Author(s): Lukka, Lauri; Karhulahti, Veli-Matti; Bergman, Vilma-Reetta; Palva, J. Matias

Title: Measuring digital intervention user experience with a novel ecological momentary assessment (EMA) method, CORTO
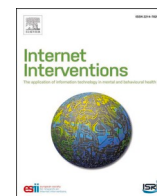
Year: 2024

Version: Published version

Please cite the original version:

Lukka, L., Karhulahti, V.-M., Bergman, V.-R., & Palva, J. M. (2024). Measuring digital intervention user experience with a novel ecological momentary assessment (EMA) method, CORTO. Internet Interventions, 35, Article 100706. https://doi.org/10.1016/j.invent.2023.100706

Measuring digital intervention user experience with a novel ecological momentary assessment (EMA) method, CORTO

Lauri Lukka [a,*], Veli-Matti Karhulahti [b], Vilma-Reetta Bergman [a], J. Matias Palva [a,c,d]

[a] *Department of Neuroscience and Biomedical Engineering, Aalto University, Finland*
[b] *Faculty of Humanities and Social Sciences, University of Jyväskylä, Finland*
[c] *Neuroscience Center, Helsinki Institute of Life Science, University of Helsinki, Finland*
[d] *Centre for Cognitive Neuroimaging, Institute of Neuroscience and Psychology, University of Glasgow, United Kingdom*

A B S T R A C T

Digital interventions often suffer from low usage, which may reflect insufficient attention to user experience. Moreover, the existing evaluation methods have limited applicability in the remote study of user experience of complex interventions that have expansive content and that are used over an extensive period of time.

To alleviate these challenges, we describe here a novel qualitative Ecological Momentary Assessment (EMA) method: the CORTO method (Contextual, One-item, Repeated, Timely, Open-ended). We used it to gather digital intervention user experience data from Finnish adults ($n = 184$) who lived with interview-confirmed major depressive disorder (MDD) and took part in a randomized controlled trial (RCT) that studied the efficacy of a novel 12-week game-based digital intervention for depression. A second dataset on user experience was gathered with retrospective interviews ($n = 22$).

We inductively coded the CORTO method and retrospective interview data, which led to four user experience categories: (1) contextual use, (2) interaction-elicited emotional experience, (3) usability, and (4) technical issues. Then, we used the created user experience categories and Template Analysis to analyze both datasets together, and reported the results qualitatively. Finally, we compared the two datasets with each other. We found that the data generated with the CORTO method offered more insights into usability and technical categories than the interview data that particularly illustrated the contextual use. The emotional valence of the interview data was more positive compared with the CORTO data. Both the CORTO and interview data detected 55 % of the micro-level categories; 20 % of micro-level categories were only detected by the CORTO data and 25 % only by the interview data.

We found that the during-intervention user experience measurement with the CORTO method can provide intervention-specific insights, and thereby further the iterative user-centered intervention development. Overall, these findings highlight the impact of evaluation methods on the categories and qualities of insights acquired in intervention research.

1. Introduction

1.1. New digital intervention evaluation methods are needed

"*Lack of user acceptance has long impeded the success of new information systems*" (Davis, 1993), writes Fred D. Davis in 1993. The problem has not since evaporated. Digital mental health interventions are expected to create scalable, cost-effective solutions to alleviate the global mental health burden (Torous et al., 2021; World Health Organization, 2022). However, new interventions often suffer from low behavioral engagement (Fleming et al., 2018; Lipschitz et al., 2022; Ng et al., 2019; Torous et al., 2020): 40 % of people drop out before completing 25 % of treatment modules (Karyotaki et al., 2015). The dropout rate is even higher in less controlled real-world settings where there is competition for the user's attention from mass media, streaming services, social media, and digital games alike (Cohen and Schleider, 2022). One study found that only 4 % of people who have a mental health app installed on their mobile phone open it on any given day, and only 3.3 % continued to use them after 30 days (Baumel et al., 2019). Such insufficient user interaction diminishes intervention effectiveness (Gan et al., 2021).

Therefore, intervention researchers and developers face a twofold objective: to design solutions that users find sufficiently engaging so that the intervention can be effective (Perski et al., 2017; Ritterband et al., 2009; Yardley et al., 2016).

Conceptually, user engagement can be divided into objective behavioral engagement and subjectively experienced engagement (Doherty and Doherty, 2019; Perski et al., 2017). Behavioral engagement refers to user-intervention interactions, which can be measured through usage data and metrics such as usage time and the number of sessions. Complementary methods are needed to illuminate the experiential aspects of engagement, the user experience. A well-designed intervention should be perceived as appealing, easy to adopt, interesting to use, and encouraging in fostering behavioral change (Baumel et al., 2017; Graham et al., 2021; Stoyanov et al., 2016), and positive perceptions may be associated with intervention effectiveness (Graham et al., 2021). "*Understanding users' opinions are critical if we aim to design effective apps that will be adopted and used by the target audience*", encapsulate Alqahtani and Orji (2020).

Crafting meaningful interventions requires user involvement, and user-centricity is recognized in practically all digital intervention development frameworks (Lukka and Palva, 2023; Mohr et al., 2017; Mummah et al., 2016; Van Gemert-Pijnen et al., 2011; Verschueren et al., 2019). To meet the implicit user needs, iterative intervention development and implementation fluctuates between design and evaluation (Mohr et al., 2017): there is no design without evaluation, and vice versa. However, the methods for evaluating the user perceptions of digital designs have not developed as rapidly as the digital technologies to be evaluated. There has been too little method development in the last decade (Kip et al., 2022), and methods do not yet make use of digital possibilities, such as collecting long-term engagement data remotely (Michie et al., 2017; Smith et al., 2023).

Ecological Momentary Assessments (EMA) have potential in this regard. They allow "subjects and patients to report repeatedly on their experiences in real-time, in real-world settings, over time and across contexts," Shiffman et al. describe (Shiffman et al., 2008). Within the digital intervention context, EMAs have typically been used to study the changes in psychiatric symptoms and their determinants (McDevitt-Murphy et al., 2018; Reichert et al., 2021). For instance, Kleiman et al. used smartphone-collected data to study the within-day changes in suicidal ideation (Kleiman et al., 2017). They found that suicidal ideation varied more frequently than earlier studies suggested, which highlighted the need for frequent measurements to capture such fluctuations. Similarly, digital intervention use may be characterized by changes in the experiential aspects of engagement (O'Brien and Toms, 2008; Short et al., 2018). Karapanos et al., for instance, described how the user experience of a new smartphone evolved over four weeks (Karapanos et al., 2009). The early stages were characterized by growing familiarity, then incorporation of the product into the daily routines, and finally identification as the product was more closely integrated with the user's life. Capturing these changes requires long-term data gathering, but earlier reviews have not found EMAs being used to study the digital intervention user experience over time (Doherty et al., 2020; Kip et al., 2022; Short et al., 2018; Smith et al., 2023). Therefore, our aim here is to introduce a new EMA method, CORTO (Contextual, One-item, Repeated, Timely, Open-ended), to facilitate the comprehensive measurement of digital intervention user experience.

1.2. The CORTO method

Here we introduce the CORTO method. We applied the method to generate qualitative data on the digital intervention user experience to facilitate its formative evaluation (Stetler et al., 2006; Van Gemert-Pijnen et al., 2011). In contrast to summative evaluation, which focuses on the study outcomes, formative evaluation examines the intervention process, feasibility, acceptability, and contextual fit, and uses the information to iteratively refine the intervention. The CORTO

method may also have applicability beyond our use case: in studying constructs such as well-being or use intentions that are related to the use of complex digital software, including entertainment video games.

The development of the CORTO method responded to our need for measuring long-term user experience within our randomized controlled trial (RCT) that spanned 12 weeks. The primary endpoints of the study included measurements of psychiatric symptoms via standardized questionnaires every four weeks. However, as our intervention was intended to be used several times a week, we were concerned that the delay could introduce challenges with the recall of the fleeting user experiences (Robinson and Clore, 2002). We concluded that we needed a way to measure such experiences within the software itself immediately after they occurred. We did not find any existing method that would have been suitable for this need (Doherty et al., 2020; Short et al., 2018), and we decided to approach the measurement of user experience with an EMA approach.

The CORTO method adapts the traditions of EMA to a new context. Shiffman et al. describe four features of EMA: the data is collected in real-world environments ("ecological"); the measurements concern current or very recent states ("momentary") as opposed to summarized recall; the measured moments are carefully selected; and multiple assessments are completed over time (Shiffman et al., 2008). We needed to specify these principles as we applied them to a particular context (digital interventions) and to study a construct (subjective engagement, i.e. user experience) that they have not, to our knowledge, been previously used to study. Earlier, EMAs have typically been used to measure what occurs outside the measurement device (de Vries et al., 2021; Doherty et al., 2020), whereas we used CORTO to measure the user experience of the software within the software itself. This required aligning the measurement with the intervention characteristics rather than the external context. We also considered it important that the measurement is brief to avoid burdening the participant (Short et al., 2018), and open-ended to qualitatively capture the subjective experiences associated with intervention use (Table 1).

1.2.1. "Contextual" means studying the experience where it happens

The CORTO method measurement occurs in situ (Reis, 2012): within the digital context studied, which facilitates answering and recall. From the participant's point of view, the research is conducted *locally* in their naturalistic everyday environment in which digital software is used, which allows for gathering data with high ecological validity. From the researcher's point of view, the study is conducted remotely, making it "*research from a distance*" (Kip et al., 2022). This diminishes the participant's reactivity and impression management, as the researchers are less salient to the respondent (Brewer and Crano, 2014).

1.2.2. "One-item" makes the questionnaire easy to answer

Previous research has raised concerns that existing structured questionnaires can be lengthy and that questionnaire brevity is essential to facilitate responding, especially in repeatedly administered EMAs (Short et al., 2018). Concerns have also been raised that the participants do not

Table 1
The five principles of the CORTO method.

Principle	Description	Rationale
Contextual	The measurement occurs within digital software	Mitigates retrospection bias and encourages answering
One-item	The measurement includes one item	Brevity encourages answering
Repeated	The measurement is done repeatedly over time	Understanding temporal changes and improving measurement coverage
Timely	The measurement is presented near the relevant user interaction	Facilitates response specificity and relevance
Open-ended	The measurement is presented as an open-ended question	Allows gathering qualitative, experiential insights

provide sufficiently elaborate answers to open-ended questions (Kip et al., 2022). Therefore, the CORTO questionnaire is short and easy to answer. In this study, the user responded via writing, but audio and video may also be used when such implementations are feasible, or such options may be offered to encourage elaborate responses.

1.2.3. "Repeated" uncovers temporal changes in user experience

Digital interventions aim to encourage behavioral and symptom change through user-software interaction (Ritterband et al., 2009). Beneficial changes occur over time (Gan et al., 2021), as does user interaction. The CORTO questionnaire allows for gathering data on how the subjective experience unfolds over time, which is particularly relevant in complex digital interventions (Skivington et al., 2021) that include numerous complementary features (Zhang et al., 2019); whose content is adaptive or non-linear; whose users have significant diversity; and whose use extends over a long period and occurs in various contexts.

1.2.4. "Timeliness" facilitates an accurate account of experiences

We align with Robinson and Clore (Robinson and Clore, 2002) who distinguish between the experiences as they are experienced and their recollections. Recall-requiring retrospective accounts may confuse experiences with subjective beliefs, rationalizations, and sense-making. Close-in-time proximity allows more detailed measurement of the immediate, present, salient experience (Doherty and Doherty, 2018; Reis, 2012), mitigates retrospective bias (Schwarz, 2012), and prevents back-filling (Smith et al., 2023). Using a metaphor, we do not ask how the participant experienced the movie they saw last week, we ask about their experience as the credits begin to roll. In complex interventions, the relevant temporal micro-context needs to be clearly defined to ensure that the self-report does not unnecessarily interrupt the usage flow. Possibilities include event-based, interactive, time-based, or randomly prompted questions (Trull and Ebner-Priemer, 2013).

1.2.5. "Open-ended" focuses on subjective experience

Finally, only the participant has access to their subjective internal state. It is information no one else knows (Reis, 2012), and open-ended self-reporting is a convenient way to gather it. This has two benefits: discovering experiences that participants share spontaneously and avoiding biasing the respondent with prompts (Reja et al., 2003). The open-ended questions produce more diverse answers allowing insights into the user's experiential realm. When measuring user experience questions such as "*What would you like to share about your experience with the intervention?*" can work well in this regard. However, the CORTO method does not include specific items, which allows it to be used to study a range of constructs.

1.3. Using the CORTO method to evaluate digital intervention user experience

We used the CORTO method to evaluate human and social factors that are vital in intervention development and consequent implementation (Enam et al., 2018). Our approach converges with the user-centered design (UCD) paradigm that is acknowledged as a best practice in digital development, including eHealth solutions (Van Gemert-Pijnen et al., 2011), psychosocial interventions (Lyon and Koerner, 2016), mental healthcare (Mohr et al., 2017), and serious games for mental health (Dekker and Williams, 2017; T.M. Fleming et al., 2016; Lukka and Palva, 2023; Verschueren et al., 2019). UCD is closely associated with participatory design (Dekker and Williams, 2017), design thinking (Mummah et al., 2016; Scholten and Granic, 2019), person-based approach (Yardley et al., 2015), and human-centered design (Kip et al., 2022), and the concepts are sometimes used interchangeably. Here, we use the concept of UCD. Central to UCD is understanding how users perceive the design which has been called subjective engagement, a concept closely associated with the notion of user experience (Doherty and Doherty, 2019; Perski et al., 2017; Yardley et al., 2016). Here we use

the latter concept to highlight how the CORTO method gathers user's verbalized experiences.

Usability and user experience are closely related to UCD. Usability focuses on functional requirements (ISO 9241-11:2018, 2018), and it is often evaluated against usability heuristics: universal design best practices. They include, for instance, facilitating user control, showing the system status, giving feedback, and preventing errors (Petrie and Bevan, 2009). Recently, specific heuristics have been described in digital health (Baumel and Muench, 2016). However, there is no consensus on which heuristics matter the most (Quiñones et al., 2018)—most likely because the interventions have considerable variance in their features, audiences, and scope. Serious games, for instance, include features rarely found in nongame interventions, such as a fictional world and a storyline (Jerzak and Rebelo, 2014; Tondello et al., 2016), and similarly, interventions taking advantage of novel technologies such as virtual reality, social media, and artificial intelligence (Torous et al., 2021) may have technology-specific features. In contrast to the feature-level usability heuristics, user experience takes a holistic view of the dynamic interaction-elicited subjective experience to improve it (Hassenzahl and Tractinsky, 2006; Law et al., 2009; Petrie and Bevan, 2009). User experience, or player experience, may be particularly important in game-based solutions that suggest excessive entertainment value beyond functional features (Sánchez et al., 2012).

The most common methods for evaluating usability and user experience include interviews, group interviews (focus groups), questionnaires, and usability tests (Hookham and Nesbitt, 2019; Kip et al., 2022; Ng et al., 2019). Recently, structured questionnaires have been developed to evaluate digital interventions specifically (Short et al., 2018). They include, for instance, the Mobile App Rating Scale (MARS) (Stoyanov et al., 2015), its end-user version uMARS (Stoyanov et al., 2016), and the mHealth App Usability Questionnaire (MAUQ) (Zhou et al., 2019). The questionnaires are valuable in structuring the intervention evaluation via rating scale questions, such as "*Is the app interesting to use? Does it present its information in an interesting way compared to other similar apps?*" (Stoyanov et al., 2016). However, the responses to these closed-ended questions do not characterize, for instance, what is and what is not interesting about the specific intervention and its features, and such qualitative data would be particularly valuable in improving the intervention.

Qualitative research can provide intervention-specific insights. Reja et al. found that an open-ended questionnaire item led to discovering 8 additional categories beyond the 10 assumed by the researchers beforehand (Reja et al., 2003). Also interviewing can be used to discover such emergent topics. Interviewing may occur concurrently with the design use, such as when the participant reacts to designed mockups or prototypes (Ospina-Pinillos et al., 2018; Pine et al., 2020; Wehbe et al., 2022) or in think-aloud usability tests (Van Den Haak et al., 2003). Interviews can also be retrospective, which allows for creating an understanding on the aggregative experience the intervention and the associated processes have elicited (Crane et al., 2017; T. Fleming et al., 2016). However, interviews can be time-consuming to conduct, transcribe, and analyze, (Short et al., 2018) and may be biased by participant selection (Kip et al., 2022). We suggest that the CORTO method could mitigate the limitations of existing methods by allowing the users to generate intervention-specific data through an open-ended question, thus increasing the relevance of the data for intervention development. In addition, the brief repeated measurements provide an avenue to gather qualitative data on scale with relative ease for both researchers and respondents.

1.4. Research aims

This research investigates the feasibility of a new method for measuring digital intervention user experience: CORTO. We conduct the study in three parts: (1) we create a ground-up categorization of user experience to facilitate its measurement with methods such as CORTO;

(2) we use the user experience categories to qualitatively describe what kind of data the CORTO method generates; and (3) we compare the data generated with the CORTO method with data generated with retrospective interview.

2. Methods

2.1. The study overview

The study included two datasets that were analyzed and reported in three phases. 250 participants receiving a new digital intervention were asked to complete the CORTO method questionnaire, which led to responses from 204 participants. In addition, 22 participants were interviewed retrospectively, with 20 participants being evaluated with both methods. The 20 interviewed participants were excluded from the CORTO method dataset. The three-phase data analysis began with the inductive coding of the dataset gathered with the CORTO method ($n =$ 184) and the dataset gathered with retrospective interviewing ($n = 22$), which created the user experience categories. Then, both datasets were coded deductively using the created user experience categories and template analysis (TA) (Brooks et al., 2015), which led to a qualitative description of the digital intervention user experience. Thirdly, the two datasets were compared with each other to understand their comparative advantages and limitations. Finally, we offered pragmatic guidelines for implementing the CORTO method.

This study was conducted as a part of a randomized, double-blind, comparator-controlled, pre-registered clinical trial, that examined the efficacy of a novel game-based digital mental health intervention, Meliora, for alleviating major depressive disorder (MDD). The Type 1 effectiveness-implementation hybrid trial design (Curran et al., 2012; Mohr et al., 2017) allowed for studying intervention efficacy while gathering qualitative insights to improve the intervention design and implementation (Cooper et al., 2014; O'Cathain et al., 2013). The RCT has received positive appraisals from the Helsinki University Hospital (HUS) research ethics committee (HUS/3043/2021) and the Finnish Medicines Agency Fimea (FIMEA/2022/002976), and it conforms with the Declaration of Helsinki. The RCT has been registered on ClinicalTrials.gov (NCT05426265) (ClinicalTrials.gov, 2022), and the interview study has been preregistered on OSF.io (Lukka et al., 2022).

2.2. Participants

The study participants were Finnish adults suffering from MDD. They were recruited through multiple channels, including healthcare partners and social media. Following the recruitment link led the possible participant to the study website with an informed consent form. After signing up, the potential candidates were sent a suite of symptom questionnaires via email and they were contacted by a clinical subject coordinator (CSC) who evaluated their eligibility in a remote interview (Appendix A). The key inclusion criteria were being 18–65-years old, having ongoing MDD, and having an ongoing mental health treatment contact. The CSC confirmed the MDD using the Mini-International Neuropsychiatric Interview (Sheehan et al., 1998). At the end of this process, 408 participants were accepted to the study. Of them, the CORTO method generated data from 184 participants and retrospective interviewing from 22 participants, and their demographic variables prior to randomization were described in Table 2.

We described the characteristics of the intervention users in depth in an earlier study (Lukka et al., 2023b) where our aim was to understand *who* the users were, considering that a novel game-based intervention may attract a self-selected population. We found that the participants had long-term psychiatric symptoms, three in four had a comorbid disorder alongside MDD, and many had previous treatment attempts that had not led to remission. The participants had close relationships with gaming and some found entertainment video game playing alleviated their psychiatric symptoms.

**Table 2**
The participant demographics.

Variable	CORTO data ($n = 184$)		Interview data ($n = 22$)	
	n	%	n	%
Intervention used				
Active (MEL-T01)	94	51.0	10	45
Comparator (MEL-S01)	90	48.9	12	55
Gender				
Female	100	54.3	17	77
Male	64	35.8	5	23
Other	15	8.1	0	0
Trans	2	1.1	0	0
Missing	3	1.6	0	0
Average age in years (SD)	32.9 (8.8)		33.5 (8.7)	
Relationship status				
Relationship or married	107	58.2	16	73
No relationship	62	33.7	5	23
Other	15	8.1	1	5
Highest education				
Primary education (9y)	12	6.5	0	0
Secondary education (12y)	108	58.7	16	73
Bachelor's	39	21.2	5	23
Master's	23	12.5	1	5
Licentiate or Doctorate	2	1.1	0	0
Life status				
Student	58	31.5	8	36
Short or long-term sick leave	44	23.9	5	23
Full-time working	36	19.6	4	18
Part-time working	20	10.9	0	0
Unemployed	15	8.2	3	14
Retired	10	5.4	2	9
Parental leave	1	0.5	0	0
Average weekly gaming hours (SD)	13.5 (12.9)		12.0 (8.7)	
Average PHQ-9 score (SD)[a]	15.3 (4.7)		15.2 (4.0)	
Average GAD-7 score (SD)[a]	10.8 (4.7)		11.8 (4.1)	
Average GAS-7 score (SD)[a]	11.8 (3.8)		11.7 (3.5)	

[a] PHQ-9 (Patient Health Questionnaire) measures depression severity (Kroenke et al., 2001), GAD-7 (Generalized Anxiety Disorder) measures anxiety severity (Spitzer et al., 2006), and GAS-7 (Game Addiction Scale) measures addictive behaviors related to gaming (Lemmens et al., 2009).

2.3. Intervention

After CSC acceptance, the participants were automatically and equally randomized to one of the three intervention arms: the active intervention arm, the active comparator arm, or the treatment as usual arm (Fig. 1). The research aimed to study the efficacy of the active intervention (MEL-T01) and the active comparator (MEL-S01) when they were used as an add-on to treatment as usual (TAU). The arms had a crossover design. The MEL-T01 arm participants first used the active intervention during a 12-week intervention period (while engaging in TAU), and then entered a 12-week period with TAU only. The MEL-S01 arm first used the active comparator during a 12-week intervention period (while engaging in TAU), and then entered a 12-week period with TAU only. The TAU group first had a 12-week period during which they only received TAU, and then they were equally randomized to use either MEL-T01 or MEL-S01 for the intervention period (while continuing to engage in TAU) (Fig. 1).

The digital intervention mechanisms of action were based on the findings that depression is associated with cognitive deficits (Rock et al., 2014), which could be alleviated through computerized cognitive training (Bediou et al., 2018). Our digital intervention aimed at measuring and training a wide range of cognitive functions, focusing primarily on cognitive deficits typically reported in MDD, such as cognitive control (Koster et al., 2017), working memory (Rose and Ebmeier, 2006), inhibition (Gohier et al., 2009), and a variety of executive functions, such as planning (Harvey et al., 2004). Training these functions has shown therapeutic potential in addressing MDD symptoms (Koster et al., 2017; Launder et al., 2021; Motter et al., 2016). We

Data gathering
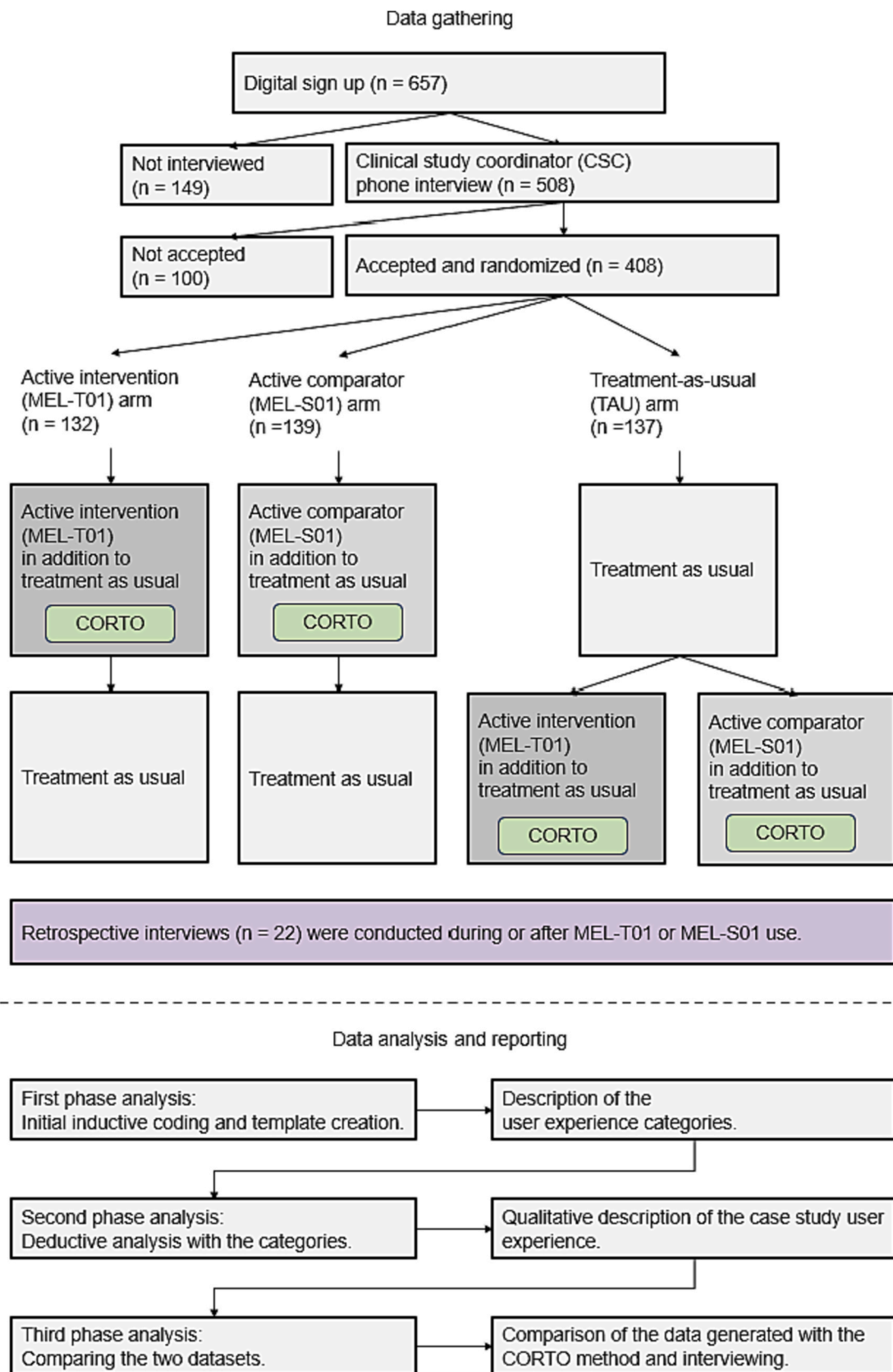
Fig. 1. The study process. The digital intervention user experience is assessed using the CORTO method ($n = 184$) and retrospective interviews ($n = 22$). The data analysis and reporting is conducted in three parts.

implemented the cognitive training as a serious game (Warsinsky et al., 2021) that included game elements ranging from high-level video-game-specific mechanics to specifically designed tasks that share similarities with tasks performed in cognitive neuroscience research. From the user's point of view, the digital intervention was designed to feel like a single-player action entertainment video game (Fig. 2).

The two intervention versions, MEL-T01 and MEL-S01, were highly similar sharing the same menu structure, audiovisuals, and core cognition-training gameplay. Both versions adapted to the participant's skill and consisted of 28 levels where the user unlocked new skills and progressed through a depression-focused narrative inspired by cognitive behavioral therapy. The key difference was that MEL-T01 incorporated additional cognition training elements implemented directly into the game mechanics and as minigames. Due to their substantial similarities, the user experience of both intervention versions was analyzed collectively.

The intervention software was developed using the Unity game engine (Unity Technologies), distributed using Steam (Valve), and installed on the participant's personal computer with a participant-specific key that allowed controlling intervention access. The participants were encouraged to interact with the intervention several times a week during the intervention period: a minimum of 24 h (2 h a week) or preferably 48 h (4 h a week). The intervention efficacy was studied with a suite of symptom questionnaires before randomization, and 4, 8, 12, and 24 weeks thereafter. The participants were compensated 50€ for meeting the lower and 120€ for meeting the higher usage goal per the Ministry of Social Affairs decree (Finlex, 2011) if they also responded to the symptom questionnaires. We report the clinical trial efficacy results elsewhere.

2.4. CORTO method implementation

The CORTO method was implemented within the intervention software. We used event-based cueing (Trull and Ebner-Priemer, 2013): each time ("Repeated"), the participant completed a level and had accumulated enough points to progress to the next level ("Timeliness"),

they were presented with an in-game ("Contextual") questionnaire in Finnish (Fig. 3). The number of points needed to progress to the next level increased per level, and therefore the frequency of questionnaires was faster in the earlier intervention levels compared to later levels. The questionnaire included one core ("One-item") open-ended ("Open-ended") question: *"Do you want to mention something regarding the game or playing it?" [Original in Finnish: "Haluatko mainita jotakin pelaamiseen tai peliin liittyen?"]*, and two supporting closed-ended questions. Prior to data gathering, we did not know whether the CORTO method implementation would lead to sufficient answers, as some researchers have suggested such approaches do not offer sufficient data (Kip et al., 2022). However, after the data gathering, we found that the open-ended question indeed provided far more detailed and useful data than the closed-ended questions, and as this study takes a qualitative approach, the two closed-ended questions remain outside the scope of the paper.

The CORTO method questionnaire was shown to 250 participants (Table 3). This number was smaller than the 408 accepted participants because the TAU group began the study with a waiting period without intervention usage, not all participants started to use the intervention, and some participants started the usage with a delay. The CORTO method generated 1017 open-ended responses from 204 (81.6 %) participants, and the questionnaire was responded to every fourth time it was shown. Of the 204 respondents, 20 were interviewed and 184 were not. We compared the CORTO data from non-interviewed and interviewed participants using an unpaired *t*-test. The average number of responses ($t(202) = 1.51$, $p = 0.13$), average response length ($t(202) = 0.22$, $p = 0.84$), and the average level after which the responses were given ($t(202) = 1.32$, $p = 0.18$) did not statistically significantly differ between two samples. We also found that the percentage of respondents and their response rates were similar, as were the participant demographics (see Table 2). These factors supported using responses only from the non-interviewed sample as the CORTO data, and not to confound the two datasets. Therefore, the CORTO method dataset included 888 open-ended responses from the 184 non-interviewed participants. Temporally, the data spans from the 27th of July 2022 to the 3rd of January 2023.
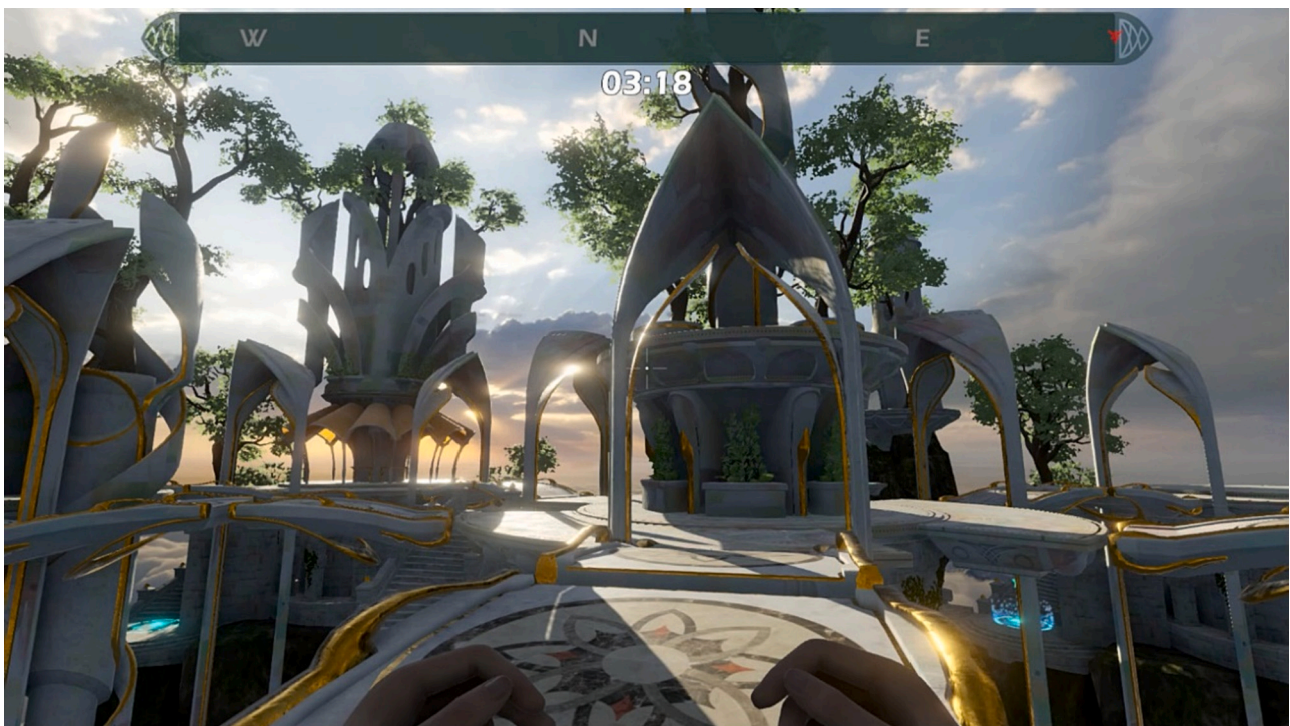

Fig. 2. The digital intervention aesthetics resembles entertainment video games. The fast-paced intervention requires the user to navigate three-dimensional environments in the first person, interact with other characters, and solve cognitively demanding tasks.

Fig. 3. The CORTO method is short and easy to answer. The questionnaire is presented to the participant after each level, up to 28 times. The translation was: "*Game experience questionnaire. Do you enjoy playing the game? (0=No answer, 1=Not at all, 7=Very much) How difficult do you consider the game? (0=No answer; 1=Way too easy; 2=Too easy; 3=Slightly easy; 4=Appropriate; 5=Slightly challenging; 6=Too challenging; 7=Way too challenging). Do you want to mention something regarding the game or playing it?*"

Table 3
The characteristics of the responses gathered with the CORTO method. The CORTO method dataset includes 888 responses from 184 non-interviewed participants.

Variable	Non-interviewed	Interviewed	Total
Number of participants shown the CORTO questionnaire	228	22	250
Number of participants giving at least one open-ended response	184	20	204
Percentage of respondents	80.7 %	90.9 %	81.6 %
Number of times the CORTO questionnaire was shown	3462	480	3942
Number of open-ended responses	888	129	1017
Response rate	25.6 %	26.9 %	25.8 %
Average responses per participant (SD)	4.8 (4.6)	6.5 (6.3)	5.0 (4.8)
Average open-ended response words (SD)	29.3 (41.8)	27.3 (29.4)	29.0 (40.4)
Average intervention level after which the CORTO questionnaire was responded	9.3 (7.0)	11.5 (7.7)	9.6 (7.2)
Total response words	25,980	3520	29,500

2.5. Retrospective interview

In October 2022, the research database was accessed for participants who had indicated their willingness to participate in a user experience interview in the signup and had interacted with the intervention for at least 1 h. 20 such participants were contacted via email, which led to 16 (80 %) responses. The first author noticed that this strategy led to the inclusion of many participants with high objective engagement. To gather experiences also from those who had less experience with the intervention use, 10 further participants, who had interacted with the intervention between 45 and 300 min, were contacted. This led to 5 (50 %) more interviews. One participant indicated their interest directly to the CSC, and they were also interviewed. In total, 22 interviews were conducted between October and November 2022.

The first author conducted the retrospective semistructured interviews remotely on Zoom (Zoom Video Communications) in Finnish with an interview guide (Appendix B). The interview consisted of two major sections. The first section explored *who* the participant was. This domain focused particularly on the participant's mental health background and their prior experiences with video games, and we reported the results in an earlier study (Lukka et al., 2023b). The second section sought to understand *how* the participant perceived the intervention, which is in focus here. The section followed the intervention process: first exploring how the participants discovered the study, why they had decided to participate, how they perceived the CSC assessment, and how the intervention was to learn and interact with. Then the interview explored participant perceptions of the intervention features, including its narrative and mechanics, and emergent themes from the participant. Further questions explored the participant's overall impression of the intervention, what they had found valuable and frustrating with it, and how they would improve it. Finally, the participants were asked about the perceived effects of use, and how the intervention would relate to other treatments, and they were asked to openly share any other thoughts.

The interviewer was a clinical psychologist, and a service and game designer, and familiar with the intervention. The interviewer was blind to the arm the participant was assigned to, and they paid particular attention to maintaining it. The interviews were recorded with explicit verbal permission from the interviewee. The interviews took, on average, 48 min (SD = 11 min, range = 30–68 min, total 1051 min), with the latter section on participant's user experience taking 51 % (536 min) of the total interview minutes. The first author transcribed the interviews verbatim which facilitated familiarity with the data. The latter interview section created a corpus of 55,025 words.

The intervention usage logs confirmed that the interviews were retrospective. Prior to the interview, the interviewees had been in the intervention period for an average of 48.1 days (SD = 22.4d, range 14–85d, median = 51.5d). During this period, they had interacted with the intervention for an average of 12.0 h (SD = 11.6 h, range 0.6–45.2 h,

median = 8.9 h). However, the participants had an average 14.0 day delay between their last use session and the interview (SD = 23.3d, range 0.1–71.9d, median = 2.1d). Thus, in contrast to the CORTO questionnaire, which was given immediately after the relevant interaction, the interviews occurred with a delay in intervention usage.

2.6. Three-phase data analysis and reporting

This mixed methods research has a pragmatist epistemological approach (Burke Johnson and Onwuegbuzie, 2004), which aligns with the research aim of furthering the user-centered digital intervention evaluation and development methods. We used two methods, the CORTO method and retrospective interviewing, to study one construct: the digital intervention user experience. The methodological triangulation was expected to offer complementary perspectives by "*exposing unique differences or meaningful information that may have remained undiscovered with the use of only one approach or data collection technique in the study*", as Thurmond describes (Thurmond, 2001). The mixed data collection (Small, 2011) was followed by sequential inductive/deductive analysis of both datasets (Proudfoot, 2022), and this uniform data analysis sought to discover the unique features of the methods used.

The first phase analysis aimed to create a template describing the key concept: user experience. The first author open coded both datasets—the 888 responses (*n* = 184) generated by the CORTO method, and the retrospective interview data (*n* = 22)—using ATLAS.TI 23 software (ATLAS.ti GmbH), while remaining open to many interpretations of the data (Saldaña, 2016). During the initial inductive coding, the researcher noticed how the CORTO method and retrospective interview data varied in their levels of abstraction (Engl and Nacke, 2013; Nacke, 2009; Nacke and Drachen, 2011): the interview data included higher-level broad observations on intervention usage, whereas the questionnaire data responses described the design in detail. This observation was reflected in two sense-making sessions with the second author, who specializes in qualitative data analysis. The inductive analysis led to the high-level template describing four categories of user experience.

The second phase analysis aimed to describe the user experience of the case study intervention using the template created. To begin this process, the initial codes were removed, and the data generated with the CORTO method and retrospective interviewing were analyzed deductively using the template. The analysis adopted an approach from a particular type of thematic analysis called template analysis that is suited to qualitative data analysis and open-ended questionnaire responses (Brooks et al., 2015). Many responses were coded into one category (e.g., "*Sometimes the glow effect of the light sources seems excessive,*" #CORTO:152:18). Other responses included several distinct meanings that were coded separately (e.g., "*For a person who plays a lot, the beginning is quite rigid. In addition, I feel that the use of the Finnish language is strange, but I understand the decision to use it. I eagerly look forward to how the game unfolds!*", #CORTO:150:1). The coding led to a three-level hierarchy: 1) the macro-level established in the framework; 2) the meso-level that pools many observations made on 3) the micro-level. After the first round of data coding, the first author performed two iteration rounds to ensure that the categories were internally homogenous and differentiated from each other, which were further reflected on twice with the second author. 10 (1.1 %) of the CORTO responses could not be categorized due to their brevity or ambiguousness (e.g., "*Meh*" #CORTO:152:7). On the other hand, numerous very accurate and elaborate accounts could include as many as 14 different micro-level codes. The interview quotes were annotated with an interview number (e.g. #intv:1) and CORTO quotes with a respondent-specific number identifier and the associated intervention level (e.g. #CORTO:1:10).

A review process was designed and implemented to ensure coding reliability. The first and third authors met in a briefing session to establish an understanding of the review, accompanied by the first authors' written instructions (Appendix C). The third author, a master's

degree student in psychology and CSC in the study, reviewed the coding focusing on coding accuracy and coding structure, and recorded their observations in detail in an Excel sheet. Their comments were reviewed in two sense-making sessions where the authors discussed the coding until agreement was found (Appendix D). Regarding CORTO data, 31 of the 46 (67 %) coding accuracy observations and 5 out of 6 (83 %) structural change observations were agreed to merit changes. Regarding the retrospective interview data, 8 of the 11 (73 %) coding accuracy observations and 6 out of 6 (100 %) structural change observations were agreed to merit changes. The first author translated the quotations, and the third author reviewed them for accuracy, which led to minor clarifications.

The third phase analysis compared the data generated with the CORTO method with data generated with retrospective interviewing to describe how they concur and differ. More specifically, the analysis aimed to answer two questions: (1) what unique data did the two methods produce, and (2) how did the data compare between the two methods. We approached the first question by examining the detection of micro-level categories in both datasets. A micro-level category was considered detected by the dataset if it had contributed at least one code to it. The second question was explored by comparing the distributions of meso-level category codes. The first author compared the two datasets quantitatively in Excel (Microsoft). Aware of the debate on analyzing qualitative data with numbers (Maxwell, 2010), we found that the quantification was compatible with the large number of descriptive coding categories produced by TA and with the nature of data that included observations grounded in the intervention software and its features. Finally, the first author reviewed the manuscript against the Standards for Reporting Qualitative Research (O'Brien et al., 2014), which led to minor clarifications.

3. Results

3.1. User experience categories

The first phase analysis resulted in a template that described the user experience in four macro-level categories across the two datasets. The categories exhibited how the user responses can "zoom out" and reflect on the intervention in the usage context, or "zoom in" and examine the intervention details. The participants considered the intervention both a game and a healthcare product that they used in their everyday setting (Context); they shared the holistic emotional experience the interaction elicited (Emotion); they examined the intervention features against existing design practices (Usability); and they described functional problems (Technical) (Table 4).

3.2. Case study user experience

The second phase analysis described the user experience of the case study intervention. This described what kind of data the CORTO method and retrospective interviewing generated, when considered together with the analysis third phase. The meso- and micro-level categories

Table 4
The user experience categories. The numbers indicate the quantity of meso- and micro-level categories identified within the macro-level category.

Analysis level	User experience			
Macro	Context Examining the intervention contextually	Emotion The interaction-elicited emotional experience	Usability Digression or compliance with existing design practices	Technical The design does not work as intended
Meso	6	2		
Micro	36	7	53	25

allowed the description of intervention-specific characteristics in detail (Table 5).

3.2.1. Contextualizing the intervention use

The context category described how the participant framed the intervention. The participants considered the intervention as a process, examining its fit with their needs and life context. The participants' initial motivations to sign up included affinity with digital games, wish to contribute to science and treatment development, and to alleviate their symptoms. During the intervention period, the participants adapted the usage to their schedules and sought moments that provided them with natural opportunities for intervention use. These included the morning, where the intervention use acted as a way to start the day, and in the evening, where it offered avenues for relaxation after the daily responsibilities. In the long term, the intervention usage was influenced by fluctuations in participant well-being: "*The winter depression becomes stronger, so it becomes more difficult to find the energy to play*" (#CORTO:10:9), explained one participant.

The participants' previous game and healthcare experiences influenced their perceptions. They used their prior media experiences to make sense of the intervention genre and mechanisms: "*This game reminds me of a modded Skyrim*" (#intv:3). They also considered playing as a skill. Prior game experience facilitated learning how to use the game-based intervention and lacking game experience could hinder the usage: "*I have played so few computer games that learning the basic keys feels clumsy*" (#CORTO:37:1). The intervention was also perceived as a treatment, and the participants found that its digital nature could improve treatment access: "*This could be for people who don't dare or want to seek help.*" (#intv:20). The intervention was viewed to augment, rather than replace, existing treatments by offering support between therapy sessions. The participants also described how the intervention had impacted them. It activated them, helped them change their negative thought patterns, and improved their cognition: "*In the beginning, the puzzles were somehow difficult, but then I noticed that I had improved in them a bit.*" (#intv:9). The negative impacts included excessive activation and negative physical experiences such as motion sickness. Overall, the intervention was viewed through complementary lenses: as an entertainment game and as a treatment.

3.2.2. The emotional experience intervention elicited

The participants frequently described the overall emotional valence the intervention elicited. This maps to the positive-negative spectrum where the person was either attracted to the intervention or turned away from it. On the negative end, the intervention was described as frustrating, annoying, and confusing. The first two emotions were associated with the intervention being perceived to be repetitive: "*The game repeats itself, even frustratingly so*" (#CORTO:145:13). This was because it did not have sufficient variation in regards to its length: "*There is too little variance*" (#intv:1). Confusion arose from not understanding the therapeutic rationale or perceiving a tension between the gameplay interactions and the intended therapeutic outcomes.

The interaction-elicited positive emotions included curiosity, interest, competence, and enjoyment. The participants explained how they enjoyed the challenge the intervention offered: "*The flow state while playing the game is unbelievable when you begin to effortlessly control ever more complex aspects*" (#CORTO:116:23). The positive feelings were associated with finding the intervention appealing and looking forward to using it again; it providing an appropriate level of challenge; and it evoking feelings of mastery and competence. Playing also inspired the participants to come up with development ideas and new features: "*It would be rewarding to see statistics, like how many meters you have moved*" (#intv:6). Overall, the emotional experiences were associated with a holistic evaluation of the intervention.

3.2.3. Usability as compliance with or digressions from heuristics

The usability category included the participants' responses regarding specific intervention features. The participants often implicitly compared the intervention to other digital software and games, and their responses reflected how the intervention features diverged from or complied with common design practices and heuristics. They included, for instance, ease of use, enjoyable visual design, and clear tutorials. The usability category included responses to all aspects of the design, which we divided here into static features and dynamic gameplay interactions.

The static features included the audiovisual design, game state indicators, instructions, and the story. The visual design established positive first impressions and encouraged playing: "*The environment is beautiful and makes playing enjoyable*" (#CORTO:2:6), although some of the enemy models could be perceived as frightening. The participants would like improvements to voice acting, and as most digital games are in English, the Finnish language was perceived as peculiar. Some game state indicators were difficult to read: "*The player health bar could be more clearly visible*" (#CORTO:114:5). In addition, the game should indicate the enemy line of sight, ability states, and environmental elements more clearly. The instructions, as well as point system descriptions, could also be improved. The story divided opinions. It motivated long-term engagement and provided insights into depression: "*The characters' thoughts resonate with my feelings and thoughts*" (#CORTO:49:5). Others found the story too straightforward, even obvious, and wished for changes to its tone of voice.

The dynamic features included gameplay interactions, which the participants found to require polishing. Some participants found the movement challenging or too fast. Navigation in the 3D environment could be confusing, and a minimap was requested. The participants found autonomy in controlling their character important: "*It feels unfair that the time progresses while the game prevents me from moving*" (#CORTO:128:3). Instructions that interrupted gameplay and locked the character in place were perceived as frustrating, and pausing and skipping the tutorials were expected. Unlocking new abilities made the game more interesting: "*It is a positive surprise how many new elements have been added*" (#CORTO:24:4). However, the abilities were found to be underpowered and challenging to use, which mitigated gameplay variety and exacerbated repetitiveness. The difficulty level was found to be variable, with a frustrating difficulty spike on level 12. Overall, the usability category illustrated numerous feature-related factors that could be iterated.

3.2.4. Technical comments allow identifying bugs

The technical category included instances where the intervention did not work as intended. The bugs spanned many aspects of the intervention: abilities, enemy encounters, movement, inputs, spelling, performance, progress, settings, and tutorials. The feedback sometimes characterized the problem very clearly: "*In the tunnels, where there is water and a slope that you have to climb, the character occasionally gets stuck there*" (#CORTO:20:10) and "*I play with inverted y-axis, and I had to reapply this every time I played because it was not saved*" (#intv:11). From a development point-of-view, this category offered the most detailed and unambiguous feedback for iterative development.

3.3. Comparing CORTO and interview data

The third phase analysis found that the CORTO method and retrospective interview provided complementary perspectives on user experience. 67 (55 %) of the 121 micro-level categories were detected by both datasets. 24 (20 %) were detected only in the CORTO data, and 30 (25 %) only in the interview data. The interview data contributed unique insights particularly into the context category, whereas the CORTO data provided unique perspectives on the usability and technical categories (Fig. 4). The CORTO data was four times denser (41.6 codes per 1000 words) compared to the interview data (10.0 codes per 1000 words) where the participants were more elaborate in their answers.

The 1081 CORTO and 550 interview data codes were unequally distributed across meso-level categories, which indicated them

Table 5

The qualitative data provides detailed and intervention-specific insights into the user experience. The data includes CORTO responses ($n = 184$) and retrospective interviewing data ($n = 22$). The 17 meso-level categories are described using the 4 macro-level categories (see Table 4).

Category	Description	Quote
Context	**Examining the intervention contextually**	
Game-based intervention research drives study participation	The participation was motivated by familiarity with digital games, contributing to science and new treatment development, and finding personal help.	*"This caught my attention, as it studies mental health, which is relevant to me. And it is a game: games have been an integral part of my life at least back in the day."* (#intv:11) *"If I can help with my small contribution, perhaps it can help those who suffer from depression later."* (#intv:3)
Intervention onboarding facilitates easy adoption	The intervention sign-up process worked fluidly, including interviews, questionnaires, and game installation.	*"It was cool to get into the study. I thought that I would not meet the criteria"* (#intv:1) *"I am used to the questionnaires because I have filled similar ones so many times"* (#intv:8)
The intervention use adapts to life context	Everyday responsibilities, daily rhythm, and well-being influenced the intervention usage.	*"There are considerable differences between days. When I have good days, playing is more meaningful and fluid. On bad days, I don't play, as I don't have the energy"* (#CORTO:145:27) *"I had an exam week last week, so I had no time to play."* (#intv:14)
Prior media experience frames the intervention	The game was easy to pick up for those with prior game experience, and it was compared to other digital games.	*"The style of the game is Playstation-like, like Call of Duty. I mostly play World of Warcraft and Diablo, so playing is somehow difficult"* (#CORTO:79:9) *"I am a pretty experienced player, so movement in games is easy for me, and I want to progress faster"* (#CORTO:15:1)
The intervention could augment existing treatments	The intervention can augment other treatments and facilitate treatment access but it was most suited for people with lower psychological knowledge.	*"I have been in therapy for a couple of years, so I am not necessarily the target audience"* (#intv:14) *"I would prefer a combination [treatment] because I have thought about going to therapy again. I miss talking about my issues."* (#intv:19)
The intervention can activate and change negative thoughts	The intervention impacts included activation, changing negative thinking, and motion sickness.	*"The game has activated me. After playing, you have the energy to do other things."* (#intv:6) *"Playing the level for a long time caused mild nausea"* (#CORTO:165:12)
Emotion	**The interaction-elicited subjective experience**	
Intervention use feels repetitive and elicits frustration	Over time, the interaction started to feel repetitive and boring because there was not enough variation in the content.	*"Unfortunately, the game starts to repeat itself. I would expect progress in the story to maintain interest in it."* (#CORTO:136:8) *"The game got boring pretty fast. The game progressed too slowly."* (#intv:8)
Intervention is interesting and engaging	Using the intervention was enjoyable and exciting.	*"Progressing through the level feels amazing"* (#CORTO:155:11)
		"I have enjoyed playing it. It has caused me enjoyment, pleasure, whatever you want to call it." (#intv:16)

Usability	Digression or compliance with existing design practices	
Audiovisual execution is appealing but needs polish	The visual design was impressive, but enemies may be frightening. The audio was appreciated, but voice acting requires improvements.	*"I was impressed by the graphics"* (#intv:7) *"The butterfly has a bad microphone and reads the story like to a child. Have some more immersion into the role, butterfly!"* (#CORTO:73:1)
The difficulty level varies too much	The difficulty level was occasionally too easy or increased suddenly.	*"This is too easy. I listen to a podcast at the same time, but it still does not stimulate enough"* (#CORTO:10:6) *"There are too many enemies and too little time"* (#CORTO:127:15)
Game state indicators need clarification	The character, enemy, and ability states were not sufficiently clearly indicated.	*"The enemy hitboxes are quite small compared to their graphics"* (#CORTO:77:10) *"It is unclear how the enemy harms the player. I see the lightning coming from the enemy, but do they hit and how much damage do they do?"* (#CORTO:73:3)
Instructions are insufficient	The gameplay instructions, including scoring, were too vague.	*"The game does not sufficiently explain how you get points."* (#intv:10) *"The instructions say that you need to hit the enemy 'at the right time', but I don't know what that time is."* (#CORTO:141:4)
Movement and navigation are challenging	The participants found movement rigid or too fast and navigating challenging.	*"The movement is not fluid enough. I fear that I fall off the cliffs."* (#CORTO:56:1) *"You need to look at the map quite often. It would be nice to see a small map all the time on the screen."* (#CORTO:46:7)
Story divides opinions	The story motivated long-term engagement, but the participants expected more of it. The story was appreciated for its insights into depression, but it may be too straightforward.	*"This story and themes are really inspiring and got me to think about my relationship with depression, and the parts in myself that are difficult to accept"* (#CORTO:116:8) *"The metaphors, shards, and other things, they have not got me to think anything in more depth"* (#intv:8)
Taking away user control frustrates	The participants would like to skip repeating tutorials, pause the game, and customize the game through settings.	*"I am still annoyed about the same repeating tutorials because you can't skip them, and time progresses during them."* (#CORTO:136:15) *"I got a phone call, and the time just progressed, and I could not pause the game. It was annoying."* (#intv:7)
Underutilized gameplay variance mitigates engagement	The participants appreciated the unlocked abilities but they were difficult to use and felt underpowered. The gameplay needed polish to avoid blockers.	*"There are so many different mechanics. There is this and that, and I don't use the majority of it. I don't need them."* (#intv:22) *"The more elements the game has, the better"* (#CORTO:66:6)
Technical	The design does not work as intended	
The intervention is not working as intended	The game had bugs such as getting stuck, misspellings, low performance, and visual glitches.	*"If you are conquering an area, and a defender spawns, the progress bar remains visible in the center of the screen."* (#CORTO:39:24) *"The mines do not work on level 14"* (#CORTO:70:14)

measuring different aspects of user experience (Fig. 5). The interview data focused more often on the context than CORTO data (39 % vs. 9 % of dataset codes): the interviews evoked considerably more responses in all meso-level categories except in the comparisons to other games. In the emotion category, the interview and CORTO data response proportions were equal (15 % vs. 14 %). However, CORTO data included proportionally more negative, and interview data more positive responses. The usability category had proportionally fewer interview responses than CORTO responses (36 % vs. 61 %), with only the story category gaining more responses in the interview data. Fewer interview than CORTO responses (3 % vs. 16 %) were found in the technical category.

4. Discussion

4.1. Overview

Here we have described how the CORTO method (see Table 1) can be used to gather qualitative data on the user experience of complex digital interventions. In the first phase analysis, we open-coded the data generated by the CORTO method and retrospective interviews, which resulted in the description of user experience categories (see Table 4). In the second phase analysis, we used the created template to describe the

digital intervention user experience (see Table 5). In the third phase analysis, we compared the two datasets, and found that the CORTO data specifically illustrated the intervention features and their technological implementation, while the retrospective interview data created unique insights into the usage context (see Figs. 4 and 5). The results exhibited how the CORTO method can extend the digital intervention evaluation methods and provide intervention-specific data to facilitate their iterative user-centered design.

4.2. Implementing the CORTO method

4.2.1. Choosing intervention evaluation methods

Previous research has emphasized the necessity of using a suite of complementary intervention evaluation methods throughout the development because one method alone cannot gather data on all aspects of the intervention (Kip et al., 2022; Ng et al., 2019; Skivington et al., 2021; Yardley et al., 2016). Here we used two methods to measure one construct—digital intervention user experience—and found significant differences in the results. Methodological awareness can be recommended because even two qualitative methods are not commensurable, but complementary (Small, 2011). Earlier, EMA methods have been used primarily to study the temporal changes in psychiatric symptoms (Reichert et al., 2021) rather than subjective
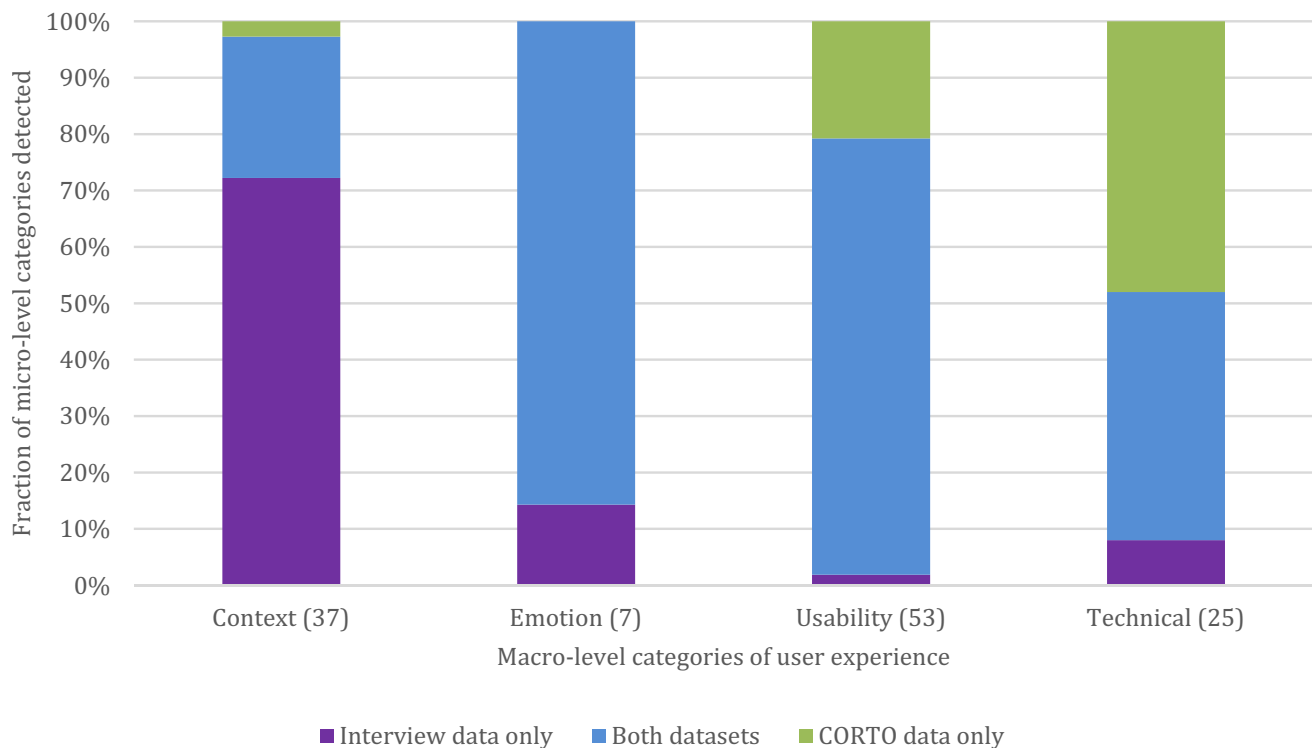
Fig. 4. CORTO method (*n* = 184) and retrospective interview data (*n* = 22) detect unique micro-categories. The stacked bars described the proportion of micro-level categories detected by either or both datasets. The total number of micro-level categories in the meso-level category is indicated in parentheses.

engagement, that is, user experience (Doherty et al., 2020). However, as low behavioral engagement remains a critical challenge in digital intervention development, new methods are needed to capture the subjective perceptions over time. Here, the CORTO method can expand and complement the existing methods, particularly interviewing, questionnaires, and usability testing.

The CORTO method, as it was implemented here, provides a complementary perspective to the retrospective interview (Table 6). We found that the CORTO data generated more detailed descriptions of intervention features, which is aligned with previous research: real-time data gathering can mitigate challenges with retrospective recall and improve data specificity (Robinson and Clore, 2002; Shiffman et al., 2008). However, there was one notable exception to this, namely the story, which gathered more responses in the interview data than in the CORTO data. We presumed that the written and verbal story elements facilitated their recall and that the participants may lack concepts for other design elements, which could have challenged remembering and describing them accurately with delay. We also found that the valence of the interview data was more positive compared with the CORTO data. This suggested the influence of interviewee selection and reactivity (Brewer and Crano, 2014): participants who view the topic more favorably may be more likely to participate in the interview, and the interview as a social situation may increase the favorability of responses. On the other hand, the CORTO method presented the user with a question immediately after the relevant user interaction, which may invite critical feedback and the participants may perceive that such feedback is more valuable to the intervention development than encouraging and validating responses.

The CORTO method implementation should take into account the development phase and intervention characteristics. Digital intervention development progresses through stages (Kip et al., 2022; Lukka and Palva, 2023; Verschueren et al., 2019). Yardley et al. describe four such stages: planning; design; development and evaluation of acceptability and feasibility; and implementation and trialing (Yardley et al., 2015). Because the CORTO method evaluates usable software, it is most feasible

in the two latter stages. Therefore, we suggest that its usage is preceded by heuristic usability evaluation with methods such as Enlight (Baumel et al., 2017) or MARS (Stoyanov et al., 2015). Interviewing, meanwhile, can provide valuable perspectives on the end-user needs, preferences, and context (Lukka et al., 2023b), as well as illuminating stakeholder perspectives that are important in the digital intervention implementation (Lukka et al., 2023a). The methods used should also take into account the intervention features: we designed the CORTO method for evaluating the user experience of digital software whose content is extensive and that is used over a long period of time in naturalistic environments. Therefore, while this case study involved a game-based intervention, we consider that the CORTO method is also applicable to evaluating complex non-game-based interventions. Because the CORTO method principles (see Table 1) allow flexible implementation, its use could also extend to domains beyond healthcare, such as evaluating the player experience of entertainment video games.

4.2.2. Planning CORTO use

Here we used the CORTO method to study the digital intervention user experience. However, the method may also be used to study other constructs and phenomena related to digital intervention engagement and effectiveness. The open-ended question can guide the participant's attention to, for instance, their emotions ("What are you feeling right now?"), symptoms ("What psychiatric symptoms do you experience now, if any?"), intention of use ("Why did you decide to use the intervention now?"), and use context ("Where are you using the intervention?"). The specific wording of the open-ended question may subtly influence the results, and testing its variations may be recommended.

However, as an EMA, the CORTO method extends beyond the item wording into its context and timing (Doherty et al., 2020; Trull and Ebner-Priemer, 2013). The researchers need to choose the moment of measurement carefully based on the construct to be investigated. For instance, the intention to use could be meaningfully asked when the user launches the intervention; the intervention user experience after completing an intervention module; and the use context at random
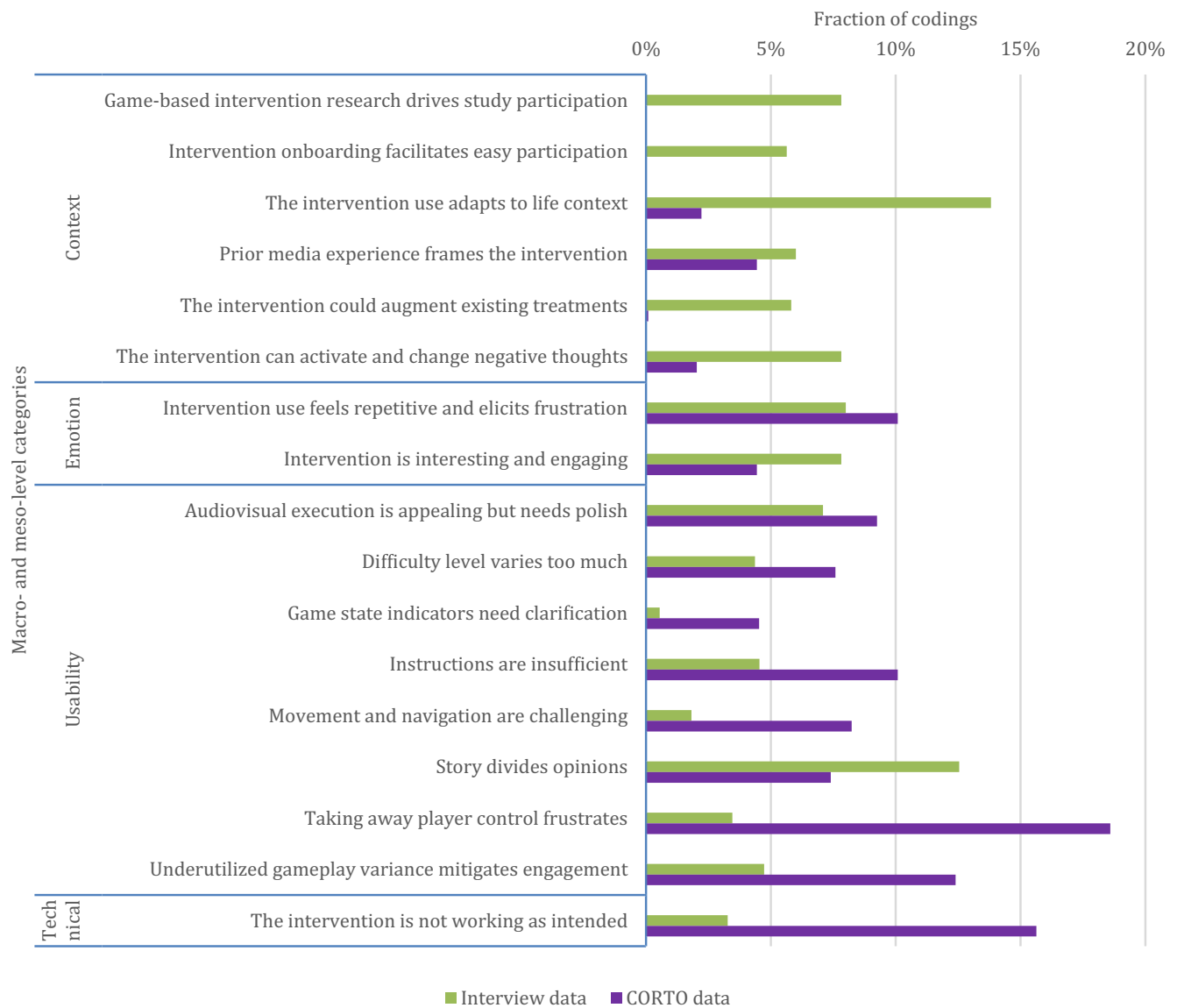
Fig. 5. The CORTO method and retrospective interviewing produce different data. The distribution of CORTO method data codes (1081, $n = 184$) and retrospective interview data codes (550, $n = 22$) differ across the macro- and meso-level categories.

Table 6
Method comparison. We compare the implementation of the CORTO method and retrospective interviews.

	CORTO method	Retrospective interview
Strengths	• High ecological validity • Produces development-relevant data • Suitable for complex interventions that are used over extensive periods of time • High data density accelerates analysis	• Illuminates interviewee context • Allows follow-up questions • Low technical requirements
Weaknesses	• Not feasible in early development stages • Not feasible for small-scale interventions • Requires technical implementation • Temporal analysis can be challenging	• Resource-intensive • Issues with recall • Interviewee reactivity • Interviewee voice as a personal identifier

intervals through the day. The prompt may also be implemented to be a reaction to certain or unusual behavior such as logging in after a long delay, a particular symptom questionnaire response, or psychophysiological measurement, and they should not unnecessarily interrupt the intervention flow.

Our implementation of the CORTO method included one open-ended item and two closed-ended rating scale items (see Fig. 3). As our study focused on the subjective experience, we only analyzed the former. However, the closed-ended questions could have subtly influenced the open-ended responses through framing (Schwarz, 1999). On the other hand, they may have also facilitated responding to the open-ended item by first giving the participant two easy questions to answer followed by a more cognitively demanding open-ended item. Complementing the CORTO method with closed-ended questions may be feasible if they are few, and as long as the focus remains on the core open-ended question.

4.2.3. Measuring and analyzing CORTO data

We found that the CORTO method gathered a broad dataset (see Fig. 5). The open-ended item was responded to every fourth time it was shown, which may be influenced by the frequency of the prompts. The questionnaire was shown as many as 28 times and the participants may not have had new thoughts to share after every level, and we did not expect full compliance with the measurement. Instead, we emphasize the depth and breadth of the qualitative data. Regarding depth, previous research has suggested that the participants may not provide sufficient

detail in open-ended questions (Kip et al., 2022; Yardley et al., 2016). In contrast, we found that the CORTO method generated a large number of candid responses, many of them exhibiting substantial effort in answering, which facilitated their analysis and offered useful insights. Moreover, we found that the data provided substantial breadth both in terms of respondents and the response content. In the CORTO data, 80.7 % of the participants gave at least one response, and the sample of 184 participants can be considered extensive for a qualitative dataset. Moreover, the CORTO responses provided insights into all aspects of the intervention content (see Fig. 5), and these responses were intervention-specific. These results exhibited how the CORTO method could help overcome the time-consuming nature of qualitative data gathering (Reja et al., 2003; Yardley et al., 2016).

The data generated by the CORTO method provides many avenues for its analysis. The macro-level categories identified in this study (see Table 4) offer one possible framework that can be used to accelerate the qualitative data analysis when user experience is measured. The framework is aligned with an earlier three-tier gameplay experience model by Engl and Nacke who described player experience occurring between the context and game characteristics (Engl and Nacke, 2013). Aligned with the model, we found that the user experiences can be categorized in a continuum from more abstract (the context category) to more concrete (the technical category). However, we found that the participant responses on the emotional category were characterized by their valence, and we contrasted the feature-level usability comments with technical comments. Overall, our ground-up methodology suggested considering user experience as holistic, contextual, dynamic, and temporally bounded (Hassenzahl and Tractinsky, 2006), and not limited to the evaluations of the intervention software alone.

In this study, we gathered the data longitudinally to enhance its precision and coverage while keeping the temporal aspects in the background in the analysis. Further studies could investigate how user experience and engagement evolve (Karapanos et al., 2009; O'Brien and Toms, 2008). Such perspectives may be particularly valuable considering that the behavioral change the digital interventions encourage also occurs over time. We look forward to further temporality-emphasizing research using the CORTO method.

4.3. Limitations

Participant self-selection and survivorship bias should be taken into account when interpreting the results. We found that the study participation was often motivated by its scientific nature, which may have encouraged the participants to share their experiences more willingly than in commercial contexts, for instance. The CORTO method and retrospective interviewing may also invite somewhat different respondents, which may influence the results: participating in a retrospective interview requires scheduling and willingness to participate in an interpersonal exchange, whereas the CORTO method can be responded to rapidly without interpersonal interaction. Also, the number of surviving participants constantly diminishes through dropout (Eysenbach, 2005; Lukka et al., 2023b). Therefore, the data only includes participants who find the intervention sufficiently interesting, have the competence to start using it, and find it meaningful in use (Levesque et al., 2013). Consequently, objective behavioral engagement is closely related to the data gathering, also with the CORTO method.

The qualitative study prompts may have influenced the study results. The retrospective interview was structured with an interview guide, as is typical for semi-structured interviews (McIntosh and Morse, 2015; Short et al., 2018). Thus, the interview themes influenced the content, but we perceived the interview guide as a feature of the method and representative of similar studies (Crane et al., 2017). As such, we considered the semi-structured retrospective interview as an example of similar interviews and their data-creation potential, an argument that is aligned with the aim of this research and its pragmatist philosophy (Johnson and Onwuegbuzie, 2007). Similarly, the CORTO prompt may have guided the

participant to focus on the intervention features rather than their current emotional state, for instance. However, these types of differences were bound to exist, because interviewing includes several questions and follow-up questions, and the CORTO method merely one. We did not expect the two methods to produce identical results, and we were particularly interested in their complementary potential, which is a common rationale for using mixed methods (Skivington et al., 2021) and engaging in methodological triangulation (Thurmond, 2001). Also, the two datasets did not include participant overlap, but we consider it unlikely that adding the 20 interviewee responses to the existing pool of 184 responses would significantly change the CORTO data examined here.

We suggested that the CORTO method contributes to the iterative intervention development. We found that the meso-level categories established in the CORTO data allowed identifying numerous features across the intervention that merit revision. On the other hand, the data also provided descriptions of what the users valued—domains that could be further articulated, developed, and emphasized in the design. However, our trial design did not enable substantial iterations to be made, and follow-up studies are needed to examine how the CORTO data can contribute to the iterations. Moreover, the two functions of iterative development—designing and evaluating (Mohr et al., 2017)—may vary substantially. The evaluation may be more standardized, compared to design that requires interpretation and synthetization of numerous data sources, contextualizing them with organizational and business perspectives, and creating new solutions that may not exist in the forward-looking digital intervention space. The intricacies of the design as a process have been underresearched, and propose an intriguing avenue for understanding how the development team plans, gathers, and translates the evaluation data into iterations.

4.4. Conclusions

In this study, we have introduced the CORTO method (Contextual, One-item, Repeated, Timely, Open-ended), a useful new EMA method for evaluating digital intervention user experience. We found that the method allows for gathering detailed, intervention-specific qualitative insights from a large pool of users with relative ease, which supports its feasibility in the user-centered intervention development. However, the scientific nature of this study may have influenced the results, and participant self-selection and survivorship bias should be taken into account when interpreting the results.

The CORTO method is aligned with the growing use of qualitative methods and proposes a shift in *what* is being studied. In psychotherapy research, qualitative studies have illustrated the experiential components of behavioral change, and therapists can use the information to support their client's change process (Levitt et al., 2016). This paradigm puts the client in the center and appreciates them as an active participant in the treatment. User-centered design has a similar ethos in the field of digital interventions, where studying the users' perceptions can be used to advance intervention design. Furthermore, our study emphasizes that *how* the users are studied matters. The CORTO method described here allows evaluating the fluctuating, contextual, and holistic user experience, and we believe that such a focus can have a substantial role in the success of digital interventions in the coming decades. Therefore, we invite measuring and designing for the critical triad in digital interventions: subjective experience, objective behavior, and intervention effectiveness.

Ethics approval and consent to participate

The study has received positive appraisals from the Helsinki University Hospital (HUS) research ethics committee (HUS/3043/2021) and the Finnish Medicines Agency Fimea (FIMEA/2022/002976) and conforms with the Declaration of Helsinki. The clinical trial has been registered on ClinicalTrials.gov (ClinicalTrials.gov, 2022), and this study has been preregistered in OSF.io (Lukka et al., 2022).

Funding

CRediT authorship contribution statement

Lauri Lukka: Conceptualization, Methodology, Investigation, Writing – original draft, Writing – review & editing, Visualization, Project administration. **Veli-Matti Karhulahti:** Writing – review & editing, Supervision. **Vilma-Reetta Bergman:** Validation, Investigation, Writing – review & editing. **J. Matias Palva:** Funding acquisition, Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no competing interests.

Data availability

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Acknowledgments

Abbreviations

EMA	Ecological Momentary Assessment
MDD	Major Depressive Disorder
TA	Template Analysis
UCD	User-Centered Design

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.invent.2023.100706.

References

Alqahtani, F., Orji, R., 2020. Insights from user reviews to improve mental health apps. Health Informatics J. 26 (3), 2042–2066. https://doi.org/10.1177/1460458219896492.

Baumel, A., Muench, F., 2016. Heuristic evaluation of Ehealth interventions: establishing standards that relate to the therapeutic process perspective. JMIR Mental Health 3 (1), e5. https://doi.org/10.2196/mental.4563.

Baumel, A., Faber, K., Mathur, N., Kane, J.M., Muench, F., 2017. Enlight: a comprehensive quality and therapeutic potential evaluation tool for mobile and web-based eHealth interventions. J. Med. Internet Res. 19 (3) https://doi.org/10.2196/jmir.7270.

Baumel, A., Muench, F., Edan, S., Kane, J.M., 2019. Objective user engagement with mental health apps: systematic search and panel-based usage analysis. J. Med. Internet Res. 21 (9), 1–15. https://doi.org/10.2196/14567.

Bediou, B., Adams, D.M., Mayer, R.E., Tipton, E., Green, C.S., Bavelier, D., 2018. Meta-analysis of action video game impact on perceptual, attentional, and cognitive skills. Psychol. Bull. 144 (1), 77–110. https://doi.org/10.1037/bul0000130.supp.

Brewer, M.B., Crano, W.D., 2014. Research design and issues of validity. In: Reis, H.T., Judd, C.M. (Eds.), Handbook of Research Methods in Social and Personality Psychology (Second Edition). Cambridge University Press.

Brooks, J., McCluskey, S., Turley, E., King, N., 2015. The utility of template analysis in qualitative psychology research. Qual. Res. Psychol. 12 (2), 202–222. https://doi.org/10.1080/14780887.2014.955224.

Burke Johnson, R., Onwuegbuzie, A.J., 2004. Mixed methods research: a research paradigm whose time has come. Educ. Res. 33 (7), 14–26.

ClinicalTrials.gov, 2022. The Effects of Videogames on Depression Symptoms and Brain Dynamics. https://clinicaltrials.gov/ct2/show/NCT05426265.

Cohen, K.A., Schleider, J.L., 2022. Adolescent dropout from brief digital mental health interventions within and beyond randomized trials. Internet Interv. 27 https://doi.org/10.1016/j.invent.2022.100496.

Cooper, C., O'Cathain, A., Hind, D., Adamson, J., Lawton, J., Baird, W., 2014. Conducting qualitative research within clinical trials units: avoiding potential pitfalls. Contemp. Clin. Trials 38 (2), 338–343. https://doi.org/10.1016/j.cct.2014.06.002.

Crane, D., Garnett, C., Brown, J., West, R., Michie, S., 2017. Factors influencing usability of a smartphone app to reduce excessive alcohol consumption: think aloud and interview studies. Front. Public Health 5 (APR). https://doi.org/10.3389/FPUBH.2017.00039.

Curran, G.M., Bauer, M., Mittman, B., Pyne, J.M., Stetler, C., 2012. Effectiveness-implementation hybrid designs: combining elements of clinical effectiveness and implementation research to enhance public health impact. Med. Care 50 (3), 217–226. https://doi.org/10.1097/MLR.0b013e3182408812.

Davis, 1993. User acceptance of information technology. Int. J. Sci. Educ. 41 (18), 2675–2695. https://doi.org/10.1080/09500693.2019.1693081.

de Vries, L.P., Baselmans, B.M.L., Bartels, M., 2021. Smartphone-based ecological momentary assessment of well-being: a systematic review and recommendations for future studies. J. Happiness Stud. 22 (5), 2361–2408. https://doi.org/10.1007/s10902-020-00324-7.

Dekker, M.R., Williams, A.D., 2017. The use of user-centered participatory design in serious games for anxiety and depression. Games Health J. 6 (6), 327–333. https://doi.org/10.1089/g4h.2017.0058.

Doherty, K., Doherty, G., 2018. The construal of experience in HCI: understanding self-reports. Int. J. Hum. Comput. Stud. 110, 63–74. https://doi.org/10.1016/j.ijhcs.2017.10.006.

Doherty, K., Doherty, G., 2019. Engagement in HCI: conception, theory and measurement. ACM Comput. Surv. 51 (5) https://doi.org/10.1145/3234149.

Doherty, K., Balaskas, A., Doherty, G., 2020. The design of ecological momentary assessment technologies. Interact. Comput. 32 (3), 257–278. https://doi.org/10.1093/iwcomp/iwaa019.

Enam, A., Torres-Bonilla, J., Eriksson, H., 2018. Evidence-based evaluation of ehealth interventions: systematic literature review. J. Med. Internet Res. 20 (11) https://doi.org/10.2196/10971.

Engl, S., Nacke, L.E., 2013. Contextual influences on mobile player experience - a game user experience model. Entertain. Comput. 4 (1), 83–91. https://doi.org/10.1016/j.entcom.2012.06.001.

Eysenbach, G., 2005. The law of attrition. J. Med. Internet Res. 7 (1), e11 https://doi.org/10.2196/jmir.7.1.e11.

Finlex, 2011. Decree of the Ministry of Social Affairs and Health on Remuneration Payable to Research Subjects. https://www.finlex.fi/en/laki/kaannokset/2011/en20110082.

Fleming, T., Lucassen, M., Stasiak, K., Shepherd, M., Merry, S., 2016a. The impact and utility of computerised therapy for educationally alienated teenagers: the views of adolescents who participated in an alternative education-based trial. Clin. Psychol. 20 (2), 94–102. https://doi.org/10.1111/cp.12052.

Fleming, T.M., de Beurs, D., Khazaal, Y., Gaggioli, A., Riva, G., Botella, C., Baños, R.M., Aschieri, F., Bavin, L.M., Kleiboer, A., Merry, S., Lau, H.M., Riper, H., 2016b. Maximizing the impact of E-therapy and serious gaming: time for a paradigm shift. Front. Psychol. 7 (65), 1–7. https://doi.org/10.3389/fpsyt.2016.00065.

Fleming, T., Bavin, L., Lucassen, M., Stasiak, K., Hopkins, S., Merry, S., 2018. Beyond the trial: systematic review of real-world uptake and engagement with digital self-help interventions for depression, low mood, or anxiety. J. Med. Internet Res. 20 (6), 1–11. https://doi.org/10.2196/jmir.9275.

Gan, D.Z.Q., McGillivray, L., Han, J., Christensen, H., Torok, M., 2021. Effect of engagement with digital interventions on mental health outcomes: a systematic review and meta-analysis. Front. Digital Health 3 (764079). https://doi.org/10.3389/fdgth.2021.764079.

Gohier, B., Ferracci, L., Surguladze, S.A., Lawrence, E., El Hage, W., Kefi, M.Z., Allain, P., Garre, J.B., Le Gall, D., 2009. Cognitive inhibition and working memory in unipolar depression. J. Affect. Disord. 116 (1–2), 100–105. https://doi.org/10.1016/j.jad.2008.10.028.

Graham, A.K., Kwasny, M.J., Lattie, E.G., Greene, C.J., Gupta, N.V., Reddy, M., Mohr, D.C., 2021. Targeting subjective engagement in experimental therapeutics for digital mental health interventions. Internet Interv. 25 https://doi.org/10.1016/j.invent.2021.100403.

Harvey, P.O., Le Bastard, G., Pochon, J.B., Levy, R., Allilaire, J.F., Dubois, B., Fossati, P., 2004. Executive functions and updating of the contents of working memory in unipolar depression. J. Psychiatr. Res. 38 (6), 567–576. https://doi.org/10.1016/j.jpsychires.2004.03.003.

Hassenzahl, M., Tractinsky, N., 2006. User experience - a research agenda. Behav. Inform. Technol. 25 (2), 91–97. https://doi.org/10.1080/01449290500330331.

Hookham, G., Nesbitt, K., 2019, January 29. A systematic review of the definition and measurement of engagement in serious games. ACM Int. Conf. Proc. Ser. https://doi.org/10.1145/3290688.3290747.

ISO 9241-11:2018, 2018. ISO.Org. https://www.iso.org/obp/ui/#iso:std:iso:9241:-11:ed-2:v1:en.

Jerzak, N., Rebelo, F., 2014. Serious games and heuristic evaluation-the cross-comparison of existing heuristic evaluation methods for games. DUXU 8517.

Johnson, R.B., Onwuegbuzie, A.J., 2007. Toward a definition of mixed methods research. J. Mixed Methods Res. 1 (2), 112–133. https://doi.org/10.1177/1558689806298224.

Karapanos, E., Zimmerman, J., Forlizzi, J., Martens, J.-B., 2009. User Experience Over Time: An Initial Framework. CHI 2009.

Karyotaki, E., Kleiboer, A., Smit, F., Turner, D.T., Pastor, A.M., Andersson, G., Berger, T., Botella, C., Breton, J.M., Carlbring, P., Christensen, H., de Graaf, E., Griffiths, K., Donker, T., Farrer, L., Huibers, M.J.H., Lenndin, J., Mackinnon, A., Meyer, B., Cuijpers, P., 2015. Predictors of treatment dropout in self-guided web-based interventions for depression: an "individual patient data" meta-analysis. Psychol. Med. 45 (13), 2717–2726. https://doi.org/10.1017/S0033291715000665.

Kip, H., Keizer, J., da Silva, M.C., Jong, N.B. De, Köhle, N., Kelders, S.M., 2022. Methods for human-centered eHealth development: narrative scoping review. J. Med. Internet Res. 24 (1) https://doi.org/10.2196/31858.

Kleiman, E.M., Turner, B.J., Fedor, S., Beale, E.E., Huffman, J.C., Nock, M.K., 2017. Examination of real-time fluctuations in suicidal ideation and its risk factors: results from two ecological momentary assessment studies. J. Abnorm. Psychol. 126 (6), 726–738. https://doi.org/10.1037/abn0000273.

Koster, E.H.W., Hoorelbeke, K., Onraedt, T., Owens, M., Derakshan, N., 2017. Cognitive control interventions for depression: a systematic review of findings from training studies. Clin. Psychol. Rev. 53, 79–92. https://doi.org/10.1016/j.cpr.2017.02.002.

Kroenke, K., Spitzer, R.L., Williams, J.B.W., 2001. The PHQ-9: validity of a brief depression severity measure. J. Gen. Intern. Med. 16 (9), 606–613. https://doi.org/10.1046/j.1525-1497.2001.016009606.x.

Launder, N.H., Minkov, R., Davey, C.G., Finke, C., Hanna, Gavelin, M., Lampit, A., 2021. Computerized cognitive training in people with depression: a systematic review and meta-analysis of randomized clinical trials. MedRxiv. https://doi.org/10.1101/2021.03.23.21254003.

Law, E.L.-C., Roto, V., Hassenzahl, M., Vermeeren, A.P.O.S., Kort, J., 2009. Understanding, Scoping and Defining User eXperience: A Survey Approach. CHI 2009.

Lemmens, J.S., Valkenburg, P.M., Peter, J., 2009. Development and validation of a game addiction scale for adolescents. Media Psychol. 12 (1), 77–95. https://doi.org/10.1080/15213260802669458.

Levesque, J.F., Harris, M.F., Russell, G., 2013. Patient-centred access to health care: conceptualising access at the interface of health systems and populations. Int. J. Equity Health 12 (1). https://doi.org/10.1186/1475-9276-12-18.

Levitt, H.M., Pomerville, A., Surace, F.I., 2016. A qualitative meta-analysis examining clients' experiences of psychotherapy: a new agenda. Psychol. Bull. 142 (8), 801–830. https://doi.org/10.1037/bul0000057.

Lipschitz, J.M., van Boxtel, R., Torous, J., Firth, J., Lebovitz, J., Burdick, K.E., Hogan, T. P., 2022. Digital mental health interventions for depression: a scoping review of user engagement. J. Med. Internet Res. https://doi.org/10.2196/39204.

Lukka, L., Palva, J.M., 2023. The development of game-based digital mental health interventions: bridging the paradigms of health care and entertainment. JMIR Serious Games 11, e42173. https://doi.org/10.2196/42173.

Lukka, L., Karhulahti, V.-M., Palva, M., 2022. Participant Perceptions of Game-based Digital Therapeutics Software for Major Depressive Disorder (Meliora). https://doi.org/10.17605/OSF.IO/9Q2KX.

Lukka, L., Karhulahti, V.-M., Palva, J.M., 2023a. Factors affecting digital tool use in client interaction according to mental health professionals: interview study. JMIR Hum. Factors 10, e44681. https://doi.org/10.2196/44681.

Lukka, L., Salonen, A., Vesterinen, M., Karhulahti, V.-M., Palva, S., Palva, J.M., 2023b. The qualities of patients interested in using a game-based digital mental health intervention for depression: a sequential mixed methods study. BMC Digital Health 1 (1), 37. https://doi.org/10.1186/s44247-023-00037-w.

Lyon, A.R., Koerner, K., 2016. User-centered design for psychosocial intervention development and implementation. Clin. Psychol. Sci. Pract. 23 (2), 180–200. https://doi.org/10.1111/cpsp.12154.

Maxwell, J.A., 2010. Using numbers in qualitative research. Qual. Inq. 16 (6), 475–482. https://doi.org/10.1177/1077800410364740.

McDevitt-Murphy, M.E., Luciano, M.T., Zakarian, R.J., 2018. Use of ecological momentary assessment and intervention in treatment with adults. FOCUS 16 (4), 370–375. https://doi.org/10.1176/appi.focus.20180017.

McIntosh, M.J., Morse, J.M., 2015. Situating and constructing diversity in semi-structured interviews. Glob. Qual. Nurs. Research 2. https://doi.org/10.1177/2333393615597674.

Michie, S., Yardley, L., West, R., Patrick, K., & Greaves, F. (2017). Developing and evaluating digital interventions to promote behavior change in health and health care: recommendations resulting from an international workshop. In *Journal of Medical Internet Research* (Vol. 19, Issue 6). JMIR Publications Inc. doi:https://doi.org/10.2196/jmir.7126.

Mohr, D.C., Lyon, A.R., Lattie, E.G., Reddy, M., Schueller, S.M., 2017. Accelerating digital mental health research from early design and creation to successful implementation and sustainment. J. Med. Internet Res. 19 (5) https://doi.org/10.2196/jmir.7725.

Motter, J.N., Pimontel, M.A., Rindskopf, D., Devanand, D.P., Doraiswamy, P.M., Sneed, J.R., 2016. Computerized cognitive training and functional recovery in major depressive disorder: a meta-analysis. J. Affect. Disord. 189, 184–191. https://doi.org/10.1016/j.jad.2015.09.022.

Mummah, S.A., Robinson, T.N., King, A.C., Gardner, C.D., Sutton, S., 2016. IDEAS (integrate, design, assess, and share): a framework and toolkit of strategies for the development of more effective digital interventions to change health behavior. J. Med. Internet Res. 18 (12) https://doi.org/10.2196/jmir.5927.

Nacke, L., 2009. From playability to a hierarchical game usability model. In: FuturePlay 2009 at GDC Canada International Conference on the Future of Game Design and Technology, pp. 11–12. https://doi.org/10.1145/1639601.1639609.

Nacke, L., Drachen, A., 2011. Towards a framework of player experience research. EPEX 11.

Ng, M.M., Firth, J., Minen, M., Torous, J., 2019. User engagement in mental health apps: a review of measurement, reporting, and validity. Psychiatr. Serv. 70 (7), 538–544. https://doi.org/10.1176/appi.ps.201800519.

O'Brien, H.L., Toms, E.G., 2008. What is user engagement? A conceptual framework for defining user engagement with technology. J. Am. Soc. Inf. Sci. Technol. 59 (6), 938–955. https://doi.org/10.1002/asi.20801.

O'Brien, B.C., Harris, I.B., Beckman, T.J., Reed, D.A., Cook, D.A., 2014. Standards for reporting qualitative research: a synthesis of recommendations. Acad. Med. 89 (9), 1245–1251. https://doi.org/10.1097/ACM.0000000000000388.

O'Cathain, A., Thomas, K.J., Drabble, S.J., Rudolph, A., Hewison, J., 2013. What can qualitative research do for randomised controlled trials? A systematic mapping review. BMJ Open 3 (6). https://doi.org/10.1136/bmjopen-2013-002889.

Ospina-Pinillos, L., Davenport, T.A., Ricci, C.S., Milton, A.C., Scott, E.M., Hickie, I.B., 2018. Developing a mental health eclinic to improve access to and quality of mental health care for young people: using participatory design as research methodologies. J. Med. Internet Res. 20 (5) https://doi.org/10.2196/JMIR.9716.

Perski, O., Blandford, A., West, R., Michie, S., 2017. Conceptualising engagement with digital behaviour change interventions: a systematic review using principles from critical interpretive synthesis. Transl. Behav. Med. 7 (2), 254–267. https://doi.org/10.1007/s13142-016-0453-1.

Petrie, H., Bevan, N., 2009. The evaluation of accessibility, usability and user experience. In: Stephanidis, C. (Ed.), The Universal Access Handbook. CRC Press.

Pine, R., Sutcliffe, K., McCallum, S., Fleming, T., 2020. Young adolescents' interest in a mental health casual video game. Digital Health 6, 1–7. https://doi.org/10.1177/2055207620949391.

Proudfoot, K., 2022. Inductive/deductive hybrid thematic analysis in mixed methods research. J. Mixed Methods Res. https://doi.org/10.1177/15586898221126816.

Quiñones, D., Rusu, C., Rusu, V., 2018. A methodology to develop usability/user experience heuristics. Comput. Stand. Interfaces 59, 109–129. https://doi.org/10.1016/j.csi.2018.03.002.

Reichert, M., Gan, G., Renz, M., Braun, U., Brüßler, S., Timm, I., Ma, R., Berhe, O., Benedyk, A., Moldavski, A., Schweiger, J.I., Hennig, O., Zidda, F., Heim, C., Banaschewski, T., Tost, H., Ebner-Priemer, U.W., Meyer-Lindenberg, A., 2021. Ambulatory assessment for precision psychiatry: foundations, current developments and future avenues. Exp. Neurol. 345 https://doi.org/10.1016/j.expneurol.2021.113807.

Reis, H.T., 2012. Why researchers should think "real-world". A conceptual rationale. In: Mehl, M.R., Conner, T.S. (Eds.), Handbook of Research Methods for Studying Daily Life. The Guilford Press, pp. 3–21.

Reja, U., Manfreda, K.L., Hlebec, V., Vehovar, V., 2003. Open-ended vs (Close-ended Questions in Web Questionnaires. Developments in Applied Statistics).

Ritterband, L.M., Thorndike, F.P., Cox, D.J., Kovatchev, B.P., Gonder-Frederick, L.A., 2009. A behavior change model for internet interventions. Ann. Behav. Med. 38 (1), 18–27. https://doi.org/10.1007/s12160-009-9133-4.

Robinson, M.D., Clore, G.L., 2002. Belief and feeling: evidence for an accessibility model of emotional self-report. Psychol. Bull. 128 (6), 934–960. https://doi.org/10.1037/0033-2909.128.6.934.

Rock, P.L., Roiser, J.P., Riedel, W.J., Blackwell, A.D., 2014. Cognitive impairment in depression: a systematic review and meta-analysis. Psychol. Med. 44 (10), 2029–2040. https://doi.org/10.1017/S0033291713002535.

Rose, E.J., Ebmeier, K.P., 2006. Pattern of impaired working memory during major depression. J. Affect. Disord. 90 (2–3), 149–161. https://doi.org/10.1016/j.jad.2005.11.003.

Saldaña, J., 2016. The Coding Manual for Qualitative Researchers, 3rd edition. SAGE.

Sánchez, J.L.G., Vela, F.L.G., Simarro, F.M., Padilla-Zea, N., 2012. Playability: analysing user experience in video games. Behav. Inform. Technol. 31 (10), 1033–1054. https://doi.org/10.1080/0144929X.2012.710648.

Scholten, H., Granic, I., 2019. Use of the principles of design thinking to address limitations of digital mental health interventions for youth: viewpoint. J. Med. Internet Res. 21 (1), 1–14. https://doi.org/10.2196/11528.

Schwarz, N., 1999. Self-reports: how the questions shape the answers. Am. Psychol. 54 (2), 93–105. https://doi.org/10.1037/0003-066X.54.2.93.

Schwarz, N., 2012. Why researchers should think "real-time": a cognitive rationale. In: Mehl, M.R., Conner, T.S. (Eds.), Handbook of Research Methods for Studying Daily Life. The Guilford Press, pp. 22–42.

Sheehan, D.V., Lecrubier, Y., Sheehan, K.H., Amorim, P., Janavs, J., Weiller, E., Hergueta, T., Baker, R., Dunbar, G.C., 1998. The Mini-international neuropsychiatric interview (M.I.N.I.): the development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. J. Clin. Psychiatry 59 (20), 22–33.

Shiffman, S., Stone, A.A., Hufford, M.R., 2008. Ecological momentary assessment. Annu. Rev. Clin. Psychol. 4, 1–32. https://doi.org/10.1146/annurev.clinpsy.3.022806.091415.

Short, C.E., DeSmet, A., Woods, C., Williams, S.L., Maher, C., Middelweerd, A., Müller, A. M., Wark, P.A., Vandelanotte, C., Poppe, L., Hingle, M.D., Crutzen, R., 2018. Measuring engagement in eHealth and mHealth behavior change interventions:

viewpoint of methodologies. J. Med. Internet Res. 20 (11), 1–18. https://doi.org/10.2196/jmir.9397.

Skivington, K., Matthews, L., Simpson, S.A., Craig, P., Baird, J., Blazeby, J.M., Boyd, K.A., Craig, N., French, D.P., McIntosh, E., Petticrew, M., Rycroft-Malone, J., White, M., Moore, L., 2021. A new framework for developing and evaluating complex interventions: update of Medical Research Council guidance. The BMJ 374. https://doi.org/10.1136/bmj.n2061.

Small, M.L., 2011. How to conduct a mixed methods study: recent trends in a rapidly growing literature. Annu. Rev. Sociol. 37 (1), 57–86. https://doi.org/10.1146/annurev.soc.012809.102657.

Smith, K.A., Blease, C., Faurholt-Jepsen, M., Firth, J., Van Daele, T., Moreno, C., Carlbring, P., Ebner-Priemer, U.W., Koutsouleris, N., Riper, H., Mouchabac, S., Torous, J., Cipriani, A., 2023. Digital mental health: challenges and next steps. BMJ Mental Health 26 (1), e300670. https://doi.org/10.1136/bmjment-2023-300670.

Spitzer, R.L., Kroenke, K., Williams, J.B.W., Löwe, B., 2006. A brief measure for assessing generalized anxiety disorder the GAD-7. Arch. Intern. Med. 166 (10), 1092–1097. https://jamanetwork.com/.

Stetler, C.B., Legro, M.W., Wallace, C.M., Bowman, C., Guihan, M., Hagedorn, H., Kimmel, B., Sharp, N.D., Smith, J.L., 2006. The role of formative evaluation in implementation research and the QUERI experience. J. Gen. Intern. Med. 21 (Suppl. 2), 1–8. https://doi.org/10.1111/j.1525-1497.2006.00355.x.

Stoyanov, S.R., Hides, L., Kavanagh, D.J., Zelenko, O., Tjondronegoro, D., Mani, M., 2015. Mobile app rating scale: a new tool for assessing the quality of health mobile apps. JMIR Mhealth Uhealth 3 (1), 1–9. https://doi.org/10.2196/mhealth.3422.

Stoyanov, S.R., Hides, L., Kavanagh, D.J., Wilson, H., 2016. Development and validation of the user version of the mobile application rating scale (uMARS). JMIR Mhealth Uhealth 4 (2), 1–5. https://doi.org/10.2196/mhealth.5849.

Thurmond, V.A., 2001. The point of triangulation. J. Nurs. Scholarsh. 33 (3), 253–258. https://doi.org/10.1111/j.1547-5069.2001.00253.x.

Tondello, G. F., Kappen, D. L., Mekler, E. D., Ganaba, M., & Nacke, L. E. (2016). Heuristic evaluation for gameful design. *CHI PLAY 2016 - Proceedings of the Annual Symposium on Computer-Human Interaction in Play Companion*, 315–323. https://doi.org/10.1145/2968120.2987729.

Torous, J., Lipschitz, J., Ng, M., Firth, J., 2020. Dropout rates in clinical trials of smartphone apps for depressive symptoms: a systematic review and meta-analysis. J. Affect. Disord. 263, 413–419. https://doi.org/10.1016/j.jad.2019.11.167.

Torous, J., Bucci, S., Bell, I.H., Kessing, L.V., Faurholt-Jepsen, M., Whelan, P., Carvalho, A.F., Keshavan, M., Linardon, J., Firth, J., 2021. The growing field of digital psychiatry: current evidence and the future of apps, social media, chatbots,

and virtual reality. World Psychiatry 20 (3), 318–335. https://doi.org/10.1002/wps.20883.

Trull, T.J., Ebner-Priemer, U., 2013. Ambulatory assessment. Annu. Rev. Clin. Psychol. 9, 151–176. https://doi.org/10.1146/annurev-clinpsy-050212-185510.

Van Den Haak, M.J., De Jong, M.D.T., Schellens, P.J., 2003. Retrospective vs. concurrent think-aloud protocols: testing the usability of an online library catalogue. Behav. Inform. Technol. 22 (5), 339–351. https://doi.org/10.1080/0044929031000.

Van Gemert-Pijnen, J.E.W.C., Nijland, N., Van Limburg, M., Ossebaard, H.C., Kelders, S. M., Eysenbach, G., Seydel, E.R., 2011. A holistic framework to improve the uptake and impact of ehealth technologies. J. Med. Internet Res. 13 (4) https://doi.org/10.2196/jmir.1672.

Verschueren, S., Buffel, C., Stichele, G. Vander, 2019. Developing theory-driven, evidence-based serious games for health: framework based on research community insights. JMIR Serious Games 7 (2). https://doi.org/10.2196/11565.

Warsinsky, S., Schmidt-Kraepelin, M., Rank, S., Thiebes, S., Sunyaev, A., 2021. Conceptual ambiguity surrounding gamification and serious games in health care: literature review and development of game-based intervention reporting guidelines (gamING). J. Med. Internet Res. 23 (9) https://doi.org/10.2196/30390.

Wehbe, R.R., Whaley, C., Eskandari, Y., Suarez, A., Nacke, L.E., Hammer, J., Lank, E., 2022. Designing a serious game (above water) for stigma reduction surrounding mental health: semistructured interview study with expert participants. JMIR Serious Games 10 (2), e21376. https://doi.org/10.2196/21376.

World Health Organization, 2022. Transforming mental health for all. https://www.who.int/publications/i/item/9789240049338.

Yardley, L., Morrison, L., Bradbury, K., Muller, I., 2015. The person-based approach to intervention development: application to digital health-related behavior change interventions. J. Med. Internet Res. 17 (1), e30 https://doi.org/10.2196/jmir.4055.

Yardley, L., Spring, B.J., Riper, H., Morrison, L.G., Crane, D.H., Curtis, K., Merchant, G. C., Naughton, F., Blandford, A., 2016. Understanding and promoting effective engagement with digital behavior change interventions. Am. J. Prev. Med. 51 (5), 833–842. https://doi.org/10.1016/j.amepre.2016.06.015.

Zhang, R., Nicholas, J., Knapp, A.A., Graham, A.K., Gray, E., Kwasny, M.J., Reddy, M., Mohr, D.C., 2019. Clinically meaningful use of mental health apps and its effects on depression: mixed methods study. J. Med. Internet Res. 21 (12) https://doi.org/10.2196/15644.

Zhou, L., Bao, J., Setiawan, I.M.A., Saptono, A., Parmanto, B., 2019. The mhealth app usability questionnaire (MAUQ): development and validation study. JMIR Mhealth Uhealth 7 (4), 1–15. https://doi.org/10.2196/11500.