Available online at www.sciencedirect.com

SciVerse ScienceDirect

Int. J. Human-Computer Studies ▮ (▮▮▮▮) ▮▮▮–▮▮▮

International Journal of Human-Computer Studies

www.elsevier.com/locate/ijhcs

Reconstructing experiences with iScale

Evangelos Karapanos[a,*], Jean-Bernard Martens[b], Marc Hassenzahl[c]

[a]Madeira Interactive Technologies Institute, Portugal
[b]Eindhoven University of Technology, The Netherlands
[c]Folkwang University of Arts, Germany

Received 13 March 2011; received in revised form 2 March 2012; accepted 17 June 2012
Communicated by K. Hornbæk

Abstract

We present iScale, a survey tool for the retrospective elicitation of longitudinal user experience data. iScale aims to minimize retrospection bias and employs graphing to impose a process during the reconstruction of one's experiences. Two versions, the *constructive* and the *value-account* iScale, were motivated by two distinct theories on how people reconstruct emotional experiences from memory. These two versions were tested in two separate studies. Study 1 aimed at providing qualitative insight into the use of iScale and compared its performance to that of free-hand graphing. Study 2 compared the two versions of iScale to free recall, a control condition that does not impose structure on the reconstruction process. Overall, iScale resulted in an increase in the amount, the richness, and the test–retest consistency of recalled information as compared to free recall. These results provide support for the viability of retrospective techniques as a cost-effective alternative to longitudinal studies.

Keywords: User experience evaluation; Retrospective elicitation; Longitudinal methods

1. Introduction

Understanding the use and acceptance of interactive products beyond initial use has always been an interest of the human–computer interaction (HCI) community (Erickson, 1996; Prümper et al., 1992). However, two recent trends make the call for a more longitudinal view more urgent (Karapanos et al., 2009). First, legislation and competition within the consumer electronics industry have led to prolonged product warranties, resulting in an alarmingly increasing number of products being returned on the basis of failing to satisfy their users' "true" needs (Den Ouden et al., 2006). Second, products have become more embedded into services. Often, products are being sold for low prices or even given away for free and revenues stem mainly from the supported service and their prolonged use (Karapanos et al., 2009). Thus, the overall focus on product quality shifts from a focus on the classic phases of pre-purchase and purchase to a more longitudinal perspective, trying to better understand use and liking over time. This is a shift increasingly taken up by the HCI community (Gerken et al., 2007; Barendregt et al., 2006; Fenko et al., 2009; Karapanos et al., 2008; von Wilamowitz-Moellendorff et al., 2006; Courage et al., 2009; Vaughan et al., 2008; Kjeldskov et al., in press).

From a methodological perspective, one may distinguish three approaches to understanding the development of usage and experience over time (von Wilamowitz-Moellendorff et al., 2006): cross-sectional, pre-post/longitudinal, and retrospective reconstruction. Cross-sectional approaches are the most popular in the HCI domain (Prümper et al., 1992; Bednarik et al., 2005). Cross-sectional studies distinguish, for example, user groups with different levels of expertise, for instance, novice and expert users. Observed variation in experience or behavior is then attributed to expertise in the sense of a quasi-experimental variable. This approach is, however, limited as it is prone to confounding variables, such as failing to control for external variation and, more importantly, falsely attributing variation across the user groups to expertise. Prümper et al. (1992) already highlighted this problem, by showing

*Corresponding author. Fax: +351 291 721 006.
E-mail address: e.karapanos@gmail.com (E. Karapanos).

1071-5819/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.ijhcs.2012.06.004

that different definitions of novice and expert users lead to different results.

Beyond the cross-sectional, one may further distinguish pre-post and true longitudinal approaches. Pre-post designs study the same participants at two points in time. For instance, Kjeldskov et al. (in press) studied the same seven nurses, using a healthcare system, right after the system was introduced and 15 months later. Karapanos et al. (2008) explored how 10 individuals formed overall evaluative judgments of a novel pointing device, during the first week of use as well as after 4 weeks of using the product. While these approaches study the same participants over an extended period of time, they cannot tell much about the exact form of change, due to the limited number of only two observations. True longitudinal designs take more measurements and employ a number of statistical techniques to track change in general and to estimate the impact of particular events on change. Because of their laborious nature, however, they are only rarely used in practice and research.

Different granularities in longitudinal studies can be distinguished (see von Wilamowitz-Moellendorff et al., 2006): a micro-perspective (e.g. an hour), a meso-perspective (e.g. 5 weeks) and a macro-perspective, with a scope of years of use. Studies with a micro-perspective assess how users' experience changes through increased exposure over the course of a single session of use. For instance, Minge (2008) elicited judgments of perceived usability, innovativeness and the overall attractiveness of computer-based simulations of a digital audio player at three distinct points: (a) after participants had seen but not interacted with the product, (b) after 2 min of interaction and (c) after 15 min of interaction. An example of a study with a meso-perspective is Karapanos et al. (2009). They followed six individuals after the purchase of a product over the course of 5 weeks. One week before the purchase of the product, participants started reporting their expectations. After product purchase, participants were asked to narrate the three most influential experiences of each day. Studies with a macro-perspective are 'nearly non-existent' (von Wilamowitz-Moellendorff et al., 2006).

A third approach is the retrospective reconstruction of personally meaningful experiences from memory. Different variants of the Critical Incident Technique, popular in marketing and service management research (Edvardsson and Roos, 2001; Flanagan, 1954), ask participants to report critical incidents over periods of weeks, months or the complete time-span of the use of a product or service. In a survey study, Fenko et al. (2009) asked participants to recall their single most pleasant and unpleasant experience with different types of products and to assess the most important sensory modality (i.e. vision, audition, touch, smell and taste) at different points in time (i.e. when choosing the product in the shop, during the first week, after the first month, and after the first year of usage). von Wilamowitz-Moellendorff et al. (2006, 2007) proposed a structured interview technique called CORPUS (Change Oriented analysis of the Relation between Product and User) for the retrospective assessment of the dynamics in users' perceptions of different facets of perceived product quality. CORPUS starts by asking participants to assess the currently perceived quality of "their" product on a number of defined facets (usability, utility, beauty, stimulation, identification, and global evaluation). Subsequently, they are asked to "go back in time" and to compare their current perception and evaluation of the product to the moment right after purchasing the product. If change has occurred, participants are further prompted to indicate the direction and shape of change (e.g., accelerated improvement, steady deterioration). Finally, participants are asked to elaborate the reasons that induced the changes in the form of short narratives, so-called "change incidents". The obtained data can be used quantitatively by constructing graphs of change (see Fig. 1 for an example) and qualitatively by exploring the reasons people give for changes.

A common critique of methods relying on memory is the degree to which recalled experiences are biased or incomplete. In the context of perceived product quality, we argue that this is of minor importance. While a given reconstruction from memory should be truthful (i.e., reflect what the participant really thinks), it seems less important, whether the reconstructed timeline as well as the reasons given are true (i.e. reflect what actually happened) as long as the participant is convinced that what she is reporting actually happened. This is because, we are foremost interested in subjective reconstructions because those (and not "objective" data) will be communicated to others as well as guide the individual's future activities. In other words, it may not matter how good a product is objectively, it is the "subjective", the "experienced", which matters (Hassenzahl et al., 2006). See also Norman (2009). To give a further example: Redelmeier and Kahneman (1996) found retrospective assessment of the pain experienced during colonoscopy to be biased. People put an extra weight on the most painful moment and the end of the examination. This has interesting consequences. One can, for example, deliberately prolong the examination (something not approved by the patients), but make sure that

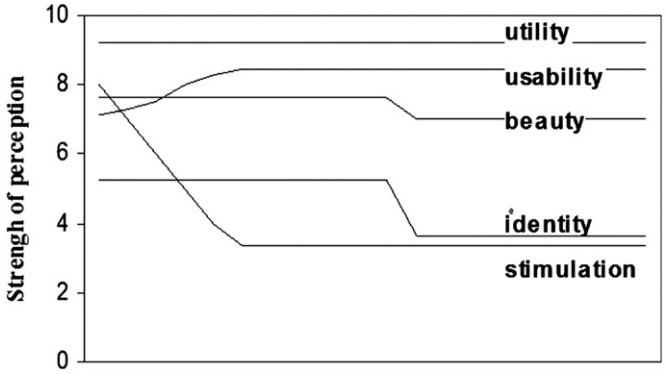

Fig. 1. Exemplary dynamics of different perceived quality dimension of mobile phones. Start and end points of each dimension are based on the mean elicited ratings of eight participants. Reprinted from von Wilamowitz-Moellendorff et al. (2006).

Please cite this article as: Karapanos, E., et al., Reconstructing experiences with iScale. International Journal of Human-Computer Studies (2012), htt p://dx.doi.org/10.1016/j.ijhcs.2012.06.004

these last, additional 2 min are not painful. The consequence is an overall assessment of the examination as less painful compared to people without the additional 2 min. While this is clearly a bias, people simply have no memory for all the moments they experience, but will remember their overall impression of the examination. The retrospective judgment is more real to them than what actually happened. While the validity of remembered experiences may not be crucial, their consistency across multiple recalls is. It seems at least desirable that participants would report their experiences consistently over multiple trials. If recall would be purely "random", the value of respective reports for design would be questionable. In other words, what we remember might be different from what we experienced; however, as long as these memories are consistent over multiple recalls, they provide valuable information.

In the area of critical incident research, interviewing techniques have been developed with the aim to assist participants in remembering more details of and contextual information around experienced critical incidents (Edvardsson and Roos, 2001). However, interviews in general, however, need substantial skills and resources. It, thus, seems desirable to create a self-reporting approach. Consequently, this paper presents iScale, a survey tool that was designed to increase participants' effectiveness in reconstructing their experiences with a product over time. iScale uses a graphical representation of change over time as a major support (i.e., time-line graphing). Other than previous approaches (von Wilamowitz-Moellendorff et al., 2006; Kujala et al., 2011), the employed procedure is more firmly grounded on theory, actually deriving variants of the procedure based on competing theoretical models of the retrospective reconstruction of episodes and experiences from memory. Graphing is assumed to support the reconstruction process through what Goldschmidt (1991) calls *interactive imagery* (i.e., "the simultaneous or almost simultaneous production of a display and the generation of an image that it triggers"). The idea of using graphing as an approach to introspecting on past emotional experiences can be traced back to Sonnemans and Frijda (1994).

We begin with laying out two different ways of obtaining retrospective reconstructions of experiences and their theoretical foundation. We then present the results of two studies. Study 1 acquired a qualitative understanding of the use of iScale in comparison to its analog equivalent (i.e. free-hand graphing). Study 2 assessed how iScale compares to an experience reporting tool without graphic support, which, can be seen as a control condition to assess the impact of iScale on participants' effectiveness and test–retest consistency in reconstructing experiences.

2. Reconstructing experiences from memory

Memory was for long understood as a faithful account of past events, which can be reproduced, when trying to remember details of the past. This idea was first challenged by Bartlett (1932). He described remembering as an act of reconstruction, which never produces the exact past event, but instead alters representation of the event with every attempt to recall. Bartlett (1932) asked participants to recall an unfamiliar story told 20 h earlier. The recalled stories differed from the original in detail, order and importance of single events. In addition, participants augmented their memories by applying rationalizations and interpretations to the original story. Each further reconstruction distorted the stories even further.

At the heart of the notion of reconstruction lies the distinction between episodic and semantic memory (Tulving, 2002). While episodic memory "is specific to a particular event from the past, semantic memory is not tied to any particular event but rather consists of certain generalizations (i.e. beliefs) that are rarely updated" (Robinson and Clore, 2002, p. 935). These two types of memory serve different needs, such as learning new information quickly – a capacity of episodic memory – or developing relatively stable expectations about the world—a capacity of semantic memory (Robinson and Clore, 2002). Reconstruction happens through the retrieval of cues stored in episodic memory. In the absence of such cues, beliefs stored in semantic memory may be used to reconstruct the past. This results in distortions, where details that actually happened are replaced by generalizations based on what we know about the world. Thus, the accuracy of remembered events depends on the degree to which contextual cues are present and active in episodic memory.

Experiences do not only consist of contextual details, but also of value-charged elements, such as emotions or overall evaluative judgments. One can distinguish two approaches to the reconstruction of value-charged experiences: the *constructive* and the *value-account* approach. The constructive approach assumes that felt emotion cannot be stored in memory but is instead reconstructed from recalled contextual cues. In contrast, the value-account approach proposes the existence of a memory structure, which stores the frequency and intensity of a person's responses to a stimulus. This in turn can be used to cue the recall of contextual details of one's experiences. In the following, we describe the two approaches in more detail.

2.1. The constructive approach

The constructive approach assumes that reconstruction happens in a forward temporal order (Anderson and Conway, 1993; Barsalou, 1988; Means et al., 1989). Barsalou (1988) asked people to recall their experiences during the last summer. Most participants started in the beginning and proceeded in a chronological order. Often, recalling one episode cues the reconstruction of further episodes and more contextual information surrounding the episode (Anderson and Conway, 1993)—just like a string of pearls.

Please cite this article as: Karapanos, E., et al., Reconstructing experiences with iScale. International Journal of Human-Computer Studies (2012), http://dx.doi.org/10.1016/j.ijhcs.2012.06.004

Robinson and Clore (2002) further argued that "emotional experience can neither be stored nor retrieved" (p. 935), but can only be reconstructed on the basis of recalled contextual cues. They propose an accessibility model that distinguishes between four types of knowledge used to (re)construct an emotion. First, *experiential knowledge* is used when an emotion is constructed "online" (i.e. as the experience takes place). When experiential knowledge is inaccessible, people will resort to *episodic information*: they will recall contextual cues from episodic memory to reconstruct the emotional experience. When episodic memories become inaccessible, people will shift to semantic memory. They first access *situation-specific beliefs*: beliefs "about the emotions that are likely to be elicited in a particular type of situation" (p. 935). If event-specific beliefs are inaccessible as well (e.g., due to rarity of the event) people will access *identity-related beliefs*: "beliefs about their emotions in general" (p. 935).

Motivated by the accessibility model of Robinson and Clore (2002), Daniel Kahneman and colleagues (2004,2008) developed the Day Reconstruction Method (DRM), an online diary method that attempts to minimize retrospection biases when recalling emotional experiences. DRM starts by asking participants to mentally reconstruct their daily experiences as a continuous series of episodes, writing a brief name for each one. This aims at eliciting contextual cues within each experiential episode as well as temporal relations between episodes. As a result, participants reconstruct the emotional experience on the basis of a more complete set of episodic information, thereby minimizing bias from overly relying on semantic information, which is detached from the actual experience.

2.2. The value-account approach

The value-account approach assumes that people recall an overall emotional assessment of an experience, but not the exact details of the experienced event. Betsch et al. (2001) proposed the existence of a memory structure called *value-account*, which stores the frequency and intensity of positive or negative responses to stimuli. Since value-account is more accessible and better retained over time than details from episodic memory, it will actually cue episodic information, when reconstructing an experience, or feeds into a potential overall evaluation even in the absence of any episodic information (Koriat et al., 2000; Neisser, 1981). This top-down process for reconstructing memories is consistent with research on autobiographical memory, where three levels are distinguished: lifetime periods, general events, and event-specific knowledge. Reconstruction takes place in a top-down fashion where knowledge stored at the level of a lifetime period cues information at the two lower levels (Conway and Pleydell-Pearce, 2000).

Both approaches, constructive and value-account, suggest specific processes for retrieving emotional experiences from memory. While the constructive approach assumes a bottom-up, chronological recall of episodic information that cues the reconstruction of experienced affect and emotion, the value-account approach suggests a top-down process, starting from affect stored in specific value-accounts to cue the recall of specific episodic information. In the following section, we will illustrate how these two processes were operationalized in two separate versions of the iScale tool.

2.3. Graphing affect as a way to support the reconstruction of experiences

Imagine being asked to provide a graphical representation of how the global evaluation of your mobile phone changed over time on a timeline, which starts at the moment of purchase and ends in the present. There are a number of ways doing this. The aforementioned CORPUS technique – somewhat atheoretically – recommended starting from the momentary perception, comparing it to the beginning of product use and then recalling specific experiences. The presented theoretical positions above, however, provide clearer, more theoretically grounded suggestions of how to proceed. The constructive approach suggests starting from the beginning, recalling a piece of experiential information (e.g., an episode, a situation) and reconstructing an affective response and the according evaluative judgment (see e.g., Hassenzahl and Ullrich, 2007 for the relation between affect and product evaluation) to this piece of information. This will then cue the next piece and so forth until change over time is reconstructed. The value-account, however, suggests starting with the affective information, that is, recalling the change of affect and evaluation over time first, and using the general shape of this to recall more specific experiential information.

Obviously, both approaches can be supported by a tool. However, using the one or the other procedure may impact the results. To study potential differences, we created two different versions of iScale, the constructive and the value-account iScale.

2.4. iScale

Experience reconstruction with iScale starts with two questions: (a) "What was your opinion about the product's [particular quality] just before you purchased it?", and (b) "How did your opinion about the product's [particular quality] change since then?" The participants are then presented with a timeline that starts at the moment of purchase and ends in the present. In general, participants graph linear segments that represent an increase or decrease in their perception and evaluation over a certain period. Each line segment is associated with an identifier, displayed below the segment (Fig. 2a). A participant can click on the segment to report one or more experienced events, which are perceived as a cause of change. For each experience report, the participant can provide a brief name

Please cite this article as: Karapanos, E., et al., Reconstructing experiences with iScale. International Journal of Human-Computer Studies (2012), http://dx.doi.org/10.1016/j.ijhcs.2012.06.004

Fig. 2. In the *constructive* iScale (top), one starts by plotting points in a serial order (feed-forward progression) and participants are asked to report details of experiences right after a line segment was added (concurrent reporting). In *value-account* iScale (bottom), a line connects the start with the end and participants may further split this in distinct periods (top-down progression). Once participants have graphed the full pattern of change over time, they are asked to report on one or more experiences for each line segment (non-concurrent reporting).

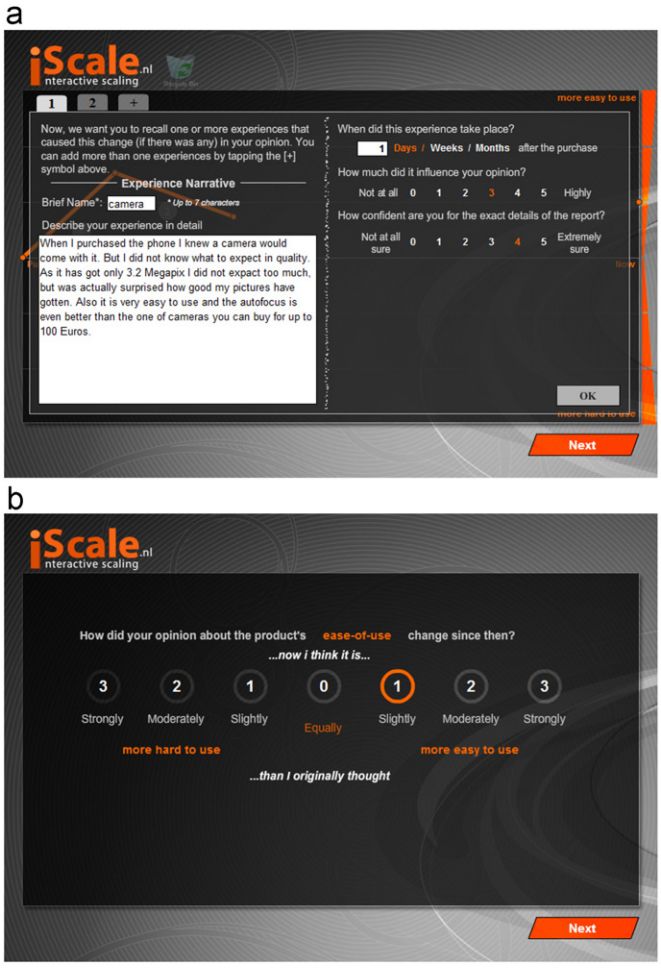

Fig. 3. (a) Interface for reporting experiences and (b) overall questions asked in the beginning of the survey.

(identifier), a more elaborate description of the experienced event (a narrative), and respond to a number of event-specific questions (Fig. 3a). For the goals of the present studies, we asked participants to recall (a) the exact time that the event took place, (b) the impact of the event on the participant's overall perception and evaluation, and (c) the participant's confidence on the exact details of the report. However, these questions can be adapted to particular research interests.

The two versions of iScale differ in the way graphing was used to support the reconstruction of experiences from memory. More specifically they differed in the progression of graphing (feed-forward versus top-down), and, the existence or absence of concurrency between graphing and reporting.

Feed-forward–Top-down progression of graphing: In the constructive iScale one starts by plotting points in a serial order; in the value-account iScale a line connects the start with the end of the timeline using the participant's response to the question asked in the first step about how product perception and evaluation has changed from the moment of purchase to the present. The participant is asked to proceed by splitting the full segment into smaller parts.

Concurrent–non-concurrent reporting: The constructive approach assumes that the affective, the value-charged component of a past experience can only be reconstructed from recalled contextual details of the event. On the contrary, the value-account approach assumes that individuals recall the value-charged component even without being able to recall the underlying contextual details. Thus, according to the constructive approach reporting should be concurrent with graphing, because reporting will increase the contextual details and, thus, result in richer recall. On the other hand, in the value-account approach, concurrent reporting might bias or hinder the process of recalling the value-charged component. Thus, in the constructive iScale the participant is asked to report details of experiences right after a line segment was added (i.e., drawn). Graphing and reporting proceeds concurrently. In the value-account iScale, this process is split into two distinct steps: the participant is first asked to graph the pattern of change over time and only after that she can report on one or more experiences for each line segment.

Please cite this article as: Karapanos, E., et al., Reconstructing experiences with iScale. International Journal of Human-Computer Studies (2012), http://dx.doi.org/10.1016/j.ijhcs.2012.06.004

In the remainder of the paper we describe two studies. Study 1 aimed at providing qualitative insight into the use of iScale and compared its performance to free-hand graphing. Study 2 compared the two versions of iScale to free recall, a control condition that does not impose structure on the reconstruction process. Overall, graphing is expected to provide temporal context for the recall of experienced events. This is expected to increase the amount and test–retest consistency of the information that the participants are able to recall. This assumption will be tested in study 2.

3. Study 1: Understanding graphing as a tool for the reconstruction of experiences

The first study aims at a qualitative understanding of graphing as a support for the reconstruction of experiences. It compares the two iScale tools to free-hand graphing.

3.1. Method

3.1.1. Participants

Twelve graduate students in HCI (seven male, median age 30 years) participated in the study. They were selected based on the diversity in their educational backgrounds. They were: five computer scientists, three industrial engineers, two linguists, one psychologist and one industrial designer.

3.1.2. Procedure

The study consisted of two parts. In the first part, each participant used three different graphing techniques: free-hand graphing (FHG) and the two iScale tools (constructive & value-account). All tasks were carried out on a Wacom Cintiq 21UX Interactive Pen Display. The order in which the tools were employed was counterbalanced across participants; FHG was always used first to avoid any bias from the iScale tools as we wished to understand users' natural behavior in graphing changes in perception, evaluation and feelings over time.

Participants were asked to graph how their opinion on three distinct qualities of their mobile phone changed over time (see Table 1). Each quality was described by a brief definition and three attributes to support the definition (Hassenzahl, 2004; von Wilamowitz-Moellendorff et al., 2006). Qualities along with their definitions were derived from von Wilamowitz-Moellendorff et al. (2006), though we rephrased the construct names utility and stimulation to usefulness and innovativeness as the former were not clear to some participants in our pilot tests. These qualities are routed in Hassenzahl's (2004) model that distinguishes *pragmatic quality*, which refers to the product's ability to support the achievement of do-goals such as making a telephone call, from *hedonic quality*, which refers to the product's ability to support the achievement of be-goals such as being stimulated or being admired.

Participants were instructed to think aloud; interactions and verbal data were captured on video. Each graphing task took approximately 4 min (min = 81 s, max = 594 s). No significant differences were found between the three tools (constructive: $M = 230$ s, SD = 147 s; value-account: $M = 205$ s, SD = 87 s; free-hand graphing: $M = 301$ s, SD = 130 s).

> "While graphing, you are asked to report experiences and events that induced these changes in your view of the product. We are interested in your exact thoughts and feelings as you perform the graphing. Why do you graph it in this particular way? What details of events, incidents, experiences do you remember? Is it just a feeling? Please think aloud while doing this."

In the second part, participants were interviewed about the differences between the three graphing techniques, using a structured interview technique, called the Repertory Grid (Fransella et al., 2003). This technique is well aligned with parallel design, and particularly in the existence of three or more alternative artifacts, and allows for inquiring into participants' idiosyncratic ways in which they differentiate the artifacts. In this way, one can inquire into the design space from a users' perspective (Hassenzahl and Wessler, 2000). This is useful not only for designed artifacts but also for methods as in the present case. Participants were given three cards, each providing a name and a screenshot of one of the three graphing techniques. Participants were first asked to identify the three techniques. Next, they were asked to "think of a property or quality that makes two of the graphing techniques alike and discriminates them from the third". They were instructed to feel free to make any combination of the three alternatives. Contrary to common practice with the Repertory Grid Technique, we did not probe participants for the exact opposite of the property they provided, but rather focused on further elaboration, when possible. This was supported by *laddering* and *pyramiding* techniques (Reynolds and Gutman, 1988). Laddering seeks to understand what "motivates" a given property and thus ladders up in an assumed means-ends-chain (Gutman, 1982)

Table 1
The three aspects, their definition and attributes.

Name	Definition	Word items
Usefulness	The ability of a product to provide the necessary functions for given tasks	Useful, practical, meaningful
Ease-of-use	The ability of a product to provide the functions in an easy and efficient way	Easy to use, simple, clear
Innovativeness	The ability of a product to excite the user through its novelty	Innovative, exciting, creative

Please cite this article as: Karapanos, E., et al., Reconstructing experiences with iScale. International Journal of Human-Computer Studies (2012), http://dx.doi.org/10.1016/j.ijhcs.2012.06.004

towards more abstract properties of the stimuli; in laddering we first asked the participant whether the mentioned property is positive, and subsequently why this property is important to him/her (e.g. "why is expressiveness important to you?"). Pyramiding, also known as negative laddering, seeks to understand the lower level properties that make up for a given property; in pyramiding we asked the participant to elaborate on what makes the given technique to be characterized with the respective property (e.g. "what makes free-hand graphing more expressive?").

3.2. Analysis and results

3.2.1. Understanding free-hand graphing

We analyzed FHG to get an idea of how people actually graph changes in product perception and evaluation. We segmented the collected graphs in discrete units. A unit was coded when two conditions were met: a semantic change in the participant's verbal report following a pause in graphing as observed in the video recorded sessions (e.g., "[pause in graphing] but then I got to the point where I got new software updates"). Pauses often suggested an initiation of a new recall. Often this was combined with a change in the slope of the curve, but this was not always the case.

Each unit was then coded for the type of curve and the type of verbal report. Curves were classified into four categories: (a) Constant (C) signifying no change in participant's opinion over a certain period, (b) Linear (L), either Increasing or Decreasing, (c) Non-linear (NL) when there were no grounds for arguing that the curve could be approximated by a linear one or when a single report was associated with two discrete linear curves of different slope (see Fig. 4b), and (d) Discontinuous (D) when the slope was approximately parallel to the vertical axis.

Table 2, line overall, shows the distribution of the four different types of curves. The majority of segments (44 of

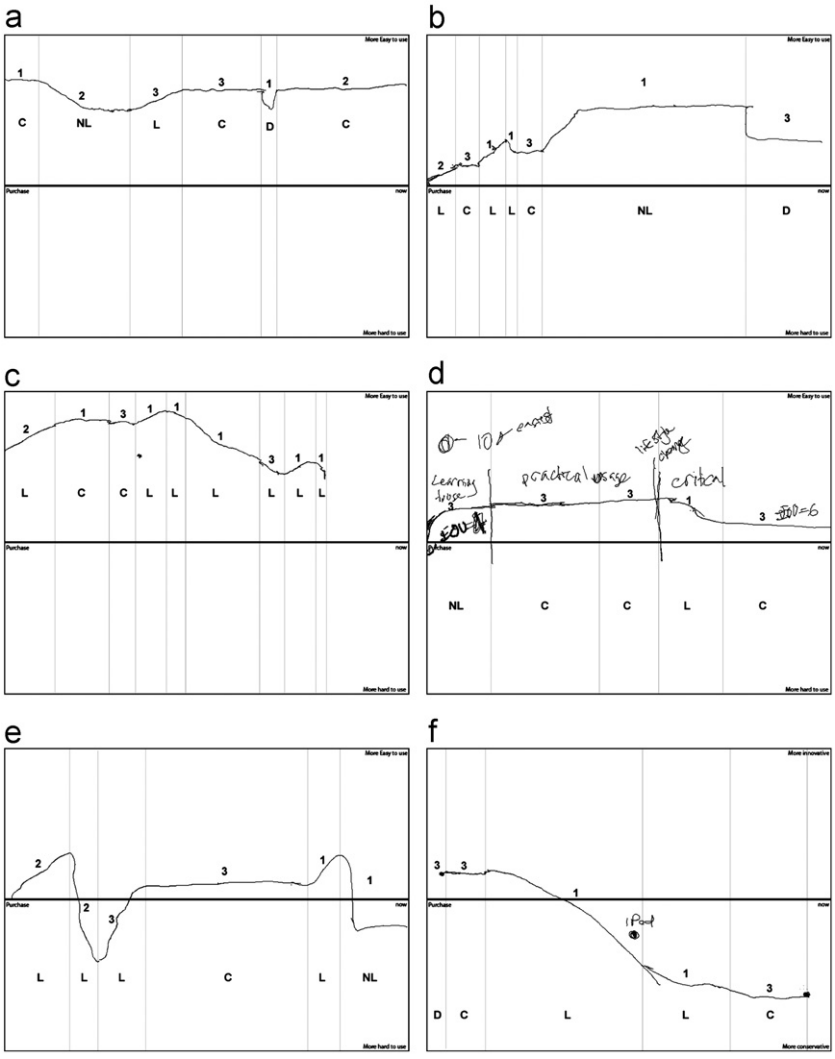

Fig. 4. Examples of free-hand graphing (left: a,c,e. right: b,d,f). Identified segments are indicated by vertical lines. Each segment is coded for the type of report (1: reporting a discrete experience, 2: reporting an overall evaluation, reasoning through experience, 3: reporting an overall evaluation with no further reasoning) and type of graph (C: constant, L: linear, NL: non-linear, D: discontinuous).

Please cite this article as: Karapanos, E., et al., Reconstructing experiences with iScale. International Journal of Human-Computer Studies (2012), http://dx.doi.org/10.1016/j.ijhcs.2012.06.004

74, 60%) were categorized as linear. Only 5% (4 of 74) of segments were non-linear. Of those, only a single report was associated with two or more linear segments with different slopes (cf. Fig. 4a segment 2, Fig. 4b segment 6, Fig. 4d, segment 1). In addition, only four of 74 (5%) instances of discontinuity were observed in the graphs. Thus, while in some cases users are inclined to draw non-linear curves, the majority of curves were linear. This allowed us to focus iScale on linear graphing, thereby reducing a number of potential problems with handling complex, non-linear types of curves in an online tool.

To get an idea of how the graphs relate to reported detailed experiences, the obtained verbal reports were classified into three broad categories. One category is the *recall of a distinct experience*: an experience that relates to a particular event with beginning and end. According reports were indicative of the constructive mode: recalling contextual information from a specific experience was followed by the reconstruction of the value judgment from the recalled facts. For example: "*The reason I got this device was to develop applications for it.* [*the company*] *has a special program for educational institutions to which provides free licenses for development. But when we contacted them, they even questioned the existence of our institution… this should have happened around here* [*points to the curve*]". Such distinct reports provided one or more contextual cues about the past experience, such as temporal information (i.e. when the event took place), changes in the context of use (e.g. "then I went on vacation…"), information related to the participant's social environment (e.g. "a friend of mine was very positive…"), etc. They constituted the most dominant type of reporting (37 of 74, 50%).

Other reports provided no contextual information about a recalled experience, but instead, the participant reported an overall evaluation without further motivation: "*after that*, [*my opinion*] *is about constant*, *it hasn't changed lately*". Such reports are typical for a pure value-account mode of recall: recalling an overall evaluation of a specific experience or period, while failing to recall contextual cues or facts about an exact experience (24 of 74, 32%).

We further found a third type of reporting that combines the two types mentioned above: "[*my opinion*] *decreased as I expected that it would be easier than that*, *for example*, *I would like to have the automatic tilting to landscape view as it has an accelerometer*". Those reports were grounded in the recall of an overall evaluation, but participants proceeded to reason about this value-judgment through reporting specific experiences (13 of 74, 18%). Most of them (10 of 13) reflected linear changes.

3.2.2. How does iScale compare to free-hand graphing?

The two iScale tools were also compared to free-hand graphing. Participants' verbal reports were transcribed and analyzed using Conventional Qualitative Content Analysis (Hsieh and Shannon, 2005). We started with *open coding* where we aimed at identifying an overpopulated list of design properties that appear to influence the design space of the three graphing techniques. Subsequently, we grouped the initial codes into overall categories through an iterative process. Each statement was coded for the property the participant mentions as well as to whether or not this property affects the graphing or the recalling process. Statements were always differentiating two of the approaches from a third (e.g., the two iScale versions as opposed to FHG) as this was imposed by the structure of the employed Repertory Grid interview technique (e.g., "think of a property or quality that makes two of the graphing techniques alike and discriminates them from the third").

Table 3 illustrates the dominant properties that were elicited in the interview sessions. For each property, it displays the number of participants mentioning it as present for a given technique, and the number of participants mentioning it as affecting the graphing or recalling process. The design properties can be distinguished into three broad groups: *expressiveness*, *control*, and *Interplay graphing-recalling*.

3.2.3. Expressiveness

As expected, the majority of users perceived the free-hand graphing approach as more expressive. This was mostly due to the *freedom in graphing* that the free-hand approach provides as opposed to the iScale tools that restrict the user in plotting points connected through line segments.

The majority of participants emphasized the effect this has on the graphing activity. While all participants expressed the perceived benefit of FHG for graphically expressing their opinions, only one participant mentioned that this also affects the recalling process as the graph provides richer feedback and therefore more cues for recalling. One participant stated freedom in graphing as a positive property for expressing impressions for which she fails to recall any supportive facts.

Table 2
Relationship between graphing and reporting in free-hand graphing. Types of graphing: C, constant; L, linear; NL, non-linear; D, discontinuous.

Type of report	Type of graph				
	C	L	NL	D	
Discrete experience	3	30	2	2	37 (50%)
Overall evaluation without motivation	17	4	1	2	24 (32%)
Overall evaluation, motivated by experience	2	10	1	0	13 (18%)
Overall (%)	22 (30)	44 (60)	4 (5)	4 (5)	74

Please cite this article as: Karapanos, E., et al., Reconstructing experiences with iScale. International Journal of Human-Computer Studies (2012), http://dx.doi.org/10.1016/j.ijhcs.2012.06.004

Next, some participants mentioned the ability to *annotate* the graph as positive , because it enhances the recalling process. Annotations helped in recollecting *contextual* and *temporal* cues from the past, such as positioning a specific experience along the timeline, splitting the timeline into periods, but also in externalizing thoughts that participants thought they might fail to recall afterwards.

3.2.4. Control

Most participants stated that the iScale tools provide more overall control. First, eight out of the 12 participants found the constrained interaction a positive aspect of iScale, providing better *interactivity*, as it consumes less resources, thus providing them *better control of the output* (four participants) and *enabling them to focus on recalling their experiences* (four participants).

Second, participants differentiated iScale from FHG in terms of the ability to modify the graph while new experiences are recalled. Some participants further differentiated between the two iScale tools in terms of modifiability.

Third, seven out of the 12 participants acknowledged that the value-account tool provides a better overview of the full timeline, and, thus, *temporal overview* enhancing their recall process.

3.2.5. Interplay graphing-recalling

Five participants in total mentioned *temporal linearity* as a property that differentiated free-hand graphing and the constructive iScale from the value-account tool. Most of those participants mentioned that recalling events in a step-by-step order helped them in recalling more events, while some of them were negative towards value-account as they felt that it constrained them when recalling events due to a focus in compiling a coherent story.

Similarly, five participants highlighted that the *concurrency* of graphing and reporting, that was a feature of the constructive version of iScale, enhances the output of both the recall and the graphing process.

3.3. Discussion

As expected, free-hand graphing was found to be more expressive than iScale due to the increased degrees of freedom in graphing as well as due to its ability to easily annotate graphs. However, participants also reported properties that were not present in the FHG, such as the two-step interactivity and modifiability of the electronic graphs that resulted in a better interoperability between the graphing and the recalling activity. Participants also reported benefits for both the constructive and the value-account iScale. The value-account version provided a temporal overview, which influenced both graphing and recall. The constructive approach provided benefits, such as the chronological order and concurrency between graphing and reporting which had a positive impact on the recall process.

The need for graphing non-linear and discontinuous curves was limited, while most non-linear curves could be approximated by two linear segments. The need for annotation was highlighted by participants in the post-use interviews and two forms of annotation were added to the tool: (a) a timeline annotation that allowed users to set the start and end date of graphed segments, thereby splitting the timeline in specified periods, and (b) a visualization of experiences along the respective line segment that they belong to, with a brief identifier for the experience (see Fig. 2a). Annotation provided users with the ability to combine an overview of the different periods as well as the experiences that defined these periods. Annotation also promotes interactivity as users have a better overview of the graphed pattern and are therefore more likely to modify it.

Table 3
Design properties elicited in the interview sessions. Number of participants using a given property to differentiate the three graphing techniques, and number of participants mentioning the given property as affecting the graphing or recalling process. FHG: Free-Hand Graphing, CON: Constructive, VA: Value-Account.

Name	Description	Tool			Impact on	
		FHS	CON	VA	Graph	Recall
Expressiveness						
Graphing freedom	Increasing the degrees-of-freedom in graphing	7			6	1
Annotation	Enabling free-hand annotation on the graph	3				3
Control						
Interactivity	Providing instant visual feedback through quick actions		8	8	4	4
Modifiability	Providing the ability to modify the graph as new experiences are being recalled		8	6	8	2
Temporal overview	Providing feedback on overall opinion across full timeline			7	3	4
Interplay graphing-recalling						
Chronological linearity	Chronological linearity in recalling	5	5		1	5
Concurrency	Concurrency of time in graphing and recalling	5	5		2	5

Please cite this article as: Karapanos, E., et al., Reconstructing experiences with iScale. International Journal of Human-Computer Studies (2012), http://dx.doi.org/10.1016/j.ijhcs.2012.06.004

4. Study 2: benefits and drawbacks of the constructive and the value-account version of iScale

While iScale appeared to be a viable alternative to free hand graphing, the comparative benefits and drawbacks of both iScale variants merited a second study. We compared the constructive and the value-account version of iScale to a control condition that entailed reporting one's experiences with a product without any support through graphing. We focused on the number, the richness, and the test–retest consistency of the elicited experience reports.

4.1. Method

4.1.1. Participants

Forty-eight individuals (17 female, median age=23, min=18, max=28) participated in the experiment. They were all students at a technical university and were rewarded for participating in the experiment; 19 of them majored in management related disciplines, 16 in design, and 13 in natural sciences and engineering. They all owned a mobile phone for no less than four and no more than 18 months; 16 participants owned a smart phone. No significant differences were found between participants in the constructive and the value-account condition in length of ownership ($M_{con}=13$ months, $M_{va}=10$ months, $t(46)=1.51$, $p=.13$) and type of mobile phone (five participants owned a smart phone in the constructive condition, 11 participants in the value-account, $\chi^2=2.3$, $p=.13$).

4.1.2. Materials

Three different versions of iScale were used in the experiment: *constructive*, *value-account*, and *no-graphing* (control). The constructive and value-account versions employed the two distinct graphing approaches described earlier. No-graphing was a stripped-down version of iScale, with the graphing interface completely removed. Thus, users were only provided with the interface to report experiences (see Fig. 3a) and this was used as a control condition to test the effect, if any, of graphing.

4.1.3. Study design

A 3×2 study design was employed with *mode of recall* (constructive, value-account, no graphing/control), and *product quality* being reported (i.e. ease-of-use versus innovativeness) as independent factors (see Fig. 5).

Mode of recall

Product quality	Constructive	Value-Account	Control
Ease of use	A	B	C
Innovativeness	D	E	F

Fig. 5. Study design.

4.1.4. Procedure

Participants joined two sessions, each one lasting approximately 40 min and separated by approximately 1 week (minimum: 7 days, maximum: 10 days). During the first session, participants used two different tools (either of the two graphing versions of iScale and the no-graphing version) to report on two qualities of their mobile phones (see Section 3.1.2 for a motivation of the chosen product qualities). For instance, participants in condition 1 (see Fig. 6) used the constructive iScale to report on ease-of-use, followed by the no-graphing tool to report on innovativeness. During the second session, participants used the same combinations of tool—product quality, but in reverse order (see Fig. 6). As such, participants' consistency across those two sessions could be used a measure of test–retest consistency of the recall process.

4.1.5. Dependent variables and expectations

Despite the fact that the study was explorative in nature, a number of predictions about the differences in performance of the three versions of the tool can be made.

Number of elicited experience reports: Based on existing evidence that the reconstruction of events in a serial chronological order cues the recall of temporally surrounding experiences and related contextual cues (Anderson and Conway, 1993), it was expected that the constructive iScale will result in an increase in the number of experiences being reported. For the value-account iScale, which makes it more difficult for participants to reconstruct their experiences in a chronological order, the difference to the control condition was expected to be smaller.

Richness of elicited experience reports: Similar to number of elicited experience reports, we expected that reconstructing in a chronological order would lead to more contextual cues, thus providing richer insight into users experiences. Such contextual information may relate to temporal (i.e. when did the event happen), factual (i.e. what happened), social (i.e. who was present) and others. To identify these different factors of richness, we submitted the experience reports to a qualitative content analysis (Hsieh and Shannon, 2005) (see Section 4.2.2 for a more elaborate description of this process).

Test–retest consistency in time estimation: As participants are expected to recall more contextual cues in the constructive iScale, this should increase the test–retest consistency in recalling factual details of the past experiences, such as temporal information (e.g. when did the experience take place) (Kahneman et al., 2004). We further predict that graphing in general (even in the value-account condition) will result in a more consistent recall of such temporal information, as graphing provides a temporal overview of the recalled experiences. To assess this, we coupled experience reports from the two sessions of the study that referred to the same experience and computed the difference in estimated time across the coupled experience reports (see Section 4.2.3 for a more elaborate description of this process).

Please cite this article as: Karapanos, E., et al., Reconstructing experiences with iScale. International Journal of Human-Computer Studies (2012), http://dx.doi.org/10.1016/j.ijhcs.2012.06.004

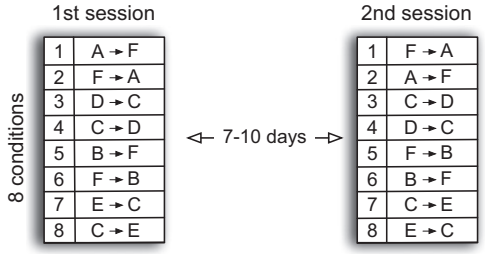

Fig. 6. Study procedure. Participants joined two sessions. In each session they used two different tools to report on two different product qualities. Tasks A to F refer to the six conditions of the study design (see Fig. 5).

Test–retest consistency of graphs: The test–retest consistency of the participants' graphs (i.e. value-charged information) was expected to be higher in the value-account version, where participants cue this directly through a hypothetical memory structure, compared to the constructive version, where participants are assumed to reconstruct this information from concrete contextual cues recalled from episodic memory. This is based on the assumption that repeated chronological reconstruction might cue a different set of experiences and, thus, lead to a different path in reconstructing the overall pattern of how one's experience with a product developed over time. As a measure of the test–retest consistency of the graphs, we computed the area between the two graphs through sampling them in 100 steps (see Fig. 9).

4.2. Analysis and results

A total of 457 experience reports were elicited. Participants provided an average of 4–6 experience reports depending on the recall condition. Ninety-five percent of all experiences were related to the first 6 months of use. We compare the two graphing tools to the no-graphing (control) condition in terms of (a) the number of elicited experience reports, (b) the richness of elicited experience reports, (c) the test–retest consistency in time estimation, and (d) the test–retest consistency of graphs.

4.2.1. Number of elicited experience reports

Fig. 7 shows the number of reported experiences as a function of the mode of recall. An average of 6.1 experience reports was elicited when using the constructive iScale, 4.6 when using the value-account iScale, and four when using the no-graphing (control) tool.

An analysis of variance with number of experience reports as dependent variable and mode of recall (constructive, value-account, control), and product quality (ease-of-use, innovativeness) as independent variables revealed a significant main effect for mode of recall, $F(2,89)=7.74$, $p<0.05$, $\eta_p^2=0.15$, but not for product quality, $F(1,89)=1.64$, $p=0.2$, $\eta_p^2=0.02$, or for the interaction between mode of recall and product quality, $F(2,89)=1.19$, $p=0.7$, $\eta_p^2=0.007$. Post-hoc tests using the Bonferroni correction revealed that participants in the constructive condition elicited a significantly higher number of experience reports than in the value-account ($p<0.05$, Cohen's $d=0.7$) and the control condition ($p<0.001$, $d=1$). No significant differences were demonstrated between the value-account and the control condition ($p=0.7$, $d=0.3$).

4.2.2. Richness of elicited experience reports

To identify the different factors of richness, we submitted the experience reports to a qualitative content analysis (Hsieh and Shannon, 2005). Open coding was performed by the first author and resulted in three main types of information present in the reports: *discrete event information* summarizes references to a concrete occurrence that influenced the experience (e.g., "I found out that there was a mail application"), *temporal information* which summarizes references to the exact time at which the experience took place (e.g., "The first week I tried to practice the touch screen"), and, *expectations* which summarize references to participants' expectations about the reported experience (e.g., "As I bought this phone in Europe I expected that at least all European languages are available for free online") (see Table 4). Each report was coded by the first author for the presence or absence of temporal information, discrete event information and expectations. Interrater agreement (Fleiss' Kappa, 2003) was computed on a small random sub-set of the reports (10%) coded by the first author and two additional researchers: temporal information ($K=0.97$), discrete event information ($K=0.71$), expectations ($K=0.77$). In all cases, interrater agreement was satisfactory.

Significant differences were observed between the constructive and the no-graphing (control) version of iScale with regard to references to discrete events ($p<0.05$, Fisher's one-tailed exact test), references to temporal information ($p<0.05$, Fisher's one-tailed exact test), but not with regard to references to participants' expectations before the experienced event. Forty-five out of 146 (31%)

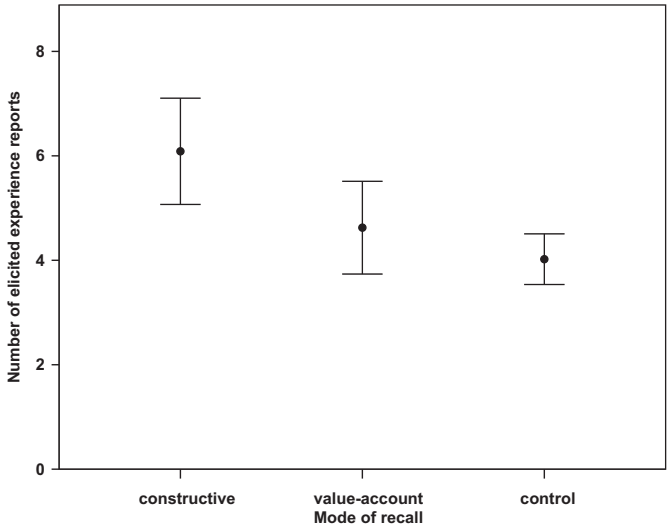

Fig. 7. Average number of experience reports, together with their 95% confidence intervals, elicited by participants using the constructive, the value-account, and the no-graphing (control) version of iScale.

Please cite this article as: Karapanos, E., et al., Reconstructing experiences with iScale. International Journal of Human-Computer Studies (2012), http://dx.doi.org/10.1016/j.ijhcs.2012.06.004

Table 4
Number of experience reports judged for the three dimensions of richness, elicited through the three different versions of iScale: constructive, value-account and no-graphing (control).

Name		Constructive	Value-Account	No-graphing (control)
Contextual information				
(a) Event: Does the participant recall one or more discrete events that lead to the realization of the reported experience?	Y:	45	27	38
	N:	101	91	154
(b) Temporal: Does the participant recall temporal information about the reported experience?	Y:	20	16	14
	N:	126	102	178
(c) Expectation: Does the participant recall his/her expectations about the reported experience?	Y:	10	11	16
	N:	136	107	176

reports elicited through the constructive iScale contained at least one cue referring to a discrete event as opposed to 38 out of 192 (20%) in the control condition, and 20 out of 146 (14%) contained at least one cue referring to temporal information as opposed to 14 out of 192 (7%) in the control condition. No significant differences were observed between the constructive and the value-account version as well as between value-account and the no-graphing (control) version in any of the three dimensions or richness.

4.2.3. Test–retest consistency in time estimation η

The two sessions of the study are expected to more or less result in the same experience reports. Thus, reports across these two sessions can be coupled. A total of 325 experience reports (71%) were coupled. We only coupled reports when we had sufficient confidence that they reported the same experience. This, inevitably, left out short reports that did not contain sufficient text to form this judgment. Seventy-two percent of reports in the constructive condition were coupled, 81% in the value-account and 69% in the control condition. Fifty-two coupled reports had incomplete time information (in either session), leaving a total of 273 complete coupled reports.

For each reported experience participants estimated the moment in time (i.e. days, weeks or months after the purchase of the product) at which the experience took place (see Fig. 3a). We used formula 1 to compute the distance (or convergence) between the recalled points in time of both sessions (for a justification of the logarithmic transformation see the appendix section)

$$\Delta = Abs(\log(t_2) - \log(t_1)) \quad (1)$$

An analysis of variance with the computed temporal distance Δ between experience reports from session 1 and session 2 as dependent variable and mode of recall (constructive, value-account, control) and product quality (ease-of-use, innovativeness) as independent variables displayed significant main effects for mode of recall, $F(2, 267)=5.42$, $p<0.01$, $\eta_p^2=0.04$, and for product quality, $F(1,267)=4.66$, $p<0.05$, $\eta_p^2=0.02$, but not for the interaction between mode of recall and product quality, $F(2,267)=1.81$, $p=0.2$, $\eta_p^2=0.01$. Post-hoc tests using the Bonferroni correction revealed that participants in the constructive condition were significantly more consistent in estimating the time that an experience took place as compared to the control condition ($p<0.01$, Cohen's $d=4.41$). No significant differences were established between the two graphing conditions ($p=0.9$, $d=1.29$), or between the value-account and the control condition ($p=0.1$, $d=3.01$). Last, a significant difference was found in the temporal consistency of reports for both product qualities, where reports of ease-of-use were temporally more consistent than reports of innovativeness, $t(271)=2.8$, $p<0.01$, Cohen's $d=3$ (Fig. 8).

4.2.4. Test–retest consistency of graphs

Fig. 9 displays example graphs by two participants in two respective sessions with the constructive iScale and two participants with the value-account iScale. The area between the two graphs is a simple measure for the inconsistency in participants' graphs. It was calculated through sampling the graphs in 100 steps.

An analysis of variance with this area measure as dependent variable and mode of recall (constructive, value-account) and product quality (ease-of-use, innovativeness) being reported as independent variables revealed a significant effect for mode of recall, $F(1, 44)=6.75$, $p<0.05$, $\eta_p^2=0.13$, but not for product quality, $F(1, 44)=0.481$, $p=0.5$, $\eta_p^2=0.01$, or for the interaction between mode of recall and product quality, $F(1,44)=0.073$, $p=0.8$, $\eta_p^2=0.002$. Graphs elicited through the constructive iScale tool were more consistent ($M_{area}=30.1$, SD=17.4) than ones elicited through the value-account iScale ($M_{area}=52.2$, SD=36.7, $t(46)=2.7$, $p<0.05$, Cohen's $d=0.8$)

4.3. Discussion

As expected, the constructive iScale tool lead to a significant increase in the number of elicited reports (with a large effect size, $d=1$, Cohen, 1992), but also the richness, of elicited reports, as compared to the control condition that involved no graphing. When using the constructive iScale, participants elicited approximately 50% more experience reports, and were more likely to recall cues referring to a discrete event, or ones referring to temporal information, such as when the experience took

Please cite this article as: Karapanos, E., et al., Reconstructing experiences with iScale. International Journal of Human-Computer Studies (2012), htt p://dx.doi.org/10.1016/j.ijhcs.2012.06.004

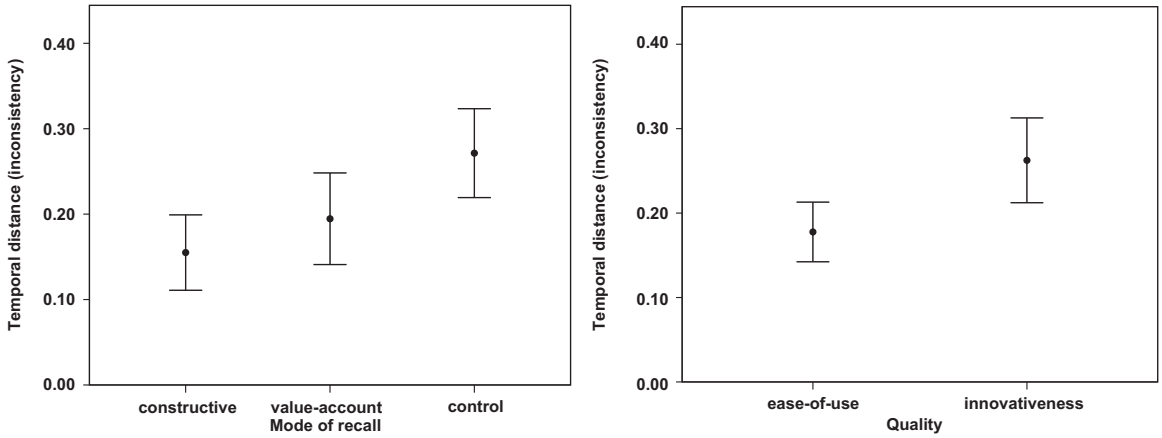

Fig. 8. Mean temporal inconsistency, together with the 95% confidence interval, between two recalls of the same experience, (left) when using the constructive, the value-account, and the no-graphing (control) version of iScale, (right) when reporting on perceived ease-of-use and perceived innovativeness.

place. Contrary to our expectations, the value-account iScale did not result in similar benefits. These findings support the idea that graphing, through imposing a chronological order in the recall process (Anderson and Conway, 1993), supports the reconstruction of the context in which experiences took place, thus forming stronger temporal and semantic links across the distinct experiences (Kahneman et al., 2004).

Next to a number and the richness of elicited reports, the constructive iScale tool also demonstrates the highest consistency across the two sessions (with a large effect size, $d=4.41$, Cohen, 1992), in recalling the exact time that each experience took place. These findings seem to be in line with the previous ones, i.e., given that the constructive iScale lead to a more effective reconstruction of the context in which experiences took place, this should also be beneficial when estimating when these experiences took place. Moreover, we found experiences relating to the ease-of-use of the product to be more reliably recalled (in terms of time estimation) in comparison to ones relating to the products' innovativeness. One possible interpretation might tap into the different nature of experiences that relate to ease-of-use and innovativeness. von Wilamowitz-Moellendorff et al. (2006, 2007) similarly observed that participants often recall with greater ease contextual cues about experiences relating to ease-of-use rather than stimulation. Ease-of-use is tied to concrete actions, whereas stimulation cannot be allocated to specific events. Thus, the effect of chronological order in reconstruction may be more salient in the case of contextually rich experiences than in case of more abstract ones.

Finally, contrary to our expectations, the constructive iScale resulted in higher test–retest consistency of the participants' graphs (i.e. value-charged information) than the value-account iScale. We expected that the value-account iScale would result in a higher consistency in the graphs, because this information is assumed to be cued directly from a hypothetical memory structure (Betsch et al., 2001). In contrast, for the constructive iScale, we assumed that this information is reconstructed from contextual details recalled from episodic memory. This would quite naturally result in lower consistency across repeated recalls, as long as the repeated chronological reconstruction is likely to cue different sets of experiences, which in turn leads to a different reconstructed shape of value-charged information over time. Our expectations were not confirmed as graphs elicited through the constructive iScale were more consistent than ones elicited through value-account iScale with a moderate to large effect size ($d=0.8$). This might relate to the finding of Reyna and Kiernan (1994) that participants may falsely recognize the overall gist of a memory, while correctly recalling its exact details.

5. Conclusion and future work

This paper has presented iScale, a graphing tool to elicit change in product perception and evaluation over time. It took the general approach of retrospective reconstruction of users' experiences as an alternative to longitudinal studies. More specifically, the tool was designed with the aim of increasing participants' effectiveness in recalling their experiences with a product.

We created two different, theoretically grounded versions of iScale. The constructive iScale tool imposes a chronological order onto the reconstruction of experiences. This will increase the contextual cues surrounding the experienced events, cues used to reconstruct value-charged information (i.e., affect). The value-account iScale aims at distinguishing the elicitation of the two types of information: the value-charged and contextual cues. It assumes that value-charged information can be recalled from a separate, specific memory structure.

Study 1 provided qualitative insights into the use of iScale and compared it to free-hand graphing. The study highlighted the importance of annotations and led to a redesign of iScale with two new forms of annotation, (a) timeline annotation that allowed users to set the start and

Please cite this article as: Karapanos, E., et al., Reconstructing experiences with iScale. International Journal of Human-Computer Studies (2012), http://dx.doi.org/10.1016/j.ijhcs.2012.06.004

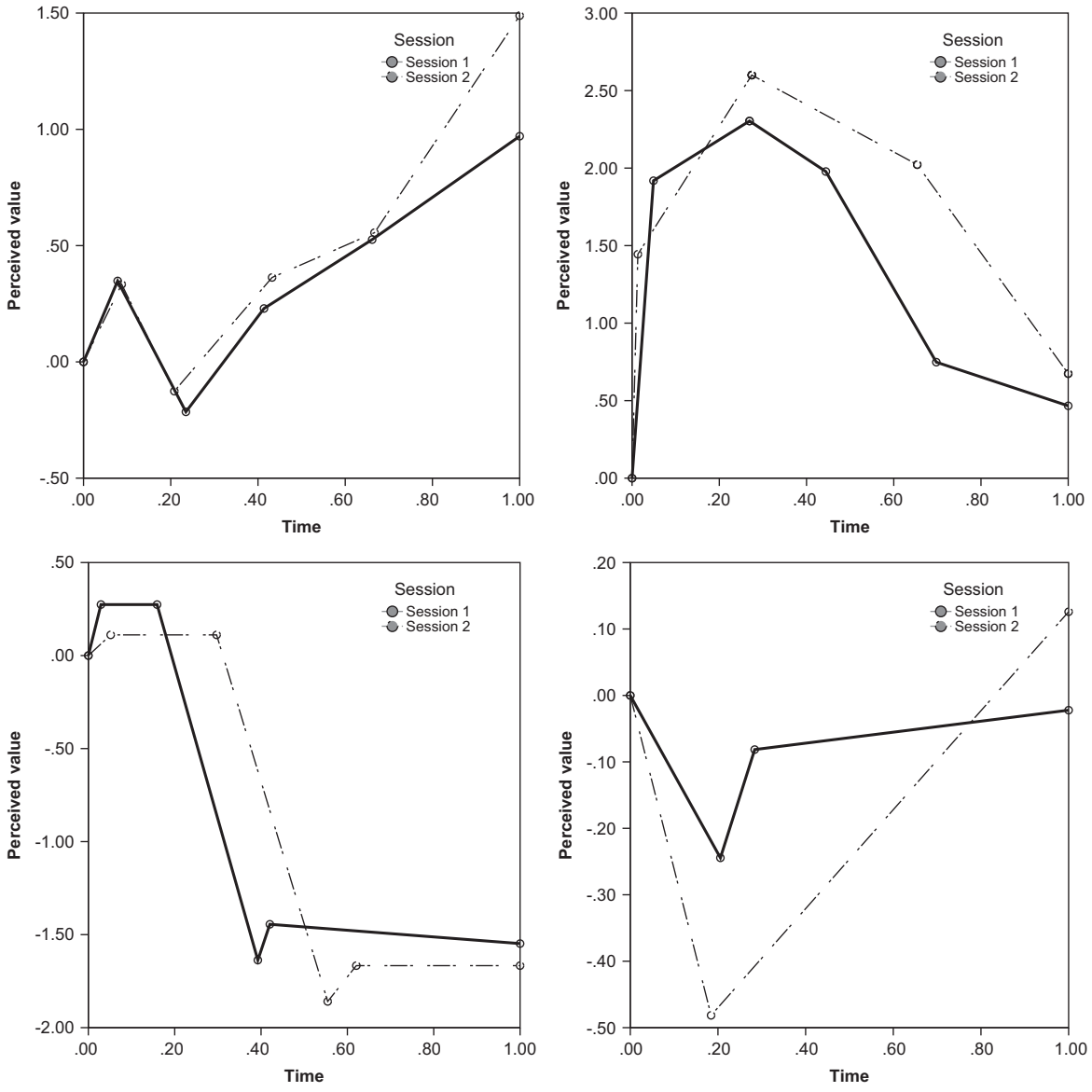

Fig. 9. Example graphs elicited in the constructive (top) and the value-account (bottom) conditions during the two sessions.

end date of graphed segments, and (b) a visualization of experiences on the main, graphing, interface that allowed an overview of one's storyline. Overall, the study confirmed our expectations that free-hand graphing is more expressive than iScale due to the increased degrees of freedom in graphing as well as due to its ability to easily annotate graphs. Nevertheless, participants reported qualities present in iScale, such as two-step interactivity and modifiability of the electronic graphs, which resulted in a better interoperability of graphing and the recalling activity.

Study 2 tested the effectiveness of graphing with the two different versions of iScale against a control condition, which allowed for the direct reporting of experiences, without any form of graphical representation. Overall, we found that the constructive iScale provided better assistance than the value-account iScale tool in the reconstruction process. When using the constructive iScale, participants elicited approximately 50% more experience reports than in the control (no-graphing) condition were more likely to recall cues referring to a discrete event, or ones referring to temporal information such as when the experience took place, and were more consistent across multiple recalls in estimating the time when each experience took place. Moreover, contrary to our expectations, participants graphed patterns were more consistent in the constructive than in the value-account condition. These findings support the idea that graphing, through imposing a chronological order in the recall process (Anderson and Conway, 1993), supports the reconstruction of the context in which experiences took place, thus forming stronger temporal and semantic links across the distinct experiences (Kahneman et al., 2004).

While iScale is promising, one crucial aspect must be kept in mind. There is a likely discrepancy between experiences elicited through longitudinal field studies

Please cite this article as: Karapanos, E., et al., Reconstructing experiences with iScale. International Journal of Human-Computer Studies (2012), htt p://dx.doi.org/10.1016/j.ijhcs.2012.06.004

composed of records of moments (e.g., through experience sampling, Hektner et al., 2007) and retrospective data elicited from memory through iScale. Retrospective reconstructions cover long periods of time and, thus, systematic biases, such as the overemphasis on especially salient moments, are likely to occur (Kahneman, 1999; Bartlett, 1932). The current studies provide no insights into how these retrospections differ from the actual experiences and future work should inquire in those differences with actual longitudinal studies.

In this paper we argued that veridicality may not be as important as the test–retest consistency of the recall process, because people actually communicate and act upon their own biased memory and not on an unbiased objective summary of what actually happened. In supporting design, understanding what users remember may be more important than what they actually experienced. Still, designers are not always interested in users memories. Often, the actual, and not the remembered, experiences should be at the forefront. Consider, for instance, the case where we might want to know the reasons that underly non-responsible driving behavior. Memories offer little understanding as to what motivates such behaviors. Retrospective techniques are not aimed to replace longitudinal field studies and in-situ methods. Instead, we propose that retrospective techniques may be a viable alternative to longitudinal studies when memories are placed at higher importance than actuality.

Next, in our study, we assessed users' test–retest consistency in recall using two different pieces of information: the exact time that an experience took place and the overall graphed pattern. These two served to indicate the two different kinds of recalled information: episodic and value-charged (i.e., affect). The former, we thought, would signify the effective recall of a substantial amount of contextual cues from episodic memory. We expected that estimating this temporal information would be an error-prone activity. Thus, a more effective recall would have a strong impact on users' consistency in time estimation, through the presence of more contextual cues. The latter, we thought, would signify the recall of experiential information (i.e., affect), either through inferring this from recalled episodic information (Robinson and Clore, 2002), or through recalling this directly from value-account Betsch et al. (2001). These two metrics of consistency are not without limitations; future work should expand to different facets of consistency.

Beyond the cost-effective elicitation of longitudinal data, this work provides support for the viability of survey methods that guide participants through a structured process of data elicitation. A wealth of such procedures exists for face-to-face interviewing. For instance, structured interview techniques such as triading (Fransella et al., 2003) and laddering (Reynolds and Gutman, 1988) imprint a particular structure onto the data elicited by participants, which makes the data computational friendly. Another example is Vermeersch' explicitation technique (see Light, 2006), which employs a particular style of interviewing that aims at supporting the interviewee to enter a state of evocation and "relive" the activity under consideration. While such techniques, when used in face-to-face interviewing, are most powerful, they are labor-intensive and require skilled interviewers (Groves et al., 2009), which always constrains the sample size of the study. Self-reporting approaches, on the other hand, can have impact because one can survey large samples and, by that, also inquire into rare experiences and atypical behaviors.

Obviously, iScale can as well produce large amount of qualitative information that will require labor-intensive analysis given traditional qualitative data analysis procedures like Content Analysis (Krippendorff, 2004; Hsieh and Shannon, 2005) and Grounded Theory (Strauss and Corbin, 1998). Novel techniques from the field of information retrieval (Landauer and Dumais, 1997; Blei et al., 2003) may prove especially fruitful in automating or semi-automating the qualitative analysis process. Finally, the interpersonal analysis of the graphs is definitely a subject for further research and was addressed here only superficially.

iScale was motivated by a need for lightweight methods that provide insights into long-term usage and related experiences. While the importance of temporality has been repeatedly highlighted in user experience research (Forlizzi and Battarbee, 2004; Hassenzahl and Tractinsky, 2006), it has rarely been systematically addressed. In two recent studies (Karapanos et al., 2008, 2009) we provided some first evidence that not only our perceptions, but also the relative weight of different product qualities change over time. So far, both academia and industry have largely focused on initial use. This has strong implications for the quality of interactive products. For instance, Den Ouden et al. (2006) found that an alarmingly increasing number of returned products, in 2002 covering 48% of all returned products, are technically fully functional (i.e. according to specifications), but they are returned on the basis of failing to satisfy users' true needs (28%), or purely on users' remorse (20%). These failures were not so much related to problems rooted in early interactions – problems that can often be overcome through learning – but to those that persist over time, pointing at failures to truly incorporate the product into daily life. We hope that iScale provides a first step towards retrospective elicitation methods as a viable, lightweight alternative to the expensive longitudinal methods.

Appendix A. Temporal transformation

In Section 4.2.3 we used users' consistency of temporal information of reported experiences across repeated recalls as a metric of reliability of the recall process. One question, however, relates to whether participants' accuracy in recalling temporal information remains constant across the full timeline, from the moment of purchase of the product to the present time. The participant's accuracy might be affected by the amount of contextual information

Please cite this article as: Karapanos, E., et al., Reconstructing experiences with iScale. International Journal of Human-Computer Studies (2012), http://dx.doi.org/10.1016/j.ijhcs.2012.06.004

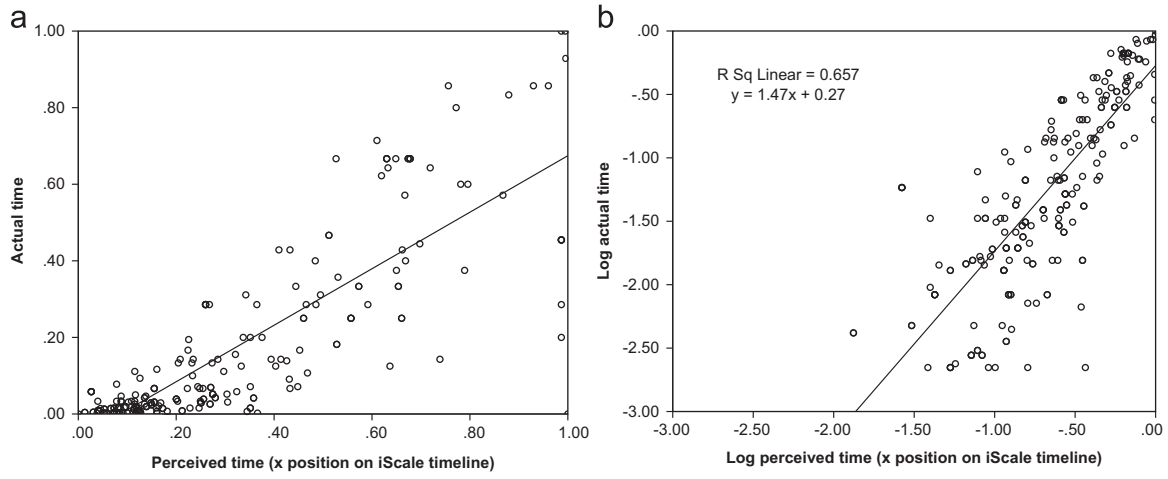

Fig. 10. Relation between actual time (reported time for a sketched node) and position of node along the iScale's *x*-axis: (a) linear relation and (b) power-law relation. Actual time (days) is adjusted to the full time of ownership of the product for each participant.

surrounding the experience that is available at the moment of recalling. Theories of recall have suggested that recent experiences (Koriat et al., 2000), or experiences associated with important milestones (e.g. the purchase of the product) (Barsalou, 1988) might be more easily accessible. If such biases exist, they will affect the reliability test as differences in the consistency of recalled information might be due to pertaining to more or less salient periods and not due to the reconstruction process. In the presence of such biases, the temporal distance between the two coupled experience reports elicited in the two distinct sessions should be transformed to account for the accessibility biases.

We attempt to assess the existence of accessibility biases through examining the way in which participants used the timescale of the tool (i.e. iScale's *x*-axis). Participants graphed linear curves through adding nodes in the graph (see Fig. 2a). Each node can be characterized by two properties: (a) the actual time (participants explicitly annotated for each node the approximate amount of days, weeks, or months after purchase that this node represents, and (b) the perceived time (the position of the node along the *x*-axis of iScale).

Fig. 10 depicts the relationship between the reported (actual) time versus the perceived time (i.e. the position of the node along iScale's *x*-axis). To enable an across-subject comparison, we normalized the reported (actual) time variable by the total time of ownership of the product for each participant, resulting to an index from 0 to 1. Given no accessibility bias, one would expect a linear relationship between these two pieces of information. One might note in Fig. 10a, however, that the variance in the dependent variable (actual time) is not uniformly distributed across the range of the independent variable (position along the x-axis of iScale). If one transforms the variables by the logarithmic function, the data become much more uniformly distributed. A linear regression on these transformed variables shows a significant prediction accounting for 66% of the variance in the dependent variable. This suggests a power law relation between the recalled actual time of the experienced event and its position along the graphing tool's timeline with a power equal to $1/1.47=0.68$ (i.e. perceived-time = actual-time$^{0.68}$). In other words, participants had a tendency to use a substantial fraction of the *x*-axis of iScale to map their initial experiences. In a similar vein, 95% of all experience reports related to the first six months of use and 75% of all experience reports related to the first month of use whereas the median time of ownership was 10 months. It thus becomes evident that experiences pertaining to initial use are more accessible in participants' memory. To account for this accessibility bias we compute the temporal distance between two events through formula 1.

References

Anderson, S., Conway, M., 1993. Investigating the structure of autobiographical memories. Journal of Experimental Psychology: Learning, Memory, and Cognition 19 (5), 1178–1196.

Barendregt, W., Bekker, M.M., Bouwhuis, D.G., Baauw, E., 2006. Identifying usability and fun problems in a computer game during first use and after some practice. International Journal of Human–Computer Studies 64, 830–846.

Barsalou, L., 1988. The content and organization of autobiographical memories. In: Neisser, U., Winograd, E. (Eds.), Remembering Reconsidered: Ecological and Traditional approaches to the Study of Memory. Cambridge University Press, Cambridge, pp. 193–243 Chapter 8.

Bartlett, F., 1932. Remembering. Cambridge, MA.

Bednarik, R., Myller, N., Sutinen, E., Tukiainen, M., 2005. Effects of experience on gaze behaviour during program animation. In: Proceedings of the 17th Annual Psychology of Programming Interest Group Workshop (PPIG'05), pp. 49–61.

Betsch, T., Plessner, H., Schwieren, C., Gutig, R., 2001. I like it but I don't know why: a value-account approach to implicit attitude formation. Personality and Social Psychology Bulletin 27, 242.

Blei, D., Ng, A., Jordan, M., 2003. Latent Dirichlet allocation. The Journal of Machine Learning Research 3, 993–1022.

Cohen, J., 1992. A power primer. Psychological Bulletin 112, 155.

Please cite this article as: Karapanos, E., et al., Reconstructing experiences with iScale. International Journal of Human-Computer Studies (2012), htt p://dx.doi.org/10.1016/j.ijhcs.2012.06.004

Conway, M., Pleydell-Pearce, C., 2000. The construction of autobiographical memories in the self-memory system. Psychological Review 107, 261–288.

Courage, C., Jain, J., Rosenbaum, S., 2009. Best practices in longitudinal research. In: Proceedings of the 27th International Conference Extended Abstracts on Human Factors in Computing Systems. ACM New York, NY, USA, pp. 4791–4794.

Den Ouden, E., Yuan, L., Sonnemans, P.J.M., Brombacher, A.C., 2006. Quality and reliability problems from a consumer's perspective: an increasing problem overlooked by businesses? Quality and Reliability Engineering International 22, 821–838.

Edvardsson, B., Roos, I., 2001. Critical incident techniques. International Journal of Service Industry Management 12, 251–268.

Erickson, T., 1996. The design and long-term use of a personal electronic notebook: a reflective analysis. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems: Common Ground, pp. 11–18.

Fenko, A., Schifferstein, H., Hekkert, P., 2009. Shifts in sensory dominance between various stages of user-product interactions. Applied Ergonomics 41, 34–40.

Flanagan, J., 1954. The critical incident technique. Psychological Bulletin 51, 327–358.

Fleiss, J.L., Levin, B., Paik, M.C., 2003. Statistical methods for rates and proportions. Wiley-Interscience.

Forlizzi, J., Battarbee, K., 2004. Understanding experience in interactive systems. In: Proceedings of the 2004 Conference on Designing Interactive Systems: Processes, Practices, Methods, and Techniques, pp. 261–268.

Fransella, F., Bell, R., Bannister, D., 2003. A Manual for Repertory Grid Technique. Wiley.

Gerken, J., Bak, P., Reiterer, H., 2007. Longitudinal evaluation methods in human–computer studies and visual analytics. In: InfoVis 2007 Workshop on Metrics for the Evaluation of Visual Analytics.

Goldschmidt, G., 1991. The dialectics of sketching. Creativity Research Journal 4, 123–143.

Groves, R., Fowler, F., Couper, M., Lepkowski, J., Singer, E., Tourangeau, R., 2009. Survey Methodology. John Wiley & Sons Inc.

Gutman, J., 1982. A means-end chain model based on consumer categorization processes. The Journal of Marketing, 60–72.

Hassenzahl, M., 2004. The interplay of beauty, goodness, and usability in interactive products. Human–Computer Interaction 19, 319–349.

Hassenzahl, M., Lai-Chong Law, E., Hvannberg, E., 2006. User experience—towards a unified view. UX WS NordiCHI 6, 1–3.

Hassenzahl, M., Tractinsky, N., 2006. User experience—a research agenda. Behaviour & Information Technology 25, 91–97.

Hassenzahl, M., Ullrich, D., 2007. To do or not to do: differences in user experience and retrospective judgements depending on the presence or absence of instrumental goals. Interacting with Computers 19, 429–437.

Hassenzahl, M., Wessler, R., 2000. Capturing design space from a user perspective: the repertory grid technique revisited. International Journal of Human–Computer Interaction 12, 441–459.

Hektner, J.M., Schmidt, J.A., Csikszentmihalyi, M., 2007. Experience Sampling Method: Measuring the Quality of Everyday Life. Sage Publications Inc.

Hsieh, H.F., Shannon, S.E., 2005. Three approaches to qualitative content analysis. Qualitative Health Research 15, 1277–1288.

Kahneman, D., 1999. Objective happiness. In: Kahneman, D., Diener, E., Schwarz, N. (Eds.), Well-being: The Foundations of Hedonic Psychology. Russel Sage Foundation, New York, pp. 3–25.

Kahneman, D., Krueger, A.B., Schkade, D.A., Schwarz, N., Stone, A.A., 2004. A survey method for characterizing daily life experience: the day reconstruction method Science, 306, 1776–1780. http://dx.doi.org/10.1126/science.1103572.

Karapanos, E., Hassenzahl, M., Martens, J.B., 2008. User experience over time. In: CHI '08 Extended Abstracts on Human Factors in Computing Systems. ACM, Florence, Italy, pp. 3561–3566.

Karapanos, E., Zimmerman, J., Forlizzi, J., Martens, J.B., 2009. User experience over time: an initial framework. In: CHI '09: Proceedings of the 27th International Conference on Human Factors in Computing Systems. ACM, New York, NY, USA, pp. 729–738.

Kjeldskov, J., Skov, M., Stage, J., 2010, A longitudinal study of usability in health care: does time heal? International Journal of Medical Informatics, 79 (6), p. 135–143.

Koriat, A., Goldsmith, M., Pansky, A., 2000. Toward a psychology of memory accuracy. Annual Review of Psychology 51, 481–537.

Krippendorff, K., 2004. Content Analysis: An Introduction to Its Methodology. Sage Publications.

Kujala, S., Roto, V., Väänänen-Vainio-Mattila, K., Karapanos, E., Sinnelä, A., 2011. Ux curve: a method for evaluating long-term user experience. Interacting with Computers 23, 473–483 (Feminism and HCI: New Perspectives).

Landauer, T., Dumais, S., 1997. A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychological Review 104, 211–240.

Light, A., 2006. Adding method to meaning: a technique for exploring peoples' experience with technology. Behaviour & Information Technology 25, 175–187.

Means, B., Nigam, A., Zarrow, M., Loftus, E., Donaldson, M., 1989. Autobiographical memory for health-related events. Vital and Health Statistics 6, 1–22.

Minge, M., 2008. Dynamics of user experience. In: Proceedings of the Workshop on Research Goals and Strategies for Studying User Experience and Emotion, NordiCHI '08.

Neisser, U., 1981. John Dean's memory: a case study. Cognition 9, 1–22.

Norman, D., 2009. THE WAY I SEE IT. Memory is more important than actuality. Interactions 16, 24–26.

Prümper, J., Zapf, D., Brodbeck, F., Frese, M., 1992. Some surprising differences between novice and expert errors in computerized office work. Behaviour & Information Technology 11, 319–328.

Redelmeier, D.A., Kahneman, D., 1996. Patients' memories of painful medical treatments: real-time and retrospective evaluations of two minimally invasive procedures. Pain 66, 3–8.

Reyna, V., Kiernan, B., 1994. Development of gist versus verbatim memory in sentence recognition: effects of lexical familiarity, semantic content, encoding instructions, and retention interval. Developmental Psychology 30 (2), 178–191.

Reynolds, T., Gutman, J., 1988. Laddering theory, method, analysis and interpretation. Journal of Advertising Research 28, 11–31.

Robinson, M., Clore, G., 2002. Belief and feeling: evidence for an accessibility model of emotional self-report. Psychological Bulletin 128, 934–960.

Schwarz, N., Kahneman, D., Xu, J., Belli, R., Stafford, F., Alwin, D., 2008. Global and episodic reports of hedonic experience. In: Calendar and Time Diary Methods in Life Course Research: Methods in Life Course Research, vol. 157.

Sonnemans, J., Frijda, N.H., 1994. The structure of subjective emotional intensity. Cognition & Emotion 8, 329–350 http://www.tandfonline.com/doi/pdf/10.1080/02699939408408945.

Strauss, A.L., Corbin, J.M., 1998. Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory. Sage Publications Inc.

Tulving, E., 2002. Episodic memory: from mind to brain. Annual Review of Psychology 53, 1–25.

Vaughan, M., Courage, C., Rosenbaum, S., Jain, J., Hammontree, M., Beale, R., Welsh, D., 2008. Longitudinal Usability Data Collection: Art Versus Science?.

von Wilamowitz-Moellendorff, M., Hassenzahl, M., Platz, A., 2006. Dynamics of user experience: how the perceived quality of mobile phones changes over time. In: User Experience—Towards a Unified View, Workshop at the Fourth Nordic Conference on Human–Computer Interaction, pp. 74–78.

von Wilamowitz-Moellendorff, M., Hassenzahl, M., Platz, A., 2007. In: Gross, T. (Ed.), Mensch & Computer 2007, pp. 49–58.

Please cite this article as: Karapanos, E., et al., Reconstructing experiences with iScale. International Journal of Human-Computer Studies (2012), http://dx.doi.org/10.1016/j.ijhcs.2012.06.004