participants in collaboration with staff from a care home. We jointly organized two information sessions for residents and relatives. Staff then approached interested residents who met our selection criteria, excluding those with dementia or unable to consent. Twelve participants completed the study (see Table 14.1); their enrollment was staggered, occurring over six months. Participants were considered frail, having visual, hearing, speech, motor and cognitive impairments.

This purposive sample was appropriate to our in-depth study, particularly of institutionalized older adults with different impairments and frailty levels.

Two participants, a married couple (P8a and P8b), wanted to share a tablet and used the same device throughout the study. Participant ages ranged from 74 to 95 (M = 82.6; SD = 5.5). Most had a college degree, but different levels of digital literacy.

14.2.2 Apparatus

The technology used during the usability study was an iPad-based accessible email client that was designed with older adults who had motor impairments and low digital literacy. The tool was based on participatory design sessions and field studies. It had four main messaging options: picture, short preset text, video, and audio. The user could also receive text, video, and picture messages. Contacts were associated with photos in an alphabetized list and the interface was icon-based.

Table 14.1 Demographics of deployment participants and usability testing method experienced: Concurrent Think-Aloud (CTA), Retrospective Think-Aloud (RTA), Co-discovery (CD)

Code	Age	Gender	Previous occupation	Impairments	Digital literacy	Method
P1	84	F	N/A	Speech & motor	Low	CTA
P2	86	F	Teacher	N/A	Medium	RTA
P3	80	F	Librarian	Motor, visual, & hearing	Low	RTA
P4	86	M	Minister	N/A	Medium	CTA
P5	95	M	Physician	Visual	Medium	CD
P6	–	–	Withdrew	–	–	CTA
P7	74	M	Engineer	Parkinson's disease	Medium	CTA
P8a	80	M	Accountant	N/A	Medium	CD
P8b	77	F	Teacher	Scoliosis	None	CD
P9	86	F	Artist	Visual & motor	None	RTA
P10	79	F	History teacher	Aphasia	None	RTA
P11	80	F	Nurse	Visual	None	CD
P12	84	F	Teacher	N/A	Low	RTA

14.2.3 Study Design

Our usability study was a part of a larger deployment study. In this section we describe the deployment and usability study design.

14.2.3.1 Deployment Study

We conducted a three-month deployment in a Canadian care home. Participants kept our iPad-based tool and used it independently every week to send messages to relatives. We performed the usability study at the end of the deployment study. The deployment included training sessions, semi-structured interviews, and field observations; used to contextualize the usability tests' findings.

14.2.3.2 Usability Study

Each participant experienced one usability testing method, as in a between-subjects design. As usability tests could cause fatigue, we didn't perform a within-subjects comparison of the methods. We had three participants in the Concurrent Think-Aloud (CTA) condition, which required verbalization while completing tasks. Five participants were in the Retrospective Think-Aloud (RTA) condition, in which they verbalized while watching a video of themselves completing the tasks. Four

participants were in the Co-discovery (CD) condition, working in pairs to complete tasks. Participants were randomized to the RTA and CTA conditions, and from these conditions, four participants were selected (based on personal circumstances) to be in the CD condition. We involved P8b in the testing session by doing CD with her husband; P5 and P11 were selected for the CD condition based on similar levels of comfort with the tool. We could not balance the conditions due to unforeseen circumstances that required sensitivity, such as P8a bringing his wife to the usability testing session.

We designed the usability test to have five tasks and take an average of 15 min. We printed the tasks on 3 × 5 inch notecards. The tasks progressed from navigating to the researcher's contact information to sending and receiving messages. We based the tasks on common and uncommon scenarios of use. The uncommon scenarios drew on what participants reported to be the least used feature (video messaging). These tasks were chosen to test for strategies, such as participants' abilities to draw inferences between tasks, given that the audio and picture features had analogous interfaces/steps. We arranged the tasks in order of difficulty, with the least difficult task being first, to ease the participant into the tasks and avoid lowering their confidence from the start. Along with testing whether they could operate some features, we wanted to know whether: (1) they could follow a longer navigation path, (2) they knew how to go back or cancel an action, and (3) they understood message history and contact order. These tasks provided insight into participants' level of system comprehension in as few tasks as possible.

14.2.4 Usability Testing Procedure

There were ten testing sessions in total (three CTA, five RTA, two CD), which lasted on average 36 min each (SD = 8.82). We began the session with warm-up questions, followed by the condition (CTA, RTA, or CD).

The researcher emphasized that the tool was being tested and not the participant. Participants were then instructed to complete the task and to ask for help only when stuck. One researcher sat next to the participants, while another researcher stood behind the participants and videotaped their hands as they interacted with the iPad.

For the CTA condition, the participants were instructed "to say what's on your mind as you are doing the task." The researcher demonstrated the CTA with a simple task on her tablet. The researcher emphasized that the participants could stop thinking aloud if it became burdensome but asked that the participants try to finish the task. For ethical reasons that may result from CTA's additional cognitive load, such as causing distress and making participants more aware of their limitations, the researcher did not prompt participants to continue to think aloud if participants fell silent.

For the RTA, the researcher gave instructions after participants completed the tasks. The instructions were for participants to verbalize their thoughts during the task while watching the video recording of themselves performing the task. The researcher also demonstrated the RTA with a video of herself doing a simple task

on her tablet. Additionally, the researcher showed participants how to pause and play the video. If participants fell silent for more than 20 s, the researcher prompted participants to think aloud by asking, "Do you remember doing the task? What were you thinking?" Because participants were not multi-tasking in the RTA, it seemed appropriate to prompt participants.

In the CD condition, participants were instructed to work together to finish tasks. The researcher also advised participants to consult one another if they got stuck before asking our assistance.

14.2.5 Analysis

We analyzed the usability tests with a tabulation of task completion, task time, errors, and challenges encountered. We also used thematic analysis (Patton 1990) to uncover general patterns, including categories and themes, within and across cases. Our thematic analysis was based on an inductive and deductive approach—i.e., codes were identified from the data based on both a priori and new insights regarding: benefits, challenges, and outcomes of the usability testing methods. Authors coded independently to identify categories and themes, following an open and then a structured coding process. An external coder determined a basic inter-rater reliability of 98% using the procedures described in (Patton 1990): this coder counted the discrepancies in category assignment between the codes and themes of the coders for half of the data. This simple procedure suits our data, sampling, and analytical technique, as recommended in the literature (Patton 1990).

14.3 Study 1: Results and Discussion

We present a comparison of the usability testing methods and three overarching themes identified from the thematic analysis: *Impression Management*, *Low Self-efficacy*, and *The Researcher's Role*.

14.3.1 Comparing the Thinking-Aloud Variants

We compared the suitability of three usability testing methods: Concurrent Think-Aloud (CTA), Retrospective Think-Aloud (RTA), and Co-discovery (CD).

14.3.1.1 Concurrent Thinking-Aloud (CTA)

We found that P1 and P4 could perform the CTA to a certain point. However, they stopped thinking aloud as soon as they became stuck/lost. Their comfort with CTA

was unexpected, especially for P1 who had speech impairment: "*I can do that, I'm a good talker. I actually talk to myself.*" It was clearly difficult for P7, as he was very frail. He spoke quietly, and it was difficult to hear his verbalizations. At the end of the session, he reported a negative impact on his physical and emotional state:

"*Tired is one thing I am, but I'm also quite tense to be trying to concentrate on this (…) which is more of a problem than being tired. I don't know, we've been doing it for 10 or 15 min now, but my hands are quite tense, and I start to shake (…) For Parkinson's disease some people get a lot of anxiety*."

P7's discomfort suggests that the frailty level could contribute to a participant's ability to perform a CTA.

14.3.1.2 Retrospective Think-Aloud (RTA)

Retrospective Think-Aloud was unsuccessful for several reasons. Three of the participants had difficulty understanding the instructions for the RTA or its purpose: *"[There's] not much action going on in the picture though…*" (P3). Of the five participants, only two completed the RTA but reported several challenges. We ended the RTA early for three participants: P3 became frustrated and stopped watching the recording, P9 reported that watching the recording made her eyes hurt, and P12 also found it difficult to watch. Watching themselves made P10 and P12 more aware of their mistakes during the test: "*I don't want to go through this again. I found it very stressful*.", P12 reports; P10 noted, "*[I completed the tasks] after a while…but [the researcher] had to fix [my mistakes]*."

Even when the participants performed the RTA, there were challenges. First, despite prompts to think aloud, participants usually watched the recording silently. Second, when they thought aloud, their verbalizations were not revealing. They commented on what they were seeing rather than what they were thinking while completing the task: "*Now I'm following the instructions and I was just told to put a short message in*" (P2). They also had difficulties remembering what the task was and what they were doing in the video, as found in prior work (Dickinson et al. 2007).

14.3.1.3 Co-discovery (CD)

Of the three conditions, Co-discovery (CD) was most appropriate for our participants. The four participants in the CD condition completed all five tasks successfully and without the researcher's assistance. Participants worked effectively as a team in the CD condition and overcame challenges together by compensating for the other participant's missing knowledge.

All four participants verbalized to keep their partner on the same page while completing a task: "*So we want a picture, and we want to turn it around (…) and now we don't want to send it*" (P11). Unlike CTA, participants continued to verbalize when they did not know how to do something by asking the other participant: "*Where's the garbage pail gone?*" (P11).

Additionally, P11 learned from P5 that she could communicate with people overseas like P5 did with his granddaughter in Iran. Sharing knowledge is a known advantage of CD over other methods: CD can uncover the overlap of the participants' knowledge while highlighting where this overlap does not exist (Miyake 1986).

Another relevant observation was the participants' interpersonal dynamic. A "pilot" and a "co-pilot" emerged: the pilot interacted with the tablet while the co-pilot suggested alternative paths, reminded the other participant of the task, and supported the other participant's decisions. This might also be because co-pilots, P8b and P11, already had a secondary role surrounding technology, which was based on deferring to their husbands for technical support. So, it may have felt natural to let the other participant lead.

14.3.2 Usability Tests: Thematic Analysis

The thematic analysis of the usability tests identified three themes: *Impression Management*, *Low Self-efficacy*, and *The Researcher's Role*. These were present across conditions and provide insight into conducting usability tests with older adults regardless of the think-aloud method.

14.3.2.1 Impression Management

Five participants engaged in impression management, i.e., behavior stemming from the desire to make a positive impression on the researcher (Goffman 2012). Many felt the need to explain why they struggled or could have performed better during the test, even when able to complete most tasks. This need may have been related to knowing the researchers over the course of the study and not wanting to disappoint them. For example, after sending/receiving a video message, P1, P2, P10, P11, and P12 told us they had never used the video feature before to justify their issues with the task. After expressing that she did not perform well on the test and apologizing to us, P10 also pointed out that she was 79 and had had two strokes. P12 also noted that she had difficulty holding the tablet because of the weakness in her hand.

P1's behavior was another example of impression management. At the beginning of the test she told us that the app was "*amazing*," that her family liked that she was using the app, and that the app was better than a phone. However, during the test it became clear that the learnability of the app was poor as she showed gaps in her understanding even after using the app for three months. Most participants (3 out of 5) who performed impression management were in the RTA condition.

14.3.2.2 Low Self-efficacy

Apologizing throughout the test was common, particularly for three participants (P1, P9, P12). These participants apologized for making mistakes and for not knowing how to perform a task. P1 and P11 verbalized their low self-efficacy by explaining they were "forgetful" and "not electronically-savvy." P10 and P12 told the researcher that they did not do well on the test saying, "*I just goofed" and "I was very hesitant about all of this*."

14.3.2.3 The Researcher's Role

Despite instructions to only consult the researcher when it was absolutely necessary, most participants engaged the researcher even when completing tasks. To avoid increasing their stress levels, we felt the need to frequently provide supportive comments and acknowledgements. Some participants needed more support from the researcher, since they were internalizing fault for making mistakes (P10, P12).

However, three participants did not want our support. Instead of asking the researcher for guidance, P2 relied on the task cards to complete the tasks. If she made a mistake, she re-read the task card before attempting another approach. P2 did not start off the deployment study with a high level of digital literacy, yet she practiced using the app, and by the time of the usability test, she had learned how to use most features and was enthusiastic about the app.

P3 struggled to complete most tasks, only completing one of five: she repeated the same steps even after they proved futile. Despite her apparent frustration, she did not ask questions and ignored the researcher's attempts to guide her. After a while she indicated that she was done with the usability test with her body language: by sitting back, with her chin on her chest. In contrast to P2, P3 showed little interest in the technology throughout the deployment study, even though she remained until its completion. She did share, however, one illuminating story: she used to work as a librarian and found the command prompt easy to use when it was first introduced at her work. When they switched to using a Graphical User Interface, she couldn't understand it. Her difficulty with the system was a major factor in her decision to retire. She carried with her this negative experience of technology, which affected her self-efficacy with technology. The researcher involvement may have been detrimental in this case, reinforcing her self-perceived incompetence.

P5 also did not ask for help during the test, partially because he could rely on P11 for support during the Co-discovery (CD). He may not have wanted assistance from the researcher to maintain his sense of autonomy (as the "pilot") and show that he could still learn something new. P5 was a highly-esteemed doctor who worked in rural areas until his retirement. He had recently begun to show signs of cognitive decline, which were difficult for him and family because of his identity as an educated and intelligent man. He enrolled in the study to slow the effects of decline. The CD may have worked well for P5 because he did not have to rely on the researcher and had a peer support (P11).

Due to participants' low self-efficacy and to minimize stress, assistance from the researcher was necessary when requested. But we also found that alternative forms of support (e.g., task cards and peer-support) were essential for keeping participants' sense of autonomy intact and bolstering their confidence.

14.3.3 Summary

Results from our deployment study suggest that Co-discovery (CD) is the most appropriate usability testing method for this group of individuals. Additionally, several insights that contribute to our recommendations: (1) Participants' impression management and low-self efficacy impacted the test more than participants' frailty, (2) the researcher should be involved in guiding and supporting the participant, but be sensitive to those who do not want assistance, and (3) for participants who do not want assistance, researchers should provide alternative means of support to boost self-efficacy and autonomy. Many of our findings were reinforced by experts in our interview study, reported below.

14.4 Study 2: Interviews with Experts

To uncover usability testing practices of HCI experts who work with older adults, we conducted an interview study. In this study, experts designed hypothetical usability tests for different personae, which were based on participants from Study 1. The usability test exercise aimed to understand experts' decision-making process and how they would account for different participant characteristics.

14.4.1 Participants

Researchers and professionals with international reputations for working with older adults while developing and evaluating technology were identified. Care was taken to identify experts from different regions and with different backgrounds. These considerations resulted in contacting 11 experts via email, six of whom agreed to be interviewed in 2015. These six HCI researchers and practitioners (three women and three men) had backgrounds in psychology, engineering, computer science, and occupational therapy. All of them had worked in centers/laboratories dedicated to studying the use of technology by older adults or by those living with aging-related impairments (e.g., tremors, speech impairments, cognitive decline). All experts had peer-reviewed publications reporting the results of usability tests with older adults. These experts were distributed across three continents and worked with older adults in English, Spanish, and French language environments.

14.4.2 Interview Procedure

The semi-structured interviews included: (1) questions about experts' use of usability testing methods with older adults, and (2) a scenario involving a usability test. The first and third authors conducted the interviews via Skype, in person, or by phone depending on the expert's preference. All interviews were recorded and conducted until theoretical saturation was reached (Morse 2012). We stopped recruiting experts when we saw overlap in their use of methods and we had an overarching understanding of their approaches.

The general question portion of the interviews elicited information about experts' experiences with different usability testing methods. It also asked about adaptive methods to meet both the participants' and their needs. Then, we gave experts a usability testing scenario; experts had to plan a usability test for a specific technology. The scenario was constrained through the specification of usability test goals and the inclusion of four hypothetical participants with varying abilities and dispositions towards technology. The personae describing these hypothetical participants were based on participants P1, P3, P4 and P7 from our deployment study (see Table 14.1). Included in the descriptions were the personae's levels of digital literacy, attitudes towards learning, and impairments. The technology to test for was similar to the one evaluated in Study 1. Experts were asked to design a usability test based on the described goals, technology, and hypothetical participants. They were also asked to justify the choices they made so that we could uncover experts' decision-making processes.

14.4.3 Analysis

An inductive thematic analysis was applied to the interviews. The procedures used were similar to those from Study 1, including calculating a basic inter-rater reliability (Patton 1990), which was also 98%.

14.5 Study 2: Results and Discussion

The thematic analysis of the interviews with experts identified four main themes: *Experts' Current Practice*, *Adapting to Participant Characteristics*, *Ethics*, and *The Researcher's Role*.

14.5.1 Experts' Current Practice

Our experts report using a variety of usability testing methods, including: task-based usability tests (E5, E6, E2), think-alouds (E1, E3, E5), observation (E1, E4), Likert-type rating scales (E1, E2), cognitive walkthroughs (E6), validated instruments such as QUEST 2.0 (E1), interviews (E1), and usability questionnaires (E4). We report on their evaluation of think-alouds and task-based usability tests.

14.5.1.1 Advantages and Disadvantages of Think-Alouds and Task-Based Usability Tests

Experts reported that the advantage of task-based tests is that they produce quantitative data and the researcher can control how participants interact with the technology, which facilitates data comparison across participants. The disadvantages of task-based tests and think-alouds are that they can cause frustration and anxiety in the participant. When the test is structured more informally, the researcher can see how participants do tasks they are already familiar with. Regarding think-alouds, E1 noticed that she had to prompt participants frequently for the think-aloud to be successful. While E3 reported that the problem with this method is that participants think aloud as long as they know what they are doing and stop when they do not know—precisely the moment at which the researcher needs to gain the most insight.

14.5.2 Adapting to Participant Characteristics

All experts except for E2 reported adapting methods to the individual and context. E5 expressed the importance of being flexible when designing the test due to the nature of the participants: "*Because we [are] dealing with a vulnerable cohort, we [are] very mindful not to have formal methods or formal processes that would start to engender fear or concern, or at all lack sensitivity to the situation (…) any method or technique had to be a servant to the context of the fragility and the vulnerability of the participants*" (E5). Although this methodological flexibility seemed common across experts' accounts, it is seldom reported on. This calls attention to the need to recognize the complexity and messiness of fieldwork and our responsibility as researchers to both our participants and scientific community. While design flexibility might affect replication or reliability of instruments across studies, it ensures validity. Reliability can then be guaranteed with a multi-methods approach (triangulation, mixed-methods) as suggested by our study and the experts.

14.5.2.1 Participant Motivation to Use Technology

The primary concern of most experts was adapting tests to maintain and enhance participants' motivation to use technology both during and after the study. Of all the hypothetical participants, experts had the most advice for P3 due to her lack of interest in the technology. First, they would interview her to understand her lack of motivation. Second, they would enhance her motivation through adapting the usability testing session so it was also training in basic computing skills to build her self-efficacy. Another way was to make clear the benefits of using the technology. For example, E5 found that putting the app in the context of communicating with overseas family was a motivator for his participants. Experts explained that while there are work-arounds for physical or cognitive frailty, there is not much the researcher can do when a participant lacks motivation. Thus, experts found it critical to address motivation, keeping participants in the study and preventing the experience from negatively affecting their self-efficacy. Similarly, Waycott et al. (2016) found that participants' experiences with technology during a study affected their self-efficacy and contributed to older adults' decision to drop out.

14.5.2.2 Low-Self Efficacy and Self-confidence

Experts reported that low self-efficacy was an issue during usability tests. E6 mentioned that participants often apologized and felt like they had let him down. To bolster their confidence, experts employed different strategies. One was to ask the participant to lead test and show the features they used most often and enjoyed using. E6 also mentioned that conducting the test from a lab can be detrimental to participants' self-confidence because "*many seniors don't have a university education [and] they get intimidated by the university setting*." E2 and E6 mentioned celebrating small successes and reassuring them "*that just by being themselves they are very valid participants*." Considering the effect of low self-efficacy on the test, E3 mentioned using a mixed approach to tease out whether participants' self-consciousness was making them perform worse.

14.5.2.3 Impression Management

Experts noted how low self-efficacy and self-perception resulted in impression management efforts, i.e., participants saying things they believe the researcher wants to hear (E1, E2, E4, E6). For instance, E1 explained that participants often say researchers "*are lovely people*" and that a given technology "*is great and wonderful (…) but then show signs of frustration and confusion.*" For this, E1 combines methods, such as behavioral coding and Likert-type scales, to identify usability issues from overly optimistic self-reports. E3 also found theatre beneficial for overcoming impression management because participants are more honest about their opinions when acting. Establishing a relationship with the participant may help reduce the conscious and unconscious need to please and be seen in a favorable light (E2, E6).

14.5.2.4 Physical and Cognitive Abilities

Experts frequently adapted usability testing sessions to their participants' abilities, primarily by doing the session from participants' homes. Not only does this allow experts to see how participants use the app in their natural environment, but their home will be set up for participants' accessibility needs. So, most experts would conduct the session with P7 (who had Parkinson's Disease) from his home. E3 also mentioned doing a remote usability test with P1 (who had a speech impairment) and P7 (who had Parkinson's disease) by asking them to do tasks and fill out a form afterwards, to avoid making them self-conscious about their impairments. For E3, the test design was sensitive not only to P1's and P7's impairments, but also to their attitudes towards their impairments.

14.5.2.5 Digital Literacy

All experts agreed that traditional usability testing methods such as think-alouds and task-based tests could be used with P4 because of his high level of digital literacy and low frailty.

14.5.3 Ethics

Experts stressed that there is a lot at stake when conducting usability tests with older adults. The usability tests can have a lasting impact on participants' perceptions of technology and on their competence and self-confidence. E6 asked, *"[If] we suspect there are already usability issues, why do we confirm our suspicions with people?*" He went on to say that in industry "*we are often just trying to get data to convince somebody else because a hunch wasn't enough*." To avoid usability tests being a negative experience for the participant, E4 emphasized first piloting the app and the usability protocol in the lab. Having an expert do a heuristic evaluation of the app is another strategy. E6 would use P4 (who had high digital literacy and no major impairments) as a control participant to test out the tasks and see if the amount of data was enough for the designers on the product team.

14.5.4 The Researcher's Role

Experts highlighted the need to communicate the purpose of the study (E1, E2, E4, E6) and the participant's role in the test (E6). Although it is important to adapt tests, participants may feel the researchers think they do not have the capacity to perform challenging tasks (E2). The researcher should be careful when adapting the test and avoid assuming low self-efficacy means low competency. Instead, adapt the

test to the level of motivation, abilities, and digital literacy. The researcher can also remind the participant that the method is not "dumbed down" but used with many different populations (E2). With the hypothetical participants, experts would be more involved in the test with P1 and P3 (both had low digital literacy). E5 also talked about having participants use the technology with one another, saying "*I would use P1 as a champion*" to facilitate relationships with the other participants through the app. He would also challenge P4 to explore assistive technologies for P7, saying that participants can contribute to the study as "*not just recipients of care, [but as] careers*." However, E2 warned facilitating peer-to-peer support should be carefully handled: peers should have similar levels of digital literacy because proficient users can disincentivize participants with low digital literacy.

14.5.5 Summary

Experts use various usability testing methods with older adults and often adapt tests to participant characteristics. Enhancing and maintaining participants' motivation to use the technology and stay in the study is crucial, followed by other issues such as self-efficacy, impression management, physical and cognitive abilities, and digital literacy. Experts focus on motivation because of usability tests' impact: these tests can have a lasting influence on participant self-confidence and perception of technology. Experts highlight the researcher's role to communicate higher level information and not assume low competence to avoid "dumbing down" the test. Finally, experts encourage peer-to-peer support to give participants a clear purpose in the study.

14.6 Recommendations

We compiled our think-aloud findings and expert insights to present four main takeaways. These recommendations gain from an approach that brings together different perspectives (end-users and experts) and an in-depth and contextualized research design.

14.6.1 Takeaway 1: Recruit and Plan for Co-discovery

We found that Co-discovery (CD) is more appropriate than Concurrent Think-Aloud (CTA) and Retrospective Think-Aloud (RTA) for our frail older adult participants. Participants in RTA struggled to see and remember what was happening in the video recording. Additionally, it reminded participants of their mistakes, influencing their self-efficacy. CTA was successful with participants who were not highly frail, and experts agreed that P4, who had no serious impairments and high digital literacy, could use traditional usability methods. One downside of CTA encountered in both studies is that participants stop thinking aloud once they did not know what they are doing.

If participants are interested in working with their peers, E5 encouraged involving more than one participant in the usability test, and we found CD effective in eliciting participant verbalizations. CD may reflect participants' existing roles toward technology, as we found a "pilot-co-pilot" dynamic with participants who deferred to their husbands for technical support. However, researchers should match participants with similar digital literacy levels. Two experts warned that grouping a high digital literacy participant with a lower digital literacy participant could be detrimental.

14.6.2 Takeaway 2: Discover and Enhance Participant Motivation to Use Technology

Experts expressed that usability tests can have a lasting impact on participants' self-efficacy and perceptions of technology. For this reason, experts emphasized ensuring the test does not demotivate participants by being too difficult; motivation to use technology appears to relate to self-efficacy, as found in prior work (Waycott et al. 2016). For P3, her lack of interest affected her performance, which may be due to negative past experiences with technology. For this participant, the usability test and study in general was an opportunity to help regain confidence with technology. P3 may have benefitted from a test including tasks that she knew and enjoyed performing. Informal testing approaches can be used for participants with low enthusiasm.

14.6.3 Takeaway 3: The Researcher's Role Includes Enhancing Participants' Sense of Autonomy and Self-confidence

How much information and guidance to provide participants in usability tests can be non-intuitive. We recommend sharing high-level information about the study, guiding participants if they ask for help, and providing alternative forms of support.

14.6.3.1 Share High-Level Information

Experts emphasized sharing high-level information such as the purpose and goals of the usability test with participants, while establishing their role as the expert in that process.

14.6.3.2 Guide the Participants When Asked

Although we instructed participants to only ask for our assistance when they were stuck, participants would often engage us in the test. Thus, we recommend a greater degree of researcher involvement when participants want it, providing encouragement and support.

14.6.3.3 Provide Alternative Forms of Support for Participants Who Do not Want Researcher Assistance

Involving two or more participants in a usability test can be beneficial for maintaining participants' sense of autonomy (as we found) and for giving them a sense of purpose in the study (according to E5). Alternative forms of support such as task cards can be helpful for participants who need some assistance but want to work through tasks independently.

14.6.4 Takeaway 4: Low Self-efficacy and High Impression Management Can Have a Greater Impact on the Usability Test Than Frailty

Results from both Study 1 and Study 2 suggest that participants' low self-efficacy and high impression management efforts can affect usability tests more than level of frailty. To offset challenges with low self-efficacy and high impression management, we recommend using multi-method approaches, allowing participants to lead, and pilot-testing.

14.6.4.1 Use Multi-method Approaches

Impression management and low self-efficacy were identified as challenges in both studies. As experts suggested, multi-method approaches can help reduce the effects of impression management and tease out whether participants' self-consciousness is making them perform poorly. Furthermore, multi-method approaches can make the best use of participants' time (e.g., combining training and usability testing in one session).

14.6.4.2 Allow the Participant to Take the Lead

For participants with low self-efficacy, experts highlighted the importance of informal approaches, such as not enforcing a list of tasks but asking participants to walk through the system with the researcher. By emphasizing participants' central role in the test, we can increase their confidence and help reduce impression management efforts.

14.6.4.3 Ensure the Prototype Works Well Before Testing It with People

Usability tests should not diminish participants' self-efficacy and confidence. Therefore, E2 and E6 emphasize piloting the tool or doing a heuristic evaluation by an expert. To conclude, by combining fieldwork with older adults and interviews with experts, we put forth recommendations to conduct usability tests with frail

older adults—a diverse population with varying needs and expectations. This practical information can assist researchers with method application, development, and refinement.

14.7 Conclusion

This study is limited by a purposive sample, so insights cannot be generalized to all usability testing contexts in which frail older adults participate. However, the breadth of experts' experience and older-adults' abilities enabled a rich understanding and comparison of a variety of issues of interest to the community.

We conducted two studies to investigate usability testing with frail older adults. Study 1 showed that Co-discovery (CD) was most suitable for our group of participants because they were able to think aloud throughout the test. Our second study with HCI experts who work with older adults showed that experts adapt to participant characteristics, such as self-efficacy, and focus on keeping participants motivated in the study. Based on these two studies, we advanced recommendations for conducting usability tests with frail older adults.

References

Anderson M, Perrin A (2017) Tech adoption climbs among older adults. Retrieved from http://www.pewinternet.org/2017/05/17/technology-use-among-seniors/

Berkowsky RW, Rikard RV, Cotten SR (2015) Signing off: predicting discontinued ICT usage among older adults in assisted and independent living. In: Zhou J, Salvendy G (eds) ITAP 2015. Springer International Publishing, Cham, pp 389–398

Dickinson A, Arnott J, Prior S (2007) Methods for human-computer interaction research with older people. Behav Inf Technol 26:343–352

Ericsson KA, Simon HA (1980) Verbal reports as data. Psychol Rev 87:215–251

Gell NM, Rosenberg DE, Demiris G et al (2015) Patterns of technology use among older adults with and without disabilities. Gerontologist 55:412–421. https://doi.org/10.1093/geront/gnt166

Goffman E (2012) The presentation of self in everyday life (1959). In: Calhoun C, Gerteis J, Moody J, Pfaff S, Virk I (eds) Contemporary sociological theory. John Wiley & Sons, pp 46–61

Miyake N (1986) Constructive interaction and the iterative process of understanding. Cogn Sci 10:151–177

Morse JM (2012) Introducing the first global congress for qualitative health research: what are we? What will we do—and why? Qual Health Res 22:147–156. https://doi.org/10.1177/1049732311422707

Neves BB, Fonseca JRS, Amaro F, Pasqualotti A (2018a) Social capital and Internet use in an age-comparative perspective with a focus on later life. PLoS One 13:e0192119

Neves BB, Baecker R, Carvalho DD, Sanders A (2018b) Cross-disciplinary research methods to study technology use, family, and life course dynamics: lessons from an action research project on social isolation and loneliness in later life. In: Neves BB, Casimiro C (eds) Connecting families? Information & communication technologies, generations, and the life course. Policy Press, Bristol, pp 113–132

Neves BB, Franz RL, Judges R, Beermann C, Baecker R (2017) Can digital technology enhance social connectedness amongst older adults? A feasibility study. J Appl Gerontol 38(1):49–72

Neves BB, Franz RL, Munteanu C et al (2015) "My hand doesn't listen to me!": adoption and evaluation of a communication technology for the 'oldest old.' In: CHI 2015, Seoul, Korea, pp 1593–1602
Patton MQ (1990) Qualitative evaluation and research methods. Sage
Torpy JM, Lynm C, Glass RM (2006) Frailty in older adults. JAMA 296:2280. https://doi.org/10.1001/jama.296.18.2280
UN (2017) World Population Ageing 2017, New York
Vines J, Blythe M, Lindsay S et al (2012) Questionable concepts: critique as resource for designing with eighty somethings. In: CHI, pp 1169–1178
Vroman KG, Arthanat S, Lysack C (2015) "Who over 65 is online?" Older adults' dispositions toward information communication technology. Comput Human Behav 43:156–166
Waycott J, Vetere F, Pedell S et al (2016) Not for me: older adults choosing not to participate in a social isolation intervention. In: CHI 2016. ACM, New York, NY, USA, pp 745–757

Rachel L. Franz is a Ph.D. student in Information Science at the University of Washington working with Jacob O. Wobbrock. Before starting her Ph.D. she graduated with an MSc in Computer Science from the University of Toronto. Her research focuses on making technologies for older adults that are accessible and enjoyable to use.

Barbara Barbosa Neves is Assistant Professor of Sociology at the University of Melbourne, Australia. Prior to that, she was Associate Director and Researcher at the 'Technologies for Aging Gracefully Lab' (TAGlab), University of Toronto, Canada. Her research has been examining social determinants and effects of adoption and non-adoption of digital technologies in a life course perspective. In particular, she is interested in the links between digital and social inequalities. More recently, Barbara has been studying how digital technologies can help combat social isolation and loneliness in later life. Her work has been published in a range of top-tier outlets in sociology and human-computer interaction.

Carrie Demmans Epp earned her Ph.D. from the University of Toronto and her MSc from the University of Saskatchewan. Since then, she has joined the Department of Computing Science at the University of Alberta where she leads the Educational Technology, Knowledge, Language, and Learning Analytics (EdTeKLA) Research Group. Her work focuses on the development of adaptive educational technologies, with much of her work focusing on supporting members of vulnerable populations.

Ronald Baecker is Chairman of famli.net Communications Inc., Emeritus Professor of Computer Science at the University of Toronto, and Founder of the Technologies for Aging Gracefully Lab (TAGlab).

Jacob O. Wobbrock is a full Professor of human-computer interaction in The Information School and an Adjunct Professor in the Paul G. Allen School of Computer Science & Engineering at the University of Washington in Seattle, WA, USA. His research focuses on input and interaction, human performance measurement and modeling, research and design methods, mobile computing, and accessible computing. He received the 2017 SIGCHI Social Impact Award for his development of ability-based design.