Recall-Grounded Scenarios: An Approach for Studying Longitudinal User Experiences in Short Sessions

Md Aseer Intiser
Tampere University
Tampere, Finland
md.aseerintiser@gmail.com

Thomas Olsson
Tampere University
Tampere, Finland
thomas.olsson@tuni.fi

Amir Pakpour Haji Agha
Tampere University
Tampere, Finland
amir.pakpourhajiagha@tuni.fi

Abstract

Many experiences that matter for good design unfold over weeks or months, yet a typical HCI research session lasts tens of minutes. Studying such an experience within a single short session is therefore a methodological problem in its own right. It is not just a problem tied to one system or population. We report a recall-grounded scenario design developed to address this challenge, using international newcomers' months-long transition into Finnish public services as a case study. Participants were invited to re-enter their first months in Finland through a structured recall prompt. They then engaged four scenario-based interactions with a conversational assistant prototype, responding from that earlier self's perspective. The framing was renewed before each scenario. The design was refined across two pilot sessions that surfaced three problems, conversational dead-ends, overly open scenarios, and session fatigue, each addressed by a specific protocol change. Across fifteen main-study sessions of approximately 35 minutes each, fourteen of fifteen participants spontaneously compared the assistant to their own remembered lack of support on arrival. Their accounts also carried detail tied to specific remembered circumstances. We also report a disconfirming case, the evidence's limits, and the design's relevance to other single-session studies of transitions that matter for accessibility and inclusion.

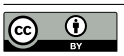
CCS Concepts

• **Human-centered computing** → **User studies**; **Empirical studies in HCI**.

Keywords

qualitative methods, scenario-based research, retrospective recall, newcomer experience, single-session studies, accessibility and inclusion, human-computer interaction

ACM Reference Format:
Md Aseer Intiser, Thomas Olsson, and Amir Pakpour Haji Agha. 2026. Recall-Grounded Scenarios: An Approach for Studying Longitudinal User Experiences in Short Sessions. In *29th International Academic Mindtrek (Mindtrek 2026), October 27–30, 2026, Tampere, Finland*. ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3843148.3859356



This work is licensed under a Creative Commons Attribution 4.0 International License.
Mindtrek 2026, Tampere, Finland
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2656-9/2026/10
https://doi.org/10.1145/3843148.3859356

1 Introduction

Many experiences that matter for inclusive design unfold over weeks or months: adjusting to a new institutional context, learning to navigate an unfamiliar system, recovering from a disruptive life event. HCI research sessions are typically measured in tens of minutes. Studying such an experience closely, in a way that can inform design, is therefore a methodological problem in itself. It comes before any problem tied to a single population or system. The core question is this: how does a researcher recover months of lived, affectively textured experience inside a session that cannot itself last months?

This paper addresses that problem through a case study in one such setting. Digital systems intended to support newcomers, whether migrants, refugees, or international students, can only work if their design is grounded in real experience. They need to reflect how those users actually go through the transition the system is meant to support. That transition is rarely instantaneous. For international students arriving in Finland, this process is not quick. Registering with public authorities, opening a bank account, finding healthcare, and locating university support unfolds across weeks or months. Each step depends on institutional knowledge the newcomer does not yet have. Digitalisation of public services has been shown to reinforce, rather than reduce, this kind of exclusion when systems assume a baseline of local knowledge that newcomers do not possess [1]. This makes the newcomer transition an accessibility and inclusion problem as much as a usability one. It is also a fitting case study for the broader methodological challenge described above.

Contextual inquiry addresses this kind of problem by having researchers observe and interview users in the context of ongoing work [2, 3]. That approach presupposes that the work is available to be observed as it happens. The lived bureaucratic work of newcomer integration is not available in this way inside the timeframe of a typical qualitative study. Waiting for a longitudinal, deployed system to accumulate real newcomer usage was also not feasible for this study.

Retrospective methods are not new to HCI and social science more broadly, and the design reported here builds on that tradition rather than working around it. The critical incident technique asks participants to recall specific past episodes that were unusually effective or ineffective [4]. The day reconstruction method has participants reconstruct a recent day in structured segments to recover experience that a general survey question would flatten [5]. Retrospective think-aloud has participants narrate a just-completed task from memory rather than as it happens [6]. Related recall-based elicitation techniques continue to be developed outside HCI. Recent work in health sciences uses stimulated recall interviews

alongside role play and other elicitation methods to reach participants' remembered experience [15], and a recent integrative review catalogues a broader family of elicitation techniques developed across qualitative interviewing more generally [16]. Each of the methods above recovers a bounded episode: a single incident, a single day, a single just-finished task. None asks a participant to sustain an earlier version of themselves across a sequence of new, current-tense tasks, renewing that stance before each one. That renewal, not the act of recall by itself, is what distinguishes the design reported here.

This paper reports a *recall-grounded scenario design*: a technique that asks participants to re-enter an earlier period of their own lived experience and to engage scenario-based tasks from that earlier vantage point. The framing is renewed at each new task. We ask two questions. First, how can a qualitative HCI method capture an affectively loaded, months-long transition within a single short session without flattening its emotional texture? Second, what evidence would distinguish genuine engagement with that earlier perspective from generic, detached role-play? We describe the technique and the pilot process that produced its final form. We also present evidence from fifteen main-study sessions bearing on the second question, including one case that complicates a simple positive reading of that evidence.

2 The Recall-Grounded Scenario Design

The technique was developed for a study of international students' interactions with Lumi, a conversational assistant prototype supporting navigation of Finnish public services. Interaction with Lumi was structured as four scenario-based tasks, one in each of four service domains covering the typical newcomer arc: student housing, banking, healthcare, and university support services.

Before the first scenario, participants were invited to think back to their first weeks and months in Finland: where they had lived, what bureaucratic encounters had been on their mind, and what had felt difficult or unfamiliar. Participants were then asked, for the remainder of the session, to respond to the scenarios from the perspective of that earlier self rather than from their current, more settled perspective. This is the recall-grounding prompt, delivered once, in full, before any interaction with the prototype began.

The framing was then renewed immediately before each of the four scenarios, rather than stated once at the outset and left to persist unsupported. Each renewal used a short, consistent script: participants were asked to think back to their first months in Finland, to imagine themselves in the situation about to be described, to interact with Lumi however felt natural, and to think aloud as they went. There was also an explicit reminder that there was no correct answer. This renewal took under thirty seconds and was delivered identically before each scenario. That consistency meant any difference in engagement across scenarios could not be attributed to a change in how the framing was worded. The recall prompt in stage one and the renewal before each scenario are the two moments where the technique is actually applied. The remaining stages are conventional interview and observation structure, and are described only briefly below.

Each session followed a fixed nine-stage structure targeting approximately 35 minutes, with observed sessions running approximately 30 to 45 minutes (Figure 1): welcome, consent, and recall grounding; a pre-interaction interview; four scenario interactions with a short breather between the second and third; and a post-interaction interview. The pre-interaction interview asked three questions: which services had been hardest to navigate on arrival, whether the participant had used any chatbot or digital assistant during that period, and what would have made a digital assistant's information feel trustworthy enough to act on. The post-interaction interview asked seven questions built around specific remembered moments rather than general evaluations: overall impression, a moment of confidence, a moment of uncertainty, whether an assistant like Lumi would have helped during the participant's first months, perceived reliability, information the participant had expected but not found, and a closing open question. Sessions were conducted online via screen share.

Fifteen international students at Tampere University took part in the main study (P01 to P15). Participants were recruited through purposive sampling [7], via researcher networks, internal Tampere University channels, and snowball referral. The inclusion criteria were that participants had moved to Finland within the last three years and had personal experience navigating Finnish public services as newcomers. No financial compensation was offered, and participation was voluntary throughout. Time since arrival varied from three months to more than two years across the sample, so the recall prompt was tested across a range of recall distances rather than a single, recent-arrival cohort. Participation was governed by informed consent, following the Finnish national guidelines on research ethics [8] and the EU General Data Protection Regulation [9] as applied under the researcher's institution. No separate ethics-committee review was required under those guidelines for this non-medical Master's research. The researcher who designed and developed the Lumi prototype was also the study's sole interviewer and analyst. This dual role is a methodological limitation, discussed further in Section 4.

The design above is the outcome of two pilot sessions conducted before the main study, using an earlier version of both the prototype and the protocol. Pilot data are excluded from the evidence reported in Section 3 on that basis. The pilots surfaced three problems, each addressed by a specific fix. First, participants who moved outside the prototype's predefined click-through paths sometimes reached conversational dead ends; this was addressed with a constrained fallback layer. Second, scenarios could often be satisfied at the homepage level without engaging a topic page; scenarios were rewritten to require a concrete sub-task inside a topic page. Third, pilot sessions of 60 to 75 minutes produced visible fatigue, particularly in the post-interaction interview; both interviews were shortened, and a one-minute breather was added between scenarios 2 and 3. This last fix bears directly on the technique itself. Fatigue is not just a duration concern, it is a data-quality concern: tired participants give shorter and less reflective answers, and that is precisely the register the recall-grounding prompt depends on.

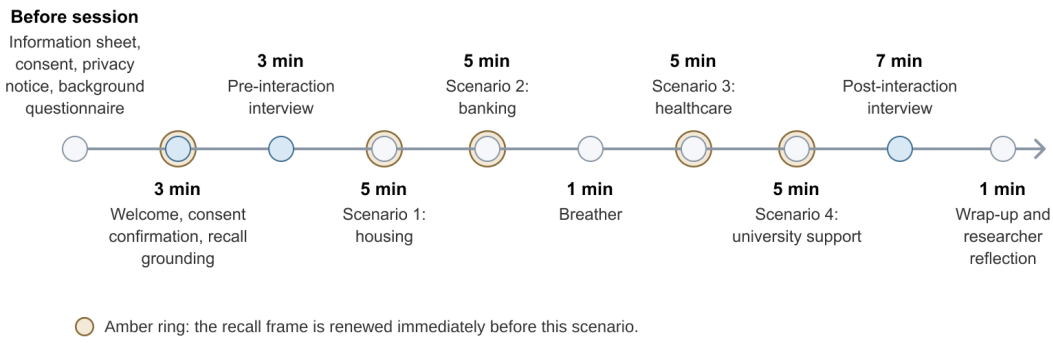


Figure 1: Session structure of the recall-grounded scenario design. The re-anchoring moment before each scenario, not just the scenario sequence itself, is the methodological device under study.

3 Preliminary Evidence of Utility

We coded an utterance as an instance of the retrospective-counterfactual pattern when a participant explicitly compared the assistant to their own remembered absence of comparable support during their arrival in Finland. Comparative or conditional language, such as "better than when I came" or "if I had had this then," counted as this pattern. General praise or dissatisfaction not framed against that remembered baseline, for example a participant simply calling a feature clear or confusing, did not qualify. Fourteen of the fifteen main-study participants produced at least one such utterance across the session. The remaining participant engaged fully with all four scenarios but did not produce an utterance meeting this criterion. We did not probe directly for the absence of such a comparison. So we cannot say whether this reflects a genuinely different disposition toward the recall frame, or simply that the topic did not come up in that particular conversation. For this reason, we report the count as fourteen of fifteen, not fifteen of fifteen. P10, describing the university-support scenario, offered a representative instance: "This is much better than when I came in and had no idea what services my university has."

A natural alternative reading is that any interview inviting reflection on a newcomer-support tool would elicit this kind of comparison, regardless of whether a recall-grounding prompt was used. This is because the post-interaction interview's fourth question directly asks participants whether an assistant like Lumi would have helped them on arrival. That reading cannot be fully ruled out from the utterance count alone. It is weakened, however, by the fact that instances of the pattern were not confined to responses to that question. P10's quotation above occurred spontaneously during the fourth scenario itself, before the post-interaction interview had even begun. That timing is harder to explain as a response to a leading question. It reads more plausibly as evidence of a frame the participant was actively holding during the task. We did not code session stage for every instance of the pattern. So we cannot report what proportion arose spontaneously during scenarios rather than in response to the interview question. We flag that as a gap a fuller report should close. This points to a more general challenge in evaluating a technique like this one. The study has no baseline condition to compare against, so the argument for the technique's contribution cannot rest on a controlled contrast. What we offer instead is a converging set of weaker signals: the timing of spontaneous instances such as P10's, the circumstantial specificity described next, the disconfirming case that follows it, and our own reading as the researchers who conducted every session. This is not a single decisive measurement, but several signals pointing the same way.

Participant accounts also carried detail tied to specific, individually remembered or anticipated circumstances rather than generic evaluative statements. In the pre-interaction interview, P04 described arriving in Finland with a fractured leg, a detail that shaped that participant's later engagement with the healthcare scenario. During the healthcare scenario, P14 said, "If I am sick, I do not want to read too much." P15 distinguished the scenario's relevance to their own situation from a family context: "Student healthcare is one thing, but family healthcare is another." We read this specificity as evidence that participants were drawing on remembered or anticipated circumstances of their own, rather than offering abstracted, hypothetical assessments. We do not treat it as proof that every participant's engagement was equally grounded. The sample includes real variation.

Engagement with the recall frame was not uniformly positive, and we treat that variation as evidence in its own right rather than noise to set aside. P13 gave the lowest reliability rating in the sample, 70 percent. He identified specific content gaps, including a missing police-card requirement in the banking scenario and location-specific information that did not match his city. He also stated he would trust the assistant only if it were trained on local, Tampere-specific data. Rather than disengaging, he proposed a specific technical fix: a different, locally-trained model backend. The recall-grounded protocol did not produce uncritical enthusiasm in his case. We take that as a reason for somewhat more confidence in the pattern reported above, not less. A design that only produced enthusiasm would be easier to dismiss as demand characteristics, or as a polite performance for the researcher.

Three of the fifteen main-study sessions were affected by technical incidents unrelated to the recall technique itself. One session experienced a prototype freeze during the first scenario, and two sessions experienced a fallback-layer failure due to token limits. We report these as constraints on the prototype rather than on the scenario design, and return to them in Section 4.

The judgments above, including which utterances count as instances of the counterfactual pattern, rest on one researcher's interpretive reading rather than an inter-rater-checked procedure. This is a genuine limit on how much independent corroboration a reader can expect. But it is not a departure from the analytic tradition the underlying study follows. This coding process applied the same reflexive thematic analysis procedure used throughout the underlying study [10]. That tradition treats inter-coder reliability as the wrong standard for interpretive work of this kind. The aim, instead, is a defensible reading grounded in the data, not a categorical count that a second coder should reproduce exactly [11]. Coding was conducted manually, without qualitative analysis software. This choice suited a dataset small enough to hold in working memory, and it supported the close engagement with participants' own language that this analytic tradition depends on [12].

The problem this design responds to, a longitudinally unfolding, affectively significant experience that must be studied within a single research session, is not specific to newcomer public-service navigation. The same recall-and-renewal structure may suit other single-session studies of transitions that unfold over an extended period and that participants can meaningfully recall. Examples include workplace onboarding, adjustment following a health diagnosis, or other life-stage transitions where a deployed, longitudinally-used system is not available to the researcher. This kind of transfer matters for accessibility and inclusion research specifically. Many of the populations most affected by exclusionary system design, newcomers, people newly navigating a health condition, people adapting to an unfamiliar institutional context, are exactly the populations whose relevant experience is distributed across months rather than available in a single sitting. The methodological contribution, not just the newcomer population, is what we position as relevant to accessibility and inclusion research: any study of a person's gradual adaptation to, or exclusion by, a system runs into the same single-session constraint, whatever produces that gradual process. This is squarely the situation facing much disability and accessibility research, where adaptation to a newly acquired impairment or to an assistive technology's learning curve unfolds over months, not minutes. Validating the design with disabled participants specifically is accordingly the most direct next step for testing that broader claim.

4 Limitations

The recall-grounded design works with reconstructed rather than observed experience. Retrospective self-report is a known source of memory distortion [13]. We cannot rule out that participants' accounts of their earlier selves were shaped by what they have since learned. We made this trade-off deliberately, choosing affective and experiential richness over behavioural realism. No comparison condition using current, first-time users of a deployed system was run alongside it. So we assert the design's advantage over that alternative here, rather than demonstrate it. A natural next step is a controlled comparison. The same four scenarios would be run against two conditions: the recall-grounded framing used here, and participants' direct, current-tense account of whatever service channel they already rely on for the same tasks. That comparison would test the recall prompt's contribution directly, rather than infer it from the pattern reported in Section 3, and would also show whether the technique adds anything beyond a standard usability comparison against the existing system.

The evidence in Section 3 comes from a single site, a sample of fifteen participants, and sessions conducted in English, a non-native language for both participants and researcher. It also rests on a single researcher who both designed the Lumi prototype and conducted the interviews and analysis. That dual role is common in research-through-design work. But it carries an obvious risk: the same person who built the system might read participants' responses to it more generously than an independent analyst would. We consider this risk managed, not eliminated. Two things support that: the negative case reported in Section 3, and our treatment of the prototype's specific failures as analytic material rather than defects to explain away.

Three sessions were affected by technical incidents in the underlying prototype. We chose to retain and report on these sessions rather than exclude them, on the view that a participant's response to a system failure is itself informative. A different, equally defensible study could have excluded them instead.

We propose the transfer described in Section 3 without having tested it. The design has been used in one domain, with one population, in one linguistic and institutional context. We would expect the pilot problems reported in Section 2, conversational dead-ends, scenario openness, session fatigue, to recur in some form in other domains. This is because they concern the interaction medium and session structure, not the newcomer context specifically. That expectation is a direction for future work, not a finding of the present study.

We report these limitations in this much detail deliberately. Qualitative research is judged in part by the sincerity and transparency with which its constraints are named, not only by the strength of its findings [14].

5 Conclusion

We have described a recall-grounded scenario design developed to study a months-long newcomer transition within a single short research session. We reported the pilot process that produced its final form, evidence bearing on whether it elicited genuine engagement with participants' remembered experience, and a case that complicates a simple reading of that evidence. We offer the technique as a transferable device for other HCI and accessibility research facing the same constraint, a longitudinal, affectively significant experience studied on a single-session budget.

Acknowledgments

We thank the fifteen participants who contributed their time and their memories of arriving in Finland. Portions of this manuscript's text were drafted and revised with the assistance of a generative AI system (Claude, Anthropic), working from the authors' own research materials and data under author direction. All ideas, analysis, interpretations, and conclusions are the authors' own; every factual claim, quotation, and statistic was independently verified against the source material during drafting. The authors reviewed the complete manuscript and take full responsibility for its content and accuracy.

References

[1] Ulla Buchert, Laura Kemppainen, Antero Olakivi, Sirpa Wrede, and Anne Kouvonen. 2023. Is digitalisation of public health and social welfare services reinforcing social exclusion? The case of Russian-speaking older migrants in Finland. *Critical Social Policy* 43, 3 (2023), 375–400.

[2] Hugh Beyer and Karen Holtzblatt. 1998. *Contextual Design: Defining Customer-Centered Systems*. Morgan Kaufmann Publishers, San Francisco, CA.

[3] Karen Holtzblatt and Hugh Beyer. 2017. *Contextual Design* (2nd ed.). Morgan & Claypool Publishers.

[4] John C. Flanagan. 1954. The critical incident technique. *Psychological Bulletin* 51, 4 (1954), 327–358.

[5] Daniel Kahneman, Alan B. Krueger, David A. Schkade, Norbert Schwarz, and Arthur A. Stone. 2004. A survey method for characterizing daily life experience: The day reconstruction method. *Science* 306, 5702 (2004), 1776–1780.

[6] Maaike J. van den Haak, Menno D. T. de Jong, and Peter Jan Schellens. 2003. Retrospective vs. concurrent think-aloud protocols: testing the usability of an online library catalogue. *Behaviour & Information Technology* 22, 5 (2003), 339–351.

[7] Michael Quinn Patton. 2015. *Qualitative Research & Evaluation Methods: Integrating Theory and Practice* (4th ed.). SAGE Publications, Thousand Oaks, CA.

[8] Finnish National Board on Research Integrity TENK. 2019. *The Ethical Principles of Research with Human Participants and Ethical Review in the Human Sciences in Finland*. TENK Guidelines.

[9] European Union. 2016. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 (General Data Protection Regulation).

[10] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2 (2006), 77–101.

[11] Virginia Braun and Victoria Clarke. 2019. Reflecting on reflexive thematic analysis. *Qualitative Research in Sport, Exercise and Health* 11, 4 (2019), 589–597.

[12] Virginia Braun and Victoria Clarke. 2021. One size fits all? What counts as quality practice in (reflexive) thematic analysis? *Qualitative Research in Psychology* 18, 3 (2021), 328–352.

[13] Norman M. Bradburn, Lance J. Rips, and Steven K. Shevell. 1987. Answering Autobiographical Questions: The Impact of Memory and Inference on Surveys. *Science* 236, 4798 (1987), 157–161.

[14] Sarah J. Tracy. 2010. Qualitative quality: Eight "big-tent" criteria for excellent qualitative research. *Qualitative Inquiry* 16, 10 (2010), 837–851.

[15] Laura Ortju, Päivi Kankkunen, Ari Haaranen, and Liisa Karlsson. 2024. Researcher-initiated role play, stimulated recall interview, storycrafting, painting and drawing as research methods to reach children's perspectives in health sciences. *Journal of Advanced Nursing* 80, 7 (2024), 2988–3001.

[16] Renate Kahlke, Lauren A. Maggio, Mark C. Lee, Sayra Cristancho, Kori LaDonna, Zahra Abdallah, Aakashdeep Khehra, Kushal Kshatri, Tanya Horsley, and Lara Varpio. 2025. When words fail us: An integrative review of innovative elicitation techniques for qualitative interviews. *Medical Education* 59, 4 (2025), 382–394.