EUD Survival "in the Wild": Evaluation Challenges for Field Deployments and How to Address Them

Daniel Tetteroo and Panos Markopoulos

Abstract This chapter discusses methodological choices facing researchers wishing to evaluate end user development technologies. While laboratory evaluations or short term evaluations are often conducted as a way to validate an end user development technology, these do not provide sufficient guarantees regarding the adoption of end user development practices and how systems should be improved to encourage such practices. The challenges pertaining to field deployments are discussed first at an operational level and second at a teleological level where we debate what should be success criteria for such studies. Discussing previous studies and our experiences from a deployment case in the healthcare sector, we propose guidelines for the evaluation of EUD technologies.

Keywords Evaluation of EUD technologies · EUD deployment success · surveys · methodological choices

1 Introduction

A recent literature survey on research methods and purposes characterizing research studies in the field of end user development (Tetteroo & Markopoulos, 2015) has shown that field evaluations of EUD systems are relatively uncommon. Mostly, these systems are evaluated in a lab setting, an approach which while useful and sometimes a necessary prerequisite to field testing, disregards the impact of the context of actual use in which such systems would be deployed in practice. Although it is worrying that there have been only a few attempts to deploy an EUD system in the field, it is also quite understandable. After all, arranging a field deployment is usually much harder and costly than arranging a lab study. Further,

D. Tetteroo (✉) · P. Markopoulos
Eindhoven University of Technology, Eindhoven, Netherlands
e-mail: d.tetteroo@tue.nl

P. Markopoulos
e-mail: p.markopoulos@tue.nl

F. Paternò, V. Wulf (eds.), *New Perspectives in End-User Development*,
DOI 10.1007/978-3-319-60291-2_9

and this is the major point throughout this chapter, evaluating the impact of an EUD deployment is far from trivial: What exactly needs to be evaluated, in order to conclude anything about the success of an EUD deployment? Which measures should be taken and which outcomes are to be expected?

This chapter discusses these methodological questions, starting with a discussion on what actually constitutes "success" in the case of EUD deployments. Then we reflect on our own deployment studies of an end-user adaptable technology for physical rehabilitation. We present a structured literature survey on previous attempts of EUD deployments, analyzing the evaluations performed and the success measures considered in those studies. Finally, we propose some guidelines for the evaluation of EUD field deployments.

2 Related Work

A recent literature survey (Tetteroo & Markopoulos, 2015) that classified research methods used in the field of End User Development (EUD), pointed out that a significant part of the work that is performed in the field of EUD (42%, of the works covered in that survey) includes an evaluation of EUD systems or parts thereof. Of course, not all of user evaluations are equal in nature; published studies apply quite diverse methods (e.g., case study, lab study) and measures. Choosing amongst these methods reflects the particular aims of the research, e.g., whether it aims to assess the usability of a system or the impact of a particular technology in the workplace, or perhaps how successful is a particular theoretical framework in guiding design, etc.

While formative evaluations are a common and, arguably, a necessary element of the design most interactive systems and therefore also EUD technologies, research articles in this field that introduce EUD technologies report summative evaluations as a means to demonstrate the success of the design effort; examples of such works are (Namoun, Wajid, Mehandjiev, & Owrak, 2012; Wong & Hong, 2007). Most often such evaluations are conducted in a laboratory setting, where test participants use the system tested on artificial tasks selected for the evaluation rather than the actual work or activities spontaneously stemming from their own interests and real life needs. Also, testing often takes place in controlled conditions rather than in the context of actual work or daily life. In these cases, success measures are often related to the usability of the tool and the efficiency with which users can complete tool-related tasks. Beyond the efficiency of the tool itself some researchers focus on the impact of specific methods, practices or functionality on the behavior of their users, e.g., (Ruthruff, Prabhakararao, & Reichwein, 2005; Tsandilas, Letondal, & Mackay, 2009). However, it is not often that research papers examine what happens if experimental EUD systems are deployed in a context of actual use. Interestingly it appears that they are also not very explicit about how they define what should be considered a successful EUD field deployment. This chapter, therefore, explores further the question of successful EUD deployments, and aims to establish a common understanding hereof amongst the members of the EUD research community.

3 Defining “Success” in Field Deployments of End User Development Technology

With regards to information technology there exists a fairly established view on how success can be defined referring to actual use and adoption of novel technologies and perhaps factors that predict it, see for example (Venkatesh, Morris, Davis, & Davis, 2003). Here we argue that transposing such concepts and criteria to EUD is not straightforward and researchers seem to hold different assumptions regarding success for EUD in the field.

3.1 What Makes EUD Special?

One could state that the deployment of an EUD system is nothing more than a specific case of software deployment in general, which brings together concerns regarding the technology, its users, and the context of deployment. For example, a successful deployment might require the technology to be functional and match the needs of its users and require it to fit the organization’s goals.

However, it is important to note that from a technological perspective EUD is often an “extra” layer, an add-on to a technology that already provides some value to its users (see the Chap. 2). After all, if the essence and main purpose of a technology would be to allow for the modification, extension and creation of software artifacts, this technology would in fact be a “regular” software development environment[1]. Similarly, from a socio-technical perspective, EUD is an additional activity that users may perform, aiding them in achieving a grander core task. After all, if development would be a person’s primary activity, the person would be a developer rather than a user of that technology.

In other words, the EUD component of a technology is per definition auxiliary to that host technology. This does not imply that the EUD functionality needs to be deployed separate from the host technology itself. In fact, it often forms an integral part of it, such as in the case of macro editors in office software, or level editors in games. Nevertheless, the core tasks that end users perform with these base technologies will not be EUD related.

Assuming this view on EUD, one can state that the adoption of EUD practices transcends regular use of the host technology. Where technology *use* implies the application of that technology for a core task, *EUD* requires end users to deviate from that task to engage in an activity that will presumably, eventually benefit the core task. As such, it creates additional challenges over and above those that come with the deployment of “traditional software.” As with most definitions

[1]The discussion here steers clear from programming environments that address novice programmers with general purpose programming languages and development environments for which success criteria are very different and more similar to information systems in general.

that imply an inclusion/exclusion criterion we expect that there will be cases that do not neatly follow this rule; however, for a large majority of cases referring to the long tail of software engineering, this definition appears like a useful departure point.

In light of this view, two important questions arise when it comes to evaluating EUD deployments:

1. Does it make sense to separate the evaluation of the *EUD-part* of a socio-technical system from the *use-part*?
2. In what way could one separately evaluate the impact of EUD?

These questions are discussed later in this chapter.

3.2 How to Define "Success" of EUD?

Often, from the perspective of the EUD researcher, "success" equals the adoption of EUD practices. The rationale adopted here is often very direct: *people are using my (EUD) tool, so it must be good*. However, such adoption is usually put forward as a means towards a higher order goal, such as increased efficiency in completing repetitive tasks through the creation of macros, or the personalization of technology, etc. Given that there can be alternative ways to achieve such higher-level goals, not necessarily involving any EUD, usage as such does not equate to success. Moreover, there might even be cases in which the adoption of EUD practices indicates failure, e.g., a system is so poorly designed that end users are forced to "fix" it through EUD. In short, simply showing that EUD is actually taking place does not represent a sufficient evaluation goal.

Consider the example of a primary school teacher who aims to increase her pupils' motivation during a math class. During the class, pupils learn arithmetic by interacting with a virtual character on their tablet computers. One way to increase their motivation is by tailoring the math exercises to the personal interests of each specific student, e.g., sports, animals, cars. In this scenario, EUD deployment would be successful if the teacher would adopt EUD practices in order to create personalized training content for her pupils, eventually increasing their motivation and performance at school. It is these latter end-goals that represent success rather than engaging in EUD as such.

Although the above scenario relates to a typical EUD case (tailorability and personalization), there are other cases in which a continuous occurrence of EUD practices are in fact a sign of failure. An example class of such scenarios is the "IKEA case": the business model of this furniture supplier requires customers to assemble their own furniture. While some customers might actually enjoy the process of assembling their newly bought furniture, most customers would probably prefer pre-assembled furniture instead and only choose to construct their furniture to save costs and facilitate transport from the shop. In a similar manner, if development tasks are "offloaded" to end users that could have been handled as well (or even better) by technology providers, one can hardly consider such end

user development practices signs of a successful software deployment (Fischer, 2011). The notion of a successful EUD deployment is thus strongly tied to the tasks it aims to facilitate and is application specific. Despite this high level of context dependency, successful EUD deployments have in common that they aim to maximize the value of EUD within their context, thus increasing the likelihood that EUD practices contribute to the achievement of end users' goals. In the words of Fischer et al. (see their chapter elsewhere in this book), users should be enabled *to participate and to contribute actively in personally meaningful problems*.

Some questions remain, however, such as: *How to best capture evidence of the success of an EUD deployment? What measures are best to be used?*, and *What methods are most likely to deliver the desired data?* In the remainder of this chapter, we first analyze and reflect on the evaluations performed during four EUD deployment studies in a healthcare setting. Then, we compare these evaluations to deployment evaluations performed by other researchers in previous studies. We discuss whether and how existing theoretical models can help design and interpret such evaluation studies, and from there we finally draw some general guidelines for future evaluations of EUD deployments.

4 Evaluating TagTrainer

In the following paragraphs we discuss and review a series of deployment studies concerning the customization and personalization of rehabilitation training technology by means of EUD. These studies and their findings relating to the quality of the therapy and the attitudes of the therapists are described extensively in (Tetteroo, Timmermans, Seelen, & Markopoulos, 2014; Tetteroo, Vreugdenhil, & Grisel, 2015); below we reflect on methodological aspects aiming to draw lessons of more general interest for evaluating EUD deployments. We start by introducing TagTiles, the host technology enabling tangible interaction, and TagTrainer, the EUD environment for constructing interactive exercises.

4.1 TagTiles and TagTrainer

TagTiles is an interactive board that supports tangible interaction with objects adorned with RFID tags; see Fig. 1. It encases a grid of RFID tag readers and a grid of RGB LED lights that provide visual stimuli and feedback for tangible interaction. The board can detect placement, lifting and movement of objects on its surface; interaction involves physical manipulations of the objects and audio/visual output.

TagTrainer is a software system that runs on a personal computer connected to the TagTiles board, which can be used to select, author, and execute interactive exercises for the board. TagTrainer supports upper extremity rehabilitation for neurological patients including stroke survivors, multiple sclerosis patients, spinal

Fig. 1 TagTiles board by SymbioTherapy, with a cup; a target for placing the mug down is highlighted with blue color

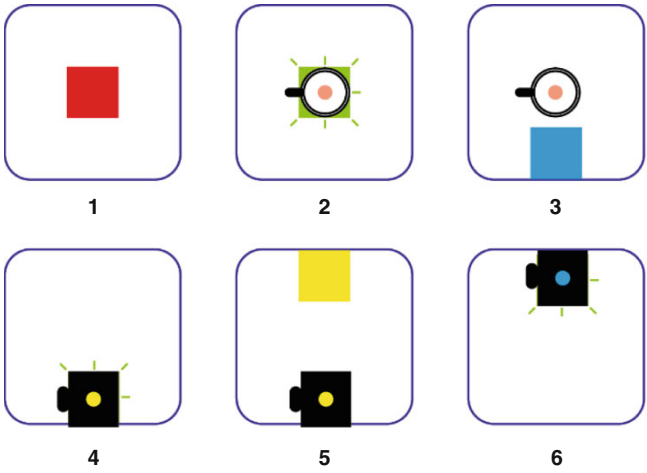

Fig. 2 Storyboard representation of a simple exercise: (1) A target (red square) lights up (2) a cup is placed on the target which turns green (3) another target lights up (blue square), (4) the cup is rotated 90° around an axis parallel to the plane (5) a final target appears (yellow square) (6) the cup is rotated the other way to touch the final target with its yellow marker. Note that the target colors correspond with the colors of the tags attached to the cup at suitable positions

cord injury patients and cerebral palsy patients. TagTrainer can help train daily living skills, e.g., opening a box, drinking from a cup, eating with knife and fork, etc., by prompting the patient to carry out relevant manipulations of such objects and by providing stimulating feedback. Typically, exercises consist of multiple iterations where a target area on the TagTiles board lights up, and the patient needs to touch this area with the appropriate side of an object (see Fig. 2).

Rather than prepackaging exercises with the interactive board, as were the first therapeutic applications for TagTiles (Lanfermann, Te Vrugt, & Timmermans, 2007; Li, Fontijn, & Markopoulos, 2008), TagTrainer provides a simple timeline based programming interface (TagTrainer Exercise Creator, see Fig. 3). Exercises can be modified or created by dragging actions (such as: “place object,” “move object,” etc.) onto the timeline, assigning RFID-tagged parts of an object (e.g., the

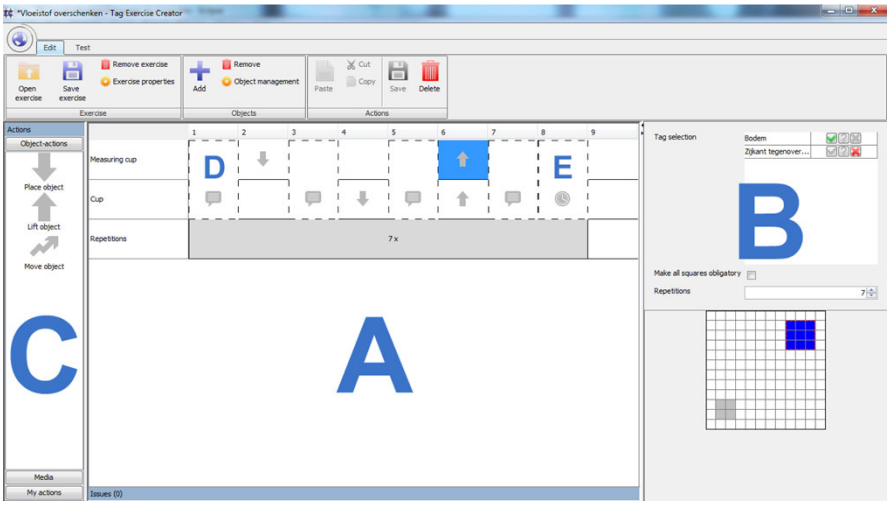

Fig. 3 The TagTrainer Exercise Creator software. The center area (A) shows the workspace with the exercise, featuring a timeline with actions associated with the objects (“measuring cup” and “cup”) involved in the exercise. Properties of the selected action (in this case “lift object”) such as position on the board are displayed to the right (B). Finally, additional actions can be dragged from the library (C) into the workspace to extend an exercise. Note that beyond actions involving manipulations on the board, other actions such as giving instructions (D) and pausing (E) can be used

bottom of a cup, the index finger on a glove) to these actions, and indicating corresponding target areas on the TagTiles board.

This TagTrainer Exercise Creator interface allows therapists to create exercises for each patient addressing their specific training needs and explicit requests. In a therapy session supported by TagTrainer the therapist might commence by inquiring what the patient wishes to train, rapidly create an exercise or retrieve an exercise created earlier, and then ask the patient to train for a certain duration or number of repetitions. This type of approach fundamentally changes the role of therapists; next to their role of caregivers – who instruct, monitor and encourage – they are also responsible for creating software content. Programming exercises for TagTiles is straightforward using TagTrainer, though creating such interactive applications without TagTrainer requires considerable effort and software development expertise.

5 Evaluation of TagTrainer

We chose to evaluate TagTrainer in several stages. First, TagTrainer was evaluated in a lab setting in order to ensure it was fit for use in the field (Hochstenbach-Waelen, Timmermans, & Seelen, 2012). After some improvements and participatory design activities carried out on location at a rehabilitation clinic, a series of four field

deployment studies was conducted. In all these studies, we were interested in whether and how rehabilitation therapists would engage in EUD practices. More specifically, we were interested in identifying and understanding factors influencing the adoption of EUD practices in the workplace, the feasibility of EUD in the context of a rehabilitation clinic, and how technical aspects of the TagTrainer influence or hinder this feasibility.

5.1 Success Criteria

The latter study goals are mainly related to an EUD research agenda. However, for EUD to occur, TagTrainer first needed to be accepted as a technology for use in physical rehabilitation. After all, without the technology being adopted by therapists, adopting EUD practices would not be possible in the first place. Therefore, the following success criteria were used during the four case studies:

SC1. Therapists accept TagTrainer as a viable technology for arm-hand rehabilitation.
SC2. Therapists use TagTrainer in daily arm-hand therapy.
SC3. Therapists are able to perform EUD activities with TagTrainer.
SC4. Therapists perform EUD activities as part of their daily work.

5.2 Methodology

Two different methodologies were used in the evaluation of TagTrainer. For the first case study, an action research methodology (Herr & Anderson, 2014) was applied. In action research the researcher has a dual agenda of effecting a change in the context of the study (here to introduce a new form of therapy which requires a different set of responsibilities for therapists) and to study the process of change. The rationale behind applying this methodology was that it would allow us to study the adoption of TagTrainer in a clinical setting, while at the same time it would allow us to perform adjustments and modifications to better fit TagTrainer into a clinical context.

The three latter deployments adopted a case study approach, see (Yin, 2003). Though "bug" fixes and minor improvements were still performed by the researcher during these studies, significant modifications of or extensions to the system were no longer undertaken, and the participants in these studies were no longer actively participating in the development of TagTrainer or in the setting of research goals, assuming the role of a test-user rather than a co-designer or co-investigator.

The case studies were performed at three different clinics in The Netherlands and Belgium. These clinics provide physical rehabilitation to patients with stroke, spinal cord injury and multiple sclerosis. In total, 24 therapists (20 female, 4 male)

participated in the studies, and both physiotherapists and occupational therapists were involved. The duration of the studies ranged from 3 weeks for the first case study, to 8 weeks for the third case study. Though from a researcher's perspective longer field studies are preferable, the study length was capped by the clinic whose business model only includes compensation for time spent with patients rather than participating in studies or creating content.

During all case studies we chose to apply a staged deployment of TagTrainer. Given that TagTrainer was new to the participants, we first introduced TagTrainer only as a technology-supported solution for providing rehabilitation training. At a later stage during the case studies, we explained to the participants the possibility to add, modify or expand upon exercises already available from the start.

The action research approach adopted in the first study helped us to quickly develop TagTrainer into a technology fit for use in a practical setting. However, the continuous presence of the first author and his active engagement with professionals on site has probably caused a compliance bias. Due to their continuous involvement in the development of TagTrainer, therapists were triggered to work with the system. Our suspicions towards this bias are strengthened by the fact that the number of EUD activities in the latter studies (where the researcher was less frequently present) was significantly lower than that of the first study. The studies and their findings are described elsewhere (Tetteroo et al., 2014, 2015), so they will not be repeated here. Rather we aim to reflect on methodological choices and limitations of the approach chosen.

5.3 Measures

In all TagTrainer deployment studies several measurements have been taken. To measure whether therapists considered TagTrainer a viable technology for arm-hand training (SC1), we administered both the UTAUT questionnaire based on the unified theory of technology acceptance by Venkatesh (2003) and the CEQ questionnaire, which measures the therapists' perception of TagTrainer as a technology suitable for arm-hand rehabilitation.

The therapists' use of TagTrainer in daily arm-hand therapy (SC2) was measured by logging all instances where TagTrainer was used, and by observing therapists during usage.

EUD activities performed with TagTrainer (SC3 and SC4) were also captured through automated logging, as all instances of exercise modification and creation were stored by the system. Additionally, a self-efficacy questionnaire constructed according to the guidelines by Bandura (2006) allowed us to capture therapists' self-confidence in performing EUD tasks, regardless of their actual performance that we captured through logging.

Finally, semi-structured interviews were used to enrich the quantitative data that was collected. They allowed us to reveal the causes of some quantitative findings and helped us to better interpret the data.

5.4 Reflection on the Case Studies

Through four case studies, we have captured large amounts of data on the deployment of TagTrainer in rehabilitation clinics. The question we consider here is, whether the methods and measures that were chosen for our evaluations have resulted in data that helps us to determine whether the implementation of TagTrainer has been successful.

Since a relative wealth of data (see Tetteroo et al., 2015) was available for measuring SC1, one would expect that it was easy to determine whether or not therapists accepted TagTrainer as a technology for rehabilitation therapy. However, the flipside of having many data sources is that these sources might support conflicting conclusions. Indeed, results from the UTAUT and CEQ questionnaires often showed a relatively favorable result for the acceptance of TagTrainer, but interview and observation data revealed a more nuanced picture. The overall picture emerging on the acceptance of TagTrainer is one of *yes, but …*: Yes, therapists do accept TagTrainer as a technology for physical rehabilitation, given that certain boundary conditions (e.g., organizational support, technical support) are met. The important question now is whether the measures used were appropriate for measuring SC1.

As far as we know, our studies are the first in the domain of EUD where the UTAUT questionnaire was applied to measure the acceptance of the deployed solution over time. The questionnaire provided us both with new insights, and data that confirmed findings obtained from other sources (e.g., interviews). Interestingly, where theoretically the UTAUT model is supposed to carry predictive value about the use of technology, in our cases it was more useful in confirming and triangulating findings obtained from other data sources. For example, though in our studies the initial results from the UTAUT questionnaire predicted fairly good levels of acceptance (and thus technology use), the use of TagTrainer declined over the duration of our studies. Eventually, at the end of the studies, the results from the UTAUT questionnaire would confirm this development. One might be inclined to question the predictive validity of survey data. Nonetheless, we think these measures used are appropriate, and the mixed results from the different data sources show the importance of longitudinal quantitative data, which can indicate a general inclination towards a particular outcome, and qualitative data, which can provide nuance and depth to this inclination.

The outcome of SC2 was mainly measured by analyzing log files that were automatically generated by TagTrainer which helped pinpoint exactly which participants were more or less actively engaged in using TagTrainer. This allowed us to query participants during interviews on their use behavior. In this respect, it was also helpful that interviews were scheduled regularly, such that changes in usage behavior over time could be tracked and explained. Again, observations and interviews provided depth to the quantitative data, explaining not only *who* was using TagTrainer, and *when*, but also *how* and *why*.

Though a previous study (Hochstenbach-Waelen et al., 2012) had already shown that, in principle, rehabilitation experts (there students) without software

expertise are able to act as creators of therapy exercises for TagTrainer, we were interested whether this finding would also hold amongst professionals in the context of a rehabilitation clinic (SC3). In this regard, especially the self-efficacy questionnaire provided useful information. Increasing self-efficacy scores on EUD related tasks aligned with actual EUD performance that was recorded in the TagTrainer log files. Once more, interviews and observations provided us with additional insights, for example as to why particular therapists seemed more (or less) skilled in EUD related tasks.

Ultimately, the question we wanted to answer through our case studies is whether therapists would adopt EUD practices as part of their daily work (SC4). Although in principle the logs combined with the data from interviews and observations provided us with the possibility to answer this question, the analysis and interpretation of this data led us to the conclusion that the success criterion may not have been well chosen in the first place.

The difficulty in measuring whether therapists adopt EUD practices *as part of their daily work* is that it is hard to define what this qualifier actually means. Taken literally, it would require therapists to perform EUD activities every single day. By the nature of the rehabilitation profession and process there are bounds to what role TagTrainer can play in therapists' daily work, so this would be an unreasonable expectation. Rather, adoption in daily work should be interpreted more broadly, meaning that therapists have embraced EUD activities as an integral part of working with TagTrainer. When EUD activities take not place on a daily basis, any evaluation on the adoption of EUD practices in this context needs to be longitudinal, before a reliable and truthful picture of therapists' EUD practices can be formed.

An additional complication to the SC4 definition is that one would expect the amount of EUD activities to decrease over time, as the set of exercises grows and the need to create even more exercises declines. So even if there would be a value for, or an understanding of therapists' engagement in EUD activities, such a value or understanding would be specific to a particular moment in time, and rather meaningless on its own.

Finally, taking a *cultures of participation*-view (Fischer, 2011) where EUD is situated in a socio-technical setting that involves multiple actors practicing various degrees of EUD, what exactly do the collected data tell us about the EUD activities taking place within the TagTrainer community in its entirety? How meaningful is it to consider the EUD activities of individual therapists, if these activities are entwined with those of other members of the community?

Concluding, in our studies we were successful in evaluating the three success conditions (SC1–3) that we regarded as instrumental for the adoption of EUD practices. Our decision to delay the introduction of EUD to the participants enabled us to record findings that may otherwise have gone unnoticed, such as the decline in TagTrainer usage after participants had been introduced to EUD. We were able to identify that it was not TagTrainer per se, but rather the organizational requirements that EUD put on our participants which hindered its usage in the later stages of our studies.

Though we were able to successfully evaluate the first three success-conditions, we were unable to get an unambiguous result regarding SC4 ("Therapists perform EUD activities as part of their daily work."). Our inability to do so is not caused by a wrongfully chosen evaluation strategy, but by a lack of clarity regarding what might constitute a successful benchmark for the adoption of EUD technology. To answer to this rather fundamental question, in the next section we present a structured literature survey on the evaluation goals, methods, measures that previous deployment studies of EUD environments have reported.

6 A Structured Literature Survey

A structured literature survey was conducted by querying the online digital libraries of ACM (dl.acm.org) and Scopus (www.scopus.com). Together these libraries include indexes of the most relevant conference proceedings and journal publications on EUD. In addition to the database searches, the proceedings of all editions of the International Symposium on End User Development (IS-EUD, (Costabile, Mussio, Parasiliti Provenza, & Piccinno, 2008; Dittrich, Burnett, Morch, & Redmiles, 2013; Pipek, Rosson, & Wulf, 2009); were manually analyzed for articles missed by the dataset search but matching the criteria of this survey. Finally, two relevant articles that had not been captured by the search were added manually. Fig. 4 provides an overview of the survey process.

6.1 Inclusion/exclusion Criteria

Articles resulting from the search were included if they were written in English, published between 1991 and June 2015 when the survey was carried out, accessible to the authors and describing actual in-the-field deployment of EUD systems. We excluded papers published before 1991, in non-English languages or describing lab-studies, usability evaluations, retrospective analyses of cases, theories, methods, etc.

6.2 Search Keywords

The databases used were queried for articles containing at least one of the following terms as keywords *end user programming, end user development, EUD*, and *meta design*. Meta-design is a *conceptual framework aimed at defining and creating social and technical infrastructures in which new forms of collaborative design can take place* (Fischer, 2007) – the term was included since work in this area is closely related to end user development.

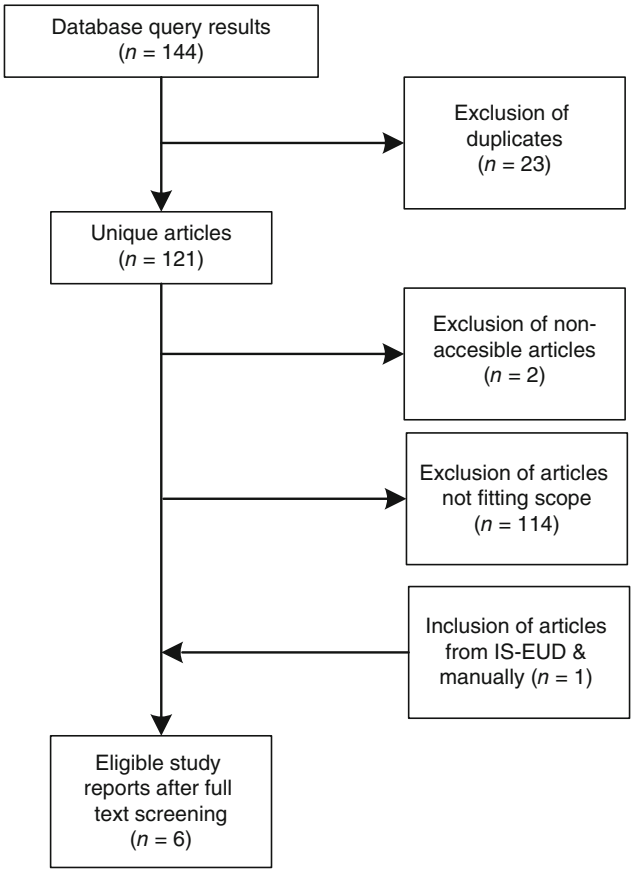

Fig. 4 Flowchart of the structured literature survey

In addition to matching one of the keywords, to filter for studies describing actual in-the-field deployments of EUD systems, articles had to include at least one of the following terms as a keyword, or in their abstracts: *deployment, case study, field study, ethnography, practices, ethnographic methods*.

6.3 Results

The combined queries returned 144 results (ACM: 31, Scopus: 113). After correcting for duplicates, 121 unique results remained. The abstracts of all candidate articles were read and compared against the inclusion/exclusion criteria. Where abstracts were found to provide insufficient information for the inclusion/exclusion decision, the full article was read. The proceedings of IS-EUD were scanned for

Table 1 Articles selected for analysis in this survey

Title	Authors	Source	Year
"Let's see your search-tool!" – On the collaborative use of tailored artifacts	Wulf, V.	Proc. GROUP'99	1999
Investigating success factors for hypermedia development tools	Bolchini, D., Garzotto, F., Paolini, P.	Proc. Hypertext'08	2008
Design, adoption, and assessment of a socio-technical environment supporting independence for persons with cognitive disabilities	Carmien, S.P., Fischer, G.	Proc. CHI'08	2008
Software development cultures and cooperation problems: a field study of the early stages of development of software for a scientific community	Segal, J.	CSCW (journal)	2009
Study of using the meta-model based meta-design paradigm for developing and maintaining web applications	De Silva, B., Ginige, A.	Proc. UNISCON'09	2009
End-user development of enterprise widgets	Spahn, M., Wulf, V.	End-user development (book)	2009
Enabling users of enterprise systems to mashup resources and develop widgets	Spahn, M. et al.		

relevant articles not covered by the database search, but no additional eligible articles were found. Most of the excluded articles, although being EUD related, did not present deployment studies. Instead, they often concerned formative (lab) evaluations and retrospective analyses of cases where EUD would already be in place. Finally, one article that was missed by the keyword search was added manually.

Six articles in total were found to be eligible for answering our research questions (see Table 1). A summarized version of these articles is given below.

Bolchini, Garzotto, and Paolini (2008) have attempted to identify *the key factors that contribute to the success of a hypermedia development tool*. They acknowledge that success factors exist on various levels, but focus on those that can be observed and are directly related to the "product." In their case study, the authors have studied the adoption of the *1001Stories* tool. This tool allows for web-based hypermedia development and has been deployed amongst Italian primary-school and high-school classes. The authors have studied the adoption of the tool through two sub-studies: the first study involved primary school children and teachers, focusing on tool and process simplicity. In this study, task-based observational user testing was used to evaluate the ease of use of the tool. Contextual inquiry was used to study the development process. Finally, a questionnaire was submitted to the participating children at the end of the study, to investigate their overall satisfaction. Their second study was *mainly devoted to investigate satisfaction, prospective adoption, and success factors on a larger statistical base* (Bolchini et al., 2008). In this study, the authors queried participating teachers through an

online questionnaire. Finally, they measure the success of their tool and its deployment by assessing the following variables: *appreciation, educational benefits, prospective adoption, tool simplicity*, and *process efficiency*. Although the results of their questionnaires are generally favorable, the authors draw no conclusions about the success of their deployments.

Carmien and Fischer (2008) have studied the use of EUD practices to enhance the independence of cognitively disabled persons. They present the Memory Aiding Prompting System (MAPS), which provides caregivers the opportunity to *create scripts that can be used by people with cognitive disabilities ("clients") to support them in carrying out tasks that they would not be able to achieve by themselves* (Carmien & Fischer, 2008).The system consists of the *MAPS-DE*, an EUD environment allowing caregivers to create and share scripts, and *MAPS-PR*, a client interface for the created scripts. They present a field study with 6 participants (3 caregiver-client dyads) who have been provided with the MAPS system. The authors used several ethnographic methods to collect data, in particular participant observation and semi-structured interviews. Their goals were to (1) learn about the client's and caregiver's world and their interactions, (2) observe and analyze how tasks and learning of tasks were currently conducted, (3) understand and explicate the process of creating and updating scripts, (4) comprehend and analyze the process of using the scripts with a real task, and (5) gain an understanding of the role of meta-design in the dynamics of MAPS adoption and use. During their field study, the authors collected audio recordings, field notes and secondary artifacts. The authors are not explicit about the success of their deployment.

Segal (2009) describes the development process of scientific software as a combined effort of several "professional end user developers" (i.e., scientists who can program). The article focuses on cooperation problems arising from the professional end user development culture associated with the development project. While her article strictly speaking does not meet the inclusion criteria for this survey (since EUD is already common practice amongst these scientists), the collaborative nature of the software project described presents new challenges to the users involved and hence can be regarded as the "deployment" of a new form of EUD. The process that Segal describes is one in which professional end user developers, as part of an organized project with multiple stakeholders that belong to several organizations, develop a laboratory information management system for use in biology research. These stakeholders are: a management board, the development team (highly heterogeneous, nine members, five locations), collaborators from research groups that have the potential to share resources with the development team, and the users. The goal of the field study that Segal conducted is to illustrate how cultural influences impact cooperation in a professional end user development project. For this, she collected data through ten observations, ten interviews, twelve phone calls, numerous emails, and by consulting project documentation.

De Silva and Ginige (2009) describe the deployment of end-user extensible websites for three small and medium-size enterprises (SMEs). They draw inspiration from Fischer's meta-design theory (Fischer & Scharff, 2000) and provide the

SMEs with a first version of a website as a seed that can later be extended by the SMEs. The goal of their study was to investigate how the tool making industry in Australia, specifically the SMEs, could benefit from end-user extensible websites. The data collected is very sparse: only data on the toolmakers' perceived ability to execute maintenance tasks and a log of the maintenance activities performed by the SMEs have been collected.

Spahn and Wulf (2009) describe the deployment of their Widget Composition Platform (WCP) – a platform that allows business users to create custom widgets, tailored to their personal information needs – to three mid-sized German companies. The goal of their evaluation was to investigate questions such as: are their end users able to create widgets using WCP? Do the widgets address practical problems in real work contexts? Are advanced end users able to create and wrap enterprise resources as new services to extend the available building blocks for widget creation? What types of end users exist with regard to widget usage and development and how do they collaborate? Spahn and Wulf used interviews, observations, and questionnaires focusing on EUD-related tasks to find answers to these questions.

Volker Wulf (1999) describes the development of a component-based search tool as an extension to an existing database system. They deployed the artifact in a German government organization and studied a small group of users in depth. The goals of their evaluation were to research (1) to which extend users without programming skills would be able to tailor the search tool, (2) which division of labor would emerge between the end users and the local experts, (3) whether end users would be able to understand the components, compound components and search tool alternatives provided to them by programmers and local experts, and (4) how to support the exchange of tailored artifacts between end users and local experts. Their deployment approach was staged: first they presented the prototype to a group of participants in a workshop and performed a formative usability evaluation, after which the tool was adopted by the designers. Then, they deployed the revised tool for a period of two weeks, starting with a workshop where the tool was presented to the participants and training was provided. During this study, participants were observed in their tailoring process and the emerging problems. Finally, semi-structured interviews were held with the participants, and the tailored artifacts were copied for analysis. During all workshops and the field study, written notes were taken and transcribed directly after each session.

6.4 Evaluating EUD Deployments

As has been demonstrated by the survey, research in which researchers attempt to *create* an environment that facilitates EUD, by implementing appropriate methods, techniques and tools, as well as by shaping facilitating conditions on a psychological and social level is far less common than retrospective ethnographic studies.

Remarkably few attempts of EUD deployments were found, and the variation between them (amongst others in terms of domain, approach, scope, authors and findings) leads to believe that this aspect of EUD is currently underexplored, and the state of the art is limited to ad-hoc attempts rather than structured and planned approaches. One possible explanation for the scarcity of deployment studies is that EUD needs not necessarily be introduced as part of an orchestrated effort - involving planned investments of time and effort by end users and commitment by some form of management. Rather, suitable environments may have evolved gradually to accommodate for EUD. Still, the articles discussed in this survey show there is a need for an orchestrated deployment of EUD technology in several cases. For example, non-information workers (such as the caregivers in Carmien & Fischer, 2008) might not be aware of the possibilities that EUD environments can provide them with to address personally relevant problems. The need for organizational support for encouraging EUD practices has been also argued on the basis of the surveys by Mehandjiev, Sutcliffe, and Lee (2006) and Kierkegaard and Markopoulos (2011).

Table 2 provides an overview of the aims of the surveyed articles, and the evaluation methods that were used in these articles. Most articles are not very clear on what they expect to find when starting their deployment. Nevertheless, by "reading through the lines" it is clear that the studies had either one or both of the following goals: to evaluate a tool for EUD, or to better understand the principles that underlie EUD (see also the "research purpose" qualifications in Kjeldskov & Graham, 2003).The methods that these studies used varied, but the use of qualitative methods such as observations and interviews is common, especially amongst

Table 2 Aims and research methods used in the surveyed studies

Article	Study aims	Methods used	New tech?
Bolchini et al. (2008)	Tool evaluation	Task-based observational user testing Contextual inquiry Questionnaires	Yes
Carmien and Fischer (2008)	Understanding EUD principles	Observations Interviews	Yes
Segal (2009)	Understanding EUD principles	Observations Interviews Documentation analysis	No
De Silva & Ginige (2009)	Tool evaluation understanding EUD principles	Usage logging Questionnaires	Yes
Spahn and Wulf (2009)	Tool evaluation understanding EUD principles	Observations Interviews Questionnaires	Yes
Wulf (1999)	Tool evaluation understanding EUD principles	Observations Interviews	No

studies that focus on creating a better understanding of EUD principles. Where questionnaires were used, their aim varied from measuring tool appreciation and usability, to measuring the participants' general opinions on the use and usefulness of EUD in their domain. Interestingly, only two studies concerned the deployment of EUD as an addition to an already existing system (e.g., an extension or plugin), while the others introduced a new technology entirely. There seems to be no correlation between the deployment type and the methods used.

7 Discussion

In order to find an answer to the question *how should deployments of EUD systems be evaluated?* we have reflected on the deployment studies of TagTrainer. Further, we presented a structured literature survey on other deployment studies of EUD systems. We discuss the results of the survey and our reflections on a number of questions that are related to the evaluation of EUD deployments.

7.1 How to Best Capture Evidence of the Success of an EUD Deployment?

The studies discussed in this chapter are characterized by a great diversity in their approaches, methods, goals and results. Therefore, it is not easy to draw a conclusion about what are suitable methods for capturing evidence of successful EUD deployments. On the other hand, if we take a step back and look at how the different studies have interpreted the evaluation task, we can make some interesting observations.

Earlier in this chapter, we stated that successful EUD deployments *maximize the value of EUD within their context, thus increasing the likelihood that EUD practices contribute to the achievement of an end user's goals*. Success, by this definition, is thus strongly related to the goals that a particular end user of the EUD technology has in a particular context. As has been shown before, these goals vary greatly between different cases, and range from "allowing patients to live more independently" (therapists, Carmien & Fischer, 2008) to "developing web-based hypermedia to pass a course" (high-school children, Bolchini et al., 2008) and "running a profitable business" (De Silva & Ginige, 2009).

As much as the goals of the end users in the contexts of the surveyed studies differ, the role that EUD plays in these contexts differs as well. For example, the relative importance of an up-to-date website for an SME in (De Silva & Ginige, 2009) might be less than the importance of a working memory prompting system for the cognitive disabled in (Carmien & Fischer, 2008). SMEs will probably primarily be focused on producing and selling goods and services. Maintaining an

up-to-date online presence can help to increase sales but is usually not amongst the core activities of such companies.

It can be argued that the evaluation approach, and the methods and measures used should be adapted to the role that EUD is expected to play for the end users. For example, adopting an action research approach where researchers collaborate with end users in the deployment and evaluation of an EUD environment over an extended period of time, might not be the right choice if the prospective adoption of EUD practices will remain low and infrequent (e.g., De Silva & Ginige, 2009). However, it is not always trivial to estimate the importance of EUD for the context in which it is being implemented. For example, in our own studies we expected the importance of EUD in the use of TagTrainer to be greater than it turned out to be. Though therapists indicated that providing patient-centered training content is an important consideration to them, in practice they often settled for readily available exercises (rather than ones tailored for a specific patient) from the library of exercises that we made available to them.

Since it is difficult to predict in advance what the rate of EUD adoption will be, it is sensible to adopt a staged approach in the evaluation of EUD systems. Rogers (2010) famously describes a five-stage model on the diffusion of innovations that provides us with sufficient theoretical guidance to propose suitable methods of evaluation for the different stages of EUD deployments. The five stages of his model are (adapted from Rogers, 2010, p. 169):

1. *Knowledge*, occurs when an individual is exposed to an innovation's existence and gains an understanding of how it functions.
2. *Persuasion*, occurs when an individual engages in activities that lead to a choice to adopt or reject the innovation.
3. *Decision*, takes place when an individual engages in activities that lead to a choice to adopt or reject the innovation.
4. *Implementation*, occurs when an individual puts a new idea into use.
5. *Confirmation*, takes place when an individual seeks reinforcement of an innovation-decision already made, but he or she may reverse this previous decision if exposed to conflicting messages about the innovation.

Importantly, the five stages of Rogers' model show us that at different moments during a deployment process, different factors become important for the end user in relation to the adoption of the technology that is being deployed. If we now turn the question with which we started this section - on the best way to capture evidence of successful EUD deployments - we can use Rogers' model to define for each stage what evidence could or should be collected in support of any statement on the success of an EUD deployment:

1. *Knowledge:* evaluate the end users' understanding of the EUD system being deployed e.g., its usability and functionality.
2. *Persuasion:* evaluate the end users' attitude towards the system, for example by using the UTAUT model (Venkatesh et al., 2003) or self-efficacy regarding EUD (Bandura, 2006).

3. *Decision:* evaluate whether, in the opinion of the end users, adopting EUD practices will lead to a positive outcome of the cost/benefit tradeoff related to the adoption of EUD practices (*relative advantage*, in Rogers' theory (2010)). Blackwell's Attention Investment model (Blackwell, 2002; Blackwell & Burnett, 2002) could be used to gauge this specifically for EUD.
4. *Implementation:* evaluate the EUD practices that end users develop, the role that EUD starts playing in the context in which it is deployed, and most importantly, the extent to which the EUD practices help the end user to achieve his or her goals.
5. *Confirmation:* evaluate whether the decision to (not) engage in EUD has sustained after a period of time, and if not, what has caused the end user to reverse his or her initial decision.

The advantage of designing evaluations in a staged approach, as outlined above, is that it is then possible to relate different evaluation studies to each other and we can pinpoint more precisely areas of improvements. It also protects us from setting up large, time-consuming and expensive evaluations that study EUD practices, if in an earlier stage we can detect threads for a successful deployment (e.g., usability flaws, acceptance issues). The first two stages can, in principle, even be evaluated in a laboratory setting. Finally, the structured and staged evaluation approach allows us to better compare different cases of EUD deployments. It provides us with terminology to discuss these cases in a context independent manner, and allows us to draw generalizations over several cases of EUD deployments, even if these cases themselves are context specific.

7.2 The Role of New Technology in EUD Deployments

Earlier in this chapter we limited our discussion to cases where EUD comes on top of an existing host technology (e.g., as a plugin to existing software), or is deployed simultaneously with another, new host technology (e.g., the case of TagTrainer). As we have experienced ourselves, evaluating the impact of introducing EUD in an organization while simultaneously introducing a new host technology can lead to difficulties. The impact of the introduction of the new technology might overshadow the impact of introducing EUD, thereby obscuring the effects that the introduction of EUD might have had. Further, the actual adoption of EUD practices might, in such a context, be hampered by the fact that the host technology introduced does not align with the existing practices within that context (i.e., what Rogers calls *compatibility with previously introduced ideas* Rogers, 2010).

In our own studies, we have tried to counter this bias caused by the introduction of a new host technology, by adopting a staged introduction of TagTrainer. First, the system was introduced as a technology for physical rehabilitation, without focusing on the possibility for therapists to modify or create exercises. Only later were the participating therapists instructed on the EUD possibilities that the system offered them.

The rationale was that therapists could first get used to working with TagTrainer as a new technology for physical arm-hand rehabilitation. Then, once they had adopted the technology for this purpose, they would be introduced to EUD. We assumed that through this approach, the novelty of the technology would no longer interfere with the introduction of EUD. Still, many of the issues that were raised by therapists in the later stages of our deployment studies were related to the system in general and not specifically to the possibility to modify or create exercises. Some of these issues would have such a negative impact on their perception of the system that they would abandon it completely, limiting our ability to study EUD adoption and practice.

The studies reviewed in the survey however reveal different results. Four of these articles report a simultaneous introduction of a new technology, as well as EUD, as part of their study. Still, they do not report on issues in the adoption of EUD practices arising from this simultaneous introduction, nor do they report on the occurrence of a results bias. It is possible that in some cases, such as (Spahn & Wulf, 2009), the technology that was introduced was compatible enough (see Rogers, 2010) to the technology their participants had been working with previously, that it did not cause any significant problems.

Earlier, we asked ourselves whether it is sensible to separate the evaluation of the EUD-part of an environment from the other parts. Unfortunately, this question cannot conclusively be answered from the results of our survey. The fact that none of the surveyed studies report on issues arising from the simultaneous evaluation of the technology and EUD practices does not mean that such issues do not occur. Moreover, since in our own studies we *did* encounter these issues, we believe that the answer to this question depends on the context in which the EUD system is being deployed.

8 Conclusion

Evaluating EUD deployments is far from trivial, since it is difficult to define the precise subject of evaluation and to determine which approach and which methods are suitable for such an evaluation. In this chapter, we have explored these questions by first defining what makes EUD deployments different from regular software deployments. Then, we discussed the evaluation of TagTrainer, after which we presented a literature survey on EUD deployment studies. One lesson we can draw from this survey is that evaluations of EUD deployments so far do not share a common framework and form a rather fragmented picture.

From this survey and from our own experiences, we discussed suitable ways to evaluate EUD deployment, and more specifically:

1. A staged evaluation approach, evaluating sequentially the end users' knowledge about, and acceptance of the deployed system, the tradeoffs that the end users face in considering to engage in EUD activities, the EUD practices and activities that end users develop, and finally whether these practices sustain after a longer period of time.

2. A staged implementation of the host technology (the technology to which support for EUD is added) and the EUD technology. Where the host technology is deployed next to EUD technology, the deployment of EUD technology should be postponed until the host technology has been accepted and incorporated by the end users.

We believe that if future EUD deployment studies take these suggestions into account, we can more effectively compare different studies and draw generalizable conclusions from their data.

References

Bandura, A. (2006). Guide for constructing self-efficacy scales. In Urdan, T., & Pajares, F. Eds. *Self-efficacy beliefs of adolescents*. IAP, 2006.

Blackwell, A., & Burnett, M. (2002). Applying attention investment to end-user programming. In *Proc. HCC 2002* (pp. 28–30). IEEE.

Blackwell, A.F. (2002). First steps in programming: a rationale for attention investment models. In *Proc. HCC 2002* (pp. 2–10). IEEE.

Bolchini, D., Garzotto, F., Paolini, P. (2008). Investigating success factors for hypermedia development tools. In *Proc. HT 2008* (pp. 187–192). New York: ACM.

Carmien, S. P., & Fischer, G. (2008). Design, adoption, and assessment of a socio-technical environment supporting independence for persons with cognitive disabilities. In *Proc. CHI 2008* (pp. 597–606). New York: ACM.

Costabile, M.F., Mussio, P., Parasiliti Provenza, L., Piccinno, A. (2008). End users as unwitting software developers. In *Proc. 4th int. workshop end-user softw. eng* (pp. 6–10). ACM.

De Silva, B., & Ginige, A. (2009). Study of using the meta-model based meta-design paradigm for developing and maintaining web applications. In *Int. united inf. syst. conf* (pp. 304–314). Springer.

Dittrich, Y., Burnett, M., Morch, A., Redmiles, D. (2013). *End-user development: 4th international symposium, IS-EUD 2013, Copenhagen, Denmark, June 10–13, 2013, Proceedings*. Berlin Heidelberg: Springer.

Fischer, G. (2007). Meta-design: expanding boundaries and redistributing control in design. In C. Baranauskas, P. Palanque, J. Abascal, S. D. J. Barbosa (Eds.). *Hum.-comput. interact. – INTERACT 2007* (pp. 193–206). Berlin Heidelberg: Springer.

Fischer, G. (2011). Understanding, fostering, and supporting cultures of participation. *Interactions*, *18*, 42–53.

Fischer, G., & Scharff, E. (2000). Meta-design: design for designers. In *Proc. DIS 2000* (pp. 396–405). New York: ACM.

Herr, K., & Anderson, G. L. (2014). *The action research dissertation: a guide for students and faculty*. Thousand Oaks: SAGE Publications.

Hochstenbach-Waelen, A., Timmermans, A., Seelen, H., Tetteroo, D., Markopoulos P. (2012). Tag-exercise creator: towards end-user development for tangible interaction in rehabilitation training. In *Proc. EICS 2012* (pp. 293–298). ACM.

Kierkegaard, P., & Markopoulos, P. (2011). From top to bottom: end user development, motivation, creativity and organisational support. In *Int. symp. end user dev* (pp. 307–312). Springer.

Kjeldskov, J., & Graham, C. (2003). A review of mobile HCI research methods. In L. Chittaro (ed). *Hum.-comput. interact. mob. devices serv.* (pp. 317–335). Berlin Heidelberg: Springer.

Lanfermann, G., Te Vrugt, J., Timmermans, A., Bongers, E., Lamber, N., Van Acht, V. (2007). Philips stroke rehabilitation exerciser. In *Tech. aids rehabil.-TAR 2007* January 25–26 2007.

Li, Y., Fontijn, W., Markopoulos, P. (2008). A tangible tabletop game supporting therapy of children with cerebral palsy. In P. Markopoulos, B. Ruyter, W. de IJsselsteijn, D. Rowland (Eds.). *Fun games* (pp. 182–193). Berlin Heidelberg: Springer.

Mehandjiev, N., Sutcliffe, A., Lee, D. (2006). Organizational view of end-user development. In H. Lieberman, F. Paternò, V. Wulf (Eds.). *End user dev* (pp. 371–399). Netherlands: Springer.

Namoun, A., Wajid, U., Mehandjiev, N., Owrak, A. (2012). User-centered design of a visual data mapping tool. In *Proc. AVI 2012* (pp. 473–480). New York: ACM.

Pipek, V., Rosson, M.-B., Wulf, V. (2009). *End-user development: 2nd international symposium, IS-EUD 2009, Siegen, Germany, March 2–4, 2009, Proceedings*. Berlin-Heidelberg: Springer.

Rogers, E.M. (2010). *Diffusion of innovations*, 4th Edition. Simon and Schuster.

Ruthruff, J. R., Prabhakararao, S., Reichwein, J., Cook, C., Creswick, E., Burnett, M. (2005). Interactive, visual fault localization support for enduser programmers. *Journal of Visual Languages and Computing*, *16*, 3–40. doi:10.1016/j.jvlc.2004.07.001.

Segal, J. (2009). Software development cultures and cooperation problems: a field study of the early stages of development of software for a scientific community. *Computer Supported Cooperative Work (CSCW)*, *18*, 581 doi:10.1007/s10606-009-9096-9.

Spahn, M., & Wulf, V. (2009). End-user development of enterprise widgets. In V. Pipek, M. B. Rosson, B. Ruyter, V. de, Wulf (Eds.). *End-user dev* (pp. 106–125). Berlin Heidelberg: Springer.

Tetteroo, D., & Markopoulos, P. (2015). A review of research methods in end user development. In Díaz P., Pipek V., Ardito C., Jensen C., Aedo I., Boden A. (Eds.), *End-user dev* (pp. 58–75). Springer International Publishing.

Tetteroo, D., Timmermans, A. A., Seelen, H. A., Markopoulos, P. (2014). TagTrainer: supporting exercise variability and tailoring in technology supported upper limb training. *Journal of NeuroEngineering and Rehabilitation*, *11*, 140 doi:10.1186/1743-0003-11-140.

Tetteroo, D., Vreugdenhil, P., Grisel, I., Michielsen, M., Kuppens, E., Vanmulken, D., et al. (2015). Lessons learnt from deploying an end-user development platform for physical rehabilitation. In *Proc. CHI 2015* (pp. 4133–4142). New York: ACM.

Tsandilas, T., Letondal, C., Mackay, W. E. (2009). Musink: composing music through augmented drawing. In *Proc. CHI 2009* (pp. 819–828). New York: ACM.

Venkatesh, V., Morris, M. G., Davis, G. B., Davis, F. D. (2003). User acceptance of information technology: toward a unified view. *MIS Q*, *27*, 425–478.

Wong, J., & Hong, J. I. (2007). Making mashups with marmite: towards end-user programming for the web. In *Proc. CHI 2007* (pp. 1435–1444). New York: ACM.

Wulf, V. (1999). "Let's see your search-tool!"—Collaborative use of tailored artifacts in groupware. In *Proc. GROUP 1999* (pp. 50–59). New York: ACM.

Yin, R. K. (2003). *Case study research: design and methods*. Thousand Oaks: SAGE Publications.