

ForesightCares

AVA at Home: A MODERATED USABILITY EVALUATION OF A SELF-ADMINISTERED, VISION-BASED MOBILITY ASSESSMENT FOR OLDER ADULTS

TECHNICAL REPORT

ForesightCares, Inc.
Charlotte, NC
info@foresightcares.com

ABSTRACT

Our earlier reports validated the measurements of AVA Mobility, an on-device computer-vision system for the Timed Up and Go, 30-second chair stand, and 4-stage balance test, when staff ran it. This report presents a moderated, in-person, formative usability evaluation of **AVA at Home**, the self-administered version for iPad and MacBook, designed for older adults to use at home without a clinician or caregiver present. Nineteen older adults (12 female, 7 male; aged 70 to 91) at a senior living community took part in July and August 2026. After a simulated clinic introduction by a physical therapist, each participant completed up to six tasks on their own, with the moderator present, while thinking aloud. We measured the System Usability Scale (SUS), an adapted five-point task-ease question, attitudes, time on task, and the frequency of usability problems coded from session recordings.

The mean SUS was 76.4 (90% CI 68.2 to 84.7) and the median 82.5 (90% CI 67.5 to 92.5). Scores varied widely: 11 participants scored 80 or above and six scored below 62. Usefulness (4.89 of 5) and trust in the results (4.84) were the highest-rated attitudes, and onboarding and the assessment received the lowest task-ease ratings. Five problems occurred in at least 11 of the 18 recorded sessions: a checklist toggle that did not read as "no" (13), a results header that showed fall risk rather than mobility (12), a login button hidden behind the keyboard (11), an unmarked scroll on the profile screen (11), and results that were hard to interpret (11). Two bugs blocked progress. Lower- and higher-scoring participants encountered similar median numbers of coded problems. High self-reported task ease sometimes co-occurred with observed assistance. Participants reported favorable usability under moderated conditions, but independent home use remains unestablished. The results identify specific interface changes for evaluation and provide a baseline for a follow-up study.

Keywords Usability · Older adults · System Usability Scale · Remote monitoring · Fall risk · Self-administered assessment · Think-aloud

1 Introduction

Falls are the leading cause of injury among older adults [1]. The CDC's STEADI initiative and the 2022 World Guidelines for Falls Prevention both recommend assessing gait, strength, and balance to identify people at risk [2, 3]. These tests are usually given in a clinic by trained staff, which limits how often they can be repeated. Repeated testing at home would let a care team see changes in mobility between visits and act earlier. Remote care for older adults has expanded quickly, but its benefits depend on whether patients can use the technology, and usability and digital literacy remain among the most frequently reported barriers [4]. Wearable sensors add their own burden, including discomfort, charging, and remembering to wear the device [5]; a camera-based assessment avoids these, but only if the app itself is easy to use.

AVA Mobility automates the three STEADI tests (the Timed Up and Go, the 30-second chair stand, and the 4-stage balance test) using a single iPad camera and on-device AI. Our first two reports validated its measurements against motion capture and physical-therapist scoring [6, 7]. Both studies, however, used staff to set up and run each assessment. **AVA at Home** is the self-administered version, designed from the outset for independent use by older adults in their own homes. A physical therapist introduces the app during a clinic visit, and the older adult then runs the assessment alone on an iPad or MacBook at home, with no clinician, caregiver, or family member needed to set it up or operate it. Results are shared back to the care team. Accuracy is necessary for this model but not sufficient. If an older adult cannot log in, answer the safety questions, position the device, and understand the results on their own, accurate measurement never happens.

This report presents a moderated usability evaluation of AVA at Home with 19 older adults. It addresses four research questions:

- **RQ1:** How usable do older adults find AVA at Home overall, and how much does this vary between people?
- **RQ2:** Which tasks are hardest, and where do breakdowns and requests for help occur?
- **RQ3:** How do participants rate the clarity, appearance, safety, and trustworthiness of the app and its results?
- **RQ4:** What should change before the next design round, and which baseline should that round be measured against?

The contributions are a detailed account of the usability of a self-administered, vision-based mobility assessment in its target population; a ranked, evidence-linked list of usability problems with the changes they imply; an exploratory analysis of how participants with high and low usability ratings differed; a check of audio-derived task timing against a live observer; and a documented baseline, including the measurement lessons, for a repeat study after the design changes.

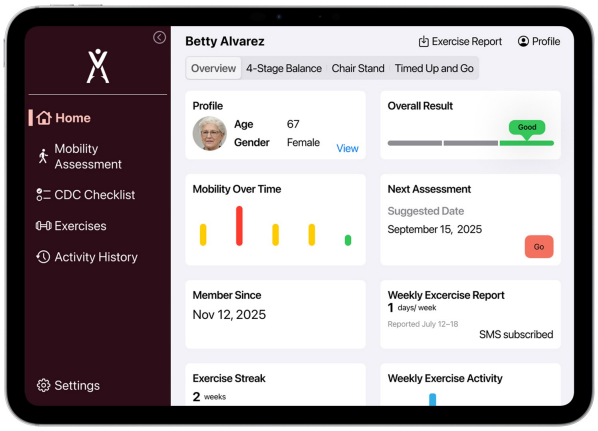
2 AVA at Home

AVA at Home is designed for one user: an older adult using the app by themselves at home. This design goal shapes the whole interface. The app guides the user through every step with on-screen text and spoken instructions; the camera measures each test automatically, so the user never has to time or count anything; results are presented in plain terms against CDC reference values; and the exercise program that follows is shown step by step with illustrations. The aim is that a person who has been introduced to the app once by their physical therapist can repeat the assessment at home, at a time that suits them, without needing anyone else present. Figure 1 shows six of the app's main screens.

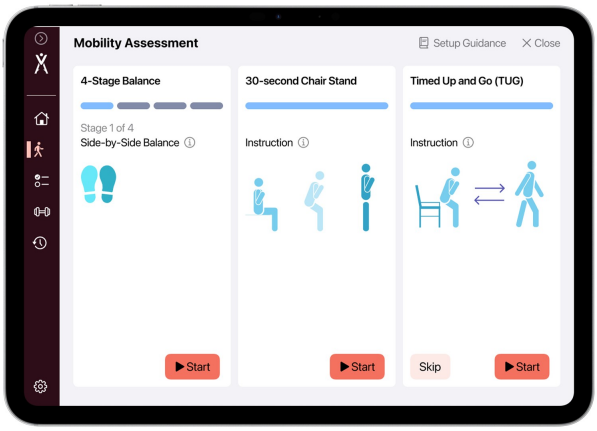
Home dashboard. After logging in, the user sees an overview of their status (Figure 1a): their profile, the overall result of their most recent assessment, a chart of their mobility over time, the suggested date of their next assessment, and summaries of their exercise activity. Tabs along the top give quick access to the results of each of the three tests.

Mobility assessment. The assessment screen (Figure 1b) presents the three STEADI tests side by side: the 4-stage balance test, the 30-second chair stand, and the Timed Up and Go. Each card shows a simple illustration of the movement and a single Start button. A progress bar shows which stage of a test the user has reached, and a *Setup Guidance* link explains how to position the device. Before the tests begin, the user completes the CDC *Stay Independent* checklist, a short set of yes/no questions about falls and balance, and a 10-item safety check.

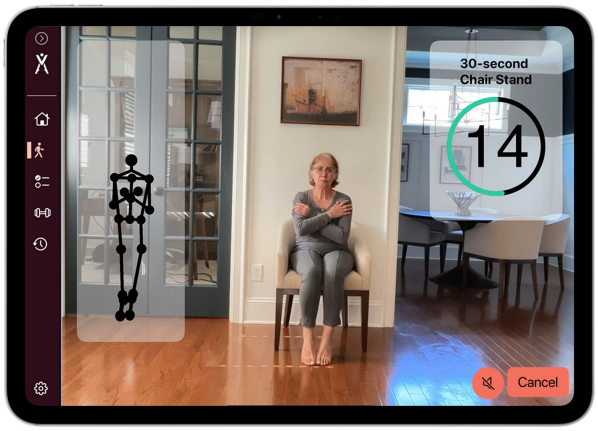
Live test. During each test (Figure 1c), the camera view shows the user with the body skeleton that AVA is tracking, so they can see that the app is following them. The app counts repetitions or keeps time automatically and displays the result in large type, here the number of completed chair stands with a ring showing the time remaining. This is



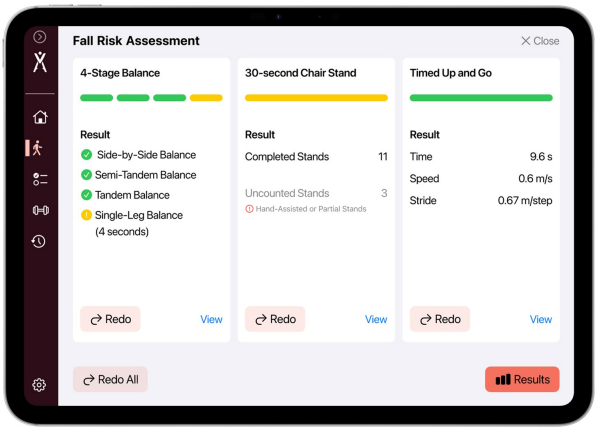
(a) Home dashboard



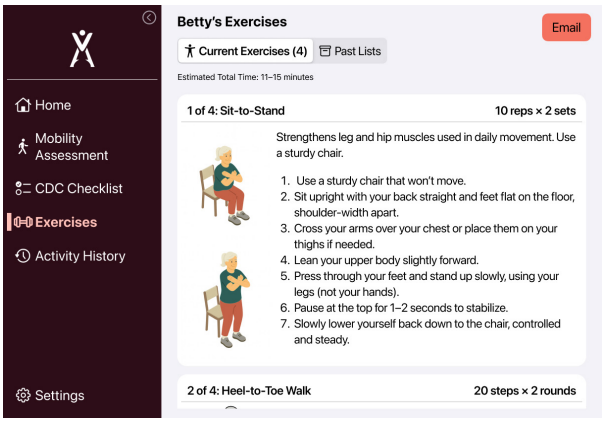
(b) Mobility assessment



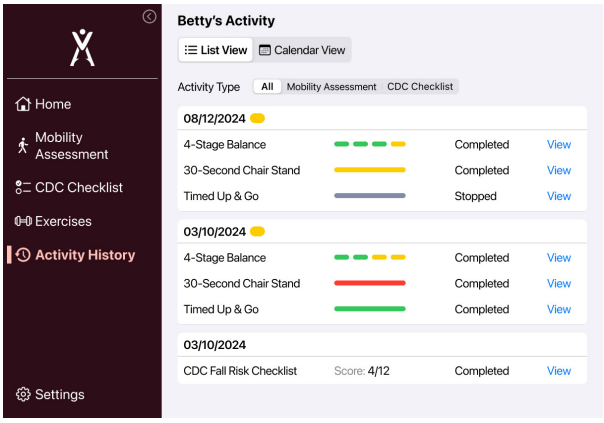
(c) Live chair-stand test



(d) Assessment results



(e) Exercise program



(f) Activity history

Figure 1: Selected screens from AVA at Home, shown with demonstration accounts and illustrative data; the names and dates on these screens (for example, membership, next-assessment, and history dates) are placeholders and do not correspond to the study. Panel (d) shows the results screen of the build tested, whose "Fall Risk Assessment" heading is the terminology mismatch reported in Section 4.5. The left-hand pane provides navigation to the home dashboard, the mobility assessment, the CDC checklist, the exercises, and the activity history. These are a subset of the app's screens.

the measurement whose accuracy we validated in our earlier reports [6, 7]. A mute button and a Cancel button remain available throughout.

Assessment results. When the tests are finished, a summary screen (Figure 1d) reports the outcome of each test: which balance stances were held, the number of completed chair stands (with any hand-assisted or partial stands shown separately), and the TUG time, walking speed, and stride length. A colored bar summarizes each result, and the user can repeat a single test (*Redo*) or the whole assessment, or open a detailed results view.

Exercise program. Based on the results, the app recommends a set of exercises (Figure 1e). Each exercise is shown with illustrations, numbered step-by-step instructions, the number of repetitions, and an estimated total time. The program can be emailed, so that users who prefer paper can print it.

Activity history. The history screen (Figure 1f) lists past assessments and checklists by date, with a colored summary of each test, so that users and their physical therapists can see changes over time.

3 Methods

3.1 Study Design and Ethics

This was a moderated, in-person, formative usability evaluation using the think-aloud method [8]. Usability was assessed in terms of effectiveness, efficiency, and satisfaction [9]. The study was reviewed and approved by Advarra, an independent institutional review board that provides ethical oversight for human-subjects research (protocol Pro00085739), and was conducted in accordance with the Declaration of Helsinki [10]. All participants gave informed consent before taking part. Sessions took place in July and August 2026. All sessions used the same protocol, and all participants are analyzed together. Several small fixes were deployed between sessions during the study period; none has yet been retested with participants, and the problem frequencies therefore describe the builds in use during the study.

3.2 Participants

Participants were older adults recruited at a senior living community as part of a scheduled data-collection session. No previous experience with AVA was required. The target user was an adult aged 65 or older with limited technology experience. Twenty-two people were scheduled, 19 attended, and 3 did not attend (Figure 2). The 19 participants (12 female, 7 male) were aged 70 to 91 years (mean 82.3, median 82) (Figure 3a). Their self-described experience with technology ranged from former physical therapists comfortable with tablets to people who called themselves "computer illiterate". Participants are identified by study code only (P094 to P115). Table 1 summarizes their age and sex.

Because AVA at Home is designed for independent use, every participant used the app by themselves. Caregivers and family members were not involved in the sessions, did not help participants with the tasks, and were not participants in the study. The only other people present were the study moderator and, in six sessions, a second researcher who observed and timed the session.

Table 1: Participant demographics.

Group	Participants	Mean age (years)	Median (years)	Range (years)
Female	12	81.5	81.5	70–89
Male	7	83.7	83.0	76–91
Total	**19**	**82.3**	**82.0**	**70–91**

3.3 Scenario and Session Procedure

Each session lasted about 30 minutes and followed a two-scene scenario read from a script (Figure 4). In **Scene 1**, the moderator played the participant's physical therapist in a simulated clinic visit and introduced AVA Mobility, the three tests, and the safety rules. Before Scene 2, participants had completed the three tests earlier the same day, which served as the clinic visit in the scenario. In **Scene 2**, the participant imagined they had gone home and used AVA at Home on their own. The moderator stayed in the room but was meant not to help, to show what the app could and could not explain by itself, and stepped in only for safety or a complete block. The script told participants that their results would be shared with their physical therapist. This design mirrors the intended real-world hand-off. Participants reached the home app already familiar with the three tests, as a real patient would after a clinic screening, so the introduction in Scene 1 is part of the scenario, not extra help. Participants were asked to think aloud throughout.

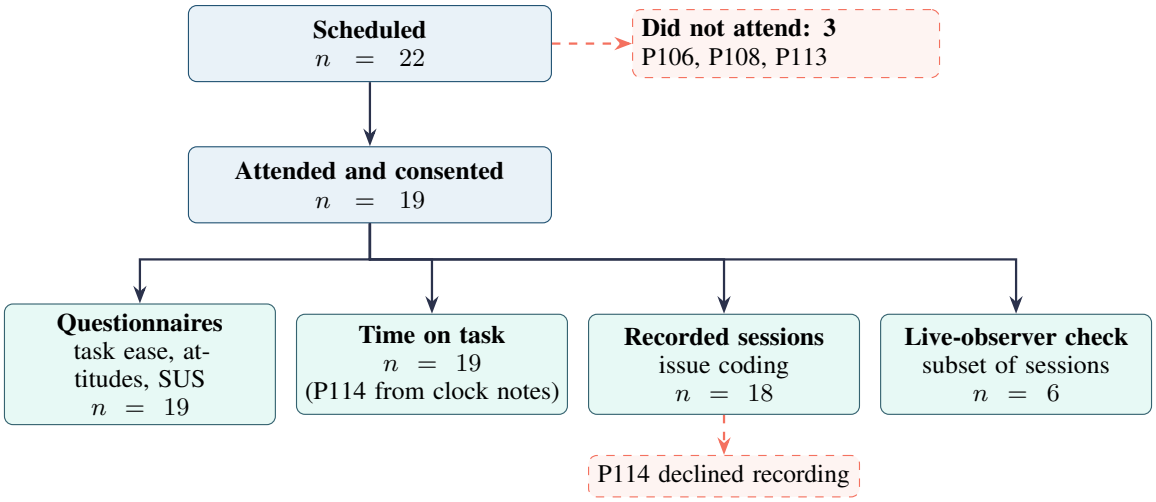


Figure 2: Participant flow and the sample available for each analysis. Questionnaire measures use all 19 participants; usability-problem frequencies use the 18 recorded sessions.

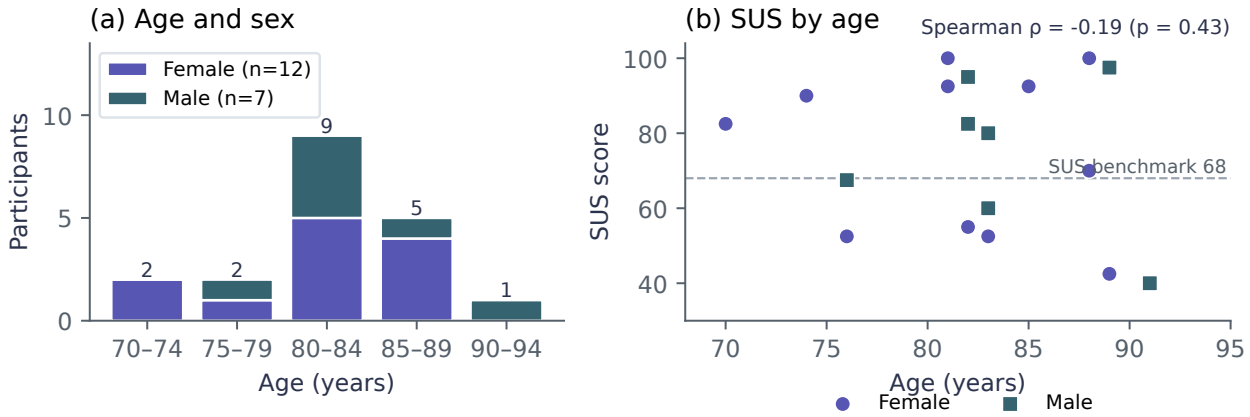


Figure 3: Participants. (a) Age distribution by sex. (b) SUS score by age; the dashed line marks the published SUS average of 68. We did not detect an association between age and SUS in this sample ($\rho = -0.19$, $p = 0.43$).

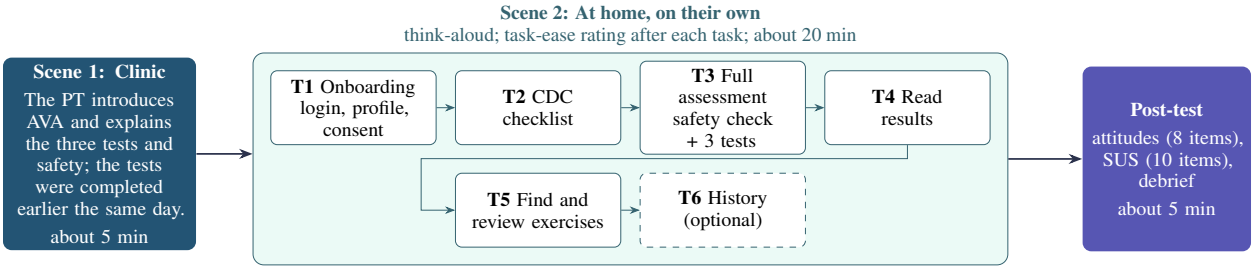


Figure 4: Session design. After a simulated clinic introduction (Scene 1), participants used AVA at Home on their own for up to six tasks (Scene 2), rating each task with the adapted five-point task-ease question, and then completed the attitude questionnaire and the SUS. The history task (T6) was dropped in a few sessions when time ran short.

3.4 Tasks

Table 2 lists the six tasks and the prompts read to participants. The history task was dropped in a few sessions when time ran short. During Task 3 the moderator was told to stay silent unless the participant was completely blocked.

Table 2: Session tasks and prompts.

#	Task	Prompt (abridged)
1	Onboarding	"Get yourself set up in the app. Check that your information in the profile looks correct, and then fill out the consent form."
2	CDC checklist	"The app will ask you some questions about your health and balance. Please go ahead and answer them."
3	Full assessment	"Start the assessment and complete all three movement tests, following the app's instructions on your own."
4	Read results	"Take a look at your results and tell me what they mean to you."
5	Exercises	"Find the exercises the app recommends for you, and tell me what you think of them."
6	History	"Imagine you did this a month ago. Can you find an earlier assessment in your history?"

3.5 Measures

System Usability Scale (SUS). The standard 10-item SUS [11] was given after the tasks. Odd-numbered items are worded positively and even-numbered items negatively. Scores range from 0 to 100; 68 is the published average, and 80 or above is usually described as excellent [12, 13, 14]. Item-level benchmarks allow individual SUS items to be interpreted as well [15]. For reporting, we grouped scores as excellent (≥ 80), around average (62 to 79), or struggling (< 62).

Adapted five-point task-ease question. After each task, participants rated how easy it was on a scale from 1 (very difficult) to 5 (very easy). This item is adapted from the Single Ease Question (SEQ) [16], which uses a seven-point scale, so published SEQ benchmarks do not apply directly. We refer to it as the adapted five-point task-ease question, or task ease for short.

Attitudes. Eight statements covering appearance, clarity of instructions, confidence and safety, trust in the results, usefulness, readability, comfort using the app alone at home, and willingness to recommend it were rated from 1 (strongly disagree) to 5 (strongly agree).

Time on task. Task times were derived from the timestamped session transcripts. Each task ran from the moderator reading the prompt to the moderator asking for the task-ease rating; where ratings were collected together at the end, the next prompt marked the end of the task. For P114, who declined recording, times came from the moderator's clock notes. A task was left untimed when recording started late, when login was blocked by account setup or a bug, or when the task was not reached.

Usability problems. The study team reviewed the session transcripts and the moderator's notes and coded each distinct usability problem. The number of coders and whether coding was performed independently were not recorded, and no inter-coder agreement was calculated, so disagreements were not formally resolved. For each problem we recorded the sessions in which it occurred, a representative quote, and a priority: P0, fix before the next evaluation round; P1, address in the next design round; P2, lower-priority refinement. Priorities reflect the team's judgment of how often a problem occurred and how much it disrupted the task; no formal severity scale was used. Each problem's frequency is the number of the 18 recorded sessions in which it was observed. Whether each participant reached every relevant feature was not recorded, so these counts are session counts, not rates among participants who were exposed to the feature. When a moderator explained a feature in advance, a problem with that feature could not be observed in that session, so the related counts are conservative.

3.6 Validation of Task Timing

Transcript timings can include pauses, repeated attempts, and conversation within a task. To check them, a second researcher observed six sessions (P102, P103, P104, P105, P107, P109) and timed each task by hand to the nearest minute. We compared the two methods for each participant and task, and used the comparison to decide how each task's timing should be reported (Section 4.4). The observer also recorded each participant's spoken task-ease rating and gave her own rating of how independently the participant had completed the task.

3.7 Analysis

We report means with 90% t-based confidence intervals for SUS, task-ease, and attitude scores [13]. Medians are reported for the SUS and for task times, whose distributions were skewed, each with a 90% bootstrap percentile interval

(20,000 resamples) [17]; wide intervals are reported as they are, because they show the uncertainty in this small sample. Problem frequencies are reported as the number of recorded sessions, out of 18, in which each problem occurred. We used Spearman correlations to explore how SUS related to age, to the number of problems each participant encountered, to their mean task ease, and to their assessment time. With 19 participants, these analyses are exploratory and the numbers should be read as indicative [18].

3.8 Deviations from the Protocol

Two planned elements could not be reported as intended. First, the protocol specified task success as the primary metric. However, assistance was not consistently recorded, preventing a defensible estimate of unaided task completion. Clinical test performance, such as inability to maintain a balance position, is distinct from successful completion of the app workflow, and the two should not be confused. We did not reconstruct task outcomes from the recordings, because doing so would require inferring assistance that was not recorded. Second, in some sessions moderators explained the checklist toggle and the 10-item safety check before participants reached them. This raises the ease ratings for those tasks and makes the frequencies of the related problems conservative. Help given during Scene 2 is treated throughout as evidence of what the app could not explain by itself.

4 Results

Table 3 summarizes the main quantitative results.

Table 3: Summary of quantitative results. Intervals are 90%. Times are transcript-derived.

Measure	Item	Value	n
SUS	Mean (90% CI)	76.4 (68.2–84.7)	19
	Median (90% CI)	82.5 (67.5–92.5)	19
	≥ 80 / 62–79 / < 62	11 / 2 / 6	19
Task ease, mean (90% CI)	Onboarding	4.00 (3.65–4.35)	19
	CDC checklist	4.47 (4.17–4.78)	19
	Assessment	4.16 (3.73–4.58)	19
	Results	4.53 (4.22–4.83)	19
	Exercises	4.68 (4.45–4.92)	19
	History	4.93 (4.82–5.00)	15[a]
Time on task, median (90% CI), min	Onboarding	4.4 (2.8–5.9)	14
	CDC checklist	3.2 (2.6–3.6)	18
	Assessment	11.0 (9.8–12.3)	19
	Find exercises	1.8 (1.5–2.5)	19
	History	0.8 (0.75–0.95)	16[a]
Exploratory timing estimates[b]	CDC checklist, adjusted	2.5	18
	Assessment	9 to 10	6
Usability problems	Distinct problems coded	21	18 sessions
	Occurring in ≥ 9 of 18 sessions	8	18 sessions
	Blocking bugs	2	18 sessions

[a] P105 reached and was timed on the history task but did not give a task-ease rating for it.
[b] Based on the six sessions with a live observer (Section 4.4); not a replacement for the transcript-derived medians.

4.1 Overall Usability (RQ1)

The mean SUS score was 76.4 (90% CI 68.2–84.7) and the median 82.5, above the published average of 68 and above the average reported for digital health apps in a recent meta-analysis, which was also close to 68 [19]. Scores varied widely (Figure 5): eleven participants scored 80 or above, three of them a perfect 100, and six scored below 62, as low as 40. Only two fell in between. The 90% interval for the median, 67.5 to 92.5, is wide, which reflects this spread and the small sample. All six low scorers described themselves as having little experience with computers, and they were also the participants who needed the most help during Scene 2.

The individual SUS items show where the split comes from (Figure 6). Nearly all participants agreed that the functions were well integrated (95%) and that the app was easy to use (89%), and most would use it frequently (84%) and

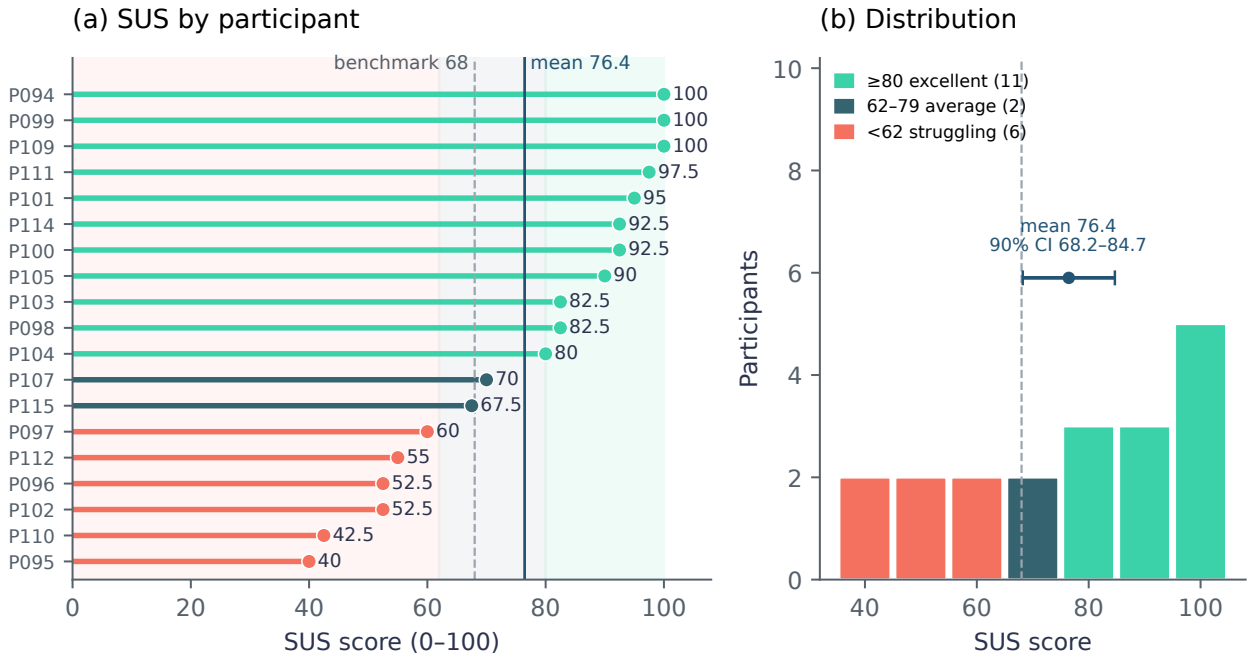


Figure 5: System Usability Scale scores. (a) Each participant's score, colored by band; the dashed line marks the published average of 68 and the solid line the sample mean. (b) Distribution, showing the wide spread of scores, with the mean and its 90% confidence interval.

expected most people to learn it quickly (84%). The two items about needing help were much weaker. Only 58% disagreed that they would need a person's support to use the app, and only 37% disagreed that they needed to learn a lot before getting going; 47% agreed with that statement. Participants found the app pleasant and coherent, but many felt they could not have started without help.

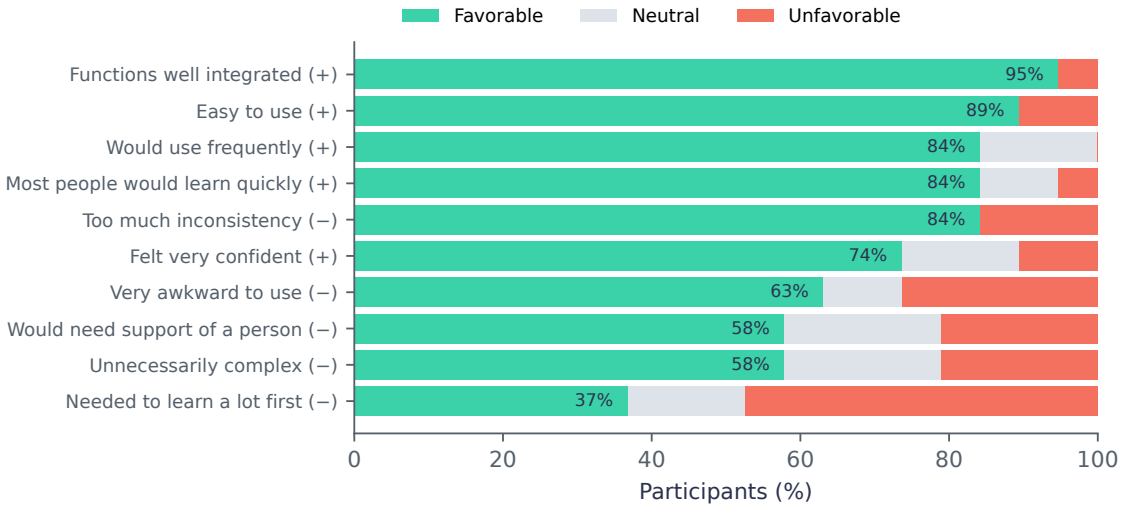


Figure 6: Responses to the individual SUS items, recoded so that "favorable" means agreement with a positive item (+) or disagreement with a negative item (−). Items are sorted by the share of favorable responses.

4.2 Task Ease (RQ2)

Mean task-ease ratings ranged from 4.00 to 4.93 (Figure 7). Onboarding (4.00) and the assessment (4.16) were the hardest tasks, and they are where the login, safety-check, positioning, and distance problems described below occur.

Exercises (4.68) and history (4.93) were the easiest once participants had found them. Results (4.53) fell in between: easy to open but, as Section 4.5 shows, often hard to interpret. Onboarding received a lower mean ease rating than exercises, and their 90% intervals did not overlap (3.65–4.35 vs. 4.45–4.92).

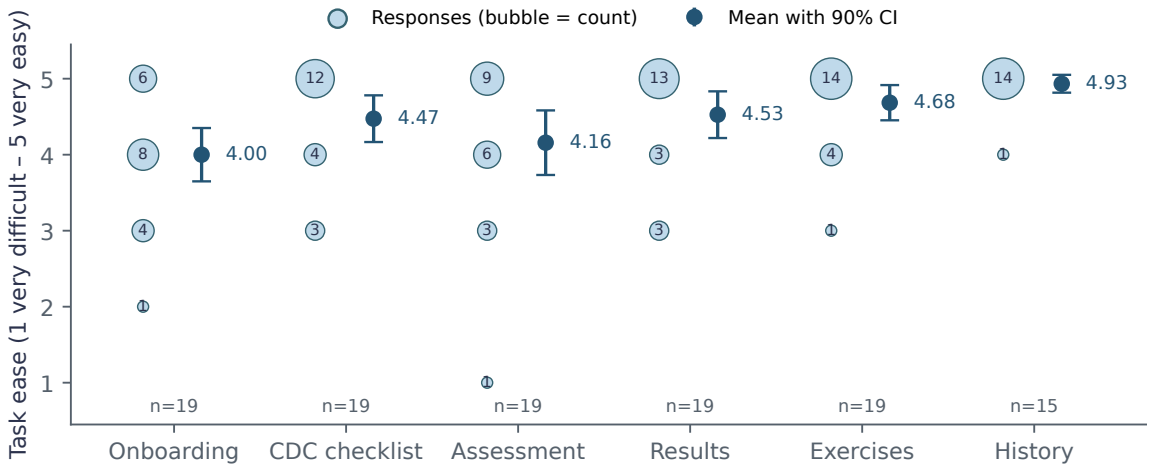


Figure 7: Adapted five-point task-ease ratings by task. Bubbles show how many participants gave each rating; dark points show the mean with its 90% confidence interval.

4.3 Attitudes (RQ3)

All eight attitude statements received mean ratings above 4.5 of 5 (Figure 8). Usefulness of the results and exercises (4.89) and trust in the results (4.84) were rated highest, and their entire confidence intervals lie above 4.65. The value of the app was clear to participants, and they believed what it told them. The two lowest-rated statements were clarity of the instructions (4.53) and feeling confident and safe during the tests (4.58). These match the problems with the distance markings, the tilt message, and audio prompts that some participants could not hear. Comfort using the app alone at home (4.63) is the measure most relevant to adoption, and it is the one to watch among participants with less technology experience.

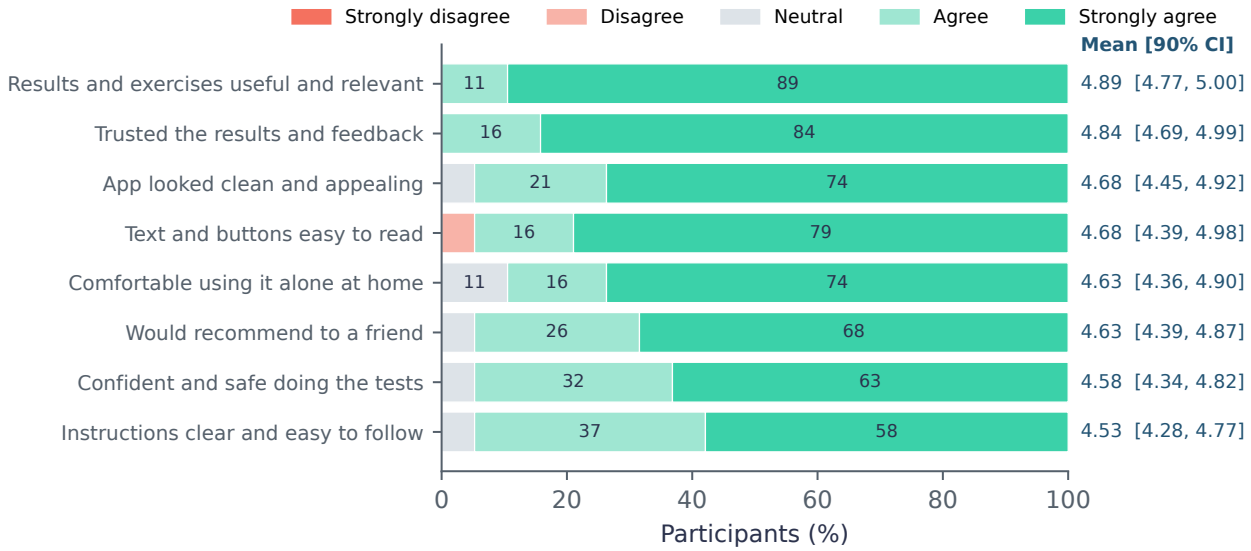


Figure 8: Attitude ratings. Bars show the distribution of responses; the right-hand column gives the mean and 90% confidence interval.

4.4 Time on Task and Timing Validation

Figure 9 shows task times for each participant, and Table 3 gives each median with its 90% interval. Finding the exercises took a median of 1.8 minutes (90% CI 1.5–2.5): 12 of 19 participants took 1.8 minutes or less, and the other seven took 2.5 to 4.7 minutes; in the recordings, the longer times reflected searching for the section rather than difficulty using it. Onboarding took a median of 4.4 minutes (90% CI 2.8–5.9; range 2.0–7.6). The transcript window for the assessment includes pauses, repeated attempts, and discussion. Its fastest clean runs took about 7 minutes (P094, P114), and its longest, 26.8 minutes (P097), was mostly time the participant was not using the app.

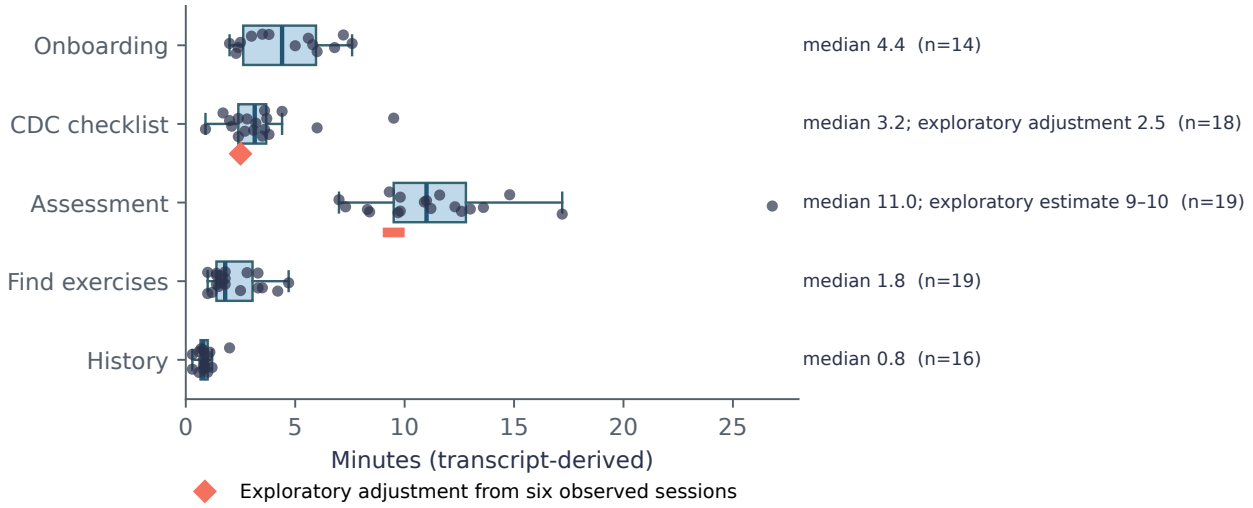


Figure 9: Time on task from the session transcripts. Boxes show the interquartile range and median; dots show individual participants. Coral markers show exploratory estimates from the six sessions with a live observer.

The live-observer comparison supported the transcript method (Figure 10). For finding the exercises and for history, the two methods agreed (median difference 0.0 and -0.2 minutes). For the CDC checklist, the transcript times ran consistently longer than the observer's, by 0.7 minutes on average, because they include reading the prompt and asking for the rating. The median transcript-derived checklist duration was 3.2 minutes. An exploratory adjustment based on the six observed sessions (subtracting 0.7 minutes) yielded an estimated duration of 2.5 minutes; this estimate does not account for uncertainty in the adjustment itself, so no interval is given for it. For the assessment, the gap varied from 0.8 to 5.6 minutes between participants, too much to correct with a fixed offset. The median transcript-derived assessment time was 11.0 minutes; in the six observed sessions the assessment took a median of 9 to 10 minutes, which we report only as an exploratory estimate. These six sessions are a reasonable anchor: their transcript median (11.7 minutes) is close to that of the whole study (11.0), and subtracting their median gap from the whole-study median gives 8.9 minutes. Onboarding was not corrected, because the observer timed login and the profile check separately. Because the observer rounded to whole minutes, about half a minute of each difference reflects rounding.

The observer's records also showed something the questionnaires could not. Her record of each participant's spoken task-ease rating matched the questionnaire for all six participants on every task, which confirms that the form was completed accurately. Her own independence ratings, however, were sometimes much lower than the participant's. P103 rated the CDC checklist 5, while the observer recorded "with help, 2"; P107 rated the assessment 5, while she recorded "with help, 3". High self-reported task ease sometimes co-occurred with observed assistance. Task-ease ratings should therefore not be interpreted as evidence of independent completion. Two explanations are plausible, and the data cannot separate them. Participants may rate the outcome rather than the process: the task was completed, so it felt easy, and the help that made completion possible goes uncounted. Self-presentation may also contribute: participants may be reluctant to report difficulty, either to avoid confirming the stereotype that older adults struggle with technology or to avoid criticizing an app whose developers were in the room. Either way, for an app meant to be used alone, self-rated ease is not a safe proxy for independent use, and an observer rating should be recorded alongside it.

4.5 Usability Problems

We identified 21 distinct usability problems. Eight occurred in at least half of the 18 recorded sessions, and the same problems recurred throughout the study. Figure 12 shows every problem, ordered by frequency, together with the participants who encountered it. Table 4 describes the most frequent ones with representative quotes.

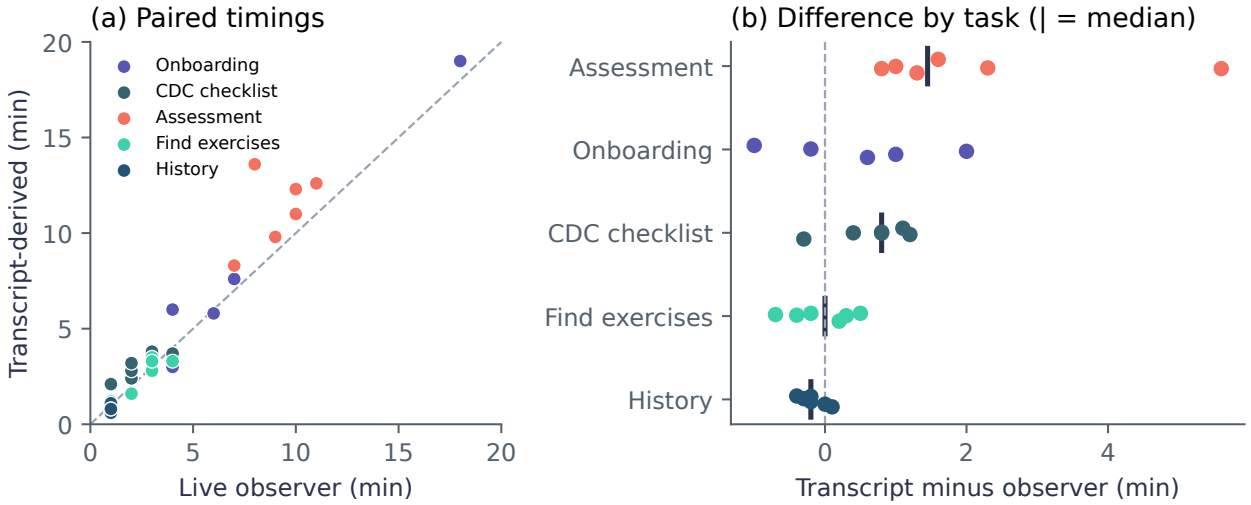


Figure 10: Transcript-derived timings compared with a live observer for the six observed sessions. (a) Paired timings; the dashed line marks perfect agreement. (b) Difference (transcript minus observer) by task.

Table 4: The most frequent usability problems, with representative quotes.

Problem	Sessions	What happened
Checklist toggle has no clear yes/no	13/18	Leaving a toggle off did not read as "no". Participants tapped it to answer no, asked whether their tap had registered, or typed "no" into the notes. P104: "I didn't know that no was automatic. It would be better if that were green instead of gray."
Results header showed fall risk, not mobility	12/18	The header still reported fall risk rather than mobility, so a good result appeared as "low" and was read as poor mobility. P096: "Is it a low risk of falling, or is my balance low?"
Login button hidden by the keyboard	11/18	The on-screen keyboard covered the login button and the page could not scroll. P104: "Where do I hit enter?"
Weight and height need an unmarked scroll	11/18	The profile screen hid these fields below the visible area with no cue. P110: "I have no idea."
Results hard to interpret	11/18	Times were read as speeds, so faster results looked worse, and the CDC-versus-current comparison was misread. P105: "So I'm a little slower?" P102, who had beaten the CDC time: "I need to practice more."
Safety check hides its length	10/18	Items appeared one at a time with no scroll indicator and no count, so participants did not know how many remained, did not realize the list continued below the visible area, and stalled when the start button stayed inactive. P105: "How would I have known that?"
Left pane not seen as navigation	9/18	The checklist and exercise sections went unfound. P104: "I thought it was a picture."
Distance and the 10 to 20 ft switch	9/18	The camera needed more than the marked 10 feet, and nothing signaled the switch to 20 feet for the TUG. P115: "moving from the 10-foot distance back to 20 feet, that was not clear."

Smaller problems included closing the profile overlay (5 of 18 sessions), chair-stand repetitions that were not counted (4), small text (4), exercise details hidden behind a chevron (4), unclear wording about falls and injuries (3), audio prompts too quiet to hear (2), and exercise doses given in unfamiliar "rounds" (2).

Problems along the task flow. Figure 11 places each problem at the step where it occurred. Most of the friction sits at the start of the experience, in onboarding, the checklist, and the safety check that precedes the movement tests, and at the results screen. Once participants reached the exercises and history sections, those tasks were easy; the difficulty was finding them.

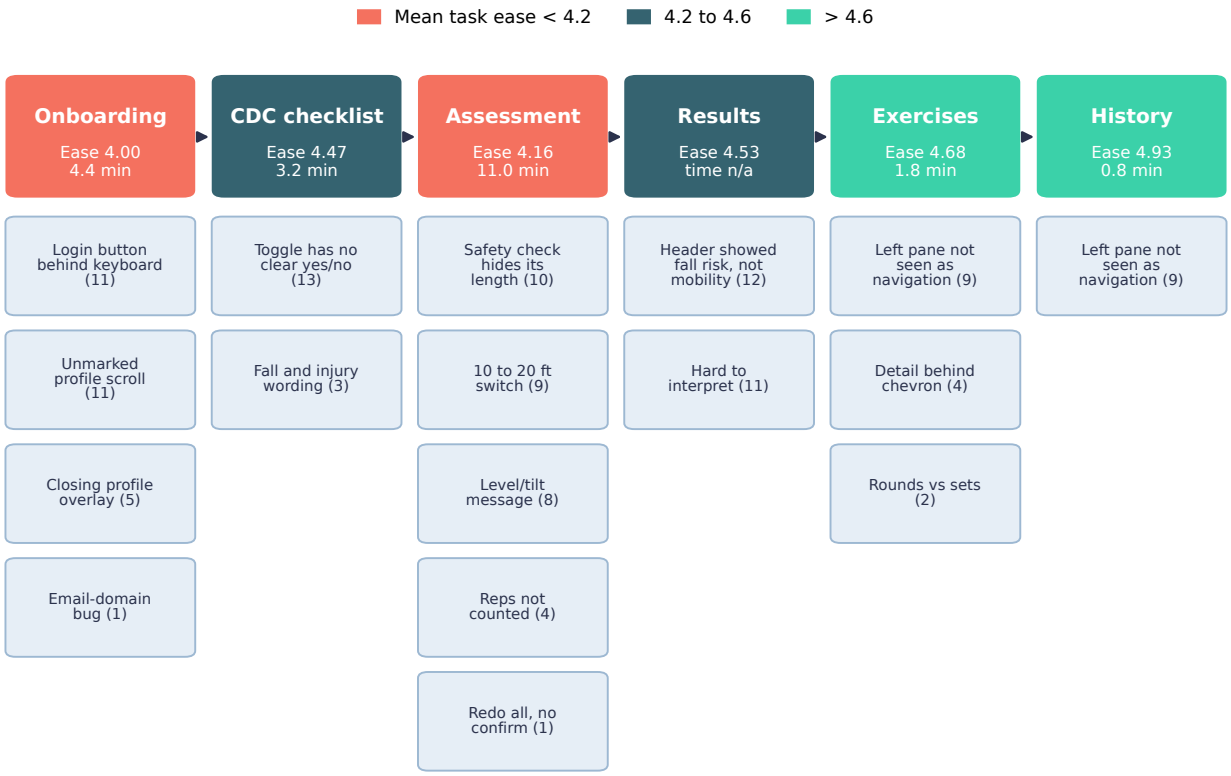


Figure 11: Usability problems mapped onto the AVA at Home task flow, with the mean task-ease rating and the median transcript-derived time for each step (Table 3 also gives exploratory estimates from the observed sessions).

4.6 Defects with Usability Consequences

Two software defects each occurred in only one session but blocked progress entirely. For these, frequency is the wrong axis: their importance lies in severity, so they are reported here rather than ranked with the problems above. First, email validation rejected short domains, locking P104 out for about 15 minutes. The rejected domains were cs.com and aol.com, providers used disproportionately by people in their eighties, so the impact of this rule falls on exactly the population the product targets; it is a finding about designing for older adults, not a generic defect. Second, *Redo all* deleted a completed assessment without asking for confirmation, so P115 lost all of their results when trying to repeat a single stage. A destructive action with no confirmation and no undo violates the error-prevention heuristic [20].

4.7 Who Struggled

Figure 12 shows which problems each recorded participant encountered, with participants ordered by SUS score. The most frequent problems affected high and low scorers alike. Participants who scored below 62 encountered a median of 7 problems, the same as those who scored 62 or above, and the number of problems a participant encountered was not related to their SUS score (Spearman $\rho = -0.16$, $p = 0.53$; Figure 13). We did not detect an association between age and SUS in this sample ($\rho = -0.19$, $p = 0.43$). In contrast, SUS was strongly related to mean task ease ($\rho = 0.78$, $p < 0.001$), and participants who took longer over the assessment tended to score lower ($\rho = -0.45$, $p = 0.06$).

Lower- and higher-scoring participants therefore encountered similar median numbers of coded problems. Session observations suggest that difficulty recovering from interface problems may contribute to lower usability ratings: some participants comfortable with tablets worked out how to dismiss a keyboard or find a hidden scroll quickly, while some with little technology experience needed the moderator. Recovery ability, however, was not systematically measured. This is consistent with the SUS items on needing support and needing to learn a lot first, and with earlier research showing that technology experience and confidence, more than age, predict whether older adults adopt new technology [21]. Frameworks of aging-related barriers to mobile health usability likewise identify unfamiliarity with interface conventions, such as hidden scrolling and keyboard behavior, as a key obstacle that design can reduce [22]. Technology

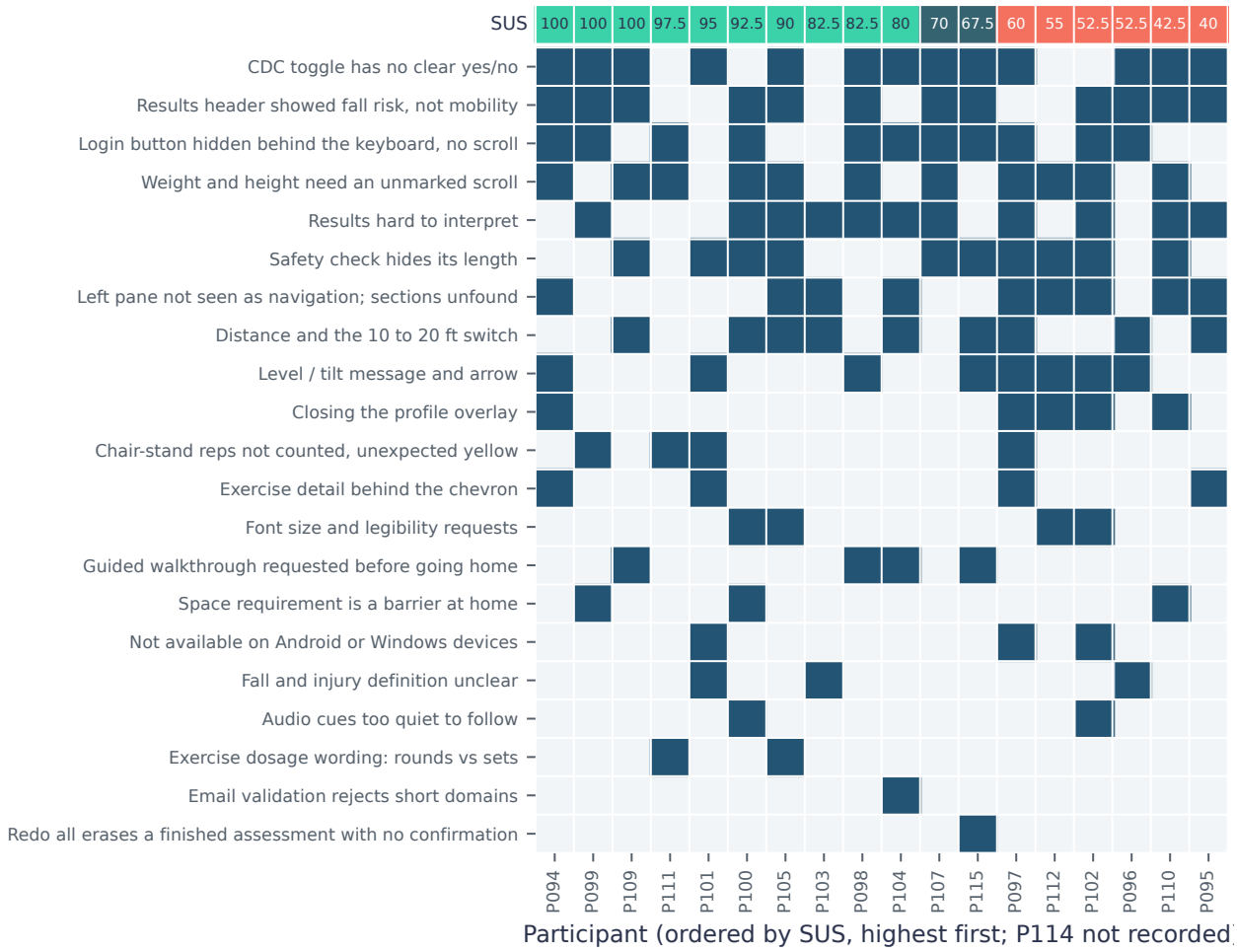


Figure 12: Usability problems encountered by each recorded participant. Participants are ordered by SUS score, shown in the top row; problems are ordered by frequency.

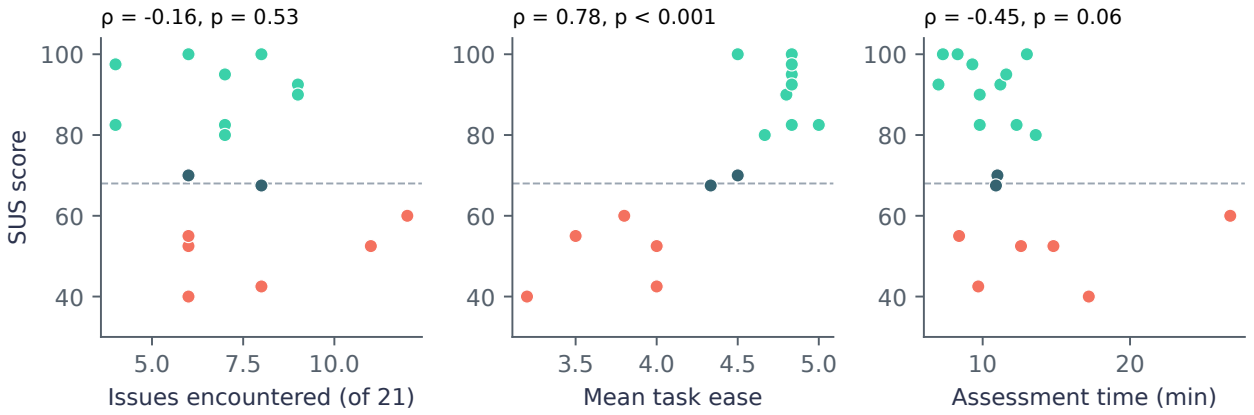


Figure 13: SUS score against (a) the number of usability problems encountered ($n = 18$), (b) mean task ease ($n = 19$), and (c) assessment time ($n = 19$). Points are colored by SUS band; the dashed line marks 68.

use among adults aged 65 and over has grown substantially in the past decade [23], but this growth is uneven, and the least experienced users remain the group most at risk of being left behind. Whether removing the common problems helps the least experienced users most is a question for the next evaluation.

4.8 What Is Working

Participants' comments and ratings identified strengths that should be preserved. The app is learnable: P097 noted it "worked well the second time through the app," and P098 said "once I understood it, I'm fine, it'd be a 5." The design was seen as credible and uncluttered (P101: "Your presentation is very clear, it's not cluttered"). The exercise content was easy to follow; P111 read a full exercise correctly from start to finish. Participants also understood the value of repeated testing. P105 said, "I have my baseline today. If I do this every month, am I getting stronger?", and P100 left the session to fetch a laptop to install the app.

4.9 Barriers Beyond the Interface

Several findings concern adoption rather than interface design. The TUG needs about 20 feet of clear floor, and some participants doubted they had that at home (P099: "I don't have 10 feet of clear space"). The app runs on iPad and MacBook; two participants had only Android devices, and one used a Windows desktop. Four participants asked, unprompted, for a guided walkthrough before using the app at home (P104: "it'd be very helpful if a therapist went through it one time with you before they sent you home"), echoing earlier findings that older adults want clear instruction and support when first using tablets [24]. Typing passwords was itself a barrier for some, which suggests biometric sign-in after the first login. Finally, printed materials appear to be the main artifact for many residents (P094 cited earlier research suggesting that most prefer printed sheets), so the emailed PDF report and the install link it contains matter more than they might seem.

4.10 Implications for Design

The findings identify a set of interface changes for evaluation. The highest-priority problems were those that most often stopped participants or required the moderator's help: the login button hidden by the keyboard, the checklist control without an explicit yes/no choice, the results header and its "lower than" wording, the safety check without visible progress, the rejection of short email domains, and the absence of a confirmation before *Redo all* discards a completed assessment. The next group concerns orientation and interpretation: making the left-hand pane read as navigation, stating that lower times are better and showing the CDC reference beside each result, and making the distance guidance consistent, including the move from 10 to 20 feet. Smaller refinements include familiar exercise terms (sets and repetitions rather than rounds), on-screen cues to accompany audio prompts, and a larger-text option. A brief guided first run at the clinical hand-off, requested by several participants, also warrants evaluation, including its staffing requirements.

Three problems need more detail, because they bear on whether a self-administered result can be trusted.

- **Results header showed fall risk, not mobility (12 of 18 sessions).** The header still reported fall risk, where low is good, inside an app that now reports mobility, where low is poor. A good result therefore appeared as "low" and was reasonably read as poor mobility; this was a terminology mismatch left over from the app's repositioning, not a participant error. This was a confirmed software defect in the app's wording, recognized by the moderator during the sessions. It did not prevent participants from completing the assessment, but it affected how they interpreted their result. It remains open.
- **Chair-stand repetitions not counted (4 of 18 sessions).** Some participants saw fewer chair stands counted than they expected, and an unexpected yellow result. The app does not count hand-assisted or partial stands, and lists them separately as uncounted (Figure 1d); the problem was therefore reported by participants as an unexplained count rather than confirmed as a software defect. It did not prevent completion, but it affected how participants interpreted their chair-stand result. It remains open; how the app explains uncounted stands has not been changed.
- **Positioning instructions (distance: 9 of 18; level and tilt: 8 of 18).** The camera needed more room than the marked 10 feet, nothing signaled the move to 20 feet for the TUG, the word "level" was read as "flat", and the tilt arrow pointed the wrong way. These were confirmed problems with the app's instructions. They delayed the assessment and, in at least one session (P115), prevented the TUG from being completed. Several related fixes were deployed between sessions, but the problems are only partly resolved.

5 Discussion

The central finding is that usability ratings were favorable for most participants and low for a substantial minority. Lower- and higher-scoring participants encountered similar median numbers of coded problems. Session observations suggest that difficulty recovering from interface problems may contribute to lower usability ratings, but recovery ability

was not systematically measured. This shapes how improvement should be judged. A higher average SUS could be achieved by making the app even easier for people who already find it easy. The more important target is the lower group: success in the next round means the six scores below 62 rising, not the top scores staying at 100. The findings identify specific interface changes for evaluation against this target.

The study also shows how self-reported ratings can mislead in assisted sessions. Participants who received help still rated the task as easy, because they had completed it. For an app meant to be used alone, the relevant question is whether the task could have been done without help. Future rounds should record an observer's independence rating (unaided, a little help, or with help) next to every self-rating, so that assisted and unaided completions can be separated.

Finally, the results point to the clinical hand-off as part of the design. The scenario already gave participants a clinic introduction, and four still asked for a walkthrough of the app itself before using it at home. A brief guided first run, led by the physical therapist, warrants evaluation, including its staffing requirements, because it targets the SUS item on needing to learn a lot first. Together with the P0 fixes, it is a candidate change for users with the least technology experience.

5.1 Limitations

The sample was opportunistic, small ($n = 19$), and drawn from a single senior living community, so the numbers are indicative. The "at home" scene took place in a community room with the moderator present, not in participants' homes, so home-specific barriers such as space and lighting were discussed rather than observed. Moderators sometimes helped or explained features in advance, which raises ease ratings and makes the frequency of some problems conservative. Task success was not reported, for the reasons given above. Task-ease ratings were sometimes collected together at the end of the session. One participant declined recording, so problem frequencies are based on 18 sessions. Several fixes were deployed between sessions during the study, so the frequencies describe the builds in use during the study, not the current one, and none of the fixes has yet been retested. Technology experience was self-described, not measured. The study was run by the team that develops AVA at Home, and the moderator also played the participant's physical therapist. Social desirability may therefore have raised the attitude and SUS scores, which is one reason we give more weight to the observed problems and task timings. Caregivers were not part of this study, so we did not evaluate how a caregiver or family member might support, or change, an older adult's use of the app.

For the next round, we recommend keeping the same codes, tasks, and timing rule so that results can be compared directly. We also recommend collecting each task-ease rating immediately after its task, starting recordings before login, avoiding interruptions during the assessment, recording an observer independence score for every task, and measuring technology experience with a brief validated questionnaire.

6 Conclusion

This study evaluated the usability of AVA at Home, a self-administered, vision-based mobility assessment, with older adults after a brief clinic introduction. Nineteen participants aged 70 to 91 gave the app a mean SUS of 76.4 and a median of 82.5, and rated its usefulness and the trustworthiness of its results above 4.8 of 5. Scores varied widely: 11 participants scored 80 or above and six scored below 62. Onboarding and the assessment received the lowest task-ease ratings, and five problems occurred in at least 11 of 18 recorded sessions: a checklist toggle that did not read as "no", a results header that showed fall risk rather than mobility, a login button hidden by the keyboard, an unmarked scroll on the profile screen, and results that were hard to interpret. Two bugs blocked progress. Lower- and higher-scoring participants encountered similar median numbers of coded problems, and high self-reported task ease sometimes co-occurred with observed assistance. AVA at Home received generally favorable usability ratings in a moderated evaluation. However, some participants required assistance, and independent task completion was not systematically recorded. These findings identify barriers to independent use but do not establish feasibility of unaided home use.

The findings set an agenda and a baseline for the next design round. The changes identified in Section 4.10, together with a guided first run at the clinic hand-off, should be evaluated in a follow-up study that records observer-rated independence for every task. That study should be compared against the baseline established here: a median SUS of 82.5 (90% CI 67.5–92.5) with six participants below 62; transcript-derived median times of 4.4 minutes for onboarding, 3.2 for the checklist, 11.0 for the assessment, 1.8 for finding the exercises, and 0.8 for history; and the five most frequent problems occurring in 13, 12, 11, 11, and 11 of 18 sessions. Together with our earlier validation of AVA's measurement accuracy [6, 7], this work identifies what must change before self-administered mobility assessment can be tested for unaided use at home.

Conflict of Interest

The authors are employed by ForesightCares, which develops AVA at Home.

References

[1] Centers for Disease Control and Prevention. Important facts about falls. `https://www.cdc.gov/falls/data-research/index.html`, 2023. Accessed: 2025-11-14.

[2] Centers for Disease Control and Prevention. STEADI: Stopping elderly accidents, deaths and injuries. `https://www.cdc.gov/steadi/index.html`, 2023. Accessed: 2025-01-15.

[3] Manuel Montero-Odasso, Nathalie van der Velde, Finbarr C. Martin, Mirko Petrovic, Maw Pin Tan, Jesper Ryg, et al. World guidelines for falls prevention and management for older adults: A global initiative. *Age and Ageing*, 51(9):afac205, 2022.

[4] Clemens Kruse, Joanna Fohn, Nakia Wilson, Evangelina Nunez Patlan, Stephanie Zipp, and Michael Mileski. Utilization barriers and medical outcomes commensurate with the use of telehealth among older adults: Systematic review. *JMIR Medical Informatics*, 8(8):e20359, 2020.

[5] Kieran Moore, Emma O'Shea, Lorna Kenny, John Barton, Salvatore Tedesco, Marco Sica, et al. Older adults' experiences with using wearable devices: Qualitative systematic review and meta-synthesis. *JMIR mHealth and uHealth*, 9(6):e23832, 2021.

[6] ForesightCares, Inc. AVAL: A validation pipeline for AVA mobility. Technical report, ForesightCares, Charlotte, NC, 2026.

[7] ForesightCares, Inc. AVAL.v2: Multi-assessment validation and error analysis of on-device mobility AI for fall-risk screening. Technical report, ForesightCares, Charlotte, NC, 2026.

[8] K. Anders Ericsson and Herbert A. Simon. *Protocol Analysis: Verbal Reports as Data*. MIT Press, Cambridge, MA, revised edition, 1993.

[9] International Organization for Standardization. ISO 9241-11:2018 ergonomics of human-system interaction, part 11: Usability: Definitions and concepts. ISO, Geneva, 2018.

[10] World Medical Association. World medical association declaration of Helsinki: Ethical principles for medical research involving human subjects. *JAMA*, 310(20):2191–2194, 2013.

[11] John Brooke. SUS: A "quick and dirty" usability scale. In Patrick W. Jordan, B. Thomas, Bernard A. Weerdmeester, and Ian L. McClelland, editors, *Usability Evaluation in Industry*, pages 189–194. Taylor & Francis, London, 1996.

[12] Aaron Bangor, Philip Kortum, and James Miller. Determining what individual SUS scores mean: Adding an adjective rating scale. *Journal of Usability Studies*, 4(3):114–123, 2009.

[13] Jeff Sauro and James R. Lewis. *Quantifying the User Experience: Practical Statistics for User Research*. Morgan Kaufmann, Cambridge, MA, 2nd edition, 2016.

[14] James R. Lewis. The system usability scale: Past, present, and future. *International Journal of Human–Computer Interaction*, 34(7):577–590, 2018.

[15] James R. Lewis and Jeff Sauro. Item benchmarks for the system usability scale. *Journal of Usability Studies*, 13 (3):158–167, 2018.

[16] Jeff Sauro and Joseph S. Dumas. Comparison of three one-question, post-task usability questionnaires. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '09)*, pages 1599–1608. ACM, 2009.

[17] Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall, New York, 1993.

[18] Robert A. Virzi. Refining the test phase of usability evaluation: How many subjects is enough? *Human Factors*, 34(4):457–468, 1992.

[19] Maciej Hyzy, Raymond Bond, Maurice Mulvenna, Lu Bai, Alan Dix, Simon Leigh, and Sophie Hunt. System usability scale benchmarking for digital health apps: Meta-analysis. *JMIR mHealth and uHealth*, 10(8):e37290, 2022.

[20] Jakob Nielsen. Enhancing the explanatory power of usability heuristics. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '94)*, pages 152–158. ACM, 1994.

[21] Sara J. Czaja, Neil Charness, Arthur D. Fisk, Christopher Hertzog, Sankaran N. Nair, Wendy A. Rogers, and Joseph Sharit. Factors predicting the use of technology: Findings from the center for research and education on aging and technology enhancement (CREATE). *Psychology and Aging*, 21(2):333–352, 2006.

[22] Gaby Anne Wildenbos, Linda Peute, and Monique Jaspers. Aging barriers influencing mobile health usability for older adults: A literature based framework (MOLD-US). *International Journal of Medical Informatics*, 114: 66–75, 2018.

[23] Michelle Faverio. Share of those 65 and older who are tech users has grown in the past decade. Pew Research Center, 2022. https://www.pewresearch.org/short-reads/2022/01/13/share-of-those-65-and-older-who-are-tech-users-has-grown-in-the-past-decade/.

[24] Eleftheria Vaportzis, Maria Giatsi Clausen, and Alan J. Gow. Older adults perceptions of technology and barriers to interacting with tablet computers: A focus group study. *Frontiers in Psychology*, 8:1687, 2017.