Quantifying Diversity in User Experience

Evangelos Karapanos
The work described in this dissertation has been carried out under the auspices of the J.F. Schouten Graduate School of User-System Interaction Research, and has been sponsored by the Dutch Ministry of Economic Affairs within the IOP-IPCR program, through the ‘Soft Reliability’ project, under the auspices of Philips and Océ.

A catalogue record is available from the Eindhoven University of Technology Library

ISBN: 978-90-386-2183-8
NUR: 964

Cover design by Ehsan Baha
Typeset using \LaTeX\, printed in The Netherlands

©Evangelos Karapanos, 2010

All rights reserved. No part of this book may be reproduced, stored in a retrieval system, or transmitted, in any form, or by any means, electronic, mechanical, photocopying, recording or otherwise, without the prior permission of the author.
Quantifying Diversity in User Experience

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de
Technische Universiteit Eindhoven, op gezag van de
rector magnificus, prof.dr.ir. C.J. van Duijn, voor een
commissie aangewezen door het College voor
Promoties in het openbaar te verdedigen
op dinsdag 23 maart 2010 om 16.00 uur

door

Evangelos Karapanos

geboren te Athene, Griekenland
Dit proefschrift is goedgekeurd door de promotoren:

prof.dr.ir. J.-B.O.S. Martens
en
prof.dr.ir. A.C. Brombacher
[Our uses of] Statistics are like a bikini.
What they reveal is suggestive, but what they conceal is vital.

Aaron Levenstein [paraphrased]
# Contents

## I Introduction

1 Introduction

1.1 From usability to experience

1.2 Two distinct approaches in User Experience research

1.3 Diversity in User Experience

1.4 Scope of the Thesis - Methodological issues in accounting for Diversity

1.5 Thesis outline

## II Understanding Interpersonal Diversity through Personal Attribute Judgments

2 Personal Attribute Judgments

2.1 Introduction

2.2 The Repertory Grid Technique

2.3 The quantitative side of RGT - some concerns

2.4 Analyzing Personal Attribute Judgments - an initial exploration

2.5 The study

2.6 Analysis Procedure

2.7 Discussion

2.8 Conclusion

3 Analyzing Personal Attribute Judgments

3.1 Introduction

3.2 The study

3.3 An MDS approach to account for diversity
CONTENTS

3.4 Discussion ......................................................... 70
3.5 Conclusion ......................................................... 73

III Understanding the Dynamics of Experience through Experience Narratives ........................................... 75

4 User Experience Over Time ........................................ 77
  4.1 Introduction .................................................. 78
  4.2 Background on Experience and Temporality .......... 79
  4.3 Study 1 ......................................................... 81
  4.4 Study 2 ......................................................... 87
  4.5 Discussion ..................................................... 103
  4.6 Conclusion ..................................................... 104

5 iScale: a tool for the elicitation of experience narratives from memory ........................................... 107
  5.1 Introduction .................................................. 108
  5.2 Sketching and memory .................................... 110
  5.3 Study 1 ......................................................... 117
  5.4 Study 2 ......................................................... 128
  5.5 Discussion & Conclusion ................................ 143

6 A semi-automated approach to the content analysis of experience narratives .................................... 147
  6.1 Introduction .................................................. 147
  6.2 Automated approaches to semantic classification .... 149
  6.3 A semi-automated approach to content analysis .. 152
  6.4 Validation of the proposed approach .................... 161
  6.5 Discussion ..................................................... 168
  6.6 Conclusion ..................................................... 170

IV Reflections .......................................................... 171

7 Conclusions .......................................................... 173
  7.1 Contributions of this Thesis ................................ 173
  7.2 Implications for the product creation process .......... 179
  7.3 Avenues for future research ............................... 181
CONTENTS

V Bibliography 185

Bibliography 187

A Appendix 203

A.1 Descriptions to participants - Day Reconstruction & Experience Narration 203

A.2 Participants’ sketches in iScale across the two sessions 205

List of Figures 215

List of Tables 220

Summary 223

Acknowledgements 225

Biography 229

Publications related to this research 231

Journals 231

Proceedings 231

Informally reviewed 232
Part I

Introduction
48% of returned products are not attributed to a violation of product specifications (Den Ouden et al., 2006). This finding was the initial motivation for this research project. Brombacher, den Ouden and colleagues (e.g., Den Ouden et al., 2006; Brombacher et al., 2005; Koca et al., 2009) found that an alarmingly increasing number of returned products, in 2002 covering 48% of 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 because of users’ remorse (20%) (Den Ouden et al., 2006). Brombacher et al. (2005) introduced the term ‘Soft Reliability’ to refer to these situations where ”in spite of meeting with the explicit product specifications, a customer explicitly complains on the (lack of) functionality of the product”.

How is this finding different from Suchman’s well-known case at XEROX in the ’80s where users were found to have problems in using a feature-rich photocopier (c.f., Suchman, 2006)? While product designers were aiming at improving the instructions for using the product, Suchman argued that learning is an inherently problematic activity, and suggested that ”no matter how improved the machine interface or instruction set might be, this would never eliminate the need for active sense-making on the part of prospective users” (Suchman, 2006, p. 9). Since the ’80s, a wealth of theories, methods, and design guidelines have been developed in the field of Human-Computer Interaction with the aim of making products more easy to learn and use in the long run. Thus, one might wonder, do Brombacher’s and den Ouden’s findings replicate what was found at XEROX almost 30 years ago, or do they introduce a new and as yet unaddressed problem? Should these consumer complaints be attributed to bad design practices, to apparently inescapable interaction flaws in first-time use, or do...
they suggest a new emerging problem in the user acceptance of interactive products?

Den Ouden et al. (2006) identified a number of trends in the consumer electronics (CE) industry that have resulted in radical differences of the current market in comparison to that in the ’90s. They argued that the emphasis in the CE industry has shifted from the production of high volumes at competitive prices to the introduction of highly innovative products at higher prices. This leads to a shift in the main uncertainty in new product development projects; while in the ’90s the uncertainty related to the technology in relation to cost-effective mass production, in the current market the dominant uncertainty relates to the attractiveness of the product and users’ expectations about the product functions.

These trends are reflected in the development of the field of Human-Computer Interaction, from the study of usability as a critical factor to the acceptance of interactive products, to a more holistic understanding of users’ experiences with interactive products, leading to the study of new concepts like pleasure (Jordan, 2000), fun (Blythe et al., 2003), aesthetics (Tractinsky et al., 2000) and hedonic qualities (Hassenzahl, 2004). While a wealth of techniques and methods exist for ensuring the usability of interactive products, research on user experience evaluation methods is only at its infancy. This thesis aims at highlighting methodological issues in user experience evaluation and proposes a number of methods for inquiring into users’ experiences with interactive products.

1.1 From usability to experience

The field of Human-Computer Interaction was for a long time identified as the field of usability engineering. Usability was seen as critical to user acceptance and a wealth of principles (e.g., Norman, 1988), design guidelines (e.g., Nielsen and Bellcore, 1992) and evaluation techniques (e.g., Dix et al., 2004) have become instrumental in the development of usable products. The field of usability engineering readily acknowledged the dual nature of the usability concept: its objective and subjective side. One of the most dominant definitions of usability, for instance, the ISO 9241-11 standard (1996) defines usability as

"the extent to which a product can be used by specific users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use”.

Effectiveness represents the accuracy and completeness with which users achieve certain goals and is typically measured through observed error rates, binary task completion, and quality of the outcome for a given task (see Frøkjær et al., 2000; Hornbæk and Law, 2007). Efficiency can be characterized as effectiveness in relation to resources spent and is typically measured through task completion time and
learning time. One can note that both these components, effectiveness and efficiency represent the objective side of usability. The third component, user satisfaction, represents users’ comfort in using and overall attitude to the product and is typically measured through psychometric scales on overall preference, product quality perceptions and specific attitudes towards the interface (see Hornbæk and Law, 2007).

An assumption underlying the distinction between the subjective and objective side of usability was that these two would strongly correlate. Nielsen and Levy (1994), in a meta-analysis of a selected set of 57 studies found that in 75% of the cases, users’ overall preference was strongly related to overall performance. These findings have, however, been repeatedly questioned in subsequent studies suggesting that subjective perceptions of usability are generally not correlated with objective measures and seem to measure something else than merely effectiveness and efficiency (Frøkjær et al., 2000; Hornbæk and Law, 2007; Kissel, 1995; Bailey, 1993). This limited view on user satisfaction as a consequence of objective performance was criticized by Hassenzahl et al. (2000):

"...it seems as if satisfaction is conceived as a consequence of user experienced effectiveness and efficiency rather than a design goal in itself. This implies that assuring efficiency and effectiveness alone guarantees user satisfaction."

Subsequently, a set of studies tried to explain these observed discrepancies between objective and subjective usability. Kurosu and Kashimura (1995) asked participants to rate several Automatic Teller Machine (ATM) designs on both functional and aesthetic aspects. They found apparent usability, i.e. subjective judgments on usability, to correlate more strongly with aesthetic judgments than with the systems’ inherent usability, i.e. objective design parameters that were expected to affects users’ performance in using the systems. Tractinsky (1997) and Tractinsky et al. (2000) replicated this study and found that these effects persisted both across different cultural backgrounds, as well as after participants had experienced the systems.

These early findings suggested that users’ experiences with products go beyond the effectiveness and efficiency in product usage. Consequently, the field of Human-Computer Interaction quested for new concepts, measures and methods in capturing a more holistic view on user experience. This development has gone hand-in-hand with a shift in the contexts of study, from professional to personal (e.g., Jordan, 2000) and social (e.g., Forlizzi, 2007; Markopoulos et al., 2004), and in the design paradigm from product to experience design (e.g., Buxton, 2007; Zimmerman et al., 2007; Forlizzi et al., 2008).
CHAPTER 1. INTRODUCTION

1.2 Two distinct approaches in User Experience research

User experience has become central to the design and evaluation of interactive products. It reflects a paradigm shift in the subject of product design and evaluation. Buxton (2007) argues the following:

“Ultimately, we are deluding ourselves if we think that the products that we design are the “things” that we sell, rather than the individual, social and cultural experience that they engender, and the value and impact that they have. Design that ignores this is not worthy of the name”

However, user experience research is often criticized for at least two things: a) for the lack of a commonly agreed definition of the notion of experience, and b) for being identical, conceptually or methodologically, to traditional usability research. Indeed, Hassenzahl (2008) and Wright and Blythe (2007), some of the strong proponents of user experience research, criticize the use of the term user experience in cases where the focus still lies in traditional usability evaluation, thus reducing the richness of experience to behavioral logs and task-focused evaluations (Wright and Blythe, 2007).

As Hassenzahl (2008) argues:

“While UX seems ubiquitous in industry, a closer look reveals that it is treated mainly as a synonym of usability and user-centered design”

Conversely, as Hassenzahl (2008) argues, academics “emphasize the differences between traditional usability and user experience”. A number of frameworks have tried to conceptualize how experiences are formed (e.g., Forlizzi and Ford, 2000; Wright and McCarthy, 2004; Norman, 2004; Hassenzahl, 2008) and tentative definitions of experience have been proposed (Forlizzi and Battarbee, 2004; Hassenzahl and Tractinsky, 2006; Hassenzahl, 2008).

Hassenzahl (2008) defines user experience as “a momentary, primarily evaluative feeling (good-bad) while interacting with a product or service” which is “a consequence of users’ internal state (e.g. predispositions, expectations, needs, motivation, mood), the characteristics of the designed system (e.g. complexity, purpose, usability, functionality), and the context within which the interaction occurs (e.g. organizational/social setting, meaningfulness of the activity, voluntariness of use)” (Hassenzahl and Tractinsky, 2006). One may argue that such a definition, while being perhaps the best of what the field of user experience can offer at the moment, is far from being mature or useful for grounding measures, methods and principles in the design and evaluation of interactive products. Yet, a common ground has been established among various disciplines and schools of thoughts in the emerging field of user experience, perhaps due to “a history of use of the term in ordinary conversation and philosophy discourse” (Wright and Blythe, 2007). A number of researchers have tried
1.2. TWO DISTINCT APPROACHES IN USER EXPERIENCE RESEARCH

to identify the dominant schools of thought and several classifications have been proposed (e.g., Battarbee and Koskinen 2005; Hassenzahl and Tractinsky 2006; Blythe et al. 2007).

We employ the distinction from Blythe et al. (2007) between reductionist approaches that have their roots in cognitive psychology, and holistic approaches that are grounded in pragmatist philosophy and phenomenology. As it will become apparent in section 1.4, we are primarily interested in distinct issues that these two approaches pose when one is concerned about methodology for understanding user experiences.

1.2.1 Reductionist approaches

Reductionist approaches in user experience maintain a similar paradigm to usability (ISO 1996) and Technology Acceptance research (see Venkatesh et al. 2003) in trying to identify distinct psychological constructs and propose and empirically test causal relations between them.

One of the first and well cited studies in user experience, grounded on reductionism, is that of Tractinsky (1997). Tractinsky was puzzled by the findings of Kurosu and Kashimura (1995) who suggested that subjective perceptions of usability relate more to the beauty of the product than to its actual, i.e. inherent, usability. This finding highlighted the importance of aesthetics in interactive products, an opinion that found Tractinsky resonant with. Tractinsky, however, predicted that this might not pertain over different cultures, taking into account that the study of Kurosu and Kashimura (1995) was conducted within the Japanese culture which is known for its aesthetic tradition. Tractinsky (1997) replicated the experiment using the same stimuli, but now in an Israeli context. His initial prediction was not confirmed as the findings of Kurosu and Kashimura (1995) were reconfirmed in this alternative setting. One possible criticism of both studies could be that the user judgments were elicited merely on the basis of the visual appearance of the interface without experiencing the systems. In a subsequent study, Tractinsky et al. (2000) elicited users’ perceptions both before and after interacting with a computer simulation of an interface of an Automatic Teller Machine (ATM). The results suggested that the aesthetics of the interface also impacted the post-use perceptions of usability. Subsequent work has supported the dominance of beauty in users’ preferences (e.g., Schenkman and Jönsson 2000; Lindgaard and Dudek 2003; Lindgaard et al. 2006; Tractinsky et al. 2006; Hekkert and Leder 2008) and have provided further insight into users’ inferences between aesthetics and usability (e.g., Hartmann et al. 2008).

Hassenzahl (2004) wanted to further inquire into the nature of beauty in interactive products. He developed a theoretical model (Hassenzahl 2005) that distin-
guishes between objective parameters, product quality perceptions and overall evaluations. Based on this model he understood beauty as "a high-level evaluative construct comparable to (but not identical with) other evaluative constructs, such as goodness or pleasantness" (Hassenzahl, 2004, p. 323) and perceived usability as a bundle of lower level judgments reflecting product quality perceptions. He distinguished between two quality perceptions: pragmatic and hedonic. Pragmatic quality, he argued, refers to the product’s ability to support the achievement of behavioral goals (i.e. usefulness and ease-of-use). On the contrary, hedonic quality refers to the users’ self; it relates to stimulation, i.e. the product’s ability to stimulate and enable personal growth, and identification, i.e. the product’s ability to address the need of expressing one’s self through objects one owns. He further distinguished between two overall evaluative judgments: goodness and beauty. Contrary to (Tractinsky et al., 2000), he found minimal correlation between pragmatic quality, i.e. usability, and beauty. Beauty was found to be a rather social aspect, largely affected by identification; pragmatic quality, on the contrary, related to the overall judgment of goodness.

In a similar vein, Tractinsky and Zmiri (2006) distinguished between satisfying and pleasant experience. They found perceptions of usability to be better predictors of satisfying rather than pleasant experience while perceptions of products’ aesthetics to be better predictors of pleasant rather than satisfying experience. Hassenzahl’s (2004) model of user experience has also been further supported by subsequent research (e.g., Mahlke, 2006; Schrepp et al., 2006; Van Schaik and Ling, 2008). Mahlke and Thüring (2007) provided a comprehensive framework linking product quality perceptions to emotional reactions and overall evaluative judgments. Their findings supported Hassenzahl’s (2004) distinction between goodness and beauty, with goodness relating primarily to instrumental qualities, e.g. usefulness and ease-of-use, and beauty relating primarily to non-instrumental qualities such as the visual aesthetics and haptic quality (Mahlke, 2006; Desmet, 2002), grounded on Russell’s (1980) model of affect, developed a tool for measuring emotional responses to products and established a framework that relates aesthetic response to meaning (see Desmet and Hekkert, 2007). Fenko et al. (2009) studied how the dominance of different sensory modalities such as vision, audition, touch, smell and taste develops over different phases in the adoption of the product such as when choosing the product in the shop, during the first week, after the first month, and after the first year of usage.

1.2.2 Holistic approaches

Holistic approaches are rooted in pragmatist philosophy and phenomenology. They criticize reductionist approaches in that they reduce the complexity and richness of
user experience to "a set of manipulable and measurable variables" and impose "ab-
stract models and classifications onto rich and complex models like affect and emo-
tion" (Wright and Blythe 2007) (see also Hassenzahl 2008). Similarly, Suri (2002)
argues that "measurement, by its nature, forces us to ignore all but a few selected
variables. Hence, measurement is useful when we are confident about which vari-
ables are relevant". She argues that designers are concerned about developing new
products and for new contexts and thus no such understanding exists about how
design attributes and contextual details interact in given contexts, and proposes al-
ternative methods, such as that of experience narratives, for inquiring into how product
meaning and value emerges in given contexts.

From a theoretical point of view, holistic approaches have contributed a number
of frameworks describing how experience is formed, adapted, and communicated in
social contexts.

Forlizzi and Ford (2000) provided a framework that attempts to describe how
experience transcends from unconsciousness to a cognitive state and finally be-
comes "an experience", something memorable that can also be communicated in so-
cial interactions. They identified four modes or dimensions of experiencing: sub-
consciousness, cognition, narrative and storytelling. Sub-consciousness represents
fluent experiences that do not compete for our attention. Cognition represents expe-
riences that require our attention, e.g. learning to use an unfamiliar product. Narra-
tive represents "experiences that have been formalized in the users’ head: ones that
force us to think about and formulate what we are doing and experiencing". Forl-
izzi and Ford (2000) suggest that a product’s set of features and affordances offers a
narrative of use. Storytelling, represents the subjective side of experience: "a person
relays the salient parts of an experience to another, making the experience a personal
story". Forlizzi and Ford (2000) argue that through these particular sense making
users attach meaning and personal relevance to situation, "creating life stories and
stories of product use". Forlizzi and Ford (2000) subsequently identify ways of shift-
ing across these four modes of experiencing. One might migrate from a cognitive
to a sub-conscious experience, for instance by learning how to use a product. Re-
versely, a fluent experience may shift to a cognitive if a user encounters something
unexpected in her interaction with the product and is forced to think about it. A nar-
rative experience can shift to a cognitive one when one "is forced to challenge his own
thinking that has been solidified in her perceptions, attitudes, and beliefs". An expe-
rience might also shift from a sub-conscious state to story-telling as she “schematizes
it, communicates it and add levels of meaning".

Forlizzi and Battarbee (2004) modified this framework to include the concept of
cO-experience proposed by Battarbee (2003). Battarbee and Koskinen (2005), further
elaborated on the social mechanisms that lift or downgrade experiences as they par-
participate in people's social interactions. They identified three distinct mechanisms: lifting up experiences, reciprocating experiences, rejecting and ignoring experiences. First, people may "lift things from the stream of events", considering them as meaningful enough to be communicated in social settings. Secondly, recipients of communicated experiences may acknowledge the described experience as personally relevant and respond to it by telling their own, similar experiences. Finally, experiences communicated in social settings may be rejected or downgraded by others, eventually altering the dominance of the given experience for the person who chose to communicate it.

McCarthy and Wright (2004) distinguished between four threads of experience: compositional, sensual, emotional, and spatio-temporal. The compositional thread concerns the way that different elements of experience form a coherent whole. It refers to "the narrative structure, action possibility, plausibility, consequences and explanations of actions". The sensual thread relates to "the concrete, palpable, and visceral character of experience that is grasped pre-reflectively in the immediate sense of a situation". The emotional thread refers to value judgments (e.g., frustration and satisfaction) that ascribe importance to other people and things with respect to our needs and desires". Lastly, the spatio-temporal thread "draws attention to the quality and sense of space-time that pervades experience". McCarthy and Wright (2004) pinpoint that while these are positioned as distinct components of experience they should be seen as intrinsically connected with each other.

Next to the four threads of experience, McCarthy and Wright (2004) described how sense-making takes place in the development of experience by decomposing it into six processes: anticipating, connecting, interpreting, reflecting, appropriating, and recounting. Anticipation refers to users' expectations and imagined possibilities that are grounded in prior experience. In connecting, users make an instant judgments referring to the immediate, pre-conceptual and pre-linguistic sense of a situation. In interpreting, users work out what's going on and how they feel about it. In reflecting users examine and evaluate what is happening in an interaction and the feelings of frustration or pleasure that are part of the experience. In appropriating, users evaluate how the new experience relates to prior experiences, and in recounting, users communicate the experienced situation to others and reinterpret the experience as it participates in storytelling.

1.3 Diversity in User Experience

While one may quickly note that holistic approaches emphasize the uniqueness of experience across different situations and people, both approaches to user experience have readily accepted that diversity in users' experiences is prevalent. A wealth of empirical studies, grounded on reductionism, acknowledge and tackle diversity
empirically (e.g. Hassenzahl and Ullrich, 2007; Mahlke and Lindgaard, 2007).

This section introduces the notion of diversity in user experience. We introduce a framework of diversity in subjective judgments and identify four different sources of diversity in users’ experiences with interactive products. Only later, in section 1.4, we will introduce the methodological differences between the two approaches, reductionist and holistic, in accounting for diversity.

1.3.1 A framework of diversity in subjective judgments

Diversity was readily accepted in the HCI field as a key issue. Not all users like the same things and different product qualities suffice in different situations (e.g. Cooper, 1999). But, other fields have been constructed on the assumption of homogeneity across different individuals. In the field of psychophysics, for example, the principle of homogeneity of perception states that different participants will more or less agree on perceptual judgments such as how much noise, or blur, an image contains, or how much friction, or inertia, one may find in a haptic control. This assumption has been instrumental in the development of respective statistical techniques; for instance, Multi-Dimensional Scaling (Green et al., 1989; Martens, 2003), motivated by this principle, assumes that judgments of different individuals may be visualized in a shared K-dimensional configuration of stimuli, for which the coordinates of the stimuli in the configuration space along different axes can be monotonically related to the observed attribute ratings of the participants.

It seems natural to accept that while different individuals might agree on low-level perceptual judgments, e.g. friction in a haptic control, a relative disagreement would be found as one moves to more cognitive judgments, e.g. the degree to which this haptic control is perceived as playful. Indeed, in an exploratory study we tried to inquire into whether people agree on product character judgments (Janlert and Stoltzman, 1997) of a physical rotary knob. A project done by Bart Friederichs in his Master’s studies, in collaboration with BMW, aimed to prove that physical controls, i.e. rotary knobs, can be designed so that they comply with the personality of a given car. But, how much do different users’ agree on high level product character judgments of haptic controls? Eighteen participants experienced fifteen different haptic scenes programmed in a haptic feedback knob and rated them in terms of four different bipolar scales: Dynamic - Static, Agile - Plump, Playful - Serious, and Luxurious - Austere. One may quickly note that the last two judgments, those of playfulness and luxury, are higher level judgments, allowing for more disagreement on what one considers playful or luxurious. We applied a uni-dimensionality test, by applying two Multi-Dimensional Scaling solutions on the four ratings: a 1D and a 2D. A $\chi^2$ test, showed that the first two judgments, which we assumed as lower level, per-
Diversity Product Character Perceptual Judgments Product Appraisal

Figure 1.1: A modification of Hassenzahl’s framework, highlighting diversity at two stages in forming overall judgments about products.

ceptual judgments, were indeed more uni-dimensional, i.e. they displayed higher consistency across different individuals.

In the field of design, Csikszentmihalyi and Rochberg-Halton (1981) asked individuals to select personally relevant objects located in their homes and describe what makes them special. Csikszentmihalyi and Rochberg-Halton (1981) found that the value of such objects did not lie in some objectively defined quality, e.g. uniformly appreciated aesthetics, but to the personal meaning that people attached to these objects and how these participated in their social lives and creation of self-identity. These results suggest that while we may all agree in perceptual judgments, e.g. color of a given product, these judgments have far less power in predicting our preferences in comparison to higher level judgments, e.g. its beauty. Hofstede (2001) suggested that human perceptions may exist at three different levels: a) some might be uniquely personal, having significance for an individual because of his or her own associations; b) others might have significance to a specific social or cultural group through shared meaning; and c) others are universal, related to human nature at an innate or fundamental level.

In Karapanos et al. (2008b) we argued that diversity may exist at two different stages in the formation of an overall evaluative judgment (see figure 1.1). Perceptual diversity lies in the process of forming product quality perceptions (e.g. novel, easy to use) on the basis of product features. For instance, different individuals may infer different levels on a given quality of the same product, e.g. disagree on its novelty. Evaluative diversity lies in the process of forming overall evaluations of the product (e.g. good-bad) on the basis of product quality perceptions. For instance, different individuals may form different evaluative judgments even while having no disagreement on the perceived quality of the product, e.g. both might think of it as a novel and hard-to-use product, but they disagree on the relative importance of each quality.

Considering the second stage, i.e. evaluative diversity, one might assume a certain universal hierarchical structure on the relative importance of different qualities. For
instance, Jordan (2000) drew on Maslow’s (1946) theory of human needs to propose a fixed hierarchical structure of the relative importance of functionality, usability and pleasure in the context of Human-Computer Interaction. According to this, a product has to provide useful and usable functionality before hedonic aspects, such as beauty and stimulation, can take effect. Similarly to the field of psychology where Maslow’s theory was readily adopted while lacking empirical evidence, Jordan’s fixed hierarchy has become widely popular in the field of user experience, but, to our knowledge, no empirical studies have attempted to confirm or disprove the framework. Contrary to Jordan (2000), Hassenzahl (2006) assumes the importance of these different qualities to vary with several contextual factors.

### 1.3.2 Four sources of diversity in user experience

In the field of user experience, a wealth of empirical studies have tried to tackle how the relative importance of different product qualities on users’ overall evaluations and preferences, is modulated by a number of contextual aspects. In figure 1.2 we highlight four different sources of diversity that user experience research has been concerned with.

First, individual differences (e.g. human values Schwartz, 1992) moderates the importance individuals attach to different qualities of an interactive product (e.g. Desmet et al., 2004; Karapanos and Martens, 2007); while some might prefer playful and stimulating products, others might value simplicity and austerity. Second, the type of the product matters (e.g. Jordan and Persson, 2007); while a playful interaction might be crucial for the success of a computer game, the same quality might be perceived as inadequate for professional software. Third, even for the same product, the way individuals use it differs across situations and this impacts the importance that they attach to different qualities (e.g. Mahlke and Lindgaard, 2007; Hassenzahl and Ullrich, 2007; Hassenzahl et al., 2008); the same mobile phone could be used for exploring the available ring tones or to make an emergency call.

A fourth aspect, which is mostly overlooked so far, is the systematic change of experience over time. As individuals use a product, their perception of the qualities of the product will change (e.g. von Wilamowitz Moellendorff et al., 2006; Karapanos et al., 2008a, 2009; Fenko et al., 2009). For example, they get used to it, which eventually changes their perception of its usability; at the same time it excites them much less than initially. Even more interestingly, at different phases of use they will evidently attach different weights to different qualities. In their first interactions with a product they may focus on its usability and stimulation. After they use it for some time, they might become less concerned about its usability, and other aspects of the product such as novel functionality or communication of a desired identity towards
Diversity in UX

Individual
Product
Situation
Time

Figure 1.2: Four different sources of diversity in user experience. These are assumed to modulate the dominance of different product qualities on users’ experience and overall evaluative judgments.

others become more important.

All these factors, the individual, the product, the situation, and time modify the importance of the qualities for a satisfying experience with an interactive product. In this view, Jordan’s (2000) hierarchy of consumer needs could be seen as a particular, context-dependent prioritization of needs (Hassenzahl, 2006).

This section introduced the notion of diversity, which mostly concerns in the user experience field the relative dominance of different product qualities while forming overall judgments about interactive products. The next section introduces the methodological debate in accounting for diversity between the two dominant approaches in user experience research, it discusses the limitations of each approach, and argues for a hybrid methodological paradigm that shares values from both approaches, the holistic and the reductionist.

1.4 Scope of the Thesis - Methodological issues in accounting for Diversity

Reductionist approaches typically employ psychometric scales (e.g. Likert, 1932; Osgood et al., 1957) in measuring latent psychological constructs that the researchers consider as relevant for a certain context. Such approaches are grounded on the assumption that people may summarize experiences in overall evaluative judgments (Hassenzahl, 2008). When a product is associated with a certain experience, the value of the experience will be partially attributed to the product.

Proponents of the holistic paradigm criticize reductionist approaches for not capturing "grey areas and moment-by-moment decision making" (Wright and Blythe, 2010).
1.4. SCOPE OF THE THESIS - METHODOLOGICAL ISSUES IN ACCOUNTING FOR DIVERSITY

Figure 1.3: A simplified framework of the design process, derived from (Martens, 2009b, personal communication) depicted in Appendix A, figure A.1. According to this framework, the designer forms beliefs about the potential consequences of use of a given product in a given context, grounded on empirical knowledge. These beliefs are in turn externalized to artifacts and the artifacts are validated on a number of criteria that have been derived from prior beliefs, or empirical insights from the previous design iteration.

Instead, they propose holistic and situated techniques that are motivated by the need to establish an empathic relation with the user (see Wright and McCarthy, 2008). In one of the most popular techniques, the cultural probes (Gaver et al., 1999), participants are typically provided with a camera and a set of abstract objects to inspire them in capturing their experiences, dreams and aspirations related to a given topic of interest. Suri (2002) proposed the elicitation of short essays, i.e. experience narratives, in understanding the frustration and enjoyment that is experienced while interacting with products. Blythe et al. (2002) proposed a technique called technology biographies in which participants are asked to reminisce how their relationship with technology has changed through their lives. In the creation of Anticipation and reflection interviews, (Blythe et al., 2006) acknowledge that experience is not limited to a given interaction instance but extend to the process of forming expectations of and reflecting on experiential episodes.

One might argue that these two approaches have distinct goals, being evaluative or inspirational. In this sense, holistic approaches to evaluation serve to inspire design solutions (Hassenzahl, 2008); evaluation is here understood as idea generation (Hornbæk, 2008). As Hassenzahl (2008) puts it "[designers] are able to build ideas from anecdotal observations and loose associations". In contrast, reductionist approaches to evaluation serve to assess the value of a design, to compare multiple designs, or to develop theory and criteria to support evaluation.

This thesis argues against this distinction between evaluative and inspirational
goals in product evaluation. It is argued that current evaluation practices that embrace this distinction lead to inefficacy in the evaluation of interactive products. This is reflected in figure 1.3 which depicts a simplification of a framework of the design process initially proposed by Martens (2009b, personal communication). The framework identifies three distinct activities in an iterative design process: a) extrapolating knowledge from context and forming beliefs about the potential consequences of use of a given product within a given context, b) transforming (implicit) beliefs to explicit knowledge through the design of artifacts, and c) validating the designed artifacts in context.

According to this view, while holistic approaches serve to extrapolate knowledge in forming or adapting the designer’s beliefs, reductionist approaches serve to validate the designed artifacts using a number of criteria that have been derived from prior beliefs, or empirical insights from the previous design iteration. This leads to a number of limitations in the evaluation process. First, reductionist approaches to evaluation through pre-defined measurement scales may miss potentially relevant concepts due to a failure of researchers to recognize their relevance, but may also result in less meaningful information when participants cannot interpret the questions with respect to their own context. Second, holistic approaches such as the experience narratives, provide rich information in all aspects that surround an experience but may create a risk of focusing on idiosyncratic experiences while failing to estimate its dominance and probability of occurrence over the total population of users.

In combination, rich situated insights derived from holistic methods may inform the validation of artifacts while reductionist approaches to evaluation may quantify the importance of a given experience and thus minimize the risk of overemphasize interesting but rare experiences.

This thesis argues for a missing link between validation and extrapolation (see figure 1.3). Below, we describe how this is addressed in relation to the two research foci of this thesis: understanding interpersonal diversity in users’ responses to conceptual designs, and understanding the dynamics of experience over time.

1.4.1 Understanding Interpersonal Diversity through Personal Attribute Judgments

Den Ouden (2006) revealed that the majority of soft reliability problems related to the concept design phase and were particularly rooted in design decisions relating to the product definition. This insight suggests that design decisions made early in the design process may not be adequately grounded on empirical user insights.

---

1The initial framework may be found in the Appendix, figure A.1.
1.4. SCOPE OF THE THESIS - METHODOLOGICAL ISSUES IN ACCOUNTING FOR DIVERSITY

Traditional approaches to measuring users’ responses to artifacts lie in the a-priori definition of the measures by the researchers. This approach is limited in at least two ways when one is concerned with capturing the richness of and diversity in user experience. First, the a-priori definition of relevant dimensions is inherently limited as researchers might fail to consider a given dimension as relevant, or they might simply lack validated measurement scales, especially in developing fields such as that of user experience where radically new constructs are still being introduced. Secondly, one could even wonder whether rating a product on quality dimensions that are imposed by the researcher is always a meaningful activity for the user, for example when the user does not consider a quality dimension as relevant for the specific product. There is increasing evidence that users are often unable to attach personal relevance to the statement provided in psychometric scales due to a failure to recall experiential information that relates to the statement or due to lengthy and repetitive questioning. Larsen et al. (2008b) reviewed a number of studies employing psychometric scales in the field of Information Systems. They found for the majority of studies the semantic similarity between items to be a significant predictor of participants’ ratings (0.00 < $R^2$ < 0.63). In such cases, they argued participants are more likely to have employed shallow processing (Sanford et al., 2006), i.e. responding to surface features of the language rather than attaching personal relevance to the question.

An alternative approach to predefined questionnaires lies in a combination of structured interviewing, that aims at eliciting attributes that are personally meaningful for each participant, with a subsequent rating process performed on the attributes that were elicited during the interview. This approach aims at increasing the diversity and relevance to the individual of concepts that are measured, thus resulting in richer insights. However, the techniques required for the quantitative analysis of such information become significantly more complex. This thesis proposes two techniques for the analysis of personal attribute judgments.

1.4.2 Understanding the Dynamics of Experience through Experience Narratives

Product evaluation practices have traditionally been focusing on early interactions. As a result, we have been mostly concerned about product qualities relating to the initial use. Den Ouden et al. (2006), however, highlighted that the reasons for product returns span a wider range of aspects than just problems related to their learnability and usability. Moreover, a number of recent trends are highlighting the importance of longitudinal evaluation practices. First, legislation and competition within the consumer electronics industry has resulted in an increase in the time-span of product warranties. This has resulted in an alarmingly increasing number of products being
CHAPTER 1. INTRODUCTION

returned on the basis of failing to satisfy users’ true needs (Den Ouden et al., 2006). Secondly, products are increasingly becoming service-centered. Often, products are being sold for lower prices and revenues are mainly coming from the supported service (Karapanos et al., 2009). Thus, the overall acceptance of a product shifts from the initial purchase to establishing prolonged use.

Traditional approaches in the study of dynamics of experience over time typically employ validated measurement and structural models across different phases in the adoption of a system (e.g. Venkatesh and Davis, 2000; Venkatesh and Johnson, 2002; Kim and Malhotra, 2005). For instance, Venkatesh and Davis (2000) employed the Technology Acceptance Model (Davis et al., 1989) at three instances in the adoption of an information system at a workplace: before the introduction of the system (inquirying into users’ expectations), right after the introduction of the system, and three months after the introduction. These approaches, while being widely validated and well-cited in the field of Information Systems, are hindered by a number of limitations, at least in developing fields such as that of user experience.

An assumption inherent in this approach is that the relevant latent constructs remain constant, but their perceived value and relative dominance might change over time. But, especially in developing fields such as user experience, substantial variations might occur over time even in what constructs are relevant to measure. Some constructs, e.g. novelty, might cease to be relevant while others, such as supporting daily rituals, enabling personalization, and communicating a positive image about one’s self (see Karapanos et al., 2009), that were not evident in studies of initial use might become critical for the long-term acceptance of a product. Firstly, this might challenge the content validity of the measurement model as relevant latent constructs might be omitted. This is often observed in studies of user acceptance and user experience where the a-priori defined constructs account for a limited amount of the variance in the predicted variable, being it preference judgments, dissimilarity judgments or attitude towards behavior (see Venkatesh et al., 2003). Secondly, it may also lead to distorted data as individuals might fail to interpret the personal relevance of a given scale item to their own context, for instance when a construct ceases to be relevant over prolonged use. Last, such approaches provide rather limited insight into the exact reasons for changes in users’ experiences. They may, for instance, reveal a shift in the dominance of perceived ease-of-use and perceived usefulness on intention to use a product (e.g. Venkatesh and Davis, 2000), but provide limited insight to the exact experiences that contributed to such changes.

An alternative approach for the measurement of the dynamics of experience over time relies on the elicitation of idiosyncratic self-reports of one’s experiences with a product, i.e. experience narratives. Each narrative provides rich insights into a given experience and the context in which it takes place. Moreover, generalized knowledge
may also be gained from these experience narratives. Such generalized knowledge may be reflected in questions like: how frequent is a certain kind of experience, what is the ratio of positive versus negative experiences and how does this compare to competitive products, how does the dominance of different product qualities fluctuate over time and what should be improved to motivate prolonged use?

This thesis makes two methodological contributions in this research problem. First, it highlights the labor-intensive nature of longitudinal studies, and proposes an alternative approach that relies on the elicitation of one’s experiences with a product from memory. iScale, a tool designed with the aim of increasing users’ effectiveness and reliability in recalling their experiences is theoretically grounded and empirically validated. Second, it proposes a computational approach that aims at supporting the researcher in the qualitative analysis of experience narratives. The proposed approach addresses two limitations of traditional qualitative analysis practices. First, qualitative analysis is a labor intensive activity which becomes increasingly a concern when qualitative data may be elicited from a large number of participants as in the case of iScale. Second, qualitative analysis has been shown to be prone to researcher bias as humans often rely on heuristics in forming judgments about the relevance or similarity of two or more data instances [Kahneman et al., 1982]. The proposed approach aims at supporting the researcher through semi-automating the process of qualitative coding, but also minimizes the risks of overemphasizing interesting, but rare experiences that do not represent users’ typical reactions to a product.

1.5 Thesis outline

This thesis argues for a hybrid paradigm between reductionist and holistic approaches to evaluation. It proposes two techniques, one grounded in personal attribute judgments and one in experience narratives. Both developed techniques aim at increasing the richness and diversity in obtained information while trying to create different levels of granularity of insight, thus enabling the researcher to move between abstracted, generalized insight and concrete, idiosyncratic and insightful information.

Part I

Understanding Interpersonal Diversity through Personal Attribute Judgments

Chapter 2 highlights the limitations of standardized psychometric scales and introduces personal attributes judgments. It introduces attribute elicitation techniques and in particular, the Repertory Grid Technique (RGT). It argues that the true value of RGT is in quantifying rich qualitative insights and highlights the limitations of relevant statistical techniques that are typically employed in the analysis of Repertory
Grid data. It presents an initial Multi-Dimensional Scaling procedure that aims at identifying diverse views in Repertory Grid data. The procedure identifies distinct user groups in a sample population and derives a two-dimensional view for each respective user group. The technique is presented through a case study where users’ views on a set of product concepts were contrasted to the ones of designers.

Chapter 3 presents a second Multi-Dimensional Scaling procedure that aims at identifying diverse views even within single individuals. The technique is applied on an existing dataset (Heidecker and Hassenzahl, 2007). It is illustrated that the traditional - averaging analysis provides insight to only 1/6th of the total number of attributes in the example dataset. The proposed approach accounts for more than double the information obtained from the average model, and provides richer and semantically diverse views on the set of stimuli.

Part II
Understanding the Dynamics of Experience through Experience Narratives

Chapter 4 presents two studies that inquired into how users experiences with interactive products develop over time. In the first pilot study, grounded on reductionism, we asked participants to rate a novel product during the first week as well as after four weeks of use. In the second study six participants were followed after the purchase of a novel product and elicited rich experience narratives over a period of one month.

Chapter 5 presents iScale, a survey tool that aims at eliciting users’ experiences with a product in the form of experience narratives. iScale employs sketching in imposing a process in the reconstruction of one’s experiences from memory. The chapter motivates the development of two distinct versions of iScale which were grounded in two opposing theoretical approaches to reconstructing one’s emotional experiences from memory. Finally, it presents two studies that compared the two different versions of iScale with traditional methods.

Chapter 6 proposes a semi-automated technique for the content analysis of experience narratives. The technique combines traditional qualitative coding procedures (Strauss and Corbin, 1998) with computational approaches for assessing the semantic similarity between documents (Salton et al., 1975). This results in an iterative process of qualitative coding and visualization of insights which enables to move quickly between high-level generalized knowledged and concrete and idiosyncratic insights.

Chapter 7 concludes the research described in this thesis by reflecting on its contributions and sketching directions for future research.
Part II

Understanding Interpersonal Diversity through Personal Attribute Judgments
Personal Attribute Judgments

Traditional approaches to measuring users’ responses to artifacts lie in the a-priori definition of the measures by the researchers. This chapter highlights the limitations of such approaches that employ standardized psychometric scales and introduces personal attributes judgments. It introduces attribute elicitation techniques and in particular, the Repertory Grid Technique (RGT). It argues that the true value of RGT is in quantifying rich qualitative insights and highlights the limitations of relevant statistical techniques that are typically employed in the analysis of repertory grid data. An initial Multi-Dimensional Scaling procedure that aims at identifying diverse views in Repertory Grid data is proposed. The procedure identifies distinct user groups in a sample population and derives a two-dimensional view for each respective user group. The technique is presented through a case study where users’ views on a set of product concepts were contrasted to the ones of designers. The technique revealed differences not only between users and designers but also between designers of different professional background and role in the design team.

2.1 Introduction

Reductionist approaches to user experience evaluation are grounded on the assumption that people may summarize experiences in overall evaluative judgments (Hassenzahl, 2008). When a product is associated with a certain experience, the value of the experience will be partially attributed to the product. Such evaluations may be substantive (Hassenzahl, 2004), e.g. perceptions of a given product quality such as usability, or verdictive, e.g. overall evaluations of goodness, appeal, or beauty.

These are all latent constructs, in that they cannot be directly measured but instead, they are estimated through a number of different measures. For instance, the Unified Theory of Acceptance and Use of Technology (Venkatesh et al., 2003) measures a latent construct termed performance expectancy through a set of four items such as "I would find the system useful in my job" and "Using the system increases my productivity". These items are assumed to be measuring different facets of the same latent construct (e.g. performance expectancy), at the same time being more or less uni-dimensional.

The development of psychometric scales is often described as a three-step process: item generation, scale development, and scale evaluation (Hinkin, 1995). The first step aims at enhancing the content validity of the questionnaire (i.e. that a complete coverage of the domain of interest is obtained through the proposed items); the latter two steps aim at enhancing the convergent and discriminant validity of the questionnaire (i.e. that each item correlates highly with other items that attempt to measure the same latent construct, and weakly with items that attempt to measure different latent constructs).

Once a substantial set of latent constructs have been developed for a given field, questionnaires may be used by researchers and practitioners to assess the value of products. Using validated questionnaires, one can measure how two or more products compare on a given quality dimension (e.g. trust), or compare two different generations of the same product to assess the impact of the redesign process.

Proponents of the holistic approach in user experience criticize the use of psychometric scales for their inability to capture the richness and diversity of experience (see Blythe et al., 2007). Below, we will try to expand on this criticism by highlighting two limitations of a-priori defined psychometric scales. We will then introduce the use of personal attribute judgments as a means to account for diversity in users’ experiences with interactive products.

Firstly, measures that are a-priori defined are inherently limited in accounting for the users’ perspective. Especially in developing fields such as that of user experience, where the development of new constructs is still in its infancy, researchers might fail in capturing a relevant experience dimension due to a failure in recognizing its im-
portance or simply due to the absence of relevant measurement scales. This issue has been repeatedly highlighted in studies of user acceptance of information systems that employ pre-defined measurement and structural models such as the Technology Acceptance Model (TAM) [Davis et al., 1989]. A number of studies have reported limited predictive power of the Technology Acceptance Model, in some cases accounting for only 25% of the variance in the dependent variable [Gefen and Straub, 2000]. Lee et al. [2003] reported that “the majority of studies with lower variance explanations did not consider external variables other than original TAM variables”. A typical case is illustrated in Figure 2.1 which displays a two-dimensional visualization of the perceived dissimilarity of three systems [van de Garde-Perik, 2008]. In the mentioned study, users were asked to judge the overall dissimilarity of the three systems as well as rate the systems on a number of pre-defined dimensions such as the perceived trust, risk, usefulness and ease of use. The configuration of the three stimuli is derived by means of Multi-Dimensional Scaling on the original ratings of dissimilarity, while the latent constructs are fitted as vectors in the two-dimensional space by means of regression. While a number of insights may be derived from this visualization, one may note that systems 3 and 1 are clearly differentiated in terms of their overall dissimilarity while none of the predefined attributes can explain this dissimilarity. In other words, the measurement model fails to capture the full set of qualities that are meaningful to the participants of the study.

Secondly, one could even wonder whether rating a product on quality dimensions that are imposed by the researcher is always a meaningful activity for the user, for example when the user does not consider a quality dimension as relevant for the specific product. A basic assumption that underlies the rating process is that the user is able to understand, interpret and position a given statement within her own context. For instance, if the user is asked to rate the product in terms of its usefulness, she is expected to interpret usefulness in her own context, to identify what aspects of her experience with the given product contribute to perceptions of usefulness and summarize them in an overall index of usefulness. First, this process is an inherently idiosyncratic process; thus, different users might reflect on different facets of the construct. Secondly, users might fail in certain occasions to interpret the personal relevance of the statement in their own context, e.g. due to a failure to recall experiential information that relates to the statement or due to lengthy and repetitive questioning. Larsen et al. [2008b] reviewed a number of studies employing psychometric scales in the field of Information Systems. They found for the majority of studies the semantic similarity between items to be a significant predictor of participants’ ratings (.00 < \( R^2 < .63 \)). In such cases, they argued participants are more likely to have employed shallow processing [Sanford et al., 2006], i.e. responding to surface features of the language rather than attaching personal relevance to the question.
An alternative approach to posing predefined questionnaires to participants lies in a combination of structured interviewing, that aims at eliciting the attributes that are personally meaningful for each individual, with a subsequent rating process performed on the attributes that were elicited during the interview. Many different interview approaches have been proposed in the fields of Constructivist and Economic Psychology. For instance, Free Elicitation, rooted in theories of Spreading Activation (Collins and Loftus, 1975), probes the participants with a stimulus and asks them to rapidly express words that come to mind. The Repertory Grid Technique (RGT), rooted in Kelly’s Theory of Personal Constructs (Kelly, 1955), provides three alternatives to the participants and asks them to define dimensions in which the three products are meaningfully differentiated. The Multiple Sorting Procedure, rooted in Facet Theory (see Al-Azzawi et al., 2007), asks the participant to sort products in a number of piles, and only later on define a label for each pile. Comparing the different techniques is not the focus of this thesis; see Bech-Larsen and Nielsen (1999), Breivik and Supphellen (2003), Steenkamp and Van Trijp (1997), van Kleef et al. (2005) for more.
information on this. While the analysis procedures that have been developed in the context of this thesis are grounded on Repertory Grid data, they may also be well applied to data derived from any of the other attribute elicitation techniques.

2.2 The Repertory Grid Technique

The RGT is one of the oldest and most popular attribute elicitation techniques. It originates from Kelly’s Personal Construct Theory (PCT) \cite{Kelly1955, Kelly1969} which suggests that people form idiosyncratic interpretations of reality based on a number of dichotomous variables, referred to as personal constructs or attributes. A personal construct is a bi-polar similarity-difference judgment. For example, when we meet a new person we might form a construct friendly-distant to interpret her character. In this process we perform two judgments: one of similarity and one of dissimilarity. Both judgments are done in comparison to reference points: people that we regard as friendly or distant.

To elicit the idiosyncratic attributes of each individual, the RGT employs a technique called triading, where the participant is presented with three products and is asked to “think of a property or quality that makes two of the products alike and discriminates them from the third” \cite{Fransella2003}. Once a bipolar construct is elicited, the researcher may further probe the participant to elaborate on the construct through the laddering and the pyramiding techniques \cite{Fransella2003}. Laddering seeks to understand what motivates a given statement and thus ladders up in an assumed means-ends-chain \cite{Gutman1982} towards more abstract qualities of the stimuli; in laddering the researcher typically asks the participant whether the mentioned quality is positive or negative, and subsequently probes the participant to motivate its importance, e.g. "why is expressiveness important to you?". Pyramiding, on the other hand, also known as negative laddering, seeks to understand the lower level attributes that make up for a given quality; in pyramiding the researcher asks the participant to elaborate on what makes the given product to be characterized with the respective attribute, e.g. "what makes this product more [easy to use]?”. This process can be repeated for all possible combinations of products and until no new attributes arise. The result is a list of attributes that the specific individual uses to differentiate between a set of products. The attributes may then be employed in rating scales, typically Semantic Differentials \cite{Osgood1957}, and each participant rates the set of products on her own elicited attributes. Participants’ ratings are subsequently analyzed with exploratory techniques such as Principal Components Analysis (PCA) or Multi-Dimensional Scaling (MDS).

With the recently increased interest in user experience \cite{Hassenzahl2006}, the RGT has become popular in the field of HCI. \cite{Hassenzahl2006}
CHAPTER 2. PERSONAL ATTRIBUTE JUDGMENTS

Wessler (2000) employed the RGT to evaluate the outcome of parallel design and analyze the perceived character of websites (Hassenzahl and Trautmann, 2001). Fallman and Waterworth (2005) elicited users’ experiences with mobile technology devices, while Davis and Carini (2004) explored player’s experience of fun in video games. Hertzum et al. (2007) studied the differences between designers’ and users’ perceptions for three diverse cultural settings while others have used the RGT as an ethnographic method in understanding how individuals organize their documents and communications (e.g. Szostek et al., 2008; Bondarenko and Janssen, 2009). It, thus, becomes evident that an increasing number of researchers in HCI, emphasize the idiosyncratic nature of subjective judgments on the quality of interactive products.

2.3 The quantitative side of RGT - some concerns

While the RGT has become popular in the field of Human-Computer Interaction and User Experience, one may note several problems in the application of RGT in practice, at least from a quantitative perspective. Below we highlight these issues and introduce the contribution of this thesis in the analysis of Repertory Grid data.

2.3.1 Are we really interested in idiosyncratic views?

The RGT originates from clinical psychology where the emphasis is on an individual’s perceptual and cognitive processes. In the field of user experience, however, the interest is not in the idiosyncratic views of an individual but rather on some more-or-less homogeneous groups of individuals. Due to this focus on idiosyncrasy within the field of personal construct psychology, interpersonal analysis of repertory grid data has received very little attention. We have come to realize two inadequacies of the RGT when used for interpersonal analysis.

First, the interpersonal classification of attributes is often performed purely on semantic grounds, i.e., without testing congruence on the rating scores. Such a practice does not take properly into account the diverse ways in which individuals construe attributes to refer to internal concepts (Kelly, 1955). It is only when two attributes agree both on semantics and in terms of ratings that one might be confident that the two participants refer to the same concept. One could argue that in certain cases two attributes refer to the same concept, but different participants value products differently on the same concept. For instance, one participant might consider product A more easy to use than product B with a second participant considering the inverse, sharing however the same notion of ease of use. Even in this extreme case, one should not group two attributes with diverse ratings. As it will become apparent in the next chapters, we place ratings at higher importance than semantics, i.e. two attributes
2.3. THE QUANTITATIVE SIDE OF RGT - SOME CONCERNS

Table 2.1: Percentages of attributes types from Karapanos and Martens (2007) (Study 1) and Hassenzahl and Wessler (2000) (Study 2) studies.

<table>
<thead>
<tr>
<th>Type</th>
<th>Study 1</th>
<th>Study 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Negation</td>
<td>67%</td>
<td>35%</td>
</tr>
<tr>
<td>Opposition</td>
<td>17%</td>
<td>26%</td>
</tr>
<tr>
<td>Non-contiguous</td>
<td>16%</td>
<td>39%</td>
</tr>
</tbody>
</table>

are checked for semantics only when there is evidence that, within the limited set of stimuli, they (quantitatively) represent the same internal concepts.

Second, techniques such as Principal Components Analysis (PCA) or Multi-Dimensional Scaling (MDS) that are typically employed in the analysis of repertory grid data assume homogeneity in the way people perceive the stimuli offered to them. This is sometimes referred to as the principle of homogeneity of perception (Martens, 2003). To our knowledge, however, all RGT approaches up to date have been employing averaging techniques for the quantitative analysis of personal attribute judgments. We believe this to be due to a lack of more advanced techniques that can account for diversity in users’ subjective judgments, eventually undermining the core value of the RGT, i.e. to account for diversity in individuals’ subjective judgments. This thesis proposes a quantitative, exploratory MDS procedure that aims at accounting for the diverse views that one or more individuals may have on a set of products. It will be demonstrated that even single participants can handle more than one view on a set of stimuli. It will be shown that by averaging interesting views are overlooked due to majorization bias.

2.3.2 On bipolarity

A central notion in the RGT is the bipolarity of the idiosyncratic constructs (i.e. attributes). Kelly, in his theory of Personal Constructs, postulated that individuals perceive the world around them through the construction of dichotomous constructs. It is our experience, however, that participants often need to be probed in order to derive a truly bipolar attribute. This raises concerns with respect to whether individuals actually do think in bipolar terms. Lyons (1977) posited that “categorizing experience in dichotomous contrasts is a universal human tendency which is only secondarily reflected in language”. He identified three different types of bipolarity: negation (i.e. practical-impractical), opposition (i.e. professional - amateurish) and non-contiguous, where the opposite pole does not constitute a negation or linguistic opposition (i.e. easy - powerful) (c.f. Yorke, 2001).

In a meta-analysis of a RGT study on early concept evaluation, we found that the majority of the elicited constructs (67%) were
negation constructs, while 17% were opposition constructs and only 16% were non-contiguous. This deviates substantially from what was observed in a study by Hassenzahl and Wessler (2000) where non-contiguous constructs constituted the major category (39%) while negation accounted for 35% of the constructs and opposition for the remaining 26%. This observed discrepancy is likely to have been influenced by two aspects: a) the instructions of the experimenter (our study had a more evaluative character aiming at eliciting attributes that can be validly used as semantic differential scales, while the latter study’s primary goal was to inform design through rich qualitative accounts) and b) the fidelity of the prototypes (in our study early concepts were communicated in sketched scenarios, while in the latter study users interacted with working prototypes).

This highlights a problem rooted in the dual nature of the RGT: it aims at eliciting rich qualitative accounts which can also be quantified. Non-contiguous attributes provide insight into the relationships that individuals perceive between design qualities (i.e. beautiful - hard-to-use) and concrete product attributes (i.e. easy to use - has many buttons). They provide rich information to design. They are however inherently problematic when used in psychometric scales as the two poles do not underly a single uni-dimensional construct, and thus they will evidently elicit distorted ratings. In our experience we have seen cases where, during rating, participants cannot recall the context in which the attributes were elicited. When both poles of a bipolar construct are not (equally) evident to the participant, ratings may very well be based mostly on one of the poles. We would thus suggest that attributes should be validated by the participant before moving to the rating phase. In this attribute validation phase, participants can be asked to remove duplicate attributes and rephrase attributes when needed.

Negation and opposition bipolarity constitute the common practice in validated questionnaires. In a small study we attempted to explore the difference, if any, between negation and opposition bipolarity in rating scales. Our interest was to test the opposition hypothesis, i.e. ratings for the negative and positive pole should have a linear correlation of -1. Fourteen participants, all students at the department of Industrial Design, rated three concepts on two negation (i.e. secure-insecure, practical-impractical) and two opposition (i.e. standard-creative, amateurish-professional) attributes, using paired comparison scales. The two poles of each scale were split in two distinct scales; ratings on these two scales (e.g. secure and insecure) should approximate a correlation of -1. All attributes except the one referring to security were selected from Attracdiff2 (see Hassenzahl, 2004), a validated user experience questionnaire. Attributes had been translated to Dutch (Frens, 2006). These three attributes were also identified in a repertory grid study with the three concepts (described in the remaining of the paper), ensuring the content validity of the scales for
### Table 2.2: Amount of variance accounted for by the latent one-dimensional construct. Original attributes were in Dutch.

<table>
<thead>
<tr>
<th>English (translated)</th>
<th>Dutch (original)</th>
<th>$R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Negation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Secure-Insecure</td>
<td>Veilig-Onveilig</td>
<td>0.88</td>
</tr>
<tr>
<td>Practical-Impractical</td>
<td>Praktisch-Onpraktisch</td>
<td>0.86</td>
</tr>
<tr>
<td>Opposition</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Creative-Standard</td>
<td>Creatief-Fantasieloos</td>
<td>0.75</td>
</tr>
<tr>
<td>Professional-Amateurish</td>
<td>Vakkundig-Als een leek</td>
<td>0.90</td>
</tr>
</tbody>
</table>

![Figure 2.2: Consistency between positive and negative poles for negation and opposition attributes. Negative scores were inverted.](image)

The opposition hypothesis was tested by attempting two different Multi-dimensional Scaling models, a one-dimensional model and a two-dimensional one. The two-dimensional model provided better fit for all four attributes ($\chi^2, p<0.001$), implying that there is significant evidence for the fact that the ratings for the positive and (inversed) negative poles of the attributes do not underlie a single latent one-dimensional construct. Table 2.2 depicts the squared multiple correlation coefficient $R^2$ between the latent one-dimensional construct and the original attribute. One can note that no clear differences between negation and opposition attributes emerged in this limited study (see figure 2.2). The attribute standard-creative (original term in Dutch fantasieloos-creatief) shows the least agreement between the two opposite poles.
2.3.3 On the measurement of meaning

Due to the notion of bipolarity, semantic differential scales [Osgood et al., 1957] are traditionally employed in RGT rating practices. In [Karapanos and Martens, 2008], we argued for comparison scales (such as paired comparisons) as an alternative, since such scales are known to be less sensitive to contextual effects than single-stimulus scales, such as the semantic differentials. de Ridder [1996] explored this issue in an experiment where quality assessments were made on a set of images. The same set of images was embedded in a larger set of images with a positively or a negatively skewed distribution in quality. The sensitivity of different scales (such as single-stimulus and comparison scales) to these contextual effects was assessed. It was found that contextual effects were negligible only when comparison scales were used. Thus, in the context of the RGT, contextual effects such as individual differences in prior experiences or order effects during the rating process will evidently make individuals’ judgments less reliable when single-stimulus scales are employed.

2.4 Analyzing Personal Attribute Judgments - an initial exploration

One of the advantages of personal attribute elicitation techniques such as the Repertory Grid, i.e. their ability to provide rich insights into individual’s perceptual and cognitive processes, is also one of their dominant limitations as this complicates the analysis phase.

Personal attribute elicitation techniques emphasize the idiosyncratic nature of perception and evaluation of objects. In other words, individuals perceive interactive products through different, individual “templates”. This in turn leads to diversity in the obtained attributes and attribute ratings. Some people may use entirely different attributes to evaluate a product, whereas others may use the same attributes but apply them differently. Researchers are then confronted with as many idiosyncratic views as participants. In practice, the consequence is either an idiosyncratic analysis with a "narrative" summarization (e.g. Hassenzahl and Trautmann, 2001) or the use of average models.

Idiosyncratic approaches often discard the quantitative structure of elicited attributes and utilize only their semantic information. The researcher attempts to summarize the semantically distinct attributes and to qualitatively identify relations among them. Techniques such as laddering and pyramiding, which probe the participant to further elaborate on the elicited attribute, provide rich qualitative insight into the elicited attribute. However, as these approaches do not properly utilize the quantitative structure of attributes, they render the Repertory Grid as yet another
2.4. ANALYZING PERSONAL ATTRIBUTE JUDGMENTS - AN INITIAL EXPLORATION

Figure 2.3: A two-dimensional MDS configuration of three websites using ratings from two attributes. Website j is perceived as more legible and colorful than websites i and k.

Quantitative analysis procedures typically rely on exploratory multivariate techniques such as Principal Components Analysis (PCA) and Multi-Dimensional Scaling (MDS). These techniques aim at modeling relations between stimuli (e.g. interactive products), attributes (e.g. "professional - unprofessional") and overall judgments (e.g. preference). More specifically, MDS looks for a K-dimensional configuration for the stimuli such that the coordinates of the stimuli in the configuration space along different axes can be monotonically related to the observed attribute ratings of the participants [Martens, 2003].

An important motivation for MDS is the principle of homogeneity of perception which states that attribute judgments from different participants are related and thus can be represented in a common configuration space [Green et al., 1989; Martens, 2003]. This view, although it often holds in perceptual judgments, has recently been challenged in more cognitive judgments where the quality dimensions of interactive products are assessed.

To our knowledge, all RGT approaches up to date have been employing an average two-dimensional configuration for the quantitative analysis of personal attribute judgments from all participants. As it will become evident in Chapter 3, such aver-
aging analysis procedures in low dimensional spaces, which constitute the common practice in the analysis of Repertory Grid data, fail in providing insight into the majority of elicited attributes.

This chapter proposes an initial Multi-Dimensional Scaling procedure that aims at inquiring into the diverse views that different individuals might have on a set of products. The procedure consists of two steps: a) identifying homogeneous groups of individuals, and b) eliciting MDS configurations for each homogeneous group to understand how they perceive and evaluate the products being studied.

2.5 The study

We applied this procedure in a study that aimed to identify the differences, if any, between users’ and designers’ perceptions of products. Previous research has emphasized that designers often fail in accurately predicting users’ preferences (Kujala and Kauppinen [2004]). Design decisions may thus be largely impacted by the implicit values shared within a design team. Den Ouden [2006] noted that most reasons for product returns are attributed to decisions made early in the conceptual phase of design, while it has been observed that often designers are unable to trace back decisions made in earlier phases of development (Koca et al. [2009]). In this study we attempt to inquire into users’ and designers’ views on conceptual designs.

2.5.1 Method

Participants

Eleven "designers" and eleven potential end-users participated in the study. Designers were employees of the R&D department of an international company developing document systems. They were all involved in the conception and realization of TouchToPrint, which is a new way of personalizing the use of a shared printer by means of fingerprint recognition. They ranged from usability experts and market experts to software engineers and documentation experts. We refer to them as ‘designers’ since they were all stakeholders in the conceptual design phase. Users were researchers and employees from our department who had no prior knowledge of the product under development.

Stimuli

The TouchToPrint concept and five alternative proposals for interacting with a shared printer were selected for the triading process. These were the Touch & Select concept, which was identical to TouchToPrint but also offering modification possibilities
2.5. THE STUDY

John sends 3 documents for print. He places his finger on the touch sensor, and all his documents are getting printed. He collects his documents and returns to his office.

Figure 2.4: An example storyboard describing one of the concepts used in the study.

at the printer (e.g. final selection of documents, stapling); The Badge concept where user identification takes place by holding an identity badge close to a sensor; the Scroll List concept where user identification takes place by scrolling through a list of usernames; The Pin Code concept where the user types her personal pin code to identify herself to the printer; and the Direct Print concept where no user identification is required, which reflects the most common interaction today. All concepts were presented in the form of storyboards describing a usage scenario of the relevant concept and highlighting the specific details of the relevant concept (see figure 2.4).

Procedure

The six storyboards were combined in three triads, i.e. combinations of three elements. These three triads were expected to elicit user perceptions related to three layers of interaction with the printer: the option of user identification at the printer (Direct print - Scroll list - Pin code), different user identification mechanisms (Scroll List - TouchToPrint - Badge), and the option of document selection at the printer (TouchToPrint - Direct print - Touch & Select). The order in which the triads were presented was counterbalanced between participants.

For every triad, participants were asked to "think of a property or quality that makes two of the products alike and discriminates them from the third". Designers were instructed to think of ways that would be meaningful to users. Users were defined as employees at a university department. This process was repeated until a maximum of six attributes were elicited for each triad.

After attribute elicitation, participants were asked to remove duplicate attributes
and rank the remaining attributes according to their importance. Finally, participants were asked to rate all products on their personal attributes, as well as on preference and dissimilarity. In contrast to the traditional Repertory Grid approach, we employed paired comparisons instead of semantic differentials, as this was a priori expected to deliver more stable results (de Ridder, 1996). While in the semantic differential technique only one product is being rated and thus being compared to an implicit reference point, in paired comparison two products are being compared on a specific attribute. Out of the six products one can form up to \(n(n-1)/2 = 15\) pairs. To reduce the number of pairs we employed a balanced incomplete design (Sandt Van de, 1970; Furlan and Corradetti, 2006) with 9 total pairs and every of the six products participating in 3 pairs. According to this design, the total number of pairs can be calculated from equation (2.1):

\[
\text{Total Number Of Pairs} = \frac{n\lambda}{2}
\]

where \(n\) is the number products and \(\lambda\) the number of pairs in which we want each product to participate in.

### 2.6 Analysis Procedure

The analysis procedure consists of three steps. First, a user segmentation map that expresses the diversity among individuals is derived from the collected dissimilarity ratings by means of Multi-Dimensional Scaling (MDS) (Martens, 2003). Homogeneous groups of users are identified within this map by means of (hierarchical) clustering. Secondly, attributes are classified into categories based on their semantic content. This semantic classification will then be contrasted to attribute ratings. Third, perceptual maps are created from the attribute, dissimilarity and preference ratings to express how homogeneous groups of participants perceive the products being studied.

#### 2.6.1 Identifying homogeneous user groups in the user segmentation map

In this step a user segmentation map that expresses the diversity among users is derived from their dissimilarity ratings by means of Multi-Dimensional Scaling. To create the user segmentation map, we define the distance \(D_{i,j}\) between participants \(i\) and \(j\) based on the correlation \(R_{i,j}\) between their dissimilarity scores. Derived distances are then visualized in two or more dimensions using the MDS tool XGms (Martens, 2003). Figure 2.5 displays a two dimensional configuration of designers and users. The closer two individuals are in the two-dimensional space, the more their ratings
of overall dissimilarity of products correlate. The dimensionality of the configuration may be judged by the experimenter using the Stress Value which is an index of goodness of fit of the model. In this case, the two dimensional visualization was judged as adequate (stress value $S=0.18$) (Clarke, 1993).

$$D_{i,j} = 1 - R_{i,j}^2$$  \hspace{1cm} (2.2)

$$R_{i,j} = \frac{\sum_k D_i(k)D_jk}{\sqrt{D_i^2(k)D_j^2(k)}}$$  \hspace{1cm} (2.3)

Hierarchical clustering performed on the reduced two-dimensional space (with minimum variance) reveals two main clusters that can be further subdivided into five more or less homogeneous participant groups. Groups 3 and 4 consist entirely of end users, while groups 1, 2 and 5 consist mostly of designers. Identification of the designers reveals that group 1 consists mostly of technically-oriented designers, while group 2 consists mostly of user-oriented designers.
CHAPTER 2. PERSONAL ATTRIBUTE JUDGMENTS

Table 2.3: Classification of attributes elicited in the study into 16 categories and three broad themes

<table>
<thead>
<tr>
<th>Effectiveness (32%)</th>
<th>Efficiency (51%)</th>
<th>Emotional Appreciation (17%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>11. Easy-to-learn</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

2.6.2 Classifying attributes for interpersonal analysis

In this step, attributes are submitted to a content analysis (Hsieh and Shannon, 2005) where key concepts, i.e. distinct categories of attributes, emerge from the data and attributes are subsequently classified into one of the elicited categories. A total of 81 attributes for designers and 95 for users were obtained in the study (6 to 11 per participant). Two rounds of analysis were performed: an exploratory semantic classification of attributes, followed by a confirmatory analysis of the classification. For the first round, sixteen semantically distinct attribute categories were first formed out of the data. To minimize the researcher’s bias, the naming of the attribute categories was restricted to choosing one of the attribute names that reflect this semantic value. Subsequently, the first author and two additional experimenters independently assigned every attribute to one of the sixteen categories (Table 2.3). Interrater agreement (Fleiss et al., 2003) of the initial classification was satisfactory (K=0.72). All sixteen categories were then classified into three overall classes: Effectiveness, Efficiency, and Emotional Appreciation (interrater agreement, K= 0.80).

During the confirmatory analysis of the classification, statistical consistency across attributes within the same category was being sought. Attribute scores were submitted to a cluster algorithm where Euclidean distances between attributes were calculated and visualized in two or three dimensions by means of Multi-Dimensional Scaling. Outlier attributes, i.e. ones that did not belong in the same or a neighbor cluster to the one that were characterized primarily by attributes of a given semantic category, were identified. The prospect of transferring the outlier attribute to one if its statistically-neighbor categories was explored. If there was no argument for a transfer to another category, the attribute was deleted.

2.6.3 Charting perceptual maps for homogeneous groups of users

In this step, a perceptual map is created for each homogeneous group of users. First, a configuration of the stimuli is established in a k-dimensional space where their dis-
2.6. ANALYSIS PROCEDURE

tance represents their degree of dissimilarity. Next, attribute and preference vectors are fitted into these spaces by performing multiple regressions between the stimulus configuration (as independent variables) and the attribute scores (as the dependent variables).

Figure 2.6 illustrates the perceptual maps of the five groups. All five perceptual spaces were adequately visualized in two dimensions ($\chi^2$ test on goodness of fit, $p<0.001$). Only significant attributes were retained in the final visualization, i.e. ones that their confidence ellipse did not include the zero point of the configuration. Attributes are represented by the bold vectors and preferences by the light vectors.

The interpretation of the perceptual maps can be done in three steps: a) identifying the similarities and dissimilarities between products, b) interpreting product dissimilarities by employing the attribute vectors, and c) inferring the values of group of participants by relating individual preferences with perceived attributes.

Identifying dissimilarities between products

One can note that all perceptual spaces are clustered in two regions, one containing TouchToPrint, Badge and Touch & Select and the other containing Scroll List and Direct Print. The perception of Pin Code in comparison to the other systems is less consistent across different perceptual spaces. It is further apparent that the configuration of products within these regions differs across the perceptual spaces. Groups 1 and 4 perceive differences between Scroll List and Direct Print while Groups 2, 3 and 5 perceive these two products as relatively similar. Further, Groups 1 and 5 perceive TouchToPrint and Badge as very similar in comparison to Touch & Select while Groups 2, 3 and 4 perceive relatively strong differences between TouchToPrint and Badge.

Interpreting product dissimilarities

As mentioned all perceptual spaces show two main clusters of products. One may notice a strong preference for products in the first cluster (i.e. only two participants prefer products from the second cluster). This can be further explained by analyzing the directions and the type of attributes. It is evident that most attributes point to product cluster 1 while only three attributes point to cluster 2. A close look at the content of these three attributes provides some more insight. Two out of the three attributes are essentially the contrast poles to a negative attribute of the comparison products. This is also evident when looking into the element differentiators. Participants indicated that "TouchToPrint is not reliable" and that "it is privacy threatening", thus other products will score better on reliability and privacy. It seemed that par-
Participants could more easily identify qualities in products of cluster 1 that were less familiar to them than in the traditional products from cluster 2.

Within product cluster 2 one can notice differences among groups. Groups 1 and 4 perceive differences between Scroll List and Direct Print while Groups 2, 3 and 5 perceive these two products as relatively similar. Group 1 seems to perceive differences between Scroll List and Direct Print in two attributes: control and personal. Apparently, for Group 1, Scroll List provides more control in the interaction as documents are only printed when the user is identifying himself to the printer. Moreover, it is more personal than Direct Print as every user has a personalized access to the printer. More interesting though is the observation that Group 3 does not regard scroll list as personal despite having this personalization feature. Apparently, users from Group 3 distinguish between the personalization features of Scroll list and those of TouchToPrint, Badge and Touch & Select. One explanation could be that the perception of personalization for Group 3 is influenced by the perception of security (i.e. with scroll list one can identify himself as another person). Further, group 4 seems to perceive differences between Scroll List and Direct Print in two attributes: environmental friendly and fast. One can notice that group 4 differs from other groups in the perception of the attribute fast. For group 4 the two extremes are Scroll List and Direct Print where the second is fast as it involves no interaction with the product and the first involves tedious interaction. It is also evident from preferences that a product quality for this group is a minimum interaction with the product. The other groups however perceive that direct print requires additional effort for identifying the documents among the pile of all printed ones, thus Direct Print is not fast. It is evident that the attribute fast is not statistically consistent between groups but it is consistent within groups. This demonstrates the value of the two-step classification process as it enables us to gain insight in the different perceptions of an indistinguishable product quality.

Further, within cluster 1 one can notice differences in the relation between Badge and TouchToPrint as compared to Touch and Select. Groups 1 and 5 perceive TouchToPrint and Badge as very similar to Touch & Select while Groups 2, 3 and 4 perceive relatively strong differences also between TouchToPrint and Badge. All three groups think that Badge provides more control in the interaction than TouchToPrint. Further, Group 4 perceives Badge as more secure than TouchToPrint while Group 2 perceives a tradeoff in TouchToPrint which is the capturing of personal biometric data (i.e. privacy) and Group 3 has concerns about the reliability of TouchToPrint due to the novelty of the technology.
2.6. ANALYSIS PROCEDURE

Figure 2.6: Perceptual maps for groups 1 to 5. The light vectors represent individual preferences, while the other vectors represent the most significant attributes.

Inferring users’ values

By relating perceived attributes to individual preferences one can infer priorities existing in users’ values as indicated within this limited set of products. One can note that for individuals in group 1, secure and fast are the most important product qualities influencing their preference judgment, while participants in group 3 value modern, personal and secure products. Participants within group 3 have negative concerns regarding the reliability of new products, such as Touch & Select, TouchToPrint and Badge, despite the fact that they prefer them. Such concerns are not reported by the participants in group 1.

To acquire richer insight into the ways in which designers and users differ we fo-
focused on a comparison between two of the six products: TouchToPrint and Badge. These two products differed only in the mechanism that was employed for the user identification process: a touch sensor and a sensor for an individual’s badge. First, a comparison of preference between the two products was performed. Four designers and eight users preferred TouchToPrint while six designers and two users had a preference for Badge. One designer and one user were neutral. One would expect designers to have a strong preference for TouchToPrint since they were recently involved in the development of this product. This was not evident though. Two possible explanations might be given: a) Badge was not yet implemented by the specific company, therefore treated as future development by designers and thus assigning to it greater value than to the already existing TouchToPrint or b) potential drawbacks of TouchToPrint were only evident after use, and hence more obvious to the designers who had actually experienced the product. The users, who had no actual experience with TouchToPrint, seemed to value it higher than designers did.

To further understand this discrepancy between designers’ and users’ preferences, we analyzed their perceptions for these two products. Figure 2.7a illustrates the reasons supporting preference for TouchToPrint over Badge, as it shows the number of attributes that are positively ranked when TouchToPrint is preferred. While users’ most frequent reason for preference for TouchToPrint was emotional attributes, for designers it was efficiency attributes. All attributes in the effectiveness category were related to security. TouchToPrint was perceived as more secure than Badge, both by designers and users. Users’ most frequent negative concerns, shown in Figure 2.7b, were related to reliability (5 out of the 7 effectiveness attributes had to do with reliability). This is also evident in Figure 2.8 where we can observe that only two designers expressed reliability concerns and ranked them as the 6th and 7th most important attributes while five users ranked reliability within their three most im-
2.7 Discussion

In this chapter a simple and effective technique was presented for acquiring diverse views on Repertory Grid data. The analysis revealed systematic differences between the product qualities that users and designers appreciated within this limited set of products. The proposed analysis procedure, however, relies on a number of assumptions.

First, it assumes the individual as a homogeneous entity. Attribute judgments of a single individual are assumed to be well predicted by a single two-dimensional configuration, and the emphasis lies in identifying individuals in a sample that form an approximately homogeneous user group. At certain occasions this might be a desired property of the technique, e.g. when the researcher wants to characterize single individuals such as identifying differences in the perceptions of different stakeholders in a design team. Diversity could however exist not only across different individuals, but also within a single individual, in the sense that different attribute judgments of a participant may reveal different, complementary, views.

Second, diversity across individuals was explored using a single measure, that of overall dissimilarity between the products. This approach assumes that individuals are able to consistently summarize all attribute judgments in a single rating of dissimilarity, an assumption that is ubiquitous in the field of psychometrics were dissimilarity serves to summarize various perceptual judgments such as the perceived blur and noise in an image (Martens 2003). One might, however, question the degree

Figure 2.8: Importance rankings for reliability attributes for designers and users. Only 2 designers reported reliability concerns in contrast to 7 users. Five users ranked reliability attributes within their three most important attributes.
Table 2.4: Number (and percentage) of attributes being adequately modeled by the resulting views of the averaging analysis and the two proposed procedures.

<table>
<thead>
<tr>
<th>Analysis Procedure</th>
<th>No of views</th>
<th>No of attributes (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Averaging analysis</td>
<td>1</td>
<td>65 (43 %)</td>
</tr>
<tr>
<td>Procedure proposed in this chapter</td>
<td>5</td>
<td>62 (41 %)</td>
</tr>
<tr>
<td>Procedure proposed in chapter 3</td>
<td>2</td>
<td>85 (56 %)</td>
</tr>
</tbody>
</table>

to which this holds in more complex settings where various cognitive judgments are made for the quality of interactive products.

Third, the technique did not explicitly optimize the goodness of fit of the diverse models for the analyzed attributes. As a result, while the technique succeeded in identifying the differences between users and designers in terms of overall dissimilarity, it might fail in accounting for more attributes than the traditional "averaging" analysis. In the next chapter, a new procedure is described that explicitly aims at increasing the number of attributes that are adequately modeled based on two criteria introduced below. In this procedure, two goodness-of-fit criteria are defined for assessing whether or not an attribute is adequately predicted by a given model: a) the amount of variance $R^2$ in the attribute ratings accounted for by a given model, i.e. MDS configuration, and b) the ratio of the maximum range of the predicted scores for an attribute divided by the standard deviation $\sigma_k$ of the estimation error in the attribute scores.

In the remaining of this section, we employ these criteria for comparing the three different analysis procedures: a) the traditional averaging analysis, b) the analysis proposed in this chapter, and c) the second analysis procedure proposed in chapter 3 that aims at optimizing the goodness of fit for the majority of the attributes.

Table 2.4 depicts the number (and percentage) of attributes being adequately modeled by the resulting views of the three different analysis procedures. The averaging analysis resulted in a total of 65 or 43 % of all attributes being adequately modeled by a single two-dimensional configuration. Surprisingly, the analysis procedure proposed in this chapter, despite resulting in five diverse views, performed worse than the averaging analysis when we are concerned about the number of attributes being adequately modeled by the resulting views. Even when analyzing together the attributes of individuals that agreed on the overall dissimilarity of the stimuli, a wealth of attributes still remained inadequately modeled by the resulting shared configuration. This challenges the validity of two assumptions made in this analysis procedure: a) that all attributes of an individual may be analyzed by a single two-dimensional configuration, and b) that dissimilarity ratings may be used to model all attributes of an individual.
### Table 2.5: Number of attributes that are adequately modeled by the resulting one or two views for each participant.

<table>
<thead>
<tr>
<th>Subj.</th>
<th>Description</th>
<th>view a</th>
<th>view b</th>
<th>remain</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>User 1</td>
<td>6</td>
<td>-</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>User 2</td>
<td>10</td>
<td>-</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>User 3</td>
<td>-</td>
<td>-</td>
<td>13</td>
</tr>
<tr>
<td>4</td>
<td>User 4</td>
<td>6</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>User 5</td>
<td>-</td>
<td>-</td>
<td>10</td>
</tr>
<tr>
<td>6</td>
<td>User 6</td>
<td>6</td>
<td>-</td>
<td>2</td>
</tr>
<tr>
<td>7</td>
<td>User 7</td>
<td>9</td>
<td>-</td>
<td>3</td>
</tr>
<tr>
<td>8</td>
<td>User 8</td>
<td>6</td>
<td>-</td>
<td>4</td>
</tr>
<tr>
<td>9</td>
<td>User 9</td>
<td>10</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>10</td>
<td>User 10</td>
<td>9</td>
<td>-</td>
<td>1</td>
</tr>
<tr>
<td>11</td>
<td>User 11</td>
<td>6</td>
<td>-</td>
<td>3</td>
</tr>
<tr>
<td>12</td>
<td>Concept Developer 1</td>
<td>11</td>
<td>-</td>
<td>1</td>
</tr>
<tr>
<td>13</td>
<td>Concept Developer 2</td>
<td>6</td>
<td>-</td>
<td>4</td>
</tr>
<tr>
<td>14</td>
<td>Concept Developer 3</td>
<td>6</td>
<td>-</td>
<td>1</td>
</tr>
<tr>
<td>15</td>
<td>Documentation Expert</td>
<td>7</td>
<td>-</td>
<td>1</td>
</tr>
<tr>
<td>16</td>
<td>Interface Designer</td>
<td>-</td>
<td>-</td>
<td>2</td>
</tr>
<tr>
<td>17</td>
<td>Market Expert</td>
<td>8</td>
<td>-</td>
<td>1</td>
</tr>
<tr>
<td>18</td>
<td>Project Manager</td>
<td>7</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>19</td>
<td>Software Expert 1</td>
<td>6</td>
<td>-</td>
<td>1</td>
</tr>
<tr>
<td>20</td>
<td>Software Expert 2</td>
<td>5</td>
<td>-</td>
<td>2</td>
</tr>
<tr>
<td>21</td>
<td>Usability Expert 1</td>
<td>-</td>
<td>-</td>
<td>9</td>
</tr>
<tr>
<td>22</td>
<td>Usability Expert 2</td>
<td>6</td>
<td>4</td>
<td>-</td>
</tr>
</tbody>
</table>

The procedure proposed in the next chapter explicitly aims at identifying whether an individual’s attributes can be modeled in one or two diverse views. Table 2.5 depicts the number of attributes being adequately modeled by the first and (in some cases) a second view. Note that only two individuals required a second view while for the majority of the individuals a single two dimensional view was sufficient for modeling the majority of his/her attributes. The attributes of four individuals resulted in no satisfactory MDS solution suggesting that their data are too noisy to be analyzed in a meaningful way.

Figure 2.9 depicts the similarity of the resulting individual views. Note that the differences between designers’ and users’ perceptions emerge also in the analysis of individual attributes. Two views, one representing the designers group and one representing the users group were derived from the configurations (see chapter 3 for the exact procedure). The resulting two views were able to adequately model a total 85 attributes (56%).
Figure 2.9: Two dimensional visualization of the dissimilarity of the individual views derived from the reanalysis of the data using the procedure proposed in chapter 3. Views connected through lines belong to the same cluster. Abbreviations used: Doc=Documentation Expert, CD=Concept Developer, UE=Usability Expert, PM=Product Manager, SOFTW=Software Expert, DES=Visual Interface Designer, MKT=Market Expert, U=User

2.8 Conclusion

This chapter highlighted two issues in the use of standardized psychometric scales in measuring users’ experiences with products. Firstly, a-priori defined measurement scales are inherently limited in missing potentially relevant measures that the researchers did not consider as relevant to a given context. Secondly, in certain cases, rating products on measures defined by the researchers is not always a meaningful activity for the user, for example when the user does not consider a quality dimension as relevant for the specific product or is unable to situate the question in her own context.

We introduced personal attribute judgments and respective attribute elicitation techniques as an alternative to standardized psychometric scales. These approaches have the potential to account for the richness of user experience and bridge the gap between formative and summative evaluation by quantifying rich qualitative insight. We highlighted one technique, namely the Repertory Grid Technique which has re-
cently become popular in the field of Human-Computer Interaction and User Experience.

We noted the lack of appropriate statistical analysis techniques that lead to practices that either treat RGT as yet another qualitative technique, or employ averaging models which undermine the core motivation of RGT and relevant techniques. We argued that the true value of RGT is in quantifying rich qualitative insights.

We identified the two dominant approaches in analyzing repertory grid data: a) a qualitative, idiosyncratic approach with a narrative summarization, and b) a quantitative approach that employs averaging procedures using exploratory multivariate techniques. We argued that these two approaches are limited in two respects. Qualitative approaches do not take properly into account the elicited attribute ratings and thus do not fully exploit the true value of the Repertory Grid Technique, which is to quantify rich qualitative insights. Averaging procedures, on the other hand, treat diversity among participants as error and thereby contradict the basic idea of The Repertory Grid and relevant personal attribute elicitation techniques.

Last, we proposed an initial quantitative technique that aims at inquiring into the diverse views that different individuals might have on a set of products. The technique employs Multi-Dimensional Scaling in a) identifying homogeneous groups of individuals, and b) eliciting a different view for each homogeneous group of individuals.

The technique was applied in a study that tried to assess the differences between designers’ and users’ perceptions on a set of early conceptual designs. The results from the study corroborated prior findings suggesting that designers may fall short in accounting for users’ views on their product and foreseeing their preferences. Surprisingly, the diversity between the designers and users was larger than the diversity across different user groups. This might be affected by the different exposure of these two diverse groups to the different concepts, thus leading to a relative agreement across the different users’ in comparison to the designers group. One has to note that designers were explicitly asked to reflect on the way that this specific user group would perceive this set of products. In this way the gap between designers’ and users’ perceptions was minimized to reflect actual design practice where assumptions and claims are made for a specific user group. These insights highlight the need for grounding design decisions that are made early in the concept design phase on quantifiable empirical data about users’ preferences.
Analyzing Personal Attribute Judgments

1 This chapter presents a second Multi-Dimensional Scaling procedure that aims at identifying diverse views even within single individuals. The technique is applied on an existing dataset (Heidecker and Hassenzahl, 2007). It is illustrated that the traditional - averaging analysis provides insight to only 1/6th of the total number of attributes in the example dataset. The proposed approach accounts for more than double the information obtained from the average model, and provides richer and semantically more diverse views on the set of stimuli.

3.1 Introduction

In the previous chapter a simple and effective technique was presented for acquiring diverse views on Repertory Grid data. This technique, however, relies on a number of assumption. First, it assumes that all attributes of an individual may be analyzed by a single 2D view, and second, that individuals are able to consistently summarize all attribute judgments in a single judgement of overall dissimilarity. As such, it discards diversity existing within a single individual and does not explicitly optimize the goodness of fit of the diverse models for the analyzed attributes. It was shown that while the technique succeeded in identifying the differences in individual’s perceptions in terms of overall dissimilarity, it failed in accounting for more attributes than the traditional "averaging" analysis.

---

Table 3.1: Attribute categories and examples

<table>
<thead>
<tr>
<th>Attribute category</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Layout</td>
<td>Graphical layout - Textual layout</td>
</tr>
<tr>
<td></td>
<td>Colorful - Pale colors</td>
</tr>
<tr>
<td></td>
<td>Professional - playful</td>
</tr>
<tr>
<td>University Image</td>
<td>Technical studies - Social studies</td>
</tr>
<tr>
<td></td>
<td>Emphasis on achievement - Average univ.</td>
</tr>
<tr>
<td></td>
<td>Refers to student life - Modern organization</td>
</tr>
<tr>
<td>Information Access</td>
<td>Fast access to information - time-intensive</td>
</tr>
<tr>
<td></td>
<td>Legible - Tangled</td>
</tr>
</tbody>
</table>

This chapter suggests a quantitative, exploratory Multi-Dimensional Scaling procedure to account for the diverse views that one or more individuals may have on a set of products. It will be demonstrated that even single individuals can handle more than one view on a set of stimuli. It will be shown that by averaging interesting views are overlooked due to majorization bias. The insights strongly advocate the view that the analysis of quality judgments of interactive products should not stop on a group level, but must be extended to the relations between the attribute judgments within an individual. The Repertory Grid combined with the suggested technique to analyze the resulting quantitative data is an important step towards the adequate account of homogeneity and especially diversity in individual quality judgments.

3.2 The study

The data for the present analysis was taken from a study of Heidecker and Hassenzahl [2007] of individuals’ perceptions of eight university websites (figure 3.1). The study was part of a larger project aiming at understanding how the Technical University of Darmstadt (TUD) is perceived in comparison to other regional competitors. Ten individuals, all students at TUD, participated in the study.

The eight university websites were presented to participants in the form of colored A4 screenshots of the main page. Using the Repertory Grid Technique, a number of attributes on which the eight websites differ, were elicited from each participant. Participants were then asked to rate the websites on their own elicited attributes, using 7-point Semantic Differential scales. The resulting data set consisted of a total of 118 attributes (10 to 14 per participant) on which ratings for the eight different stimuli were elicited.
3.2. THE STUDY

Figure 3.1: The eight stimuli used in the study
3.3 An MDS approach to account for diversity

The starting point of the proposed approach is that of identifying the different views that each participant has on the set of products. In this step, an average model is attempted for each participant. However, attributes that are not adequately predicted by the average model (see Table 3.2) are removed and used in deriving a second model, i.e. a second view for the specific participant (Figure 3.2 illustrates two diverse views derived for one participant).

Once the diverse views of all individuals have been identified, the similarity among them is assessed and views are clustered into groups of increased homogeneity.

A final set of diverse configurations is formed by grouping the similar views, which are then used to model the attributes from all participants.

3.3.1 Identifying the different views

In identifying the different views that an individual might hold, one tries to model the individual’s perceptions in one or more non-trivial $K$-dimensional models, each explaining adequately a part of his/her attribute judgments. Individual views should provide close matches between the measured and the modeled attributes that are associated with that view. Therefore, defining a goodness-of-fit will be an essential step in creating views.

The maximum dimensionality $K$ is limited by the number of degrees of freedom in the data, but may also be set a priori by the data analyst. For the example data set considered below the dimensionality was fixed to $K=2$ so that different visualizations can be easily presented on paper. Note that models of degree higher than 2 need multiple 2D views to be assessed anyhow. However, in this latter case, the views are different 2D projections of a shared multi-dimensional configuration. The 2D views that result from the analysis presented in this chapter, on the other hand, can be independent. Views of 2 or higher dimensionality provide relations not only about the stimuli but also about the relations between the different attributes and are thus preferred over 1D views.

A two-step procedure is proposed to establish whether zero, one or two models with dimension $K=2$ can adequately model the attribute scores of a single observer. In the first step, all attributes of a participant are modeled together, as is common practice in MDS (average model). However, only the attributes that satisfy a particular goodness-of-fit criterion are considered to be adequately modeled. These attributes are analyzed to form the first model, i.e. the individual’s most dominant view on the set of products.

In the second step, the attributes that displayed the least fit to the average model
3.3. AN MDS APPROACH TO ACCOUNT FOR DIVERSITY

Table 3.2: Goodness of fit Criteria. Attributes that are adequately predicted are employed in model 1. A second model is attempted only on attributes that display the least fit, to ensure diversity between the two models.

<table>
<thead>
<tr>
<th>Criteria</th>
<th>$R^2$</th>
<th>$R_k$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Adequate fit</td>
<td>$R^2 &gt; .5$</td>
<td>$R_k &gt; 6$</td>
</tr>
<tr>
<td>2. Average fit (Excluded)</td>
<td>$4 &lt; R_k &lt; 6$</td>
<td></td>
</tr>
<tr>
<td>3. Least fit (attempt 2nd model)</td>
<td>$R_k &lt; 4$</td>
<td></td>
</tr>
</tbody>
</table>

are grouped and used to attempt a second model. By selecting the least-fit attributes, instead of all remaining attributes, we promoted the diversity between the two models. The same goodness-of-fit criteria are applied for the second model to select the attributes that are retained.

3.3.2 Defining goodness-of-fit criteria

We suggest a combined goodness of fit criterion. First, for an adequately predicted attribute, a substantial amount of its variance should be accounted for by the model. This proportion of explained variance is the $R^2$ statistic (i.e., the squared multiple correlation coefficient). A threshold $R^2 > 0.5$ was set, implying that only attributes are retained for which at least 50% of their variance is accounted for by the model. A limitation of this criterion is that it is insensitive to the range of the ratings for the different stimuli on a given attribute. An attribute might make no meaningful differentiation between stimuli (e.g. if all stimuli are rated as 4 or 5 on a 7-point scale) but can nevertheless be well-predicted by a model. To account for this limitation, we combine it with a second criterion.

This second criterion is a modification of a measure originally proposed by [Draper and Smith (1998)]. It is the ratio of the maximum range of the predicted scores for attribute $k$ divided by the standard deviation $\sigma_k$ of the estimation error in the attribute scores (1a).

$$R_k = \frac{\hat{A}_{k,\text{max}} - \hat{A}_{k,\text{min}}}{\sigma_k}$$  \hspace{1cm} (3.1)

$$R_k^* = \sqrt{\frac{1}{n^2} \sum_{i,j} \frac{[\hat{A}_{ki} - \hat{A}_{kj}]^2}{k^2}}$$  \hspace{1cm} (3.2)

A combined criterion thus takes into account both the accounted variance in the attribute as well as the range of the scores for the different stimuli (i.e. the attribute’s strength). The obvious limitation of the second measure is its sensitivity to outlier scores. However, in single-stimulus scales such as the semantic differential scales,
CHAPTER 3. ANALYZING PERSONAL ATTRIBUTE JUDGMENTS

Table 3.3: Goodness of fit statistics for the two diverse models of participant one. Attributes (2,4,6,8,12,13) were adequately predicted ($R^2 > .5$ & $R_k > 6$) by model 1. Attributes (1,5,7,9,10,11) displayed the least fit ($R_k < 4$) and were used to derive a second model. Attributes (5,7,9,10) were adequately predicted by model 2.

<table>
<thead>
<tr>
<th>No</th>
<th>Attribute Variance ($\sigma^2$)</th>
<th>Avg. Model $R^2$</th>
<th>Model 1 $R^2$</th>
<th>Model 1 $R_k$</th>
<th>Model 2 $R^2$</th>
<th>Model 2 $R_k$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2.6</td>
<td>0.47</td>
<td>2.2</td>
<td>0.47</td>
<td>2.2</td>
<td>0.36</td>
</tr>
<tr>
<td>2</td>
<td>3.8</td>
<td>0.89</td>
<td>7.3</td>
<td>0.89</td>
<td>7.3</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.6</td>
<td>0.73</td>
<td>4.1</td>
<td>0.73</td>
<td>4.1</td>
<td>0.56</td>
</tr>
<tr>
<td>4</td>
<td>1.9</td>
<td>0.98</td>
<td>18.6</td>
<td>0.98</td>
<td>18.6</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>3.7</td>
<td>0.49</td>
<td>2.3</td>
<td>0.49</td>
<td>2.3</td>
<td>0.95</td>
</tr>
<tr>
<td>6</td>
<td>2.2</td>
<td>0.99</td>
<td>40.5</td>
<td>0.99</td>
<td>40.5</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>1.7</td>
<td>0.48</td>
<td>2.4</td>
<td>0.48</td>
<td>2.4</td>
<td>0.99</td>
</tr>
<tr>
<td>8</td>
<td>6.3</td>
<td>0.93</td>
<td>9</td>
<td>0.93</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>4.1</td>
<td>0.63</td>
<td>4.8</td>
<td>0.63</td>
<td>4.8</td>
<td>0.99</td>
</tr>
<tr>
<td>10</td>
<td>4.5</td>
<td>0.26</td>
<td>2.1</td>
<td>0.26</td>
<td>2.1</td>
<td>0.61</td>
</tr>
<tr>
<td>11</td>
<td>3.9</td>
<td>0.08</td>
<td>0.9</td>
<td>0.08</td>
<td>0.9</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>1.9</td>
<td>0.88</td>
<td>6.8</td>
<td>0.88</td>
<td>6.8</td>
<td>0.48</td>
</tr>
<tr>
<td>13</td>
<td>5.6</td>
<td>0.99</td>
<td>50.4</td>
<td>0.99</td>
<td>50.4</td>
<td>0.85</td>
</tr>
</tbody>
</table>

these outlier scores may actually be very valuable, since they point at the stimuli that most strongly influence the existence of the attribute scale in the first place. When using more sensitive scales such as paired comparisons [Karapanos and Martens, 2007], one might consider adopting the modified measure [3.2] that averages across differences in predictions. [Draper and Smith, 1998] proposed a minimum ratio value of four, meaning that any attribute predictor with a ratio below four hardly makes any distinction between the stimuli and is pretty useless. Predictors with a ratio value above ten are considered to be excellent. We decided to use an acceptable ratio of six for the data analysis reported in Table 3.3.

### 3.3.3 Two diverse views for one participant

Table 3.3 illustrates the analysis process on the attribute judgments of a single participant. A first (average) model was attempted on all attributes of the participant. Attributes (2,4,6,8,12,13; in bold) were adequately predicted by the average model, using the two criteria that were discussed before, i.e. $R^2 > .5$ & $R_k > 6$. Model 1 was then derived by optimizing the average model only for the attributes that were adequately predicted by the average model.

Note that the $R^2$ and $R_k$ values are identical (at least in this exact decimal point) for Model 1 and the average model. This implies that when removing the attributes that are not adequately predicted (even with these arbitrary criteria), the 2D configuration space (which is represented in the model parameters) displays virtually no change. In other words, the attributes that were removed (according to the arbitrary criteria)
3.3. AN MDS APPROACH TO ACCOUNT FOR DIVERSITY

had no contribution to the configuration space. Thus, the information contained in these attributes is not modeled when attempting an averaging analysis and therefore it is lost.

Out of all the attributes that were not adequately predicted, attributes (1,5,7,9,10,11; in italics) displayed the least fit by model 1, i.e., $R_k<4$. These were used to derive a second model. Out of them, only attributes (5,7,9,10) turned out to be adequately predicted by model 2, using the same goodness of fit criteria as used in model 1.

Figure 3.2 illustrates the different insights that the two diverse views bring. One can note that the two views highlight semantically different attributes. Each attribute is visualized as an arrow, i.e. a dimension, on which the relative positions of the websites can be compared. The length of each arrow indicates the strength of the attribute, reflecting the variance in the predicted attribute ratings for the different stimuli; on some attributes all websites might be rated as 4 or 5 on a 7-point scale, while others might make strong differentiations between sites, i.e. across the whole range of the scale.

The first view provides overall three different insights. First, that the universities of Frankfurt, Manheim and Mainz are perceived as putting less emphasis on achievement, as compared to the remaining five universities. This may be induced by the websites but may also reflect prior beliefs of the individual. Second, the websites of the universities of München, Aachen, Karlsruhe and Heidelberg have a more profes-
personal layout as opposed to the remaining four which have a more playful one. Last, the participants perceive this same group of websites as legible as opposed to the remaining four in the upper part of the figure that are perceived as having no clear structure.

The second view partly provides overlapping information (emphasis on achievement), but also gives three new insights. First, the website of the University of Heidelberg is differentiated from all others by having a less colorful layout. Second, the Universities of Darmstadt, Aachen and Karlsruhe are differentiated as universities that provide primarily technical studies, as opposed to the universities of Mainz, Mannheim and Frankfurt that are referred to as universities of average quality, and third, as opposed to the university to Heidelberg that is perceived as a university offering primarily medical studies.

Note that an attribute may range from being purely descriptive, i.e. referring to specific features (e.g. allows searching), to having an evaluative tone, e.g. referring to the perceived quality of the product (e.g. easy to use) or the product’s overall appeal (e.g. good). This enables the researcher to gain a better understanding of the inferences individuals make as they form evaluative judgments of products.

3.3.4 Assessing the similarity between different views

Table 3.4 summarizes the results of the above analysis for all ten participants. For two of the ten participants (7, 8), no substantial agreement between their attribute judgments is observed, i.e., no satisfactory MDS-model can be derived. This implies that they either have as many different views as their attribute judgments, or more likely, that their ratings are too noisy to be analyzed in a meaningful way. For another three participants (3, 5, 9) only one satisfactory model is determined, which accounts for roughly half of their attributes (17 of 37). The remaining five participants (1, 2, 4, 6, 10) have two different, complementary models, i.e., the number of attributes in the first model (26) is comparable to the number of attributes in the second model (19). This shows that diversity is prevalent. Half of the participants even hold two different views, explaining subgroups of attributes.

All together, 13 different views emerged from the ten individuals. These views may partly overlap, which motivated us to group similar views and identify the major diverse of this user group.
3.3. AN MDS APPROACH TO ACCOUNT FOR DIVERSITY

Table 3.4: Number of attributes explained by the two views for the ten participants of the study.

<table>
<thead>
<tr>
<th>Subj.</th>
<th>Total</th>
<th>View a</th>
<th>View b</th>
<th>Remain</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>13</td>
<td>5</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>13</td>
<td>6</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>14</td>
<td>5</td>
<td>-</td>
<td>9</td>
</tr>
<tr>
<td>4</td>
<td>10</td>
<td>5</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>12</td>
<td>8</td>
<td>-</td>
<td>4</td>
</tr>
<tr>
<td>6</td>
<td>11</td>
<td>6</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>13</td>
<td>-</td>
<td>-</td>
<td>13</td>
</tr>
<tr>
<td>8</td>
<td>11</td>
<td>-</td>
<td>-</td>
<td>11</td>
</tr>
<tr>
<td>9</td>
<td>11</td>
<td>4</td>
<td>-</td>
<td>7</td>
</tr>
<tr>
<td>10</td>
<td>10</td>
<td>4</td>
<td>3</td>
<td>3</td>
</tr>
</tbody>
</table>

of the configuration space. The distance between configurations \( X_n \) and \( X_m \) can be calculated using the MATFIT procedure, developed by Ramsay (1990). MATFIT seeks for a transformation matrix \( M \) that minimizes the distance measure:

\[
d^2 = \text{trace}[(X_n M - X_m)^T (X_n M - X_m)]
\]  

An arbitrary \( K \times K \) transformation matrix \( M \) was applied. The procedure was repeated with the matrices in reverse order as a means to calculating both distances: with \( X_n \) as independent and \( X_m \) as dependent, and vice versa. The resulting distances were visualized in three dimensions using the program XGms (Martens 2003). A hierarchical (minimum variance) clustering algorithm was applied to the 3-D configuration (a cluster is denoted by the lines connecting the different views). Figure 3.3 represents a 2-D perspective on the 3-D configuration of individual models. Note that the distances in this 2D perspective do not necessarily reflect the true distances in 3D, which is why one should rely on the lines that visualize the clusters (clustering was performed in 3D). Participant 7 and 8 are excluded, because no individual model could be fitted. In case of two fitting models per participant (1, 2, 4, 6, 10) the first model is denoted as a, the second as b.

Three clusters of models emerged. Cluster 1 summarizing 6 of the 13 single models (1a, 2b, 4a, 6b, 9, 10b), cluster 2 summarizing 4 models (1b, 2a, 4b, 10a) and cluster 3 summarizing the remaining 3 models (6a, 5, 3). The complementary models (a & b) for these five participants appear to be quite dissimilar as illustrated in figure 3.3 by the fact that they belong to different clusters. These clusters represent homogenous views, which can subsequently be mapped out.
3.3.5 Grouping the homogeneous views

In the last phase we establish a final set of configurations that represent the major diverse views across all participants and all attributes, on the set of stimuli. Views that belong in the same cluster are analyzed together and a shared MDS configuration is sought. Attributes that are not adequately predicted by the model are eliminated with the same criteria as in phase 1. The resulting 'averaged' views are then used for modeling the attributes from all participants. Attributes are allowed to exist in more than one configuration if they are adequately explained by all of them. When attributes in the same semantic category are not significantly different (which can be deduced from the fact that they have overlapping confidence ellipses in the K-dimensional configuration space), they are grouped. Attributes that cannot be grouped (have no replicates) are eliminated since no evidence exists that they contain reliable information.

3.3.6 How do the diverse views compare to the average view?

This question will be addressed in three ways. Firstly, it will be illustrated that the average model predicts less than half of the attributes predicted by the three diverse models together (attributes that are adequately explained by more than one model
are only counted once for the model that explains them best). Secondly, it will be illustrated that, for the attributes that are predicted by the three diverse models, these models provide a better fit than the average model, as demonstrated by the amount of explained variance in the attribute data and the values of the well-established Akaike Information Criterion (AIC) for model selection. Thirdly, by exploring the resulting views, it will be illustrated that the diverse models, not only account for more attributes and with a better fit, but that they also result in semantically richer insights, i.e., introduce more semantically different attributes.

Surprisingly enough, the average model could only predict 1/6th of all the attributes from the ten participants, i.e., 18 out of the 118 attributes. This means, that when deriving an average configuration to understand how individuals distinguish between these websites, only 1/6th of the attributes are taken into account. This is illustrated by the high correlation between the two resulting configurations ($R = .99$), the one derived using all 118 attributes and the one derived using only the 18 attributes that are well predicted. Thus, the consequence of averaging is that we account only for 1/6th of the information available. The three diverse models predict 12, 10, and 16 attributes respectively (attributes predicted by more than one model were excluded from the ones that displayed the least fit). Thus, by accounting for diversity, even with our clearly heuristic and therefore sub-optimal procedure, we account for more than double the number of attributes than in the case of the average model.

Table 3.5 illustrates the goodness of fit of the average and the three diverse models for the 38 in total attributes resulting from models 1 to 3. As expected, a significant increase in the accounted variance ($R^2$) of the attribute data is observed as we move from the average to the specific (i.e., diverse) model. But, does this increase in the goodness of fit of the model outweigh the increase in model complexity, i.e., going from one 2D to three 2D models? One of the most widely used criteria for model selection is the Akaike Information Criterion (AIC) (Burnham and Anderson, 2004) which is a function of the log likelihood value reflecting the goodness of fit of the model and the M degrees of freedom in the model reflecting its complexity:

$$AIC_c = -2 \log(L(\hat{\theta})) + 2M \frac{n}{n - M - 1}$$

Burnham and Anderson (2004) proposed a set of heuristics when comparing the AIC values of two models. $\Delta_i$ reflects the difference between the AIC of the simpler (i.e., average) model and the AIC of the more complicated one (i.e., consisting of three sub-models). $\Delta_i \leq 2$ provides significant evidence for the simpler model, $4 \leq \Delta_i \leq 7$ provides weak support for the heterogeneous model, while $\Delta_i \geq 10$ provides strong support for the heterogeneous model. In our case $\Delta_i = 354 \gg 10$, providing signifi-
Table 3.5: Goodness of fit of the average and the three diverse models for the 38 in total attributes resulting from models 1 to 3

<table>
<thead>
<tr>
<th>Attribute No</th>
<th>Variance (R²)</th>
<th>Average R²</th>
<th>Model 1</th>
<th>Model 2</th>
<th>Model 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>0.36</td>
<td>0.99</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>4.3</td>
<td>0.86</td>
<td>0.99</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>1.7</td>
<td>0.91</td>
<td>0.95</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>4.6</td>
<td>0.91</td>
<td>0.94</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>1.7</td>
<td>0.32</td>
<td>0.91</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>2.7</td>
<td>0.76</td>
<td>0.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>4.2</td>
<td>0.91</td>
<td>0.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>6.6</td>
<td>0.7</td>
<td>0.87</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>1.7</td>
<td>0.21</td>
<td>0.87</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>6.1</td>
<td>0.83</td>
<td>0.86</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>4.2</td>
<td>0.75</td>
<td>0.85</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>3</td>
<td>0.68</td>
<td>0.69</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>2.2</td>
<td>0.39</td>
<td>0.99</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>5.6</td>
<td>0.78</td>
<td>0.98</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>1.9</td>
<td>0.7</td>
<td>0.94</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>6.3</td>
<td>0.82</td>
<td>0.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>4.7</td>
<td>0.68</td>
<td>0.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>2.8</td>
<td>0.54</td>
<td>0.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>1.8</td>
<td>0.89</td>
<td>0.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>3.8</td>
<td>0.68</td>
<td>0.89</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>2.3</td>
<td>0.76</td>
<td>0.89</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>1.9</td>
<td>0.54</td>
<td>0.88</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>4.8</td>
<td>0.75</td>
<td>0.99</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>1</td>
<td>0.95</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>7</td>
<td>0.67</td>
<td>0.94</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>17</td>
<td>0.91</td>
<td>0.94</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>7</td>
<td>0.93</td>
<td>0.94</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>2.6</td>
<td>0.99</td>
<td>0.93</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>4.3</td>
<td>0.89</td>
<td>0.93</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>1.6</td>
<td>0.83</td>
<td>0.92</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>3.7</td>
<td>0.91</td>
<td>0.91</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>5.9</td>
<td>0.91</td>
<td>0.91</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>2.6</td>
<td>0.88</td>
<td>0.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>7.4</td>
<td>0.8</td>
<td>0.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>6.1</td>
<td>0.83</td>
<td>0.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>17</td>
<td>0.8</td>
<td>0.87</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>4.7</td>
<td>0.59</td>
<td>0.85</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>2.7</td>
<td>0.57</td>
<td>0.83</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\[ \text{AIC}_{\text{avg}} = 1480 \quad \text{AIC}_{123} = 1126 \quad \text{Delta} = 354 \]
3.3. AN MDS APPROACH TO ACCOUNT FOR DIVERSITY

Figure 3.4: The average model and the three diverse models corresponding to clusters 1 to 3 in figure 3.3.

cant evidence that the diverse models, despite the increase in the model complexity, perform better than the average model.

Figure 3.4 illustrates the insights gained by the average and the three diverse models that are derived from the views corresponding to clusters 1 through 3. A significant overlap exists between models 1 and 3 (five common attributes), while model 2 provides a completely different view.

The average model, although it accounts for more attributes than each of the
diverse models, fails to predict semantically similar attributes. Thus, replicate attributes (i.e. attributes pointing towards the same direction with overlapping confidence ellipses) exist only for two attribute categories, namely "Fast access to information" and "Supports search". The websites of the university of München and Aachen are differentiated from the remaining ones as websites that provide fast access to information, while the second attribute differentiates mainly the websites of the universities of Mainz and Frankfurt as the ones that do not support searching.

These two attributes are present also in two of the three diverse models, model 1 and model 3. Model 1 further differentiates the websites of Aachen and München as having a "graphical layout", the website of the university of Aachen is mainly differentiated from all others as a website that does not "refer to student life". On the contrary, model 2 provides a different insight. It reveals that the websites of the Universities of Mannheim, Frankfurt and Mainz put "less emphasis on achievement". The set of websites can also be split in two groups based on the amount of information that they provide to the user.

3.4 Discussion

This chapter challenged traditional "averaging" practices in the analysis of repertory grid data. It was illustrated that only 15% of the attributes were adequately predicted by the average model and that the remaining attributes had minimal contribution to the configuration space, as shown by the high correlation (r=0.99) of the two spaces, i.e. one employing all attributes and one employing only the 15% of attributes being adequately predicted. The diverse models accounted for more than double the information accounted by the average model (38 attributes out of all 118), however a substantial amount of attributed remained non-modeled. One might wonder whether the information contained in the remaining attributes is truly non-modeled, or whether these attributes just do not differentiate strongly between the stimuli and thus do not contain any substantial information.

Figure 3.5a displays the variance in the attribute’s ratings versus the accounted variance by the three diverse models for two groups of attributes: the ones that are adequately predicted by the three diverse models (38) and the remaining ones. One may note that the hypothesis stated above does not necessarily hold; no substantial differences may be found in the variance of the modeled and the non-modeled attributes. It thus becomes evident that while the diverse models result in a substantial increase in the information being modeled, still a substantial amount of attributes containing high variance remain non-modeled.

One potential reason for this might be a limitation that we pointed earlier, the use of heuristics in judging the goodness of fit of a model for a given attribute’s data.
3.4. DISCUSSION

An alternative procedure proposed by Martens (2009a) employs an iterative process, grounded on Singular Value Decomposition (Eckart and Young, 1936), that aims at optimizing the overall accounted variance ($R^2$) for all attributes in the data. Figure 3.6 illustrates a comparison of the performance of the heuristic technique versus the algorithmic technique which explicitly optimizes the accounted variance $R^2$. As expected, the algorithmic techniques provides an improvement in terms of the accounted variance $R^2$ - in total 109 out of the 118 (92%) meet the explicit criterium $R^2>.5$ as compared to the 87 attributes (74%) of the heuristic technique. Nevertheless, the technique provides no improvement in terms of $R_k$ (38 attributes meet $R_k>6$), which is somewhat expected as it does not explicitly take this criterion into account in the optimization process. Thus, an ideal technique, should take into account both $R^2$ and $R_k$ in optimizing the goodness of fit of a model for a given set of attributes. While $R^2$ reflects the amount of variance in the attribute’s ratings accounted for by the model, some attributes might contain only limited variance, and thus display high $R^2$ while providing limited meaningful differentiation between the stimuli. While our heuristic technique explicitly took into account the $R_k$, it has a number of limitations that could be addressed in the future. First, while $R_k$ was used as a criterion in the determining whether an attribute is adequately predicted by a given model, it was not employed in the Multi-Dimensional Scaling optimization procedure for determining the best configuration of stimuli when trying to model a set of attributes. Thus, the majority of the attributes resulted in low $R_k$ values. Second, as $R_k$ is the range in the predicted attribute ratings divided by the estimated noise in the attribute’s ratings,
the denominator, i.e. estimated noise, will obviously have a significant impact on the $R_k$, and thus may render it as unreliable if we cannot reliably estimate the noise. In other words, if noise is unrealistically low, this will result in unrealistically high values of $R_k$. This issue pertains to the nature of the dataset as they lack repeated measures, thus noise cannot be directly estimated within a single participant and for a given attribute. Eliciting repeated measures in a repertory grid study would add substantial effort which might render the technique infeasible in several contexts. Figure 3.5b depicts the range $\hat{A}_{k,max} - \hat{A}_{k,min}$ versus the estimated noise ($\sigma_k$) for the 38 attributes that are adequately modeled by the heuristic technique. One may note that for the majority of the attributes $\sigma_k<1$, which is rather unrealistic given the quantization error when transforming the discrete data into continuous. Thus, an ideal technique should not allow for $\sigma_k<1$:

$$R_k = \frac{\hat{A}_{k,max} - \hat{A}_{k,min}}{\min(1, \sigma_k)}$$

(3.5)
Given the limitations of the proposed $R_k$ measure one might then question the validity of our initial result, that only 18 out of all 118 attributes are adequately modeled by the averaging analysis, since $R_k$ was critical for determining goodness of fit. We earlier noted however a high correlation ($r=.99$) of the two configuration spaces, one attempting to model all 118 attributes and one attempting to model only the 18 best predicted attributes. This high correlation means that when excluding the 100 least-fit attributes, there was virtually no change in the MDS configuration space. In other words, when including these 100 least-fit attributes in the MDS optimization procedure, they did not contribute in the creation of this space. We also saw earlier that these attributes were not substantially different from the 18 best-fit in terms of the variance in their ratings. Thus, the averaging analysis results in a configuration space that is optimized for a small fraction of the dataset. We thus argue that despite the limitations of the $R_k$ measure, averaging analysis fails to model a substantial fraction of the dataset.

3.5 Conclusion

This chapter argued against averaging in the analysis of personal attribute judgments. It was illustrated that when using averaging only 1/6th of the attributes in our study, i.e. 18 out of 118, were taken into account. A new MDS procedure that can better account for diversity in judgments was developed and its added value was illustrated through the reanalysis of published data. The analysis resulted in three diverse views on the data which were directly compared to the average view that is the common practice in RGT studies. The diverse models were found a) to account for more than double of the attributes accounted for by the average model, b) to provide a better model fit even for the attributes that were adequately predicted by the average model, and c) to result in semantically richer insights, since the diverse models can account for more semantically different attributes.

It was further illustrated that diversity exists not only across different individuals, but also within a single individual, in the sense that different attribute judgments of a participant may reveal different, complementary, views. At any point in time individuals can have different, seemingly conflicting views. For instance, individuals may regard one car as beautiful, but at the same time expensive. Individuals’ overall evaluations of the car might thus be modulated by contextual aspects such as their motivational orientation (whether they just saw it in a newspaper on a Sunday morning or they are in the process of purchasing it, see Hassenzahl et al. 2008). Thus, being able to understand individuals’ conflicting views is crucial for understanding how individuals infer the overall value of a product.

These insights strongly advocate the view that the analysis of quality judgments
of interactive products should not stop on a group level, but must be extended to the
relations between the attribute judgments within an individual. The Repertory Grid
combined with the suggested technique to analyze the resulting quantitative data
is an important step towards the adequate account of homogeneity and especially
diversity in individual quality judgments.
Part III

Understanding the Dynamics of Experience through Experience Narratives
1 Product evaluation practices have traditionally been focusing on early interactions. Recent studies and trends in the consumer electronics industry, however, highlight the importance of understanding prolonged use. This chapter presents two studies that inquired into how users’ experiences with interactive products develop over time. The first study assessed the ways in which ten individuals formed overall evaluative judgments of a novel interactive product at two moments in time during the adoption of a product, more precisely, in the first week and after four weeks. The second study followed six individuals through an actual purchase of a novel product and inquired into how their expectations and experiences developed from 1 week before until 4 weeks after the purchase of the product.

The chapter attempts two contributions. Firstly, it provides some initial insights into the differences between initial and prolonged experiences in terms of the way users form overall evaluative judgments about a product. Secondly, it raises a number of methodological issues in the assessment of the dynamics of user experience over time. This chapter ends with a proposal for a new approach to the study of the dynamics of experience over time and raises two research questions that will be addressed in chapters 5 and 6 respectively.

1This chapter is partially based on
CHAPTER 4. USER EXPERIENCE OVER TIME

4.1 Introduction

Product evaluation practices have traditionally been focusing on early interactions. As a result, we have been mostly concerned about product qualities relating to initial use. Den Ouden et al. (2006), however, highlighted that the reasons for product returns span a wider range of aspects than just problems relating to the learnability and usability of interactive products. Moreover, a number of recent trends are highlighting the importance of longitudinal evaluation practices. First, legislation and competition within the consumer electronics industry has resulted in an increase in the time-span of product warranties. This has resulted in an alarmingly increasing number of products being returned on the basis of failing to satisfy users’ true needs (Den Ouden et al., 2006). Secondly, products are increasingly becoming service-centered. Often, products are being sold for lower prices and revenues are mainly coming from the supported service (Karapanos et al., 2009). Thus, the overall acceptance of a product shifts from the initial purchase to establishing prolonged use.

This chapter attempts two contributions. First, it attempts to inquire into the differences between initial and prolonged experiences in terms of the way users form overall evaluative judgments about products across time. In the field of user experience, a number of studies have attempted to inquire into how users form overall evaluative judgments of products on the basis of quality perceptions (Hassenzahl, 2004; Mahlke, 2006; Tractinsky and Zmiri, 2006; Hartmann et al., 2008; Van Schaik and Ling, 2008). An aspect that has been largely overlooked so far is that of temporality, i.e. how users’ experiences develop over time. As we use a product, our perception of the qualities of the product will change (von Wilamowitz Moellendorff et al., 2006). For example, we get used to it, which eventually changes our perception of its usability; at the same time it excites us much less than in our first moments with it. Even more interestingly, at different phases of use we will evidently attach different weights to different qualities. In our first interactions with a product we may focus on its usability and the stimulation that it provides to us. After we use it for some time, we might become less concerned about its usability, and other aspects of the product such as novel functionality or communication of a favorable identity to others become more important. Two studies are being reported. In study 1 we provided 10 participants with an innovative pointing device connected to an Interactive TV set top box, and elicited their perceptions during the first week as well as after four weeks of use. In study 2 we followed six participants through the purchase of an Apple iPhone and the first month of use. Both studies provide empirical findings on the differences between initial and prolonged experiences. A conceptual model of temporality of experience is attempted and implications for HCI practice are suggested.
Second, it raises a number of methodological issues in the assessment of the dynamics of user experience over time. It questions the value of reductionist approaches where a-priori defined measurement models are employed in measuring the user experience and suggests an alternative methodological approach that relies on a) eliciting the experiences that are personally meaningful to each participant, in the form of experience narratives, and b) employing content analysis techniques in creating multiple levels of abstraction, from concrete idiosyncratic insights to abstracted and generalized knowledge.

4.2 Background on Experience and Temporality

This section reviews some holistic and reductionist models in user experience research and discusses their relevance for the study of the temporality of users’ experiences.

4.2.1 Temporal aspects in frameworks of Experience

The holistic thread of user experience has contributed a number of frameworks describing how experience is formed, adapted, and communicated in social contexts. Forlizzi and Battarbee (2004) described how experience transcends from unconsciousness to a cognitive state and finally becomes "an experience", something memorable that can also be communicated in social interactions. Battarbee and Koskinen (2005) elaborated on the social mechanisms that lift or downgrade experiences as they participate in our social interactions. McCarthy and Wright (2004) described how sense-making takes place in the development of experience by decomposing it into six processes, from anticipation to reflection and recounting.

Although one can note that these frameworks approach temporality through a micro-perspective, i.e. how experiences are formed, modified and stored, one could also raise a number of macro-temporal issues. For instance, does the distribution between unconscious and cognitive experiences remain stable over time or do cognitive experiences reduce as users’ familiarity increases? Next, what motivates the process of lifting up experiences and communicating them in social contexts? Do these underlying motivations change over time, e.g. as users’ initial excitement fades out? A framework of temporality of experience, proposed in this chapter, aims at providing answers to these questions by conceptualizing the missing dimension of time.

4.2.2 Beauty, Goodness and Time

Reductionist approaches to user experience have contribute a wealth of new measurement and structural models. Hassenzahl (2004) distinguished between two qual-
### Figure 4.1: Hassenzahl’s framework distinguishing between two overall evaluative judgments, i.e. goodness and beauty, and three quality perceptions, i.e. pragmatic quality, hedonic quality - stimulation, and hedonic quality - stimulation.

<table>
<thead>
<tr>
<th>&quot;Objective&quot;</th>
<th>Perceptions</th>
<th>Evaluations</th>
</tr>
</thead>
<tbody>
<tr>
<td>mental effort (indicates actual usability problems)</td>
<td>hedonic - identification</td>
<td>goodness</td>
</tr>
<tr>
<td>pragmatic (perceived usability)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Pragmatic quality** refers to the product’s ability to support the achievement of behavioral goals (i.e. usefulness and ease-of-use). On the contrary, **hedonic quality** refers to the users’ self; it relates to **stimulation**, i.e. the product’s ability to stimulate and enable personal growth, and **identification**, i.e. the product’s ability to address the need of expressing one’s self through objects one owns. Tractinsky and Zmiri (2006) drew on the work of Rafaeli and Vilnai-Yavetz (2004) to propose three distinct product quality attributes: usability, aesthetics, and symbolism. Forlizzi (2007) extended this model to further account for the emotional and social aspects of product use.

An interesting question relates to how these quality perceptions are combined to form an overall evaluation of the product (Hassenzahl, 2004; Mahlke, 2006; Tractinsky and Zmiri, 2006). Hartmann et al. (2008) and Van Schaik and Ling (2008) suggested two distinct overall evaluative judgments of the quality of interactive products: **beauty** and **goodness**. He found goodness to be affected primarily by pragmatic aspects (i.e. usefulness and usability). On the contrary he found beauty to be a rather social aspect, largely affected by identification (i.e. the product’s ability to address the need of self-expression). In a similar vein, Tractinsky and Zmiri (2006) distinguished between satisfying and pleasant experiences. They found perceptions of usability to be better predictors for a satisfying rather than pleasant experience while perceptions of the products’ aesthetics to be better predictors for a pleasant rather than satisfying experience.

But, how stable are such relations over time? Study 1 attempts an initial insight in the differences between initial and prolonged experiences in the way users form overall judgments about products. In Hassenzahl’s study the users’ experience with
the products was limited to short interaction episodes, where users were asked to carry out a number of predefined tasks. Study 1 attempts to explore how the user experience and the subsequent evaluative judgments develop over a longer period of time, and in less controlled interactions.

4.3 Study 1

The objective of study 1 is to understand how users form evaluative judgments during the first experiences with a product and after prolonged use. Hassenzahl’s (2004) distinction between the two evaluative judgments of goodness and beauty was employed in an effort to replicate existing findings and extend them over prolonged use. Given previous results one would expect beauty to be relatively stable over time (Lindgaard et al., 2006; Tractinsky et al., 2006; Hassenzahl, 2004) and related to the self-image (Hassenzahl, 2004; Tractinsky and Zmiri, 2006; Tractinsky et al., 2006) that the product communicates to relevant others. Further, one would expect judgments of goodness to be primarily affected by the product’s pragmatic aspects (i.e. utility and usability) (Hassenzahl, 2004; Mahlke, 2006).

4.3.1 Method

Participants

A total of ten individuals (four female) participated in a four weeks study of an Interactive TV set-top box (STB). They all responded to an invitation that was placed on the website of a multinational consumer electronics company. Their age varied from 22 to 35 years (mean 26y). Their likelihood to recommend the brand ranged from 3 to 9 on a 10-point scale (mean 7.8, std. 2.3). Participants were classified to respective market segments based on demographic information. A bias towards innovator consumers was observed, as it was expected. The study focused on a particular part of the set-top box, uWand. uWand is a novel pointing device for interactive TV contexts. It uses led technology to identify where the user points at within the content that appears on the TV.

Procedure

During the four week testing period participants were asked to rate uWand at two different times, during the first week of use as well as at the end of the 4th week. The AttracDiff 2 questionnaire (Hassenzahl, 2004) was employed for the assessment of three distinct aspects of the quality of interactive products: pragmatics (e.g. simple, practical, clear), stimulation (e.g. innovative, exciting, new) and identification (e.g. unique, familiar, personal).
in inclusive, classy, presentable). Each quality aspect is measured with seven bipolar attributes, employed in Semantic Differential scales (Osgood et al., 1957). Beauty and goodness were measured with single items (taken from AttracDiffs appeal construct). Both are evaluative judgments with Goodness focusing on the complete product, while Beauty is rather restricted to visual features. Note that for evaluative, high level summary judgments single item measurements are appropriate and commonly used (e.g., to measure subjective well-being).

4.3.2 Results

Since we were interested in a detailed picture of the relationship between product attributes, we decided to analyze every attribute (in Table 4.1) separately. The distance $D_{ij} = 1 - |R_{ij}|$ between the individual attributes $i$ and $j$ was derived from the correlation $R_{ij}$ between the ratings by different participants on the two attributes. The obtained distances were subsequently visualized in three dimensions (Stress value $S=0.19$ 1st week; $S=0.15$ 4th week) using the Multidimensional Scaling tool XGms (Martens, 2003). Hierarchical clustering (with minimum variance) was applied to the predicted distances in the 3-dimensional space. Figure 4.3 illustrates a 2D projection of the 3-dimensional visualization of the distances between the quality attributes. The obtained clusters are denoted by the connecting lines. The left figure reflects the users’ ratings during the first week of use while the right figure reflects the users’ ratings after four weeks of use. All in all, clusters derived from the measurement in the first week reflect the assumed underlying constructs, with a close knit groups of mainly pragmatic and hedonic stimulation attributes and a looser rest of hedonic.
### Table 4.1: Bipolar attributes measuring pragmatic quality, stimulation and identification (Hassenzahl, 2004)

<table>
<thead>
<tr>
<th>Pragmatic Quality</th>
<th>Stimulation</th>
<th>Identification</th>
</tr>
</thead>
<tbody>
<tr>
<td>Technical-Human</td>
<td>Typical-Original</td>
<td>Isolating-Integrating</td>
</tr>
<tr>
<td>Complicated-Simple</td>
<td>Standard-Creative</td>
<td>Amateurish-Professional</td>
</tr>
<tr>
<td>Impractical-Practical</td>
<td>Cautious-Courageous</td>
<td>Gaudy-Classy</td>
</tr>
<tr>
<td>Cumbersome-Direct</td>
<td>Conservative-</td>
<td>Cheap-Valuable</td>
</tr>
<tr>
<td></td>
<td>Innovative</td>
<td></td>
</tr>
<tr>
<td>Unpredictable-Predictable</td>
<td>Lame-Exciting</td>
<td>Noninclusive-Inclusive</td>
</tr>
<tr>
<td>Confusing-Clear</td>
<td>Easy-Challenging</td>
<td>Takes me distant from people -</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Brings me closer to people</td>
</tr>
<tr>
<td>Unruly-Manageable</td>
<td>Commonplace-New</td>
<td>Unpresentable-Presentable</td>
</tr>
</tbody>
</table>

identification attributes. After four weeks, the perceptions seem much more differentiated, and relationships among attributes lost some of their strength.

During the first week, Beauty judgments relate mostly to attributes reflecting the quality of stimulation (i.e. original, creative, new, innovative) and to one attribute reflecting identification (i.e. classy). This close relationship between stimulation and beauty seems to have disappeared after four weeks of use; beauty now seems to be a disjoint judgment only related to one identification (i.e. classy) and one stimulation (i.e. challenging) attribute. While the relation between “classy” and beauty remained over the period of four weeks, stimulation seemed to influence beauty judgments mostly during the first experiences. Surprisingly, stimulation seemed to be a more important determinant of beauty than identification in the first experiences.

A similar pattern can be observed for judgments of Goodness. During the first week judgments of goodness relate mostly to pragmatic quality attributes (i.e. practical, direct, manageable, predictable, clear) and to one attribute reflecting identification (i.e. presentable). After four weeks of use goodness appears to be related mostly to identification (i.e. professional, inclusive, valuable, integrating, brings me closer to people) while a weaker relation can be observed with attributes relating to stimulation (i.e. creative, courageous, original, creative). Pragmatic aspects seem to be relevant for goodness judgments only for the first experiences with a product. Over time, identification (i.e. what the product expresses about its owner) becomes a more prominent aspect of the goodness of the product.

### 4.3.3 Discussion

Two questions were of interest in the current study: what makes a product good or beautiful, and how does this develop over time.

As far as goodness judgments are concerned, we partially replicated Hassenzahl’s
Figure 4.3: 2D view on the 3-dimensional visualization of distances between quality attributes, beauty and goodness. Users’ perceptions during the 1st week of use (left) and after 4 weeks of use (right).
4.3. STUDY 1

(2004) results. During the first experiences facets of pragmatic quality were the qualities being most related to goodness. Users are still exploring the functionality of the product, trying out new things and experiencing usability problems. As people get used to using the product they learn to handle usability problems; at the same time they restrain themselves only to part of the product’s functionality that is most attractive to them. The value of the product is now derived on a different basis, being ownership-based rather than usage-based. Social aspects (i.e. identification) became more prominent.

For beauty judgments, however, the results seemed more divergent. While Hassenzahl (2004) found identification to be the quality being most closely related to beauty judgments, we found stimulation to be even more prominent than identification in the first experiences. In both cases, beauty seems to be related more to hedonic than to pragmatic aspects.

This different finding can possible be attributed to the product sample. While the current study employed a novel consumer electronics product, Hassenzahl’s study focused on different variations of mp3 player skins; such variations in aesthetics and usability of the same product (i.e. a virtual mp3 player) may not affect the perceived stimulation (i.e. innovation, novelty). Mahlke (2006), for instance, observed a quality called expressive aesthetics (Lavie and Tractinsky 2004b), arguably comparable to the quality here called stimulation, to have an impact on goodness but not on beauty judgments, during the first experiences with tangible mp3 players. The nature of the experience also differed significantly in this study. Both Hassenzahl (2004) and Mahlke (2006) asked participants to carry out a number of tasks in a laboratory context; in this study participants had the opportunity to use the product at their homes over an extended period. The first evaluation took place during the first days of use.

After four weeks of use, stimulation seemed to loose its dominance on beauty judgments. Eventually, users were not any more surprised by the product’s stimulating character and the product’s novelty lost its power to make the product more beautiful in the users’ eyes.

Overall, despite the exploratory character of the study, it seems that we came across some interesting results. The proposition that novelty and stimulation impact beauty judgments resonates with Berlyne’s work on stimulation and surprise as arousal-heightening attributes and their impact on the appraisal of art (Machotka 1980). Furthermore, time seems to have an impact on the importance we attach to different qualities of the experience with interactive products. For example, despite the crucial importance of usability in a product’s initial acceptance, aspects of product ownership (and not use) are essential for a user to resonate with a product and value it in the long term.
4.3.4 Limitations of the study

A number of limitations of the study have to be noted before proceeding to general conclusions. First, an inherent limitation of the study was the evaluation of only one product and only by a small number (10) of participants. As a result, these findings cannot be generalized to a wider population and range of products; they provide, however, some initial insights that can be falsified or corroborated in subsequent studies with different products and participants. However, it highlights an inherent limitation of longitudinal studies in producing generalized knowledge due to their labor intensive nature which evidently restricts the sample of products, participants, or time studied. Alternative techniques such as the one we will propose in the next chapter, and others proposed by von Wilamowitz Moellendorff et al. (2006) and Fenko et al. (2009) are essential for overcoming the labor-intensive nature of longitudinal studies.

Second, the two measurements (1st and 4th week of use) assessed users’ perceptions at each time (current state) rather than directly assessing how their perceptions changed over time. One could be concerned about the sensitivity as well as the reliability of such absolute measures where judgments do not take place in contrast to some specified reference point. As discussed in chapter 2, paired comparison scales that define explicitly both reference points, have been shown to be less sensitive to contextual effects than single-stimulus scales. Thus, one would assume that by explicitly asking participants to assess how their opinion on a given quality has changed over the course of time, may increase the technique’s sensitivity in assessing this change.

Third, one major limitation of the study was the use of a-priori defined measurement model in a pre-post design (with only two measures being taken). Although in such a study design, measurement has the least influence on users’ attitudes, it provides a rather limited insight into the actual patterns of change over time. First, as only two measures were taken, one could not know whether the identified changes were an effect of time, or of random contextual variation. Second, as we employed an a-priori defined measurement model (Hassenzahl 2004), we might have evidently missed constructs that dominate prolonged use and are not apparent in this model. Secondly, due to the quantitative nature of the study, we gained no insight as to the reasons for these changes in users’ experiences. In study 2, we attempt a qualitative understanding into the differences between initial and prolonged use.
4.4 Study 2

Study 1 provides some evidence for the point of view that prolonged experiences with products can differ substantially from initial experiences in terms of the way that different product qualities relate to each other.

The question that was raised then was: what causes these changes? Can we describe the adoption of a product in terms of distinct phases? And what qualities would dominate each of these phases? While longitudinal studies on product adoption are scarce in the field of HCI, much work has been performed in the field of cultural studies of technology [Du Gay et al., 1997; Silverstone and Haddon, 1996], trying to understand how technology is being adopted and incorporated in specific cultures. We agree with McCarthy and Wright [2004] that cultural studies have a tendency to downplay the role and diversity of individual experience, yet, we believe that much can be learned from examining the relevance of cultural studies frameworks for the study of user experience.

A promising framework for the study of prolonged user experiences is the one from Silverstone and Haddon [1996] on the dimensions of adoption. They suggested three dimensions, but also moments, in the process of technology adoption: commodification, appropriation and conversion. Commodification, they argued, refers to all activities from both producers and users that result in specific claims for a function and an identity for a new product. As users participate in the commodification process, they form expectations about ways in which the product could become relevant to their lives. In appropriation, users accept enough of the relevance of the product and they gradually incorporate it into their life routines. Finally, in conversion, users accept the product as part of their self-identity and employ it in their social interactions.

Silverstone and Haddon’s framework, however, approach product adoption from a cultural and macro-temporal perspective, thus undermining the details that describe how individuals’ experiences develop over time. For instance, commodification is conceived as an iterative process where both users and producers make claims for new functions, eventually resulting in new products in the market. They are less concerned about how expectations impact users’ experience with a product. Next, how exactly does appropriation happen? As it will become evident later, we distinguish between two aspects of appropriation, namely orientation and incorporation.

This study, inspired by the framework of Silverstone and Haddon [1996], uses the iPhone to validate distinct phases in users’ experience, and understand what differentiates them, how users’ experience changes across these phases, and how this impacts users’ evaluative judgments about the product. More specifically, it addresses the following questions:
CHAPTER 4. USER EXPERIENCE OVER TIME

1. Can users’ experiences be articulated in distinct phases in the adoption of the product?
2. What motivates the transition across these phases?
3. How does each phase contribute to the overall perceived quality of the product?

4.4.1 The Study

Product

We selected the iPhone as a product of study because of its proven success not only during initial but also over prolonged use. This was considered important as we wanted to elicit experiences relating to the successful adoption of a product in the different phases as identified by Silverstone and Haddon (1996). The iPhone was moreover considered an interesting example as it highlights some non-instrumental aspects of experience (e.g. stimulation & identification (Hassenzahl, 2004)) that are currently discussed in the field of user experience.

Participants

We recruited 6 participants through a prescreening virtual advert of an iPhone sale. Our motivation was to recruit participants that were at that time considering the purchase of the product before motivating them to participate in the study with a monetary incentive. After responding to the advert, a second email was sent, introducing the details of the study and inviting them to participate. We observed a strong bias for participants with technical background. In the final selection we aimed for a homogeneous participant sample; only one participant did not previously own a smart phone. Their age ranged from 28 to 33 years (mean 31y). Two out of six were female.

Method

Our criteria for choosing a method were a) its ability to retrieve accurate recalls on the product’s perceived quality within single experiential episodes, and b) its ability to elicit rich qualitative accounts on the experienced episodes. We chose the Day Reconstruction Method (DRM) (Kahneman et al., 2004; Schwarz et al., 2008) over the more popular Experience Sampling Method (ESM) (Hektner et al., 2007) and event-contingent diaries (Bolger et al., 2003), as it enables capturing rich qualitative accounts offline.

The DRM is typically conducted at the end of a reported day or at the beginning of the next day. In an effort to minimize retrospection biases, DRM asks participants
to mentally reconstruct their daily experiences as a continuous series of episodes, writing a brief name for each one. Experiential episodes are thus being recalled in relation to preceding ones, which enables participants to draw on episodic memory when reporting on the felt experience (Schwarz et al., 2008). Hence, participants are better able to reflect on the perceived quality of the product within a single experiential episode, avoiding inferences from their global beliefs about the product. As demonstrated by Kahneman et al. (2004), the DRM combines the advantages of an offline method with the accuracy of introspective approaches such as the Experience Sampling.

Process

One week before the purchase of the product, participants were introduced to the study. During this week, participants were asked to capture their major expectations about the product in the form of short narratives. The perceived importance of each expectation was assessed, using a 7-point Likert scale, both before the purchase as well as at the end of the study. After purchase, participants captured their daily experiences at the end of each day. This process consisted of two main activities: day reconstruction, and experience narration. In day reconstruction, participants listed all activities of the day that somehow related to their iPhone. A brief name and an estimation of time spent were recorded for each activity. In experience narration, participants were asked to pick the three most impactful, either satisfying or dissatisfying, experiences of that day. They were explicitly instructed to "use [their] own feeling or a definition of what 'satisfying' and 'dissatisfying' experience means". For each of the three experiences, participants were asked to write a story that describes in detail the situation, their feelings and their momentary perceptions of the product.

Finally, for each experience narration, participants rated the product as perceived within that specific situation (see figure 4.4). A shortened version of the Attrakdiff 2
CHAPTER 4. USER EXPERIENCE OVER TIME

<table>
<thead>
<tr>
<th>Latent construct</th>
<th>Measurement item</th>
</tr>
</thead>
<tbody>
<tr>
<td>Usefulness</td>
<td>useful - useless</td>
</tr>
<tr>
<td>Ease-of-use</td>
<td>simple - complicated</td>
</tr>
<tr>
<td>Stimulation</td>
<td>innovative - conservative</td>
</tr>
<tr>
<td>Identification</td>
<td>positive message about me - negative message about me</td>
</tr>
<tr>
<td>Goodness</td>
<td>good - bad</td>
</tr>
<tr>
<td>Beauty</td>
<td>beautiful - ugly</td>
</tr>
</tbody>
</table>

[Hassenzahl, 2004] questionnaire was employed, that identifies two overall evaluative judgments, i.e. beauty and goodness, and three distinct product qualities: pragmatics (i.e. utility and ease-of-use), stimulation (i.e. the product’s ability to address the human need of stimulation, novelty and challenge) and identification (i.e. the product’s ability to address the need of expressing one’s self through objects one owns). Each construct was measured with one single item (see table 4.2) that displayed the highest loading on the latent construct during a prior study [Karapanos et al., 2008a]. Pragmatic quality was split in two distinct components, usefulness and ease-of-use. Identification was measured with a single item derived from Tractinsky and Zmiri (2006).

4.4.2 Data analysis

A total of 482 experience narratives were collected during the four weeks of use. These were submitted to a conventional qualitative Content Analysis (CA) [Hsieh and Shannon, 2005; Krippendorf, 2004]. Conventional CA is appropriate when prior theory exists but the researcher wishes to stay open to unexpected themes and only at a later stage relate findings to existing theory, whilst it shares a similar analytical approach with Grounded Theory. Our approach consisted of three steps:

Open coding - A detailed coding aimed at identifying key themes in the data without imposing pre-conceived categories. The process resulted in about 70 loosely connected codes referring to about 700 instances in the data.

Axial coding - In the second step, the initial set of phenomena described by open codes was categorized using axial coding. Open codes were grouped into categories which were subsequently analyzed in terms of properties and dimensions. This resulted in a set of 15 main categories reflecting aspects like the aesthetics of interaction, learnability and long-term usability (see table 4.3).

Quantitative analysis - All experience narratives were classified as being primarily related to one of the fifteen categories. This process was independently conducted by the first author and an additional researcher (Interrater agreement K=.88). Both
4.4. STUDY 2

Table 4.3: The coding scheme of the analysis

<table>
<thead>
<tr>
<th>Orientation</th>
<th>Incorporation</th>
<th>Identification</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Satisfying experiences</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>7. Capturing momentary information</td>
<td>12. Sense of community</td>
</tr>
<tr>
<td></td>
<td>8. Avoiding negative social situations</td>
<td></td>
</tr>
<tr>
<td><strong>Disatisfying experiences</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>15. Usefulness problems</td>
<td></td>
</tr>
</tbody>
</table>

Researchers were already immersed in the data as they both participated in the axial coding process. Narratives for which no agreement was attained were excluded from the subsequent analysis. We avoided clarifying disagreements to ensure high uniformity within experience groups. The distribution of experience narratives over the four weeks of the study was then identified for each of the 15 categories. Based on the resulting temporal patterns and semantic information, the 15 categories were then mapped into 3 broad themes reflecting distinct phases in the adoption of the product: *Orientation*, *Incorporation* and *Identification*. An additional theme, called *Anticipation*, was added to reflect users’ a priori expectations that were captured during the first week of the study. Finally, separate regression analyses with the two overall evaluative judgments, i.e. *goodness* and *beauty*, as dependent and the four quality attributes, i.e. *usefulness*, *ease-of-use*, *stimulation* and *identification*, as independent variables, were run for the three main groups of experiences, i.e. Orientation, Incorporation, and Identification, to understand what product qualities dominate in each phase of use.

4.4.3 FINDINGS

All in all, three phases were identified in the adoption of the product, i.e. *Orientation*, *Incorporation*, and *Identification*. These phases reflected different qualities of the product, which were found to display distinct temporal patterns. We conceptualized temporality of experience as consisting of three main forces, i.e. an increasing *familiarity*, *functional dependency* and *emotional attachment*. These forces motivate the transition across the three phases in the adoption of the product (figure 4.5).
Figure 4.5: Temporality of experience, consisting of three main forces, an increasing familiarity, functional dependency and emotional attachment, all responsible for shifting users’ experience across three phases: orientation, incorporation and identification. In each phase, different product qualities are appreciated.

Anticipation, i.e. the act of anticipating an experience resulting in the formation of expectations, happens prior to any actual experience of use. Micro-temporality, i.e. the emergence of a single experiential episode, is thus visualized as the transition from the core of the circle towards its outer radius. Our interactions are typically filled with a multitude of such experiential episodes. Each of these experiences highlights different qualities of the product such as its aesthetics or its daily usefulness. While many different experiences may co-exist in a single time unit (e.g. day), their distribution changes over time, reflecting the transition across different phases in the adoption of the product.

Orientation refers to users’ initial experiences that are pervaded by a feeling of excitement as well as frustration as we experience novel features and encounter learnability flaws. In Incorporation we reflect on how the product becomes meaningful in our daily lives. Here, long-term usability becomes even more important than the initial learnability and the product’s usefulness becomes the major factor impacting our overall evaluative judgments. Finally, as we accept the product in our lives, it participates in our social interactions, communicating parts of our self-identity that serve to either differentiate us from others or connect us to others by creating a sense of community. This phase we call Identification.
Next, we illustrate how this framework was developed from the actual study by addressing our three overall questions:

**Can users’ experiences be articulated in distinct phases in the adoption of the product?**

**Anticipation**

Participants formed an average of six pre-purchase expectations. Expectations related to opportunities for positive experiences (76%) such as the performance of the multi-touch screen, the incorporation of a mobile agenda and mobile internet in daily life, the aesthetics of packaging and product, as well as friends’ and colleagues’ reactions,

”...I bought my iPod not only as a music player but also as an organizer. But synchronizing iPod with my iCal was not that easy and I could not even add anything to my agenda using iPod (very bad of Apple). The iPhone will make my life much easier because of its seamless integration with Mac’s iCal. I can add events using both devices and they will talk to each other as two natives talk...”

but also to fears of negative implications (24%) such as battery endurance, typing efficiency, as well as reliability and tolerance in daily accidents (e.g. drop on the ground):

”My last phone had a QWERTY keyboard that I liked very much. I am curious how the virtual keyboard will be working on the iPhone. I hope it’s not going to have too small keys and it will be really responsive.”

**Orientation**

Orientation refers to all our early experiences that are pervaded by a feeling of excitement as well as frustration as we experience novel features and encounter learnability flaws. These experiences displayed a sharp decrease after the first week of use (see figure 4.6).

Satisfying experiences (N=71) related to Stimulation (N=33) induced by the product’s visual aesthetics (N=12) and the aesthetics in interaction (N=21), but also to positive surprises regarding the simplicity with which certain initial tasks could be carried out, i.e. learnability (N=38):

[Visual aesthetics, day 1] “my first impression when I saw the box was WOW!, very nice!!”, [Aesthetics in interaction, day] “when I clicked on the album, I just loved the way it turned around and showed all the songs in it”, [Learnability, day 2], “I tried to set up my iPhone’s WiFi which I expected would be a little bit difficult...
it was just 3 steps away! amazing! 3 steps away! It automatically detected the WLAN and then connected to it. My iPhone was ready for internet browsing in less than a minute. Just cool!!!"

Dissatisfying experiences reflected learnability problems (N=50) induced by unexpected product behavior:

“[day 3] “I started typing an SMS in Polish and the dictionary tried to help me by providing the closest English word. There was no button to switch the dictionary off, no easy option to edit my preferences about it.”

Incorporation

As participants gradually incorporated the product in their lives, their experiences increasingly reflected the ways in which the product was becoming meaningful in diverse use contexts (see figure 4.6).

Satisfying experiences (N=113) related to design aspects that enhanced users’ efficiency over time, i.e. long-term usability (N=43), but also to the product’s usefulness (N=70), reflecting ways in which the product supported participants’ daily activities. These related to providing fast access to information (N=33) when mobile, or at home, by alleviating boredom in idle periods (N=18) through activities such as browsing the web, browsing photos or playing games, by enabling capturing momentary information (N=11) when mobile, either probed by external stimuli or during introspection, and by avoiding negative social situations (N=8), e.g. when identifying typed phone numbers from contact list before establishing a call, enabling easy access to destination time when calling abroad, or allowing a fast mute of all sounds when being in a meeting:

[Long-term usability, day 3] “turning the iPhone sideways not only turns the page but also magnifies it, so text is easily readable. Truly well done! I don’t see this kind of attention to details too often”, [fast access to information, day 3] “it’s so easy to just pick up the phone to check the web rather than having to switch the computer on - I am becoming a great fan of it. It’s simply saving time”, [alleviating boredom in idle periods, day 7] “I like playing - I find it a nice activity when waiting, traveling and at any point when I can’t really do anything else”, [capturing momentary information, day 12] “Now I tend to go joking when I want to think of my work as I can easily write down whatever comes to my head”, [avoiding negative social situations, day 22] “It was so nice that iPhone recognized a phone number from my contacts list and showed it to me before I started calling. Thanks to that I didn’t leave yet another voice message that would be staying there for another week or two.”
Dissatisfying experiences (N=130) related to long-term usability problems (N=69), and to usefulness problems (N=61), i.e. expected but missing features,

"[Long-term usability problems, day 23] "When I wear gloves I am not able to work with iPhone. It is really impractical when I am cycling or riding a motorcycle", [day 23] "...carrying iPhone in one hand and then pressing the button at the very bottom to take a picture was quite difficult. It is difficult to balance it", [usefulness problems, day 3] "... I could not believe it had no zoom! I messed around for a while but all in vain. Why someone should zoom while taking pictures from iPhone? Right? Simplicity is key...make products simple and do not even give those features which people actually want!!!"

Identification

Finally, identification reflected ways in which participants formed a personal relationship with the product as it was increasingly incorporated in their daily routines and interactions.

Identification was found to have two perspectives: personal and social. Participants were increasingly identifying with the product as they were investing time in adapting
and personalizing it (N=23), but also as the product was associated with daily rituals (N=8):

[personalization, day 14] “I downloaded a new theme ... It looks very beautiful. Now my iPhone looks much much better than before”, [day 27] “Today I tried this application again to categorize application icons on the screen... Now my screen looks so nice and clean, just the way I wanted it to be”, [daily rituals, day 9] “I put a lot of pictures of my daughter on the iPhone... I like that functionality very much, and I look at the pictures at least a few times a day.”

Next, identification experiences related also to the social aspects of product ownership, in two ways: enabling self-expression and creating a sense of community. Self-expressive (N=18) experiences addressed participants’ need to differentiate themselves from others:

[Day 8] “...I had the chance to show off my iPhone to some of my colleagues. I showed them some functions that are rather difficult to operate in other phones... I felt good having a BETTER device. I still have some cards to show which I will in do due time to surprise them even more.”

Often, such experiences were initiated as an ice-breaker to initiate a conversation. Especially when meeting friends who also owned an iPhone, participants reported that this was always a topic of discussion. These conversations addressed individuals’ need to feel part of a group with shared values and interests (N=13), creating in this way a sense of community:

[Day 25] “Yet another friend of ours has an iPhone. It’s a guaranteed subject of conversation if you see another person having it... we chatted about how many applications and which we have on it. It is nice to get recommendations for new cool stuff you could use”

Experiences relating to Identification displayed a more complex trend (figure 4.6). While experiences reflecting the personal side of identification increased over time, social experiences displayed an initial decrease, but also a gradual and sustaining increase. These two patterns were found to be rooted in distinct aspects of social identification. Experiences relating to self-expression (median day=8), e.g. announcing the recent possession in social contexts, wore off along with users’ initial excitement. Experiences relating to the feeling of being part of a community sharing similar values and interests, however, displayed an increasing and sustaining effect (median day=24).
What motivates the transition across these phases?

These temporal patterns were found to relate to three underlying forces: familiarity, functional dependency and social and emotional attachment. First, as users' familiarity with the product increased, the number of experiences relating to learnability problems, but also stimulation and self-expressive identification decreased:

[Day 15] "My typing speed on iphone is gradually improving... now I am a big fan of this keyboard and I find it very comfortable and easy to use", [Day 20] "With today's busy schedule I didn’t even remember I had an iPhone. I think the toy becomes just a nice daily use item - usable and good to have but the initial excitement seems to be gone."

Second, as users incorporated the product in their daily lives, they were experiencing an increasing functional dependency, resulting in experiences relating to the product’s usefulness and long-term usability:

[day 10] "...I am becoming a great fan of it. It’s simply saving time", [Day 15] "...I’ve slowly started adapting to those things and I must say it feels like my phone-life got a little bit easier."

Last, as the product is incorporated in users’ lives, it not only provides the benefits that were intended by the designers but also becomes a personal object, participating in private and social contexts, resulting in an increasing emotional attachment to the product:

[Day 18] "My daughter seems to be attracted to everything that shines, and whenever she spots the iPhone she grabs it. I try to distract her, by giving her the iPhone’s case. Unfortunately she is smarter than that : I find it very funny to see that she likes the same things as me", [Day 2] "In the evening we had friends over for dinner. They are also quite technology freaks. Quite quickly I told them that I’ve got an iPhone and showed it to them. I really liked watching them playing with it...”

How does each phase contribute to the overall perceived quality of the product?

[Hassenzahl] distinguished between two overall evaluative judgments of the quality of interactive products, namely judgments of Goodness and of Beauty. While prior work suggests goodness to be a goal-oriented evaluation, relating to the pragmatic quality of the product (usefulness and ease-of-use), and beauty a pleasure-oriented evaluation, relating to hedonic quality (stimulation and identification) (Hassenzahl, 2004; Tractinsky and Zmiri, 2006; Mahlke, 2006; Van Schaik and Ling, 2008), we saw something different. In each phase, different qualities of the product were crucial for its gradual acceptance.
98

CHAPTER 4. USER EXPERIENCE OVER TIME

Table 4.4: Multiple Regression analysis with usefulness, ease-of-use, stimulation and identification as predictors and Goodness as predicted (β values and significances * p<.05, ** p<.001) for both satisfying and dissatisfying experiences.

<table>
<thead>
<tr>
<th>Orientation</th>
<th>Incorporation</th>
<th>Identification</th>
</tr>
</thead>
<tbody>
<tr>
<td>Usefulness</td>
<td>.49**</td>
<td></td>
</tr>
<tr>
<td>Ease-of-use</td>
<td>.43**</td>
<td>.19**</td>
</tr>
<tr>
<td>Stimulation</td>
<td>.43**</td>
<td>.22**</td>
</tr>
<tr>
<td>Identification</td>
<td>.14**</td>
<td>.53**</td>
</tr>
<tr>
<td>Adjusted R²</td>
<td>.63</td>
<td>.79</td>
</tr>
</tbody>
</table>

Table 4.5: Multiple Regression analysis with usefulness, ease-of-use, stimulation and identification as predictors and Beauty as predicted (β values and significances * p<.05, ** p<.001) for both satisfying and dissatisfying experiences.

<table>
<thead>
<tr>
<th>Orientation</th>
<th>Incorporation</th>
<th>Identification</th>
</tr>
</thead>
<tbody>
<tr>
<td>Usefulness</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ease-of-use</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stimulation</td>
<td>.22*</td>
<td>.27**</td>
</tr>
<tr>
<td>Identification</td>
<td>.51**</td>
<td>.47**</td>
</tr>
<tr>
<td>Adjusted R²</td>
<td>.47</td>
<td>.44</td>
</tr>
</tbody>
</table>

Tables 4.4 and 4.5 display regression analyses with usefulness, ease-of-use, stimulation and identification as predictors and Goodness or Beauty as predicted variables. Each variable is measured through a single item. Each case depicts users’ ratings within a single experience. Cases were categorized in three groups based on whether an individual experience was classified as relating to the orientation, the incorporation, or the identification phase.

While during Orientation the Goodness of the product was primarily derived on the basis of its ease-of-use (Regression analysis: β=0.43, t=4.79, p<.001) and stimulation (β=0.43, t=4.79, p<.001), in Incorporation, the product’s usefulness (β=0.49, t=10.84, p<.001) became the primary predictor of Goodness, and in the phase of Identification the qualities of identification (β=0.53, t=3.57, p<.01) and ease-of-use (β=0.44, t=2.96, p<.01) became the most dominant qualities impacting the overall goodness of the product.

As expected, Beauty appeared to be highly related to the quality of identification, i.e. the social meanings that the product communicates about its owner (Orientation: β=0.51, t=4.32, p<.001, Incorporation: β=0.47, t=8.17, p<.001, Identification: β=0.78, t=5.73, p<.001), and stimulation (Orientation: β=0.22, t=1.89, p=.06, Incorporation: β=0.27, t=4.69, p<.001).

Next, we found a priori expectations to have surprisingly limited impact on the actual experience with the product. Based on earlier research, one would expect a priori expectations to have a major role in forming overall evaluative judgments
Confirming a priori expectations has been seen as the major source of satisfaction both in HCI (Lindgaard and Dudek, 2003) and Information Systems (Oliver, 1980) research. The comparison standards paradigm (Oliver, 1980), which dominates user satisfaction research, posits that individuals form stable expectations to which the actual product performance is compared, to derive a satisfaction judgment. In this study, we saw a priori expectations to evolve in a number of ways.

For 72% of a priori expectations, participants reported a change in their perceived importance. 19% of participants’ expectations exhibited a decrease in their importance over time. Although these expectations were on average disconfirmed (i.e., median=3 on a 7-point scale), they did not lead to dissatisfaction (median=5). This was attributed to two major phenomena: transition from fantasy to reality, and post-purchase situational impact variations. First, participants reflected that these expectations were unrealistically high, i.e., “[they] hoped for, but not expected”. As a result, disconfirmation of these expectations was not attributed to the product as a failure but rather to their own perceptions as a ‘loss of illusion’. Second, as users were incorporating the product in their routines, the design space was adapting. For example, some participants became less concerned about the coverage of mobile internet through the cell network as they found themselves having access to internet mostly over WiFi networks, while others became less concerned about the ease with which 3rd party applications are being installed as they found themselves satisfied with the pre-installed ones.

53% of a priori expectations exhibited an increase in their importance over time. The majority of these expectations (87%) were either confirmed or exceeded. The major source of the increase in their perceived importance was participants’ initial inability to judge the impact of the expected feature in the long run. As participants incorporated the feature in their daily lives, they were becoming more dependent on it and its perceived importance was increasing. These expectations mostly related to the use of mobile internet, mobile agenda, and to the effectiveness and efficiency of the virtual keyboard.

4.4.4 Discussion

Overall, we showed time to be a significant factor altering the way individuals experience and evaluate products. We identified distinct phases in product adoption and use, which we summarize here.
CHAPTER 4. USER EXPERIENCE OVER TIME

From orientation to incorporation

The impact of novelty in users’ experience displayed a sharp decrease after the first week of use. Contrary to common belief that iPhone’s success is largely due to its aesthetics and novel interaction style, these aspects were found to play a minimal role in providing positive prolonged experiences.

Next, we found a shift in users’ concerns over time from ease-of-use to usefulness. While ease-of-use was the primary predictor of goodness during early orientation experiences, usefulness became an even more dominant predictor during the incorporation phase. This resembles recent research in the field of ubiquitous computing urging for a shift in the emphasis from efficient use to meaningful incorporation (Hallnäs and Redström 2002; Davidoff et al. 2006). Moreover, the types of interaction problems that users experienced shifted over time, in support of the findings of Mendoza and Novick (2005) and Barendregt et al. (2006). While early use was described by learnability problems induced by unexpected product behavior, prolonged usability related to repeating problems, often rooted in unanticipated use.

From incorporation to identification

Participants were found to develop an emotional attachment to the product as they increasingly incorporated it in their daily life. We found emotional attachment to be closely tied to the type of product. The iPhone is a very personal product as it connects users to loved persons, allows adaptation to personal preferences, and is always nearby. It is also a very social product as it communicates qualities of self-identity and connects to others by displaying shared values and interests. It is unknown how emotional attachment will develop with products that do not participate in users’ personal and social interactions.

Actual experience more influential than expectations

While earlier work (Lindgaard and Dudek 2003) would suggest that a priori expectations play a major role in the formation of satisfaction judgments, we found them to evolve during the actual experience with the product. Often, this was induced by lack of knowledge. As users’ experience with certain features exceeded their a priori expectations, these features became increasingly important to overall satisfaction with the product.

At the same time, disconfirmed expectations seemed to become less relevant to users’ satisfaction over time. A possible explanation for this could be supported by the theory of Cognitive Dissonance (Festinger 1957), which postulates that after a purchase there is a certain degree of psychological discomfort rooted in the discrep-
ancy between the desired and the actual outcome of the choice. The most likely outcome of dissonance is attitude spread, namely, participant’s efforts in justifying their choice by adapting their a priori expected outcome, or in our context, the perceived importance of their expectations.

All in all, the actual experience with the product seemed to be more influential to participants’ satisfaction judgments than their a priori expectations. Note, that we do not claim that forming expectations about a future possession does not influence experience; instead, we believe the act of anticipation to be a crucial part of our experience. Often, anticipating our experiences with a product, becomes even more important, emotional, and memorable than the experiences per se. It is only when conflicting with actual experience that a priori expectations appear to adapt in an effort of reducing the discrepancy between expected and actual experience.

Finally, what makes a product good and beautiful? Most studies suggest that goodness is a goal-oriented evaluation, related to pragmatic quality perceptions and beauty a pleasure-oriented evaluation related to hedonic quality perceptions (Hassenzahl, 2004; Mahlke, 2006; Tractinsky and Zmiri, 2006; Van Schaik and Ling, 2008).

The current study has diversified this view. While goodness was on average related to pragmatic quality perceptions, it was significantly affected by stimulation during orientation and by identification during social experiences. These findings suggest that the overall value, or the goodness of a product is contextually dependent, e.g. a novel product will be better than a more traditional one during our initial interactions but not necessarily in our prolonged experiences. Overall, we showed time to be a significant factor altering the way individuals experience and evaluate products.

4.4.5 Implications for design

What does this work suggest to HCI practice? HCI has up to now mostly focused on early interactions. As a consequence we have been mostly concerned about the product qualities that dominate in early use. We argue that the focus of HCI practice should expand from the study of early interactions to the study of prolonged experiences, understanding how a product becomes meaningful in a person’s life. We therefore promote three interesting avenues for further research.

Designing for meaningful mediation

What contributes to the successful appropriation of products? When does a product become useful in one’s life? We found usefulness to be much broader than the functionality of the product, relating to the impact of the functionality in participants’ lives. iPhone’s usefulness emerged through its appropriation in specific contexts and the changes this brought to participants’ lives. For instance, the reflection of one of
CHAPTER 4. USER EXPERIENCE OVER TIME

The participants on the Notes™ functionality was that it provided the freedom of going for jogging whenever she wanted to think of her work, as she could easily write down notes while being mobile (c.f. "capturing momentary information"). Usefulness, in this case, was derived from supporting her need for autonomy, being able to combine physical exercise and progress in her work.

On one hand, this provides hints that the product’s usefulness emerges in a process of appropriation in certain contexts of use, and thus may not become evident in early use and user tests involving minimal exposure to the product. On the other hand, one could speculate that this context of use was most likely not anticipated during the design of the iPhone. The question raised then is, how can we design for contexts that we cannot anticipate? We believe iPhone’s success here to be rooted in what Taylor and Swan (2005) call designing for artful appropriation, i.e. designs that are specific enough to address one single need, but flexible enough to enable the artful appropriation in diverse contexts.

Designing for daily rituals

People love parts of their daily lives and the products that are associated with them. Drinking a cup of coffee after waking up, listening to one’s favorite songs while driving home, drinking a glass of wine in the evening; these are some examples of activities that become habituated and cherished. We found activities mediated through the iPhone, like checking for new emails after waking up, or looking at a daughter’s photos several times during the day gradually becoming daily rituals that people love to perform. But, how can we design for new daily rituals? How can we identify the activities that people love in their daily lives if these are habituated and perhaps not apparent to the individual? It is crucial to follow the appropriation of products in participants’ lives, but also to understand the impact of the forfeiture of these products once these have been embedded in habituated activities.

Designing for the self

People become attached to products that support a self-identity they desire to communicate in certain settings. The iPhone supported two needs in participants’ social experiences: self-expression and differentiation from others (e.g. showing off to friends and colleagues), as well as a need for integration and feeling part of a group.

Products and self-identity have been a major part of consumer behavior research, but remain largely unexplored in HCI and design research. How can we understand the social meanings that users communicate through the possession of products? And how can we adapt our task-focused HCI methods to design for the more experi-
ential aspects of product use and ownership like the social meanings of products? One example could be the work of Zimmerman (2009) who proposes techniques for understanding and designing for the dynamics of self-identity where individuals have to reinvent themselves in a new role.

4.5 Discussion

This chapter has discussed two studies that tried to assess how users’ experiences with products develop over time. Besides having distinct goals, they employed diverse methodology for “measuring” the user experience: a reductionist and a holistic one.

The first study was reductionist in nature. It employed a validated measurement model (Hassenzahl, 2004) and sampled users perceptions across two points in time. It then tried to identify variations in the structural relationships between the latent constructs, i.e. the individual quality perceptions and the overall judgments of goodness and beauty. While longitudinal studies in user experience are scarce, similar methodological approaches can be found in the field of Technology Acceptance (Venkatesh and Davis, 2000; Venkatesh and Johnson, 2002; Kim and Malhotra, 2005). Such studies typically employ validated structural models across different phases in the adoption of a system. For instance, Venkatesh and Davis (2000) employed the Technology Acceptance Model (Davis et al., 1989) over three points in the adoption of information systems at work settings: before the introduction of the system (inquiring into users’ expectations), right after the introduction of the system, and three months after the introduction.

An assumption inherent in this approach is that the relevant latent constructs remain constant, but their perceived value and relative dominance might alter over time. But, especially in developing fields such as that of user experience, substantial variations might occur over time even in what constructs are relevant to measure. Some constructs, e.g. novelty, might cease to be relevant while others that were not evident in studies of initial use might become critical for the long-term acceptance of a product. Note for instance, the wider spectrum of experiences relating to daily rituals and personalization that could not be captured by the measurement model that we employed in the first study. This challenges the content validity of the measurement model as relevant latent constructs might be omitted, but may also lead to distorted data as the participant might fail in interpreting the personal relevance of a given scale item to her own context, for instance when a latent construct and its individual scale items cease to be relevant. In such cases participants may shallowly process the statement of the scale and ratings may reflect superficial language features of the scales rather than participant’s perceptions (Larsen et al., 2008b).
Moreover, such approaches provide rather limited insight into the exact reasons for changes in users’ experiences. They may, for instance, reveal a shift in the dominance of perceived ease-of-use and perceived usefulness on intention to use a product (e.g. Venkatesh and Davis [2000]), but provide limited insight into the exact experiences that contributed to such changes, the underlying motivations for changes in users’ preferences, and the contextual variations in product use.

Study 2 was more effective in providing rich insights into the exact reasons for the dynamics of users’ experiences over time. Beyond eliciting anecdotal reports on users’ experiences, through content analysis of the narratives, we were able to quantify the dominance of different product qualities on users’ overall evaluative judgments. It enabled capturing aspects of experience that were not identified a-priori, but also quantifying their significance and understanding how these relations changed over time.

In this view, experience narratives provide at least two kinds of information. Firstly, each narrative provides rich insights into a given experience and the context in which it takes place. Secondly, generalized knowledge may also be gained from these experience narratives. Such generalized knowledge may be reflected in questions like: how frequent is a certain kind of experience, what is the ratio of positive versus negative experiences and how does this compare to competitive products, how does the dominance of different product qualities fluctuate over time and what should we improve to motivate prolonged use?

This leads to two research problems. Firstly, how can we elicit experience narratives efficiently? Longitudinal designs such as the one employed in study 2 are labor intensive and, consequently, they are often restricted in terms of user population, product population, and studied time. In chapter 5 we review existing methodological approaches for studying temporal changes in user experience and present a novel survey technique that aims at assisting users in self-reporting their most impactful experiences with a product.

Secondly, how can we aggregate the idiosyncratic experiences into generalized knowledge? In chapter 5 we propose a novel technique for the semi-automatic analysis of experience narratives that combines traditional qualitative coding procedures (Strauss and Corbin [1998]) with computational approaches for assessing the semantic similarity between documents (Salton et al. [1975]).

4.6 Conclusion

This chapter presented two studies that aimed to assess the differences between initial and prolonged experiences with interactive products.

The first study assessed the ways in which 10 individuals formed overall evalu-
ations of a novel pointing device across two points in the adoption of the product: during the first week and after four weeks of use. Findings suggested judgments of the overall goodness of the product to shift from a use-based evaluation dominated by the pragmatic quality of the product, i.e. usefulness and ease-of-use, to an ownership-based evaluation dominated by aspects of identification, i.e. what the product expressed about their self-identify in social contexts. Judgments of beauty seemed to be dominated by perceptions of novelty during initial interactions, but this effect seemed to disappear after four weeks of use.

The second study followed six individuals through an actual purchase of the Apple iPhone and inquired into how their expectations and experiences developed one week before and four weeks after the purchase of the product. The study revealed that the product qualities that provided positive initial experiences were not very crucial for motivating prolonged use. Product adoption contained three distinct phases: an initial orientation to the product dominated by the qualities of stimulation and learnability, a subsequent incorporation of the product in daily routines where usefulness and long-term usability became more important, and finally, a phase of increased identification with the product as it participated in users’ personal and social experiences. We conceptualized temporality of experience as consisting of three main forces, an increasing familiarity, functional dependency and emotional attachment, all responsible for shifting users experiences across the three phases in the adoption of the product. Based on the findings, we promoted three directions for future HCI practice: designing for meaningful mediation, designing for daily rituals, and designing for the self.

Next to providing empirical insights into the dynamics of experience over time, these two studies raised a number of methodological issues in the study of time in the user experience field. we highlighted a number of limitations of traditional reductionist approaches where a-priori defined measurement models are employed in measuring the user experience. We suggests an alternative methodological approach that relies on a) eliciting the experiences that are personally meaningful to each participant, in the form of experience narratives, and b) employing content analysis techniques in creating multiple levels of abstraction, from concrete idiosyncratic insights to abstracted and generalized knowledge. We concluded by raising two research questions that will be addressed in the chapters and respectively.
iScale: a tool for the elicitation of experience narratives from memory

1 We present iScale, a survey tool for the retrospective elicitation of longitudinal user experience data. iScale employs sketching in imposing a process in the reconstruction of one’s experiences with the aim to minimize retrospection bias. Two versions of iScale, the Constructive and the Value-Account iScale, were motivated by two distinct theories on how people reconstruct emotional experiences from memory. The constructive iScale tool imposes a chronological order in the reconstruction of one’s experiences. It assumes that chronological reconstruction results in recalling more contextual details surrounding the experienced events and that the felt emotion is constructed on the basis of the recalled contextual details. The Value-Account iScale tool explicitly distinguishes the elicitation of the two kinds of information: value-charged (e.g. emotional) and contextual details. It assumes that value-charged information can be recalled without recalling concrete contextual details of an experienced event due to the existence of a specific memory structure that stores the frequency and intensity of one’s responses to stimuli.

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 sketching. Study 2 compared the two versions of iScale to free recall,

---

1 This chapter is partially based on
a control condition that does not influence the reconstruction process. Significant differences between iScale and free recall were found. Overall, iScale resulted in an increase in the amount, the richness, and the test-retest reliability of recalled information. These results provide support for the viability of retrospective techniques as a cost-effective alternative to longitudinal studies.

5.1 Introduction

The previous chapter established the importance of the study of prolonged use. From a methodological perspective, one could distinguish between three dominant approaches in understanding the development of users’ behavior and experience over time (von Wilamowitz Moellendorff et al., 2006).

Cross-sectional approaches are the most popular in the HCI domain (e.g. Prumper et al., 1992; Bednarik et al., 2005). Such studies distinguish user groups of different levels of expertise, e.g. novice and expert users. Differences between the user groups are then attributed to the manipulated variable, e.g. expertise. Such approaches are limited as one may fail to control for external variation and may falsely attribute variation across the different user groups to the manipulated variable. Prumper et al. (1992) already highlighted this problem, by showing that different definitions of novice and expert users lead to varying results.

Beyond the cross-sectional, one may further distinguish pre-post and longitudinal approaches in repeated sampling designs. Pre-post designs study the same participants at two points in time. For instance, Kjeldskov et al. (2008) studied the same 7 nurses, using a healthcare system, right after the system was introduced to the organization and 15 months later, while Karapanos et al. (2008a) studied how 10 individuals formed overall evaluative judgments of a novel pointing device, during the first week of use as well as after four weeks of using the product. While these approaches study the same participants over an extended period of time, they cannot inquire much into the exact form of change, due to the limited number of measurements. Longitudinal designs take more than two measurements. Because of their laborious nature, however, they are only rarely used in practice and research. von Wilamowitz Moellendorff et al. (2006) distinguished different "resolutions" in those studies: 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 and the idea to map the whole product lifecycle. Studies with a micro-perspective assess how users’ experience changes through increased exposure over the course of a single session. 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
minutes of interaction and c) after 15 minutes of interaction. An example of a study with a meso-perspective is Karapanos et al. (2009). They followed 6 individuals after the purchase of a single product over the course of 5 weeks. One week before the purchase of the product, participants started reporting their expectations. After product purchase, during each day, participants were asked to narrate the three most impactful experiences of the day using a retrospective diary method, the Day Reconstruction Method (Kahneman et al., 2004). Studies with a macro-perspective are "nearly non-existent" (von Wilamowitz-Moellendorff et al., 2006).

A third approach is the retrospective recall of personally meaningful experiences from memory. Variants of the critical incident technique, popular in the fields of marketing and service management research (c.f. 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), for example, 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 named CORPUS (Change Oriented analysis of the Relation between Product and User) for the retrospective assessment of the dynamics in users’ perceptions of product quality. CORPUS starts by asking participants to compare their current opinion on a given product quality (e.g. ease-of-use) to the one they had right after purchasing the product. If change has occurred, participants are asked to assess the direction and shape of change (e.g., accelerated improvement, steady deterioration). Finally, participants are asked to elaborate on the reasons that induced these changes in the form of short reports, the so-called "change incidents".

One may wonder about the degree to which these recalls are biased or incomplete. However, we argue that the veridicality of one’s remembered experience is of minimal importance, as these memories (1) will guide future behavior of the individual and (2) will be communicated to others. In other words, it may not matter how good a product is objectively, its quality must also be "experienced" subjectively to have impact (Hassenzahl et al., 2006). See also Norman (2009).

Although the validity of remembered experiences may not be crucial, their reliability is. It seems at least desirable that participants would report their experiences consistently over multiple trials. If recall is random in the sense that different experiences are perceived to be important at different recalls, then the importance of such elicited reports may be questioned. 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 of helping the participant in cueing more contextual information surrounding an experienced critical incident (c.f. Edvardsson and Roos, 2001). Interviews may however elicit only a limited number of reports. Self-reporting approaches, for instance through online surveys, have far more impact because one can survey large samples and, thus, also inquire into rare experiences. Such approaches, however, are less controlled than face-to-face interviews. Thus, the question at hand is: How can a survey procedure support a participant in recalling her experiences with a product in a reliable way?

This chapter presents iScale, a survey tool that was designed with the aim of increasing participants’ effectiveness and reliability in recalling their experiences with a product. iScale uses sketching to impose specific guiding procedures, assumed to improve the participant’s ability to recall experiences from memory. In the following, we will describe the theoretical motivations for the development of iScale and present the results of two studies. Study 1 aimed at acquiring a qualitative understanding of the use of iScale in comparison to its analog equivalent, i.e. a free-hand sketching, and aimed at informing its redesign. Study 2 aimed at assessing how iScale compares to an experience reporting tool that provides no sketching, and, thus, can be seen as a control condition to assess the impact of iScale on participants’ effectiveness and reliability in recalling.

5.2 Sketching and 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 in Barlett’s (1932) seminal work. He suggested that remembering is an act of reconstruction that can never produce the exact past event, but instead, every attempt to recall results in a new, often altered representation of the event. Bartlett (1932) asked participants to recall an unfamiliar story that they were told 20 hours before. Recalled stories differed from the original one in missing details, altering the order and importance of events, or in applying rationalizations and interpretations to the original story. Stories were further distorted through repeated reconstruction.

The notion that remembering is an act of reconstruction instead of mere reproduction has received wide support. At the heart of reconstruction lies the distinction between episodic and semantic memory (c.f. 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). These two types of memory serve different needs such as learning new information quickly - a capacity of episodic memory -
5.2. SKETCHING AND 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 from episodic memory. In the absence of contextual cues in episodic memory, beliefs found in semantic memory may be used to reconstruct the past, resulting in distortions such as the ones found in Barlett’s study. Thus, overall, the accuracy of one’s remembered events lies in the degree to which contextual cues are still present in the person’s episodic memory.

But, how do we reconstruct emotional experiences that contain not only contextual details of the experienced event, but also value-charged information such as emotions or overall evaluative judgments on the event? One can distinguish between two distinct approaches to the reconstruction of value-charged experiences. The first one, the Constructive approach, assumes that felt emotion cannot be stored in memory but is instead reconstructed from recalled contextual cues. The second approach, the Value-Account approach, proposes the existence of a memory structure that is able to store the frequency and intensity of one’s responses to a stimulus. This information may in turn be used to cue the recall of contextual details of one’s experiences. In the next sections we describe the two approaches in more detail.

5.2.1 Two approaches to experience reconstruction

The Constructive Approach

The constructive approach assumes that reconstruction happens in a forward temporal order (e.g. Anderson and Conway, 1993; Barsalou, 1988; Means et al., 1989). Barsalou (1988) asked people to recall their experiences during the summer. Most participants started in the beginning of the summer and proceeded in a chronological order. Often, the recall of an event cues the reconstruction of more events and contextual information surrounding the event (Anderson and Conway, 1993) - like a string of pearls.

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 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, i.e. recall contextual cues from episodic memory in reconstructing the emotional experience. When episodic memories become inaccessible, people will shift to semantic memory. People will first access situation-specific beliefs, i.e. “a belief about the emotions that are likely to be elicited in a particular type of situation”. If event-specific beliefs are inaccessible, e.g. due to rarity of the event, people will access identity-related beliefs,
Motivated by the accessibility model of Robinson and Clore (2002), Daniel Kahneman and colleagues (2004, 2008) developed the Day Reconstruction Method (DRM), an offline 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 but also the temporal relations between the episodes. As a result, participants reconstruct the emotional experience on the basis of sufficient episodic information, thus avoiding retrospective biases that most offline methods suffer from as participants draw on semantic information to reconstruct the emotional experience. Kahneman et al. (2004) demonstrated that DRM may achieve an accuracy close enough to that of online reporting as in the case of the Experience Sampling Method (Hektner et al., 2007).

The Value-Account Approach

The Value-Account approach assumes that reconstruction happens in a top-down fashion. It assumes that people may recall an overall emotional assessment of an experience without recalling the exact details of the experienced event. Betsch et al. (2001) proposed the existence of a new memory structure called Value-Account, that is able to store the frequency and intensity of positive or negative responses to stimuli. Since Value-Account is assumed to be more easily accessible than concrete details from episodic memory, it may cue episodic information in reconstructing the experienced event, or inform the construction of an overall evaluation even in the absence of episodic information (c.f. Koriat et al., 2000; Neisser, 1981).

While some studies have shown that value-account may be falsely recalled even in the presence of accurate episodic information, it is generally accepted that value-account information is better retained over time than episodic information (c.f. Koriat et al., 2000) and may be used in cuing episodic information. In a related field of memory research, that of autobiographical memories, researchers distinguish between three levels of specificity in memory: lifetime periods, general events, and event-specific knowledge. Reconstruction has been found to take place in a top-down fashion where knowledge stored at the level of a lifetime period may cue information at the two lower levels (Conway and Pleydell-Pearce, 2000).

Both approaches, constructive and value-account, suggest specific processes of retrieving emotional experiences from memory. While the constructive approach suggests a chronological order in recalling episodic information that subsequently cues the reconstruction of the experienced emotion, the value-account approach suggests a
5.2. SKETCHING AND MEMORY

*top-down progression* where the affective information stored in value account is used to cue the recall of episodic information. In the following section, we will illustrate how these two processes were operationalized in two distinct versions of the iScale tool.

5.2.2 Can sketching affect the reconstruction process?

Imagine being asked to "sketch" how the perception of the usability of your mobile phone changed over time; you are given a timeline that starts at the moment of purchase and ends in the present. How would you start? One may go back to the past, right after the purchase of the product, and try to recall her first experience with the product. What was it about? Was it positive or negative? What else happened after that? Reconstruction is assumed to take place in a chronological order and the sketch, the overall evaluation of one’s experiences, is constructed from the recalled details of the experiences. Another person may start the sketching exercise by thinking of the overall experience, the change over time. Did my opinion about the product’s ease-of-use increase overall? If so, was it mainly in the beginning or at the end? This might then cue the recall of experienced events that caused these changes (see von Wilamowitz-Moellendorff et al. 2006, 2007).

Sketching, in the above scenario, provides 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". This imagery is an incomplete, reconstructed representation of the experienced past. It consists of two sources of information: a) contextual details of experienced events such as the temporal, factual, and social context of an experience, and b) value-charged information such as emotions and evaluations of the experienced event. Product evaluations are here seen as carriers of affective information, that is, affect that is attributed to the product (Hassenzahl and Ullrich, 2007).

The veridicality of this reconstructed representation, i.e. the convergence between the representation and the past, is likely to be influenced by the process that the participant follows in reconstructing it from memory. Sketching may, thus, impose a certain process on the reconstruction of the past and by that crucially influence the way experiences are remembered. In the remainder of this section we describe iScale, a survey tool that elicits experiences with a product through sketching how one’s opinion changed over time. We will introduce two different versions of iScale, each trying to lead to a different experience reconstruction process. For reasons of simplicity we will call them the Constructive and the Value-Account iScale.

Interacting with iScale is done in three steps. First, the participant is asked to respond to two questions (figure 5.2b): a) "What was your opinion about the product’s
[certain quality] just before you purchased it", and b) "How did your opinion about the product’s [certain quality] change since then". While one could use participants’ ratings to elicit a between-subjects estimation of how the perceived quality of a certain product develops over time, with the participants’ time of ownership of the product as the independent variable, this is not the primary information that we are interested in. Instead, we assume that these questions can help the participant in positioning herself in the past and in recalling contextual details before the start of the sketching activity.

Second, the participant is presented with a timeline that starts at the moment of purchase and ends at the present time. The participant is asked to sketch how her opinion about a certain quality of the product has developed over time. The two distinct modes of sketching will be discussed below. Overall, the participant may sketch linear segments that represent an increase or decrease in her opinion over a certain period. Each period can be annotated along the time by specifying the time that has passed from moment of purchase.

Third, each line segment is associated with a line identifier that is displayed below the segment (figure 5.1b). A participant may click on the segment and an interface is presented for reporting one or more experienced events that are perceived to have caused the sketched change in the participant’s overall opinion (figure 5.2a). For each experience report, the participant may provide a brief name (identifier), a more elaborate description of the experienced event - experience narrative, and respond to a number of event-specific questions. For the goals of the specific study we present in this chapter, 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 opinion, and c) the participant’s confidence on the exact details of the narrative.

The Constructive and the Value-Account iScale

The two versions of iScale differ only in the second component, that of sketching. These aim at imposing distinct modes of reconstruction of experiences from memory. More specifically: the existence or absence of concurrency between sketching and reporting, and feed-forward or top-down progression of sketching.

**Feed-forward - Top-down progression of sketching:** The constructive approach to reconstruction suggests that recalling experiences in a chronological order will cue more contextual details surrounding the experience and that this will in turn lead to a better reconstruction of the experienced emotion as well as a recall of further temporally aligned experiences. On the other hand, the value-account approach assumes that participants follow a top-down approach where a participant may first form an overall evaluation of the change
Figure 5.1: (a) Constructive iScale, (b) Value-Account iScale
CHAPTER 5. ISCALE: A TOOL FOR THE ELICITATION OF EXPERIENCE NARRATIVES FROM MEMORY

Figure 5.2: (a) interface for reporting experiences, and (b) overall questions asked in the beginning of the survey
over the full time of ownership, and proceed by forming subsequent judgments on the sub-parts of each change. Thus, 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 regarding how the overall opinion has change from the moment of purchase to the present. The participant may then proceed by splitting the full segment into parts.

**Concurrent - Non-concurrent reporting:** The constructive approach to reconstruction assumes that the affective component of a past experience, in other words the valuecharged information, can only be reconstructed from recalled contextual details of the event. On the contrary, the value-account approach assumes that individuals may recall an overall emotional assessment of an experience even without being able to recall the underlying contextual details. Thus, according to the constructive approach reporting should be concurrent with sketching as reporting would 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 this value-charged information. Thus, in the constructive iScale the participant is asked to report on experiences right after a line segment is sketched. Sketching and reporting is thus proceeding concurrently in a step-by-step process. In the Value-Account iScale, this process is split into two distinct steps: the participant is urged to first sketch the desired pattern before proceeding to the next step where she may report one or more experiences for each sketched line segment. Both methods however retain flexibility, as the sketched pattern can be modified even after experiences have been reported for the existing pattern.

Overall, sketching is expected to provide temporal context for the recall of experienced events. This is expected to increase the amount and test-retest reliability of the information that the participants are able to recall. This assumption will be tested in study 2.

### 5.3 Study 1

The first study attempts a qualitative understanding of sketching as a process for supporting the reconstruction of one’s experiences. First, it questions some of the core assumptions that underlie the design of iScale through the observation of users’ behavior when employing free-hand-sketching (FHS), the paper version of Constructive iScale. Secondly, it compares the two iScale tools to the FHS approach and identifies the design qualities that a sketching tool should have in order to support the
Table 5.1: The three product qualities that participants reported on, along with definitions and word items.

<table>
<thead>
<tr>
<th>Name</th>
<th>Definition</th>
<th>Word items</th>
</tr>
</thead>
<tbody>
<tr>
<td>Usefulness</td>
<td>The ability of a product to provide the necessary functions for given tasks.</td>
<td>Useful, Practical, Meaningful</td>
</tr>
<tr>
<td>Ease-of-use</td>
<td>The ability of a product to provide the functions in an easy and efficient way.</td>
<td>Easy to use, Simple, Clear</td>
</tr>
<tr>
<td>Innovativeness</td>
<td>The ability of a product to excite the user through its novelty.</td>
<td>Innovative, Exciting, Creative</td>
</tr>
</tbody>
</table>

5.3.1 Method

Participants

A convenience sample of 12 graduate students in HCI (7 male, median age 30 years) participated in the study. They were chosen due to the diversity in their educational backgrounds. They were: five Computer Scientists, three Industrial Engineers, two Linguists, one Psychologist and one Industrial Designer.

Procedure

The study consisted of two main parts. In the first part, each participant used the three different sketching techniques, i.e. free-hand sketching and the two iScale tools. All tasks were carried out on a Wacom Cintiq 21UX Interactive Pen Display. The order in which the tools and qualities were employed was counterbalanced across participants; FHS was always used first to avoid any bias from the iScale tools as we wished to understand users’ natural behavior in free-hand sketching.

Participants were asked to sketch how their opinion on three distinct product qualities of their mobile phone developed over time (see table 5.1). Each quality was described by a brief definition and three words to support the definition (Hassenzahl [2004] Von Wilamowitz-Moellendorff et al. [2006]). Participants were instructed to think aloud; interactions and verbal data were captured on video.

"While sketching, you are asked to report experiences and events that induced these changes in your opinion about the product. We are interested in knowing your exact thoughts as you perform the sketching activity. What makes you sketch something? Do you remember something? Is it just a feeling? We ask you to **think aloud** while doing this task."

In the second part, participants were interviewed about the differences between the three sketching techniques, using a structured interview technique, called the
5.3. STUDY 1

Repertory Grid [Fransella et al., 2003]. Participants were given three cards, each providing a name and a screenshot of one of the three sketching 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 sketching 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 in providing a bipolar construct (see Karapanos and Martens, 2008) while we instructed them to elaborate when possible.

Participants were further probed using the laddering and pyramiding techniques (Reynolds and Gutman, 1988). Laddering seeks to understand what motivates a given statement and thus ladders up in an assumed means-ends-chain (Gutman, 1982) towards more abstract qualities of the stimuli; in laddering we first asked the participant whether the mentioned quality is positive, and subsequently why this quality is important to him/her, e.g. "why is expressiveness important to you?". Pyramiding, on the other hand, also known as negative laddering, seeks to understand the lower level attributes that make up for a given quality; in pyramiding we asked the participant to elaborate on what makes the given technique to be characterized with the respective attribute, e.g. "what makes free-hand-sketching more expressive?".

5.3.2 Analysis and results

Understanding free-hand sketching

Throughout the development of iScale, FHS acted as our reference for testing assumptions and gaining confidence for the respective design decisions. Often, compromises had to be made. For instance, enabling users to create non-linear curves would increase the expressiveness in sketching but would on the other hand also increase the complexity of the task by either increasing the number of actions needed in enabling users to define all parameters of a curve, or minimizing users’ control over the result by imposing restrictions to the curve in an effort to minimize the number of actions required. Thus, it is reasonable to explore the value of non-linearity in such sketches. In other terms, do users sketch non-linear curves in free-hand-sketching and if so, what does this non-linearity express?

Next, iScale assumes that sketching supports and is supported by the reconstruction of discrete experiences. It is thus assumed that users will associate changes in their opinion to one or more discrete experiences. This notion can however be questioned. First, the degree to which a sketched change is associated to one or more discrete experiences will depend on the mode of reconstruction, being it either constructive or value-account. In the constructive mode, users recall contextual cues about the discrete experience and reconstruct the overall value-judgment based on the recalled
facts. In the value-account mode users may recall this overall evaluative judgment first and may or may not further reason to associate the recalled evaluative judgment to underlying reasons for this, e.g. one or more experiences. Thus, to what extent do users succeed in recalling discrete experiences? Second, assuming that a user recalls a discrete experience: Is it associated to a continuous change or a discontinuous one? In other terms, if a user thinks in terms of discontinuities, i.e. in terms of discrete events instead of overall opinion, does he/she relate these discontinuous recalls to a continuous graph?

These questions were explored by observing users’ reported experiences and sketched patterns while employing FHS. Users’ free-hand sketches were segmented in discrete units based on cues available in participants’ verbalized thoughts as well as users’ pauses in their sketching behavior as observed in the video recorded sessions. A new unit was coded when both conditions were observed: a semantic change in the participant’s report following a pause in sketching. 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 under four categories: a) Constant (C) signifying no change in participant’s opinion over a certain period, b) Linear (N), either Increasing or Decreasing, c) Non-linear (NL) when there were no grounds that the curve could be approximated by a linear one or when a single report was associated with two discrete linear curves of different curvature (see 5.3b), and d) Discontinuous (D) when the slope was significantly higher than on average.

What kinds of curves do users sketch?

Table 5.2 illustrates the distribution of users’ sketches across the four types of curves. The majority of curves (44 of 74, 60%) were categorized as linear, signifying a change that can reasonably be approximated by a linear curve. Only 5% (4 of 74) curves were non-linear. For these curves, a single report was associated with two or more linear curves with different slopes (cf. figure 5.3a segment 2, figure 5.3b segment 6, figure 5.3d, segment 1). Thus while in certain cases users draw non-linear curves, the majority of curves will be linear ones, and therefore the overall value of non-linearity in a sketching tool appears to be limited.

In a similar vein, only 4 of 74 (5%) instances of discontinuity were observed in users’ sketches. One might expect that the recall of discrete events increases users’ tendency to sketch discontinuous curves, as the mode of recalling, being it continuous or discontinuous, should relate to the mode of sketching. This expectation is not supported by the data.
Figure 5.3: Examples of free-hand sketching. 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 sketch (C: Constant, L: Linear, NL: Non-Linear, D: Discontinuous).
Table 5.2: Relationship between sketching and reporting in free-hand sketching. Types of sketching: C=Constant, L=Linear, NL=Non-Linear and D=Discontinuous

<table>
<thead>
<tr>
<th>Type of report</th>
<th>Type of sketch</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reporting a discrete experience</td>
<td>C  L  NL  D</td>
</tr>
<tr>
<td>Reporting an overall evaluation with no further reasoning</td>
<td>3  30 2  2  37 (50%)</td>
</tr>
<tr>
<td>Reporting an overall evaluation, reasoning through experience</td>
<td>17  4  1  2  25 (34%)</td>
</tr>
<tr>
<td>Overall</td>
<td>22 (30%) 44 (60%) 4 (5%) 4 (5%) 74</td>
</tr>
</tbody>
</table>

How do curves relate to experiences?

Participants’ reports were classified into three broad categories. First, reports rooted in the recall of a discrete experience. Such reports represented the constructive mode of reconstruction: recalling contextual information from a specific experience followed by the reconstruction of the value judgment from the recalled facts. Such 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%).

Recall of a discrete experience: "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]"

On the contrary, other reports provide no contextual information about a recalled experience, but instead, the participant reports an overall evaluation without further reasoning. Such reports represented the value-account mode of reconstruction: recalling an overall evaluation of a specific experience or period, while failing to recall contextual cues or facts about an exact experience.

Recall an overall evaluation without further reasoning: "after that, [my opinion] is about constant, it hasn’t changed lately"

Table 5.2 illustrates a strong association of discrete experiences with linear curves (10/13), while reporting merely an overall evaluative judgment takes place mostly when participants sketch a constant line (17/24), signifying no change in their perception. Reporting only evaluative judgments may thus be a side effect of sketching, rooted in the fact that participants are asked to sketch a continuous timeline.
5.3. STUDY 1

Last, we found a third type of reporting that combines the two core types. Those reports were grounded in the recall of an overall evaluation, but participants proceeded to reason about this value-judgment through reporting discrete experiences.

Recall an overall evaluation followed by reasoning about an experience: "[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"

How does iScale compare to free-hand sketching?

The two iScale tools were also compared to free-hand sketching. Participants’ verbal reports were transcribed and analyzed using Conventional Qualitative Content Analysis (Hsieh and Shannon, 2005). We started with open coding (Strauss and Corbin, 1998) where we aimed at identifying an overpopulated list of design qualities that appear to influence the design space of the three sketching techniques. Subsequently we grouped the initial codes into overall categories through an iterative process. Each statement was coded for the type of quality that the participant refers to as well as to whether or not this quality affects the sketching or the recalling process. Statements were always differentiating two of the approaches from a third as this was imposed by the structure of the interview technique.

Table 5.3 illustrates the dominant qualities that were elicited in the interview sessions. For each quality, it displays the number of participants mentioning it as present for a given sketching technique, and the number of participants mentioning the given quality as affecting the sketching or recalling process. The design qualities can be distinguished in three broad groups: expressiveness, control, and concurrency. Meaningful differences emerged across the three different techniques; these should only be considered as tentative qualitative insights, however, due to the small number of observations.

Expressiveness

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

"[p6] it is easier to convey my thought into paper, the other ones I could only draw a straight line and that constrained me", "I think this allows you more to relate to your opinion, because you can really sketch what you want to... you can give slight ditches on the curve"
## Table 5.3: Design qualities elicited in the interview sessions. Number of participants using a given quality to differentiate the three sketching techniques, and number of participants mentioning the given quality as affecting the sketching or recalling process.

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
<th>FHS</th>
<th>CON</th>
<th>VA</th>
<th>Sketch Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tool Impact on</td>
<td>Concurrency of time in sketching &amp; recalling</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>Chronological linearity in recalling</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>Temporal Overview</td>
<td>4</td>
<td>3</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Mobility</td>
<td>2</td>
<td>6</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Interactivity</td>
<td>4</td>
<td>8</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Interplay sketching-recalling</td>
<td>3</td>
<td>6</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Expressiveness</td>
<td>1</td>
<td>9</td>
<td>7</td>
<td></td>
</tr>
</tbody>
</table>

Techniques, and number of participants mentioning the given quality as affecting the sketching or recalling process.
The majority of participants emphasized the effect this has on the sketching activity. While all participants expressed the perceived benefit of FHS as it enables them communicating their thoughts, only one participant mentioned that this also affects the recalling process as the sketch provides richer feedback and therefore more cues for recalling. One participant stated freedom in sketching as a positive quality for expressing impressions for which she fails to recall any supportive facts.

"[p1] free-hand sketching provides the freedom to sketch something without having to justify the exact reason for this... I might not recall something specific; still I can communicate my impression through the curve... with the electronic ones you select points, therefore you feel obliged to justify them"

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

"[p12] I can annotate things... it helps as I might forget the reasoning in the beginning of the graph and when I am at the end, ...like in ease of use, I looked back and it helped me remember what my reference point meant, I didn’t find it very intuitive at the zero point, and I was able to put my scale on diagram", "[p6] free-hand sketching provides the opportunity to add notes, you can sketch and then split the x-axis in terms of periods.", "[p3] in free-hand sketching you can make notes in the meanwhile, bring other thoughts you have..."

Control

Most participants stated that the iScale tools provide more overall control, related to three core aspects of the interaction. Firstly, eight out of the twelve participants found the constrained interaction a positive aspect of iScale, providing better interaction, as it consumes less resources, thus providing them better control of the output (4 participants) and enabling them to focus on recalling their experiences (4 participants):

"[p8] the digital ones in some way restrict your freedom... I think it’s a good thing because you think before you put the dot, so elaborate more before you sketch, it is a more thoughtful account", "[p5] It’s difficult to just draw on a free line whatever you think... it feels like a drawing, it doesn’t feel like making a realistic graph", "[p10] in free-hand sketching I didn’t have the control of the output as much as with this one", "[p12] it constrains me to draw line segments, at the same time focusing my efforts on just creating the events... I wouldn’t have to worry about how neat the line looked..."
"the electronic ones are more playful, you add points, move them, it has a certain dynamic, with free-hand sketching you cannot play and see the result... with the other ones, you may with quick actions alter the output, they are more interactive. I think I am more fond of them as I can communicate what I have in mind"

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

"[p7] it’s easier to adapt these lines in that you just move part of the line, in free-hand sketching you have to clean some parts of the line", 
"[p6] in the value-account tool, you had to decide on the overall slope right from the start... This constrained me in..., which is not actually true because I could select the point in between and move it, ... but in any case it was harder."

Third, seven out of the twelve participants acknowledged that the value-account tool provides a better overview of the full timeline; approximately half of the participants were concerned about the accuracy of the graph, while the rest thought that this temporal overview enhances their recall process.

"[p8] these ones [free-hand, constructive] surprise you at the end, you can end up somewhere you didn’t expect, the value-account gives you a better overview of totality.", 
"[p9] it provides an explicit timeline, you are thinking ‘I have to do some clustering...’ it is easier for recalling, it helps in rethinking"

Concurrency

At the same time some participants perceived the temporal overview as a negative quality of the value-account tool. As one participant mentions:

"[p4] while I don’t know where the story ends, I can add whatever event that fits to the story, it is more open ended... with the end point, I had to structure my story more towards a coherent story"

Five participants in total mentioned temporal linearity as a quality that differentiated free-hand sketching 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:

"[p6] it allows you to visualize things step by step and also historically... you ascertain that the decision that you make in each step is correct... you just make one step, you don’t have to think about the whole curve", 
"[p5] in free-hand sketching you are
really forced to think about the experience from the moment of purchase till now”, 
“[p5] the more you do the more you remember... I think with the value-account in 
that sense you can relate to less experiences, you first draw an overall picture and you 
are less inclined to relate to experiences..., in both other cases, you are more focused on 
the spot...”, 
“[p4] value-account felt more limited, I had to think where I was going to 
go, I would disregard things that didn’t go where I wanted to go”
	Similarly, five participants highlighted that the concurrency of sketching with re-
porting, that was lacking from the value account tool, enhances the output of both 
the recall and the sketching process:
	“[p7] you go step by step with what you want to report and the visualization. ...It 
aligns the action that you are making with the thinking process that you want to 
make”, 
“[p7] [in value-account iScale], you have to really judge... I was trying to 
compare the purchase point to my overall impression now and this changes, there are 
different periods, I am trying to average over different time periods and experiences, 
which is not actually the purpose... the decisions I made at first were wrong decisions 
because I have to change it, the other tools are more accurate”

5.3.3 Conclusion

Overall, the study provided support for the iScale tool. The need for sketching non-
linear and discontinuous curves was limited. In addition, most non-linear curves 
could be approximated by two linear segments. The need for annotation was high-
lighted 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 sketched 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 figure 5.1h). 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 sketched pattern and are therefore more likely 
to modify it. The interviews confirmed that free-hand sketching is more expressive 
than using iScale due to the increased degrees of freedom in sketching as well as 
due to its ability to easily annotate sketches. Participants also reported qualities that 
were not present in the free-hand sketching, such as the two-step interactivity and 
modifiability of the electronic sketches that resulted in a better interoperability be-
tween the sketching and the recalling activity. Last but not the least, participants 
also reported benefits for both the constructive and the Value-Account over FHS. 
The Value-Account provided a temporal overview which influenced both the sketch-
ing and the recalling process. The constructive approach provided benefits such as
chronological order and concurrency between sketching and reporting which had a positive impact on the recall process.

5.4 Study 2

The second study aimed at comparing the two different versions of iScale, the constructive and the Value-Account version, to a control condition that entails reporting one’s experiences with a product without performing any sketching. It aims at testing whether sketching impacts the number, the richness, and the test-retest reliability of the elicited experience reports.

5.4.1 Method

Participants

48 participants (17 Female), ranging from 18 to 28 years (median=23 years), took part in the experiment. They were all students at a technical university; 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 eighteen months; 16 participants owned a smart phone.

Materials

Three different versions of iScale were used in the experiment: Constructive, Value-Account, and Experience Report. The Constructive and Value-Account versions employ two discrete sketching approaches aiming at inducing two distinct modes of reconstruction from memory as described earlier. Experience Report constitutes a stripped down version of iScale where the sketching interface is removed; instead the user may only provide report experiences. It uses the same interface as the two sketching tools in reporting experiences.

Experimental design

The experiment used a balanced incomplete design with three factors: presence or absence of sketching (the two sketching tools versus Experience Report), mode of reconstruction (i.e. Constructive or Value-Account), and product quality being reported (i.e. ease-of-use versus innovativeness). Mode of Reconstruction was between-subjects with the other two variables being within-subject.
Table 5.4: Experimental design: manipulated variables (rows) and process (columns). Each participant joined two sessions, each consisting of two tasks. Participants in group A, for instance, which belong in the Constructive condition, in the first session, they used the sketching tool to report on ease-of-use followed by the non-sketching tool to report on innovativeness. In the second session the order was counterbalanced but the two coupling of the two variables, sketching and quality, remained the same.

<table>
<thead>
<tr>
<th></th>
<th>Session 1</th>
<th>Session 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Task 1</td>
<td>Task 2</td>
</tr>
<tr>
<td><strong>Constructive</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sketching</td>
<td>Ease-of-use</td>
<td>A</td>
</tr>
<tr>
<td></td>
<td>Innovativeness</td>
<td>B</td>
</tr>
<tr>
<td>No-sketching</td>
<td>Ease-of-use</td>
<td>C</td>
</tr>
<tr>
<td></td>
<td>Innovativeness</td>
<td>D</td>
</tr>
<tr>
<td><strong>Value-Account</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sketching</td>
<td>Ease-of-use</td>
<td>E</td>
</tr>
<tr>
<td></td>
<td>Innovativeness</td>
<td>F</td>
</tr>
<tr>
<td>No-sketching</td>
<td>Ease-of-use</td>
<td>G</td>
</tr>
<tr>
<td></td>
<td>Innovativeness</td>
<td>H</td>
</tr>
</tbody>
</table>

Procedure

Participants joined two sessions, each one lasting approximately 40 minutes. The second session took place roughly one week after the first one (minimum: 7 days, maximum: 10 days).

In each session, participants used two different tools for reporting on two different qualities of their mobile phones: a tool that allowed both sketching and reporting and a tool that allowed only reporting. We did not follow a complete design as that would mean that participants would have to report on each quality twice. This however restricts the power of the experiment as only between-subjects analyses may be performed on the data.

Participants were split into two groups; the first group used the Constructive iScale while the second used the Value-Account iScale. The product qualities used in this study were ease-of-use and innovativeness. Both qualities were introduced to participants through the definitions used in study 1. The order of tool and product quality was counterbalanced within each group using a latin square design, and reversed across the two sessions so that each participant would use, in the second session, the same tool for reporting on the same quality but in the reverse order (see table 5.4).

Assumptions

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.

First, based on existing evidence that the reconstruction of events in a serial chronological order cues the recall of both a) more temporally surrounding experi-
riences and b) more contextual cues about the experience being reported \cite{Anderson and Conway, 1993}, it is expected that within the constructive condition, sketching will result in an increase of a) the number of experiences being reported and b) the richness of reported experiences. For the Value-Account iScale which obstructs participants in reconstructing their experiences in a chronological order, the difference to experience reporting is expected to be smaller.

Second, as long as participants are expected to recall more contextual cues in the constructive iScale, this is expected to influence their test-retest reliability in recalling factual details of the past experiences, such as temporal information (e.g. when did the experience take place). In a similar vein, this is expected to happen only in the constructive condition, but not in the value-account condition.

Third, the consistency of the participants’ sketches, i.e. value-charged information, is 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 an 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.

5.4.2 Analysis and results

A total of 457 experience narratives were elicited. Participants provided an average of 3.7 to 6 experience reports depending on the recall condition. 95% of all experiences related to the first six months of use. We compare the two sketching tools to the no-sketching condition in terms of a) number of elicited experiences, b) qualitative richness of elicited experiences, c) reliability of recalled event-specific information, and d) reliability of overall value-charged information.

**Does sketching impact recalling?**

**Does sketching result in an increase in the number of reported experiences?**

Figure 5.4a shows the number of reported experiences as a function of the mode of elicitation. In the constructive mode of recalling, an average of 6 experience reports was elicited when using iScale with an average of 4.4 when only reporting was used. In the value-account condition, the number was 4.6 when reporting was combined with sketching and 3.7 when participants only reported, without sketching, on their experiences.
5.4. STUDY 2

Figure 5.4: Average number of experience reports elicited (a) in the presence or absence of sketching in the constructive and in the Value Account modes of reconstruction, and (b) for the two respective product qualities in relation to mode of reconstruction.

An analysis of variance with number of experience reports as dependent variable and mode of reconstruction (constructive, value-account), sketching (no-sketching, sketching) and quality (ease-of-use, stimulation) as independent variables displayed significant effects for mode of reconstruction, $F(1,88) = 7.16, p<.05$, and the presence of sketching, $F(1,88) = 10.55, p<.05$. As expected, sketching resulted in a higher number of elicited experience reports compared to non-sketching (sketching: $\mu = 5.3$, no-sketching: $\mu = 4.0$) as it provided a temporal context during recall, which cued the recall of further experiences. This did not only happen for the constructive sketching tool, which imposes a chronological order of events as existing retrospective diary methods (Kahneman et al., 2004), but also for the Value-Account sketching tool. However, the constructive sketching tool resulted overall in more experience reports than the Value-Account sketching tool (constructive: $\mu = 5.2$, value-account: $\mu = 4.2$).

Further, while there was no overall effect of the product quality being reported, we found a marginal interaction effect between mode of reconstruction and quality, $F(1, 88) = 2.92, p=.09$. In the constructive condition participants elicited a significantly higher number of experience reports when reporting on ease-of-use ($\mu=5.9$) than when reporting on innovativeness ($\mu=4.5$), $t(46)=-2.14, p<.05$. This was not observed in the Value Account condition. One possible interpretation might tap into the different nature of events that relate to ease-of-use and innovativeness. von Wilamowitz-Moellendorff et al. (2007) 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 action, whereas stimulation cannot be allocated to specific events. Thus, the effect of chronological order in reconstruction may be more salient in case of contextually rich experiences than in case of more abstract ones. All other main effects and interactions remained insignificant.

**Does sketching result in richer reported experiences?**

Based on existing knowledge ([Barsalou, 1988](#)), one would expect that reconstructing in a chronological order would lead to more contextual cues in the elicited experience reports, thus provide 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) etc. To identify these different factors of richness, we submitted the experience reports to a qualitative content analysis. Open coding ([Strauss and Corbin, 1998](#)) was performed by the first author and resulted in three main types of information present in experience reports: **temporal information** summarizes references to the exact time at which the experience took place, **discrete event information** summarizes references to a concrete occurrence that influenced the experience, and lastly, **expectations** summarize references to participants’ expectations about the reported experience.

Sketching was found to have a significant effect on the number of contextual cues referring to a discrete event in the constructive mode of reconstruction ($\chi^2=4.07$, $p<.05$) where 45 out of 146 reports contained at least one cue referring to a discrete event in the sketching condition as opposed to 20 out 103 in the no-sketching condition (see table 5.5 line b), but not in the value-account condition ($\chi^2=0.21$, $p=.650$) where 27 out of 118 reports referred to discrete event in the sketching condition as opposed to 18 out of 89 in the no-sketching condition. No significant effect was found on the number of temporal cues present in experience reports both for the constructive (Pearson’s $\chi^2=2.13$, $p=.14$), and for the value-account mode of reconstruction ($\chi^2=2.48$, $p=.12$). Last, no significant differences were found on the number of experience reports containing cues about participants’ expectations before the experienced event.

In a similar vein, one might expect that recalling more contextual cues about the experienced event will result in an increase in design-relevant information. We distinguished four groups in the design relevance of an experience report. First, an experience report might provide **no design relevant information**, e.g. the participant only reflects on her attitude towards the product without any further reasoning. Second, the participant may **pinpoint a specific feature** of the product that an evaluative judgment is made for without any reasoning motivating her attitude. Third, the report might also **contain reasoning about the reported attitude** such as the nature of an
Table 5.5: Number of experience reports judged for the three dimensions of richness and for the four levels of design relevance, for the four conditions resulting from presence or absence of sketching and constructive or value-account mode of reconstruction

<table>
<thead>
<tr>
<th>Name</th>
<th>Constr.</th>
<th>V-A</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>S</td>
<td>NS</td>
</tr>
<tr>
<td><strong>Contextual information</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>a. Temporal: Does the participant recall temporal information about the reported experience?</td>
<td>Y: 20</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>N: 126</td>
<td>95</td>
</tr>
<tr>
<td>b. Event: Does the participant recall one or more discrete events that lead to the realization of the reported experience?</td>
<td>Y: 45</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>N: 101</td>
<td>83</td>
</tr>
<tr>
<td>c. Expectation: Does the participant recall his/her expectations about the reported experience?</td>
<td>Y: 10</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>N: 136</td>
<td>91</td>
</tr>
<tr>
<td><strong>Design Relevance</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0. The report contains no design relevant information</td>
<td>12</td>
<td>5</td>
</tr>
<tr>
<td>1. The report pinpoints a product feature that the user is positive/negative about</td>
<td>67</td>
<td>41</td>
</tr>
<tr>
<td>2. The report contains reasoning about the users’ attitude towards the feature</td>
<td>66</td>
<td>55</td>
</tr>
<tr>
<td>3. The report suggests a design solution</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>146</td>
<td>103</td>
</tr>
</tbody>
</table>

interaction problem or the exact actions that the user performed and which lead to the realization of the problem. Fourth, the participant might suggest a design solution for solving the reported problem. Overall, no significant differences were found in the design relevance of the experience reports across the different conditions. Thus, while sketching seems to impact the richness of elicited experience reports, it does not necessarily lead to significantly more design-relevant information.

**Does sketching result in more reliable recall of temporal structure?**

In the introduction of this chapter we argued that while the veridicality of reconstructed experiences, i.e. the convergence of memory and actuality, is not that important, their reliability, i.e. the convergence of the same memory across multiple recalls, is. In other words, we argued that even if what we remember might be different from what we experienced, as long as these memories are consistent over multiple recalls, they provide valuable information.

In this section we assess the test-retest reliability of the recalled experiences based on their temporal information. Participants joined two distinct sessions with the second session being no less than seven and no more than 10 days after the first. In both sessions participants were asked to perform the same task, to report their experiences with the product. The conditions (i.e. presence or absence of sketching and product quality) remained the same across the two sessions, but the order of
reporting was counterbalanced. For example, participants that used the iScale tool (sketching condition) to report on ease-of-use followed by the Experience Reporting tool (no-sketching condition) to report on innovativeness, in the second session they first used the Experience Reporting tool to report on innovativeness followed by the iScale tool to report on ease-of-use. These two sessions are expected to more or less result in the same experiences, thus experiences across two sessions may be coupled. For each reported experience participants estimated the time (i.e. days, weeks or months after the purchase of the product) at which this experience took place. In this test we use this temporal information, the convergence of the two reported times of two coupled experience reports elicited from the two sessions, as a metric of the reliability of the reconstruction process.

One question, however, relates to whether participants’ accuracy in recalling temporal information remains constant across the full timeline, i.e. 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 surrounding the experience that is available at the moment of recalling. Theories of recall have suggested that recent experiences (see 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 reliability of experiences reports 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 sketched linear curves through adding nodes in the graph (see figure 5.1a). 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).

Figure 5.5 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 figure 5.5, 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 sketching 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 narratives related to the first six months of use and 75% of all experience narratives 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 the following formula:

\[ \Delta = \text{Abs}(\log(t_2) - \log(t_1)) \]  

(5.1)

where \(t\) is the reported time that has elapsed from the moment of purchase of the product.

We used formula 5.1 to compute the temporal distance between the recalled time of occurrence of an experience, across the two sessions that each individual participated in. In the second session participants were asked to repeat the same tasks as in session 1. Participants were expected to elicit approximately the same experiences. The elicited experience reports from the two sessions were coupled based on their textual descriptions. The temporal distance was then calculated between each experience report elicited in session 1 and its coupled report elicited in session 2.
CHAPTER 5. ISCALE: A TOOL FOR THE ELICITATION OF EXPERIENCE NARRATIVES FROM MEMORY

Following existing evidence, we assumed that serial reconstruction, i.e. reconstructing in a chronological order, helps participants in eliciting the surrounding context of each experience [Anderson and Conway, 1993], and this positively affects participants’ ability to recall more reliably contextual cues of the experience, such as when exactly did the experience take place [Kahneman et al., 2004]. Similarly, we assumed that sketching (even in the value-account condition), when compared to the no-sketching condition, would result in more reliable recall of such temporal information as sketching provides a temporal context for all recalls, i.e. each experience report is related to others in a single timeline.

An analysis of variance with temporal distance \( \Delta \) between experience reports from session 1 and session 2 as dependent variable and mode of reconstruction, presence or absence of sketching and quality of reporting as independent variables displayed significant effects for the presence or absence of sketching, \( F(1, 265) = 9.61, p<.01 \), where the sketching condition resulted in significantly lower temporal distance (higher consistency) than the no-sketching condition (sketching: \( \mu_{\Delta}=0.178 \), no-sketching: \( \mu_{\Delta}=0.272 \)), and the quality of reporting, \( F(1,265) = 7.43, p<.01 \), where narratives of ease-of-use were temporally more consistent than narratives of innovativeness (ease-of-use: \( \mu_{\Delta}=0.183 \), innovativeness: \( \mu_{\Delta}=0.266 \)). All other main effects and interaction remained insignificant.

The results supported our initial expectations that sketching provides a temporal context and thus results in higher test-retest reliability of the reconstruction process, at least in terms of temporal information. Moreover, similarly to the analysis of the total number of experience narratives, experiences relating to the ease-of-use of the product were more reliably recalled in comparison to ones relating to the products’ innovativeness. This may be attributed to the same reasoning that experiences of ease of use are tied to concrete action, whereas experiences of innovativeness cannot be allocated to specific events. Thus, participants might be better able to reconstruct such temporal information in experiences relating to concrete events where more contextual information is available than in more abstract reports of innovativeness. Last, contrary to our initial expectation, no significant difference was found between the constructive and the value-account condition. A marginal interaction effect between the presence or absence of sketching and the quality of reporting was found, \( F(1,265) = 3.32, p=.069 \). The presence of sketching had a stronger impact when participants were reporting on innovativeness than when reporting on ease-of-use (figure 5.6a). For experience reports that related to the products’ ease-of-use, sketching had an impact in the constructive condition, but not in the value-account condition. This was not observed in experience reports of innovativeness for which sketching had a strong impact in both the constructive and the value-account conditions.
5.4. STUDY 2

Figure 5.6: Temporal inconsistency in the sketching and the no-sketching condition for the two respective qualities of reporting.

Does Value-Account result in higher sketched pattern consistency?

In the previous section we tested the reliability of recalled temporal information of experience reports across the two sessions of the study. In this section we will test the consistency of the sketched patterns, the overall evaluative judgment on participants’ experiences over time. We expect that the value-account condition will result in a higher consistency in the sketched pattern. It is sensible to assume that given that this value-charged information is assumed to be cued directly from a hypothetical memory structure \cite{Betsch2001} in the value-account condition and thus would be more consistent across repeated recalls. Contrary, in the constructive condition we assume that this information is reconstructed from contextual details that are recalled from episodic memory and thus will result in lower consistency across repeated recalls, as repeated chronological reconstruction might cue a different set of experiences and thus follow a different path in reconstructing the overall pattern of how one’s experience with a product developed over time.

Figure 5.7 displays example graphs sketched by two participants in two respective sessions in the constructive condition and two participants in the value-account condition. The area A between the two sketches is a simple measure for the inconsistency in participants’ sketched pattern over repeated recalls. It was calculated through sampling the graphs in 100 steps.
An analysis of variance with this area measure as dependent variable and mode of reconstruction and product quality being reported as independent variables displayed significant effects for the mode of reconstruction, $F(1, 45) = 6.05, p<.05$, while all other main effects and interactions remained insignificant. Contrary to what we expected, participants were more inconsistent across the two sessions when employing the value-account tool (constructive: $\mu_A = 30.2$, value-account: $\mu_A = 50.8$). 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.
How does the perceived quality of mobile phones change over time?

While presenting a detailed analysis of the obtained data is not within the scope of this chapter, this section provides a short overview of how users’ reported experiences and perceptions developed over time. We use these data to validate the framework of temporality proposed in chapter 4. While the initial study was restricted to the first month of use and one particular product, the Apple iPhone, the present dataset extends the findings to the first six months of using mobile phones. While we limited the focus of the study to one product category, we did not opt to control further the type of product.

A number of tentative assumptions were formed on the basis of the study presented in chapter 4. The majority of users’ experiences with iPhone relating to the product qualities stimulation and learnability displayed a sharp decrease over time: 58% of all experiences relating to stimulation and 59% of experiences relating to learnability took place during the first week of use. We expected this finding to be apparent also in the case of a wider spectrum of products such as the ones used in the current study, and not to be tied to the nature of iPhone, the product used in the original study. We further expected this pattern to continue beyond the first month of use, with only a small number of experiences relating to learnability or stimulation being reported after the first month. On the contrary, long-term usability and usefulness became the dominant source of satisfying and dissatisfying experiences over time. We expected this to hold also in the current study and extend beyond the first month of use. Note that we did not distinguish between satisfying and dissatisfying experiences relating to a given product quality as they both tend to display a similar pattern over time (Karapanos et al., 2009).

As discussed earlier in this chapter, 95% of all experience narratives related to the first six months of use, while the median ownership time was 10 months. One would assume that recent experiences are more easily accessible in episodic memory and would thus be over-represented in a retrospective elicitation method (see Koriat et al., 2000). This hypothesis was not supported by the data. Instead, the temporal proximity of the experience to important milestones, such as the purchase of the product, seemed to affect participants’ ability to recall or experience impactful events (Barsalou, 1988). Note, that this decrease in the number of recalled experiences over time was not affected by participants’ motivation in reporting which is often apparent in longitudinal studies as the one reported in chapter 4, instead we are more inclined to attribute this to participants’ varying ability to experience and recall impactful events over time. Consequently, the mental representation of time is expected to be affected by the variant accessibility of events pertaining to different periods. This was indeed found in our analysis where the time-scale (i.e. x-axis in the
Figure 5.8: Number of narratives in a certain period relating to learnability, stimulation, long-term usability and usefulness as a function of time of ownership.

iScale tool) was found to relate to actual time\(^2\) through a power-law. In other words, participants had a tendency to use a substantial fraction of the x-axis of iScale to map their initial experiences. Based on these two insights, it was decided to restrict the analysis of this section only to the first six months of usage; the remaining experience narratives were discarded. In addition, we chose to divide the time in a non-linear way. Specifically, figure 5.8 distinguishes three time intervals: a) Purchase - 1\(^{st}\) Week, b) 1\(^{st}\) Week - 1\(^{st}\) Month and c) 1\(^{st}\) Month - 6\(^{th}\) Month. In the same vein, figure 5.9 represents participants’ sampled perceptions of ease-of-use and innovativeness at 1\(^{st}\) Day - 6\(^{th}\) Day, 1\(^{st}\) Week - 3\(^{rd}\) Week and 1\(^{st}\) Month - 6\(^{th}\) Month.

Experience narratives were submitted to a qualitative content analysis (Hsieh and Shannon, 2005) using an a-priori defined classification scheme imposed by the temporality framework (Karapanos et al., 2009). We found experiences to tap to a wider spectrum of product qualities and not only to the ones that participants were asked to report. Overall, we identified four categories out of the initial six categories of the temporality framework. These were: usefulness, long-term usability, learnability, and innovativeness.

In support of prior work (Karapanos et al., 2008a, 2009; von Wilamowitz Moellendorff et al., 2006), the dominance of learnability and stimulation experiences decreased over time (see figure 5.8). Out of all experiences relating to learnability and pertaining to the first month of use, 74% of them took place during the first week of use. This is even higher than the 59% of the original study in chapter 4. An example of such an early learnability problem is as follows:

\(^2\)Each node in the iScale graph is annotated by the exact time which differentiates the two aligning periods (see figure 5.1a)
“[learnability, 1 week] I had to make a new contact list. This took a lot of my time. It wasn’t possible to copy my contact list to my new phone. I had to type every contact separately”

Only a small number of experiences, 7% (12) of all experiences relating to learnability, were estimated to have taken place after the first month of use. These mostly related to participants’ reflection on their increasing familiarity with the product:

“[learnability, 1 month] ... it took one month to find out how to activate the keypad lock because it was not indicated in the manual and it was not obvious on the phone”

“[learnability, 1 month] After some weeks, using my phone went better and easier. I figured out how I could use the Menu properly and the time I spend on making a text message was shorter. So I was glad about the purchase again”

In a similar vein, 75% of experiences relating to stimulation took place during the first week of use, as compared to the figure of 58% that was found in the original study. This difference between the two studies, which is apparent in both product qualities, might be an effect of the different products being studied, but also an effect of the different method, being longitudinal or retrospective. In the original study participants were asked to report daily, over a period of one month, the three most impactful experiences. Contrary, the current study asked participants in a single session to retrospect throughout the full time of ownership of the product and narrate the most impactful experiences throughout this whole period. In other words, the first study informs us about the relative dominance of experiences within a single day; the latter one informs us about the relative dominance of remembered experiences throughout the whole lifespan of a product. Thus, this difference between the two studies might pertain to the discrepancy between what we experience on a daily basis and what we remember after some time.

Next, while experiences relating to stimulation displayed a decreasing pattern, a considerable number of experiences (17%) took place after the first month. We found this to be routed in a number of different reasons.

Firstly, the novelty of a feature was often not apparent in initial use, but only after participants used it in a real context, or appropriated it in different contexts:

“[innovativeness, 4 months] With this mobile I can take pictures of the faces and then edit them in a funny way, such as making them look like an insect or an alien, which is a feature of the mobile. It also has some funny frames in which I can put peoples faces have fun”

Secondly, in certain cases participants had low expectations about the product, leading to feature discovery at a later stage in product adoption:
“[innovativeness, n/a] When I first bought the mobile I did not know that it had web access, also because I did not care much about it..."

Thirdly, after some time participants started reporting a decrease in perceived innovativeness due to new products coming out in the market or acquired by others in their social circles.

“[innovativeness, n/a] After 2 months there were coming more and more gadgets on the market. Like iPhones and Blackberrys. The number of functions of those phones is increased strongly. So when I compare my phone to those phones, it seems to be not as innovative as I thought before. Having a mobile phone with a color screen, radio and music, is normal in today’s life"

As expected, users’ experiences increasingly related to aspects of usefulness and long-term usability, and the value of the product was derived through its appropriation in different (and often unexpected) contexts. While the number of these experiences decreased over time, their relative dominance over experiences relating stimulation and learnability increased.

“[usefulness, 2 months] I found a new use for an option on my phone; you can leave a spoken message on it. So I use it for short messages to myself", [long-term usability, 2 months] “The button at the side of the phone allows easy access to the camera function... This is extremely convenient when we want to take photos while wearing gloves.”

Similarly, users were often dissatisfied by features that did not sustain their value or through long-term usability flaws that became apparent after some usage.

“[usefulness, 2 months] Although at the beginning I enjoyed taking pictures with my mobile phone, later when I saw that the quality is really low, i did not use it as much as I did before"

“[usefulness, n/a] I thought I would get used to the menu but after such a long time I still do not like it. My phone remembers where I left the menu the last time, so when I open it, it starts there. But ending there once doesn’t mean for me that I want to start there the next time. That’s a pity and I can’t adjust the settings of it”

Overall, out of all experiences that took place in the first month of use (n=305), we found a significantly higher percentage of these experiences to relate to stimulation (20% as opposed to 9% in the original study, $\chi^2, p < .001$), while a significantly lower number of experiences related to usefulness (26% as opposed to 36%, $\chi^2, p < .001$). One tentative explanation of this difference might be grounded on the retrospective experience elicitation method used in this study as compared to the longitudinal method of the original study. In other words, while our reality might be dominated by “dull” experiences relating to the product’s usefulness, we might
have a bias for remembering the most emotionally aroused ones such as experiences relating to stimulation (Christianson, 1992).

Next, an overall pattern of the perception of the two qualities, i.e. ease-of-use and innovativeness, can be obtained by averaging participants’ sketches across time. Participants’ sketched patterns were sampled according to actual time (i.e. the reported day that each experience took place), to construct average graphs of how two product qualities, ease-of-use and innovativeness, changed over time. In this sampling, each participant’s sketched pattern is transformed to an actual timeline through the use of the self-reported exact time for each node of the graph (see figure 5.1a).

The resulting averaged pattern suggests that users’ perception of the innovativeness of mobile phones increased during the first month and then remained approximately stable. On the contrary, users seemed to experience usability problems mostly in the first week of use; after this period usability displayed a sharp increase over the course of the first month while this increase continued more moderately till the end of the studied six-month period.

5.5 Discussion & Conclusion

This chapter has presented iScale, a sketching tool intended for longitudinal online surveys of user experiences. It took the general approach of retrospective elicitation of users’ experiences as an alternative to longitudinal studies. More specifically, the tool was designed with the aim of increasing participants’ effectiveness and reliability in recalling their experiences with a product. Two different versions of iScale were created based on two distinct theoretical approaches to the reconstruction of one’s emotional experiences. The constructive iScale tool imposes a chronological order in the reconstruction of one’s experiences. It assumes that chronological reconstruction
results in recalling more contextual cues surrounding the experienced events and that the felt emotion is constructed on the basis of the recalled contextual cues. The value-account iScale tool aims at distinguishing the elicitation of the two kinds of information: value-charged (e.g. emotions) and contextual cues. It assumes that value-charged information can be recalled without recalling concrete contextual cues of an experienced event due to the existence of a specific memory structure that can store the frequency and intensity of one’s responses to stimuli.

Study 2 tested the effectiveness of the two iScale tools against a control condition, that of reporting experiences without any form of sketching. We observed significant differences between iScale and the control condition. First, both sketching tools resulted in a higher number of experience reports when compared to the control condition, while the constructive sketching tool elicited significantly more experience reports than the value-account sketching tool. These results support the idea that sketching assists in 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). In addition, imposing a chronological order in the elicitation of experiences seems to have a positive impact in the effectiveness of recalling (Anderson and Conway 1993).

Second, sketching was found to result in recalling significantly more contextual information regarding an experienced event when the participants followed a chronological order in the reconstruction of their experiences, i.e. in the constructive condition, but not when they followed a top-down reconstruction approach, i.e. in the value-account condition. These results provide further evidence for the claim that chronological reconstruction seems have a positive impact in the effectiveness of recalling (Anderson and Conway 1993). This improved recall of contextual cues, however, did not impact participants’ ability to report design relevant information that was approximately similar in all four conditions. Thus, while iScale does not influence the design relevance of elicited information, it may provide richer details about the context in which a given event was experienced.

Third, sketching seemed to have a significant impact on participants’ test-retest reliability in recalling concrete contextual cues of the experienced events such as temporal information. Contrary to the initial expectations, the constructive sketching tool resulted in significantly higher consistency across two repeated recalls in the formation of overall evaluation, i.e. the sketched pattern, than the Value-Account sketching tool.

These results suggest that the constructive iScale tool outperforms the Value-Account iScale tool and offers a significant improvement in the amount, the richness and the reliability of recalled information when compared to conventional recall of experiences that does not involve any techniques for improving participants effec-
tiveness and reliability in recalling their experiences. iScale is thus a promising tool for the elicitation of longitudinal data from memory. It provides a cost-effective solution, enabling the elicitation of large amounts of information from the field while increasing the reliability of the elicited information. One has to be aware, however, of the potential discrepancy between the actual experiences as elicited through longitudinal field studies and retrospective data elicited through iScale. These retrospections may span long periods of times and thus one would expect systematic biases to occur (Kahneman 1999; Bartlett 1932). In this chapter we argued that veridicality may not be as important as the reliability of these data. This also implies that it is upon the researcher to define what she or he is interested to explore. In some cases, understanding what users remember from the temporal development of their experiences may be more important than the actual experiences, while in others the actual experiences may be what the researchers are interested in understanding.

Next, iScale will evidently result in large amounts of qualitative information that will require labor-intensive analysis given traditional human-performed qualitative data analysis procedures like Content Analysis (Krippendorff 2004; Hsieh and Shannon 2005) and Grounded Theory (Strauss and Corbin 1998). Novel techniques found in 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 sketched graphs is definitely a subject for further research and was addressed here only superficially.
A semi-automated approach to the content analysis of experience narratives

iScale will typically result in a wealth of experience narratives relating to different stages of products’ adoption. The qualitative analysis of these narrative is a labor intensive, and prone to researcher bias activity. This chapter proposes a semi-automated technique that aims at supporting the researcher in the content analysis of experience narratives. The technique combines traditional qualitative coding procedures (Strauss and Corbin [1998]) with computational approaches for assessing the semantic similarity between documents (Salton et al. [1975]). This results in an iterative process of qualitative coding and visualization of insights which enables to move quickly between high-level generalized knowledge and concrete and idiosyncratic insights.

The proposed approach was compared against a traditional vector-space approach for assessing the semantic similarity between documents, the Latent-Semantic Analysis (LSA), using a dataset of a study in chapter 4. Overall, the proposed approach was shown to perform substantially better than traditional LSA. However, interestingly enough, this was mainly rooted in the explicit modeling of relations between concepts and individual terms, and not in the restriction of the list of terms to the ones that concern particular phenomena of interest.

6.1 Introduction

The previous chapter proposed a survey technique for the elicitation of experience narratives from a large sample of users. The researcher is then faced with an over-
wholesome amount of idiosyncratic experience narratives. Each narrative may provide a rich insight into the experience and the context in which it takes place. However, generalized knowledge may also be gained from these experience narratives. Such generalized knowledge may be reflected in questions like: how frequent is a certain kind of experience, what is the ratio of positive versus negative experiences and how does this compare to competitive products, how does the dominance of different product qualities fluctuate over time and what should we improve to motivate prolonged use?

Contrary to idiosyncratic insights, generalized knowledge requires a preprocessing step: that of assessing the similarity between different experiences. While it is well-acknowledged that every experience is unique and non-repeatable (Forlizzi and Battarbee 2004; Hassenzahl and Tractinsky 2006; Wright and McCarthy 2008), different experiences share similar features: some might refer to the same product feature; others might be motivated by the same human need despite the seemingly different nature (Hassenzahl 2008); others might refer to the same product quality; some are filled with positive emotions while others are dominated by negative ones. These features may form the basis for assessing the similarity of different experience narratives.

In assessing the similarity between different narratives, content analysis techniques (Krippendorff 2004; Hsieh and Shannon 2005) may be employed in identifying key concepts in the data, identifying a hierarchical structure among the concepts and classifying narratives into broad categories. Such approaches are laborious as the researcher needs to process all narratives in identifying the key concepts in the data.

A number of automated approaches to semantic similarity assessment have been proposed in the field of Information Retrieval and can potentially assist the analyst in this task. Such approaches typically rely on vector space models (Salton et al. 1975) in which the degree of semantic similarity between documents is related to the degree of term co-occurrence across the documents. As we will argue in this chapter, these approaches exhibit a number of limitations when one is concerned about analyzing self-reported experience narratives. First, they assume a homogeneity in the style of writing across documents which does not hold in this context as the vocabulary and verbosity of documents might substantially vary across different participants. Second, similarity is computed based on the co-occurrence of all terms that appear in a pool of documents while in the qualitative analysis of experience narratives the researcher is typically interested only in a limited set of words that refer to the phenomena of interest. As a result, words that are of minimal interest to the researcher may shadow the semantic relations that researchers are pursuing at identifying. Third, these automated approaches lack an essential part of qualitative
research, that of interpretation. As different participants may use different terms or even phrases to refer to the same latent concept, an objectivist approach that relies purely on semantics will evidently fail in capturing the relevant concepts.

In this chapter, we propose a partially automated approach which combines traditional content analysis techniques (Strauss and Corbin, 1998) with computational approaches to assess the semantic similarity between documents (Salton et al., 1975). In identifying the concepts which will form the basis for computing the similarity between narratives, the proposed approach combines existing domain-specific knowledge with open-coding procedures that aim at identifying constructs that are not captured by existing measurement models. Domain-specific knowledge, on the other hand, for instance within the field of user experience, may be found in psychometric scales measuring perceived product qualities and emotional responses to products. Thus, only concepts that are relevant to the given context are considered in computing the similarity between narratives. At the same time, the process is automated as concepts identified in a small set of narratives may be used to assess the similarity among the full set of narratives. This results in an iterative process of coding and visualization of obtained insights.

Next, we describe the application of a fully-automated procedure that relies on a vector-space model (Salton et al., 1975), the Latent Semantic Analysis (LSA) and motivate the proposed adaptations of this approach towards a semi-automated approach.

6.2 Automated approaches to semantic classification

A number of automated approaches exist for the assessment of semantic similarity between documents (for an extensive review see Kaur and Hornof, 2005; Cohen and Widdows, 2009). These approaches rely on the principle that the semantic similarity between two documents relates to the degree of term co-occurrence in these documents (Deerwester et al., 1990). In this sense, every document may be characterized as an n-dimensional vector where each element of the vector depicts the number of times that a given term appears in the document. The similarity between documents may then be computed in a high-dimensional geometrical space defined by these vectors.

Latent-Semantic Analysis (LSA) (Deerwester et al., 1990), also known as Latent-Semantic Indexing within the field of Information Retrieval, is one of the most popular vector-space approaches to semantic similarity measurement. It has been shown to reflect human semantic similarity judgments quite accurately (Landauer and Dumais, 1997) and has been successfully applied in a number of contexts such as that of identifying navigation problems in web sites (Katsanos et al., 2008) and structuring
and identifying trends in academic communities (Larsen et al., 2008a).

LSA starts by indexing all $n$ terms that appear in a pool of $m$ documents, and constructs a $n \times m$ matrix $A$ where each element $a_{ij}$ depicts the number of times that the term $i$ appears in document $j$. As matrix $A$ is high-dimensional and sparse, LSA employs Singular-Value Decomposition (SVD) in reducing the dimensionality of the matrix and thus identifying the principal latent dimensions in the data. Semantic similarity can then be computed on this reduced dimensionality space which depicts a latent semantic space. Below, we describe in detail the procedure as applied in this chapter.

### 6.2.1 LSA procedure

#### Term indexing

Term-indexing techniques may vary from simple "bag-of-words" approaches that discard the syntactic structure of the document and only index the full list of words that appear in a document, to natural language algorithms that identify the part-of-speech, e.g. the probability that a term is a noun or a verb, in inferring the essence of a word (Berry et al., 1999). LSA typically discards syntactic information and treats each document as a pool of terms. However, it applies two pre-processing procedures in order to enhance the quality of the indexing procedure.

Firstly, a number of words, called stop-words, such as prepositions, pronouns and conjunctions, are commonly found in documents and carry no semantic information for the comprehension of the document theme (Fox, 1989). Such words are excluded from further analysis as they do not provide meaningful information and are likely to distort the similarity measure. We used a list stop-words provided by Fox (1989).

Secondly, the remaining terms are reduced to their root words through stemming algorithms. For instance, the terms "usability" and "usable" are reduced to the term "usabl", thus allowing the indexing of multiple forms of a word under one dimension in the vector-space model. We employed Porter’s (1980) algorithm for stemming.

#### Normalizing impact of terms

The first step in the procedure has resulted in a $n \times m$ matrix $A$ where each element $a_{ij}$ depicts the number of times that the stemmed term $i$ appears in document $j$. The frequencies of different terms across different documents will vary substantially. This results in undesired impacts of terms that are more frequent across a larger set of documents as they receive higher weight than terms that appear in only a small set of documents. However, these terms that appear in many documents have limited discriminatory power and are thus not very informative. One term-weighting criterion
that counterbalances for this inherent bias is the term-frequency inverse-document frequency (TFIDF) [Salton and Buckley, 1988]:

$$a_{i,j}\_\text{weighted} = a_{i,j} \times \log\left(\frac{nDocs}{nDocs_i}\right)$$ \hspace{1cm} (6.1)

which weights the frequency $a_{i,j}$ by the logarithm of the ratio of the total number of documents $nDocs$ by the number of documents $nDocs_i$ in which the term $i$ appears. Thus, frequent terms that appear in a large amount of documents and thus have little discriminatory power receive lower weight in the final matrix.

Dimensionality reduction

Matrix A is sparse and high-dimensional. Moreover, certain groups of terms may display similar distributions across the different documents, thus underlying a single latent variable. LSA attempts to approximate A by a matrix of lower rank. Singular Value Decomposition is used to decompose matrix A in three matrices U, S, V in that

$$A = USV^T$$ \hspace{1cm} (6.2)

Matrices U and V are orthonormal matrices and S is a diagonal matrix that contains the singular values of A. Singular values are ordered in decreasing size in matrix S, thus by taking the first $k \times k$ submatrix of S, we approximate A by its best-fit of rank $k$.

$$A_k = U_{nk}S_{kk}V_{mk}^T$$ \hspace{1cm} (6.3)

Computing document similarity

The similarity between different documents or different terms may then be computed on the reduced dimensionality approximation of A. Matrices 6.4 and 6.5 constitute $m \times m$ and $n \times n$ covariances matrices for the documents and terms, respectively. The proximity matrices for the documents and terms are then derived by transforming 6.4 and 6.5 to correlation matrices.

$$S_R = A_k^T A_k = V_{m \times k}S_{k \times k}^2 V_{m \times k}^T$$ \hspace{1cm} (6.4)

$$A_k A_k^T = U_{n \times k}S_{k \times k}^2 V_{n \times k}^T$$ \hspace{1cm} (6.5)

Each element $s_{i,j}$ represents the similarity between documents, or terms $i$ and $j$. The proximity matrix is normalized to a range $(0,1)$ and transformed to a distance matrix with each element $d_{i,j} = 1 - |s_{i,j}|$. 
6.2.2 Limitations of LSA in the context of qualitative content analysis

Latent-Semantic Analysis has been shown to adequately approximate human judgments of semantic similarity in a number of contexts (Landauer et al., 2003; Katsanos et al., 2008; Larsen et al., 2008a). However, one may expect a number of drawbacks when compared to traditional content analysis procedures as applied by researchers.

First, LSA assumes a homogeneity in the style of writing across documents. Thus, the extend to which different words occur in one document over a second one denotes a difference in content across the two documents. This assumption has been shown to hold in contexts of formal writing such as web pages (Katsanos et al., 2008) or abstracts of academic papers (Larsen et al., 2008a), but it is not expected to hold in qualitative research data such as interview transcripts or self-provided essays in diary studies as the vocabulary and verbosity of documents might substantially vary across different participants.

Second, LSA computes the similarity between documents based on the co-occurrence of all possible terms that may appear in the pool of documents. In the analysis of qualitative data, however, one is interested only in a small set of words that refer to a phenomenon that the researchers are interested in. As a result, words that are of minimal interest to the researchers may shadow the semantic relations that researchers are pursuing at identifying.

Third, LSA lacks an essential part of qualitative research, that of interpretation. As different participants may use different terms or even phrases to refer to the same latent concept, an objectivist approach that relies purely on semantics will evidently fail in capturing the relevant concepts. Ideally, automated vector-space models could be applied to meta-data that have resulted from open coding qualitative procedures (Strauss and Corbin, 1998). In the next section we propose such a semi-automated approach to semantic classification.

6.3 A semi-automated approach to content analysis

In this section we propose a semi-automated approach that aims at addressing the limitations of Latent-Semantic Analysis in assessing the similarity between self-reported experiences. The approach is different from Latent-Semantic Analysis in the following respects.

First, only terms relevant to the phenomena that the researcher is interested in are employed in the similarity measure. The approach exploits existing domain-specific knowledge in identifying relevant concepts, but also acknowledges that such knowledge will always be incomplete; thus the intervention of the researcher in extracting additional relevant concepts from the data is crucial.
Second, through manual coding the researcher is able to define explicit relations between concepts, i.e. latent dimensions, and individual measurement items, i.e. terms of phrases. For instance, an researcher may wish to code when the participant refers to a relevant person in a social context. The researcher may then create a concept named ‘relevant others’ that is identified in the text through various individual terms such as ‘friend’, ‘brother’, ‘colleague’ etc. Latent-Semantic Analysis would not be likely to identify these terms as semantically similar as these would not likely occur in the same documents. In contrast, through manual coding the researcher is able to explicitly relate these terms as different manifestations of the same latent concept.

Third, the approach utilizes visualization techniques in assisting the researcher in identifying the relevant concepts. For instance, an interacting visualization of the dissimilarity between narratives enables a systematic comparison of diverse narratives.

### 6.3.1 Incorporating existing domain-specific knowledge

The procedure starts by incorporating existing domain-specific knowledge. In characterizing users’ experiences with interactive products, relevant knowledge might be found in psychometric scales measuring subjectively perceived product qualities (e.g. Hassenzahl 2004; Lavie and Tractinsky 2004b; Hornbæk and Law 2007) and emotional responses to products (e.g. Desmet and Hekker 2007).

Table 6.1 illustrates the five latent concepts that were used in the current study. The concepts were derived from Hassenzahl (2004) and Lavie and Tractinsky (2004b). Each construct is measured through a number of individual (bi-polar) semantic differential scales. Hassenzahl (2004) distinguishes pragmatic and hedonic quality in interactive products, while Lavie and Tractinsky (2004b) differentiate between classic and expressive aesthetics. Both poles of all individual scales are used in defining each respective concept, after being stemmed to their root form (Porter, 1980).

### 6.3.2 Iterative open coding

Next to existing domain-specific knowledge, the researcher may want to annotate the data using traditional coding practices (Strauss and Corbin, 1998). Coding consists of two steps. In open coding the researcher derives concepts from terms or phrases in the raw data. A concept may be defined through a direct relation to a specific term in the raw data (in-vivo coding; Glaser and Strauss, 1967), for instance the researcher may create a concept termed ‘friend’ referring to all terms ‘friend’ in the raw data), or may relate to terms through an intermediate step of interpretation (for instance, the researcher might desire to group terms ‘friend’, ‘family’ and ‘colleague’ under the con-
Table 6.1: Domain-specific knowledge employed for the analysis of the 329 narratives. The domain-specific concepts were derived from Hassenzahl (2004) and Lavie and Tractinsky (2004b) psychometric scales.

<table>
<thead>
<tr>
<th>Concept</th>
<th>Individual terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pragmatic Quality</td>
<td>technical, human, complicated, simple, impractical, practical, cumbersome, direct, unpredictable, predictable, Confusing, clear, Unruly, manageable</td>
</tr>
<tr>
<td>HQ-Stimulation</td>
<td>typical, original, standard, creative, cautious, courageous, conservative, innovative, lame, exciting, Easy, challenging, Commonplace, new</td>
</tr>
<tr>
<td>HQ-Identification</td>
<td>isolating, integrating, amateurish, professional, gaudy, classy, cheap, valuable, non-inclusive, inclusive, unpresentable, presentable</td>
</tr>
<tr>
<td>Classic aesthetics</td>
<td>aesthetic, pleasant, clear, clean, symmetric, artistic</td>
</tr>
<tr>
<td>Expressive aesthetics</td>
<td>creative, fascinating, special effects, original, sophisticated</td>
</tr>
</tbody>
</table>

cept named 'relevant others'). In a second step, termed axial coding, the researcher may define a hierarchical structure between concepts. The creation of this hierarchical structure is realized through the definition of a superordinate concept that includes all terms that appear in its subordinate concepts.

For instance, table 6.2 displays the full list of concepts along with coded examples, derived from the analysis of 329 narratives of the study described in chapter 4. Figure 6.1 displays a two-dimensional visualization of the similarity between all concepts including those deduced from domain-specific knowledge (table 6.1) and those derived from coding the data (table 6.2). Distances between concepts were derived from equation 6.7 and submitted to Multi-Dimensional Scaling. A two-dimensional solution was extracted. A-priori defined concepts are depicted in italics, while superordinate concepts are denoted in bold. Note that similar concepts (e.g. Novelty - Aesthetics in Interaction) often co-occur in the same narratives as displayed by the high similarity in the two-dimensional space.

One problem often met in coding procedures is the fixation of the researcher onto a certain perspective leading to increased biases in the interpretation process, what Kahneman et al. (1982) call anchoring bias. Strauss and Corbin (1998) proposed a number of techniques aimed at supporting the researcher in taking different perspectives in the interpretation process. They argued that the very same comparison that can be performed on two objects (e.g. two flowers) resulting in the identification of the objects’ properties (e.g. size, shape, color), may also be performed on theoretical concepts. For instance, the flip-flop technique involves a systematic comparison between a concept and its extreme opposite in identifying the concept’s properties, e.g. what is meant by "immediate" access to information: time, effort or ubiquity.
Table 6.2: Concepts that were derived from the analysis of the 329 narratives along with examples of coded data. Next to existing domain-specific knowledge, the researcher annotates data by relating individual terms or phrases to a set of latent concepts (linked concepts are denoted in [brackets]).

<table>
<thead>
<tr>
<th>Concept</th>
<th>Individual terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Self-expression</td>
<td>proud, show off, impress, status, jealous</td>
</tr>
<tr>
<td>Relevant others</td>
<td>friend, colleague, dad, daughter, brother</td>
</tr>
<tr>
<td>Sense of community</td>
<td>technology-freak, discuss, chat</td>
</tr>
<tr>
<td>Simplicity</td>
<td>easy, simple, straightforward, one click, seamless</td>
</tr>
<tr>
<td>Understandability</td>
<td>very logical, no clue how, i could not understand</td>
</tr>
<tr>
<td>Efficiency</td>
<td>almost instantly, no boot time, had to go through all the list</td>
</tr>
<tr>
<td>Effectiveness</td>
<td>organizes my messages nicely, simply not allowed, was not able to see</td>
</tr>
<tr>
<td>Fits context</td>
<td>use only one hand, wear gloves, while driving</td>
</tr>
<tr>
<td>Adaptable</td>
<td>adapts the screen brightness, totally customized, dynamically change</td>
</tr>
<tr>
<td>Familiarity</td>
<td>more used to, I am finding out, after using it for a while</td>
</tr>
<tr>
<td>Novelty</td>
<td>cool and intelligent design, genius idea, hay look at that its magic</td>
</tr>
<tr>
<td>Visual aesthetics</td>
<td>everything that shines, beautiful colors, shaky icons look beautiful</td>
</tr>
<tr>
<td>Aesthetics of interaction</td>
<td>glow effect, i just loved the way it turned around, so easy and fun</td>
</tr>
<tr>
<td>Fast access to information</td>
<td>immediate access, overhead of booting a computer, would start up my computer</td>
</tr>
<tr>
<td>Alleviate boredom in idle periods</td>
<td>in case I get bored, playing a bit every time, I can’t really do anything else</td>
</tr>
<tr>
<td>Capturing momentary info</td>
<td>offload my own memory, have no means to note it down, added quite a text</td>
</tr>
<tr>
<td>Missing functionality</td>
<td>does not have, not supported, all phones can do it</td>
</tr>
<tr>
<td>Stimulation</td>
<td>[Novelty], [Visual aesthetics], [Aesthetics of interaction]</td>
</tr>
<tr>
<td>learnability</td>
<td>[simplicity], [Understandability]</td>
</tr>
<tr>
<td>Long-term usability</td>
<td>[Efficiency], [Effectiveness], [Fits context], [Adaptable], [Familiarity]</td>
</tr>
<tr>
<td>Useful</td>
<td>[fast access to information], [alleviate boredom in idle periods], [capturing momentary information], [missing functionality]</td>
</tr>
</tbody>
</table>
Figure 6.1: Multi-Dimensional Scaling on concept dissimilarity (eq. 6.7). Concepts connected through lines belong to the same cluster. A-priori defined concepts are depicted in italics, while superordinate concepts are denoted in bold.

Strauss and Corbin (1998) emphasize the role of diversity in the cases in the process of deriving new interpretations. This process may be supported through the use of interactive visualizations. For instance, figure 6.2 illustrates an interactive scatter plot that aims at enabling the researcher in browsing through diverse narratives. It computes the similarity between narratives from the co-occurrence of concepts (i.e. equation 6.6) and visualizes the first two dimensions derived from Multi-Dimensional Scaling. As more narratives are analyzed, the plot is updated. By selecting one of the points in the plot, the narrative is displayed on the right part of the interface and the coded data are highlighted and annotated by the respective concept. The researcher is further able to visualize other aspects of the experience narratives such as self-reported satisfaction for a given experience, day of occurrence for the reported experience, as well as the number of codes within each narrative.
6.3. A SEMI-AUTOMATED APPROACH TO CONTENT ANALYSIS

Figure 6.2: Interactive scatter plot depicting the similarity between narratives derived from the co-occurrence of coded concepts and processed by means of Multi-Dimensional Scaling. The researcher may browse through different data instances, or identify the least coded narratives. By selecting a narrative the full text appears with the coded part highlighted and annotated by the respective concept. The analyst can further code the data, which will in turn influence the visualizations.

show off
in the evening we had Friend [Relevant Others]'s over for dinner, they are also quite technology freak [sense of community]'s quite quickly i have told them that i've got an iphone and showed it [self-expression] to them. i really liked watching them playing wit it we started browsing web - i was showing them how tiling works, then we tried to download a youtube movie but it failed, we tried another time and it failed again. it is probably not a problem with the iphone but with the wi-fi connection but at this point we all lost interest in playing with it anymore.

Day: 2
Satisfaction: 6
Figure 6.3: Interactive scatter plot depicting the number of codes within each narrative. The analyst may browse through different data instances, or identify the least coded narratives. By selecting a narrative, the full text appears with the coded part highlighted and annotated by the respective concept. The analyst can further code the data, which will influence the visualizations.

**Satisfaction**

- Highly productive [simply easy] when even looks perfect!
- Never feels like being overloaded with a complex interface

**Day 15**

- Good long drop-down [long-term memory] [tolerance]
- Need to express my point of view [long-term memory] [tolerance]
- Need to express my points of view [long-term memory] [tolerance]

**Satisfaction high again**
6.3.3 Computing narrative similarity

The similarity between narratives is computed based on an explicitly defined latent space of concepts that the researcher is interested in. Using the vector-space model (Salton et al. 1975), a \( n \times m \) matrix \( A \) is defined where each element \( a_{ij} \) depicts the presence or absence (or number of times) of concept \( i \) in narrative \( j \). Each narrative is thus described by an \( n \)-dimensional vector and the similarity between the narratives can be equated to the cosine of the mutual angle between the two vectors, as in Latent-Semantic Analysis.

One limitation, however, of this measure is that concepts that do not appear in both narratives are taken into account in computing the similarity between the narratives. These concepts are likely to dominate as only a small set of the concepts is likely to appear in 2 out of the 329 narratives. Figure 6.4 displays a two-dimensional configuration of all possible states for a given concept: appearing in none of the narratives (\( N_{00} \)), appearing in both narratives (\( N_{11} \)), or appearing in one of the narratives (\( N_{01} \) and \( N_{10} \)). If all concepts are plotted for a given pair of narratives, the \( N_{00} \) would dominate due to their high frequency. These concepts, however, bear no useful information about either of the narratives.

An alternative measure can be defined by discarding concepts that do not appear in either narrative. More specifically, we propose to equate the similarity \( S_{i,j} \) between two narratives to the ratio of the concepts that appear in both narratives \( N_{11} \) over the sum of the concepts that appear in one of the narratives. The distance is then derived in 6.7 as \( D_{i,j} = 1 - S_{i,j} \):

\[
S_{i,j} = \frac{N_{11}}{N_{01} + N_{10}} \tag{6.6}
\]

\[
D_{i,j} = \frac{N_{01} + N_{10} - N_{11}}{N_{01} + N_{10}} \tag{6.7}
\]
CHAPTER 6. A SEMI-AUTOMATED APPROACH TO THE CONTENT ANALYSIS OF EXPERIENCE NARRATIVES

Table 6.3: Three most dominant concepts in each cluster

<table>
<thead>
<tr>
<th>Cluster No.</th>
<th>terms.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cluster 1</td>
<td>Stimulation, Novelty, Aesthetics in Interaction</td>
</tr>
<tr>
<td>Cluster 2</td>
<td>Missing functionality, Useful, HQ-Stimulation</td>
</tr>
<tr>
<td>Cluster 3</td>
<td>Long-term usability, Efficiency, HQ-Stimulation</td>
</tr>
<tr>
<td>Cluster 4</td>
<td>Learnability, Simplicity, HQ-Stimulation</td>
</tr>
<tr>
<td>Cluster 5</td>
<td>Useful, Fast access to info, HQ-Stimulation</td>
</tr>
<tr>
<td>Cluster 6</td>
<td>Relevant Others, Self-expression, HQ-Stimulation</td>
</tr>
<tr>
<td>Cluster 7</td>
<td>Long-term usability, Fits context, Adaptable</td>
</tr>
<tr>
<td>Cluster 8</td>
<td>Understandability, HQ-Stimulation, Pragmatic Quality</td>
</tr>
<tr>
<td>Cluster 9</td>
<td>Long-term usability, Effectiveness, simplicity</td>
</tr>
</tbody>
</table>

Likewise, the similarity between two concepts may be computed from the ratio of the number of narratives in which both concepts appear over the sum of the number of documents in which one of the concepts appears.

6.3.4 Hierarchical Clustering

Once the similarity between all pairs of narratives has been computed, the distance matrix can be submitted to a clustering algorithm and the emerging clusters can be characterized by the three most dominant terms or concepts that appear in the narratives that are grouped together.

For instance, table 6.3 presents the three most dominant concepts for each of the first nine clusters that resulted from Ward’s hierarchical clustering with a minimum variance criterion on the distance matrix of the 329 narratives. One may note that cluster 1 is dominated by experience narratives relating to Stimulation, clusters 4 and 8 mostly by narratives relating to learnability, clusters 3, 7 and 9 by narratives relating mostly to long-term usability, clusters 2, 5 by narratives relating to usefulness and narratives in cluster 6 relate to social experiences. This information may assist the researcher in further grouping the resulting clusters.

6.3.5 Visualizing insights

Once narratives have been classified into categories that represent similar meanings, a number of generalized insights may be gained through the visualizations and exploration of the interrelations of meta-data such as the day in which a reported experience occurred (out of the 28 days of the study), and a self-reported satisfaction value \cite{Osgood et al., 1957}.

For instance, figure 6.5 illustrates a visualization of frequency and average impact of four types of experiences over the four weeks of the study. The Y axis corresponds to the number of experiences being reported for each respective experience type per
week, while the average impact of experiences (as reported by the participants) for a given type are visualized through the size of each circle. Four overall types of experiences are distinguished: experiences relating to stimulation, learnability, long-term usability and usefulness. A 4-point (0-3) impact index of each reported experience is derived from the self-reported satisfaction elicited through a 7-point (1-7) semantic differential scale\textsuperscript{6.8} Satisfying experiences are differentiated from dissatisfying ones.

\[ Impact_{0-3} = |Satisfaction_{1-7} - 4| \] (6.8)

One may note that experiences related to learnability are the most dominant experiences during the first week. These experiences sharply decrease in the second week while the number of experiences relating to long-term usability and usefulness constantly increase over the four weeks of the study. Experiences related to stimulation seem to be the most impactful experiences during the first week followed by experiences of learnability. Surprisingly, experiences relating to long-term usability seem to be the least impactful despite their dominance over time. This finding did not become apparent in the initial analysis described in chapter 4 as the relation between time and impact of experiences was not explored. Visualization tools like the one illustrated in figure 6.5 enable rapid hypothesis construction and testing in the exploration of such data.

A different visualization, displayed in figure 6.6, can provide insights into the relations between product qualities (rows) and product features (columns, see table 6.4). Each cell in the matrix depicts the percentage of experiences referring in at least one instance to the respective product quality and product feature out of the total list of experience narratives that refer to the given feature (i.e. every column adds up to 100\%). The researcher may for instance obtain insights into the overall experience of a given feature through the distribution of experiences over the five types, i.e. product qualities. This may lead, for instance, to the identification of the features that induce desired experiences such as the ones related to self-representation or the features that are dominated by learnability and long-term usability problems. By clicking on a cell, the full set of narratives relating to the respective product quality and product feature appear.

### 6.4 Validation of the proposed approach

Three distinct approaches were tested on a subset of the data in chapter 4. These were (a) Computing similarity on explicitly defined dimensions using the approach described in this chapter (Concept Analysis), (b) applying LSA on a restricted list
CHAPTER 6. A SEMI-AUTOMATED APPROACH TO THE CONTENT ANALYSIS OF EXPERIENCE NARRATIVES

162

Figure 6.5: Visualizing frequency of experiences (y axis) and average impact (size of circle) for four types of experiences colored differently over the four weeks of the study.

Figure 6.6: Visualizing relations between product qualities and product features. Each column depicts the distribution of experiences over the five product qualities for a given product feature.
Table 6.4: Concepts relating to product features derived from the analysis of the 329 narratives along with examples of coded data. Inclusion of subordinate concepts is denoted through brackets.

<table>
<thead>
<tr>
<th>Concept</th>
<th>Individual terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interaction</td>
<td>[Sensors], [Interactive widgets]</td>
</tr>
<tr>
<td>Sensors</td>
<td>accelerometer, proximity, sensor, rotation, guitar</td>
</tr>
<tr>
<td>Interactive widgets</td>
<td>interact, button, top bar, finger, scroll, touch screen, zoom, rotation, double-touch, sketch, draw, navigate</td>
</tr>
<tr>
<td>Hard components</td>
<td>[Packaging], [Form], [Battery]</td>
</tr>
<tr>
<td>Packaging</td>
<td>packaging, case</td>
</tr>
<tr>
<td>Form</td>
<td>form, color</td>
</tr>
<tr>
<td>Battery</td>
<td>battery</td>
</tr>
<tr>
<td>Interface</td>
<td>interface, menu, folder</td>
</tr>
<tr>
<td>Synchronization</td>
<td>sync, computer, laptop</td>
</tr>
<tr>
<td>Installation</td>
<td>[Language], [Install], [Connectivity], [Contactlist]</td>
</tr>
<tr>
<td>Language</td>
<td>language, english, dutch, dictionary</td>
</tr>
<tr>
<td>Install</td>
<td>install, configure, volume</td>
</tr>
<tr>
<td>Connectivity</td>
<td>WiFi, wi-fi, wifi, WIFI, wlan, wireless, network, blue-tooth, blue tooth</td>
</tr>
<tr>
<td>Contactlist</td>
<td>contact</td>
</tr>
<tr>
<td>Apps</td>
<td>[Note], [Calendar], [Alarm], [Games], [Photo], [Calculator], [Internet], [Video], [Communication]</td>
</tr>
<tr>
<td>Note</td>
<td>note, todo, To-Do, to-do</td>
</tr>
<tr>
<td>Calendar</td>
<td>calendar, appointment</td>
</tr>
<tr>
<td>Alarm</td>
<td>alarm</td>
</tr>
<tr>
<td>Games</td>
<td>game</td>
</tr>
<tr>
<td>Photo</td>
<td>photo, picture, camera, webcam</td>
</tr>
<tr>
<td>Calculator</td>
<td>calculator</td>
</tr>
<tr>
<td>Internet</td>
<td>internet, safari, brows, web</td>
</tr>
<tr>
<td>Video</td>
<td>video, youtube, movie</td>
</tr>
<tr>
<td>Communication</td>
<td>[Instant messaging], [Email]</td>
</tr>
<tr>
<td>Instant messaging</td>
<td>sms, chat, messenger, facebook</td>
</tr>
<tr>
<td>Email</td>
<td>email</td>
</tr>
<tr>
<td>Audio</td>
<td>[Ringtone], [Music]</td>
</tr>
<tr>
<td>Ringtone</td>
<td>ringtone</td>
</tr>
<tr>
<td>Music</td>
<td>piano, equalizer, album, song, ipod, itunes</td>
</tr>
</tbody>
</table>
of terms, and (c) applying LSA on all terms (traditional LSA). Thus, by splitting the proposed approach in two distinct procedures (a and b), we are able to distinguish between the impact of i) restricting the list of terms, and ii) defining explicit relations between concepts and the observed terms.

6.4.1 Preparing the dataset

These amounted to a total of 347 experience narratives classified under five main categories: stimulation, learnability, long-term usability, usefulness and social experiences. Figure 6.7 displays the frequency and cumulative percentage of narratives across the number of words contained in each narrative. Large variances in the number of words across different narratives is likely to impact the computation of the semantic similarity of narratives with long narratives receiving higher similarity scores to other narratives as the probability of occurrence of words of interest is greater. This bias may be counterbalanced by weighting frequencies $a_{ij}$ by the length of the document $j$. However, narratives below a certain number of words are not likely to contain adequate semantic information and may be excluded from further analysis. For this dataset, narratives in the lower end of the graph (5%) with less than 24 words were excluded from further analysis, resulting in a total of 329 narratives.

6.4.2 Concept Analysis

The first approach used the procedure proposed in this chapter for the analysis of the 329 narratives. This analysis resulted in a total of 26 concepts referring to 539 terms or phrases in the narratives. Five concepts were derived from existing domain-specific knowledge (see table 6.1) while the remaining 21 concepts were derived from the data through the qualitative coding procedure discussed in this chapter (see table 6.2).

The dissimilarity between narratives was then computed using the 26 concepts (equation 6.7), resulting in a 329x329 distance matrix.

6.4.3 Latent-Semantic Analysis on restricted terms

In the second approach, the explicit relations between concepts and terms were discarded. Instead, Latent-Semantic Analysis was applied using the restricted list of terms (539) that were identified by the researcher. Singular Value Decomposition was applied to the 539x329 matrix to extract the 26 most dominant latent dimensions. The optimal dimensionality in LSA is an ongoing research question, with some suggesting a dimensionality between 100 and 300 (Landauer and Dumais 1997), while others suggest that most variance can be captured in the first 10 dimensions (Kon-
6.4. VALIDATION OF THE PROPOSED APPROACH

<table>
<thead>
<tr>
<th>Number of narratives</th>
<th>50</th>
<th>40</th>
<th>30</th>
<th>20</th>
<th>10</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of words per narrative</td>
<td>250</td>
<td>240</td>
<td>230</td>
<td>220</td>
<td>210</td>
<td>200</td>
</tr>
</tbody>
</table>

![Histogram on the length (number of words) of experience narratives. 5% of the narratives, i.e. the ones containing less than 24 words, were excluded from further analysis.](image)

We applied a shared dimensionality of 26 in all three approaches to minimize any effects induced by differences in dimensionality when comparing the three approaches.

### 6.4.4 Traditional LSA

The third approach was a traditional Latent-Semantic Analysis as described in this chapter. It involved two pre-processing steps: a) extracting a list of stop-words, and b) stemming terms to their root form. This resulted in a total of 1873 unique terms that were used to characterize the 329 narratives. The resulting 1873x329 matrix was submitted to a Singular-Value Decomposition and the dominating 26 latent dimensions were extracted.

### 6.4.5 Cluster analysis on Dissimilarity Matrices

All three procedures resulted in a 329x329 matrix depicting the dissimilarity between the narratives. The three dissimilarity matrices were then submitted to hierarchical cluster analysis using a minimum variance criterium and the first *nine* clusters were extracted.
The performance of the three approaches is compared by contrasting the output of each method to the output of the hand-coded classification in the original study (chapter 4). The original hand-coded classification resulted in the identification of five overall categories: stimulation, learnability, long-term usability, usefulness and social experiences. Traditional content analysis, as applied in the original study, is considered as an optimal classification and used as reference for the three automated procedures.

To enable the comparison between the output of the three approaches with the output of the content analysis of the initial study, a mapping needs to be created between the 9 clusters generated by each of the three approaches and the five categories of the traditional content analysis. Once all 9 clusters are assigned to one of the five overall categories, interrater agreement indices such as the Kappa statistic (Fleiss et al., 2003), or the overall percent of correctly classified narratives may be computed in assessing the agreement between the three automated approaches and the traditional content analysis.

We employ two approaches for assigning each of the nine clusters to one of the five identified categories. First, this may be performed based on the distribution of narratives within a cluster over the five categories. The distribution for all nine clusters may be visualized in a 9x5 matrix where each cell of the matrix $m_{i,j}$ depicts the number of narratives that are classified to the cluster $i$ (out of the 9 overall clusters that resulted from the automated analysis procedure) and to the $j$ category (out of the 5 categories that resulted from the manual coding procedure in the initial study). According to this criterium, each cluster is assigned to that category that contains the highest number of narratives. This approach minimizes the error induced by the mapping process, and results in the best possible value for the agreement between the automated methods and traditional content analysis.

However, this best possible value may not be obtained in real settings where human interpretation is required to further classify the narratives. Thus, a second approach involves human raters. Each cluster, as proposed earlier in this chapter, can be characterized by the three most dominant terms in the experience narratives that are clustered together. This information may assist the researcher in interpreting the theme of each cluster and proceed in further grouping of the clusters.

Using the three most dominant terms for each cluster, six individuals were asked to assign each of the nine clusters to one of the five overall categories: stimulation, learnability, long-term usability, usefulness and social experiences.

Figure 6.8 displays the percentage of narratives being correctly classified, i.e. assigned to the same category as in the hand-coded analysis of the initial study Karapanos et al. (2009), both for the optimal approach that employs the distribution of narratives across the five categories and for the approach that employs human raters.
6.4. VALIDATION OF THE PROPOSED APPROACH

![Graph showing validation results]

Figure 6.8: Percentage of narratives being correctly classified, i.e. assigned to the same category as in the hand-coded analysis of the initial study Karapanos et al. (2009), both for the optimal approach and for the approach that employs human raters in assigning the 9 clusters to five categories. The 95% confidence intervals depict the uncertainty derived from the human classification of the nine clusters into the five overall categories.

A number of insights are gained from this analysis. First, note the substantial difference in interrater agreement between the optimal mapping process (based on narratives distribution) and the process that involved human raters. This difference is larger in the two approaches that involved LSA as these procedures perform less favorably and are more ambiguous in characterizing the clusters through the three most dominant terms.

Next, the proposed method performs substantially better than traditional LSA, i.e. 66% of narratives were correctly classified versus 24% in the case of LSA when human raters are involved, while the optimal mapping process results in respective scores of 88% (Interrater agreement K=0.85 [Fleiss et al., 2003]) for the proposed method and 63% (Interrater agreement K=0.54) for traditional LSA.

Surprisingly, restricting the list of terms when applying LSA provides minimal improvement over the output of traditional LSA, i.e. on the full list of terms. Only when the explicit relations between terms and latent concepts are modeled there was a substantial improvement in the agreement with the hand-coded classification in the initial study.

Tables 6.5, 6.6 and 6.7 depict confusion matrices for the three semi-automated approaches (using the best possible value approach when mapping the nice clusters
CHAPTER 6. A SEMI-AUTOMATED APPROACH TO THE CONTENT ANALYSIS OF EXPERIENCE NARRATIVES

6.5 Discussion

Overall, the proposed approach, was shown to display a substantially closer fit to the results of manual clustering of narratives in comparison to Latent-Semantic Analysis. However, interestingly enough, this was mainly rooted in the explicit modeling of relations between concepts and individual terms, and not in the restriction of the list of terms to the ones that concern particular phenomena of interest.

At the same time, the proposed approach displays a number of advantages over traditional manual coding procedures as concepts already defined through the analysis of a subset of narratives, are indexed and used to characterize all narratives. This has two main implications. First, in the case of an established body of knowledge in the domain, the researcher can start from an initial classification. Second, the

\[
\begin{array}{cccccc}
\text{Stim.} & \text{Learn.} & \text{LT Usab.} & \text{Usef.} & \text{Social} \\
\hline
C1. & 0 & 0 & 0 & 0 & 0 \\
C2. & 24 & 51 & 23 & 1 & 1 \\
C3. & 1 & 23 & 79 & 22 & 20 \\
C4. & 1 & 1 & 1 & 78 & 2 \\
C5. & 0 & 1 & 2 & 3 & 4 \\
\end{array}
\]

Table 6.5: Confusion Matrix for traditional LSA

\[
\begin{array}{cccccc}
\text{Stim.} & \text{Learn.} & \text{LT Usab.} & \text{Usef.} & \text{Social} \\
\hline
C1. & 0 & 0 & 0 & 0 & 0 \\
C2. & 22 & 53 & 8 & 10 & 7 \\
C3. & 2 & 13 & 95 & 19 & 15 \\
C4. & 3 & 9 & 0 & 72 & 1 \\
C5. & 0 & 0 & 0 & 0 & 0 \\
\end{array}
\]

Table 6.6: Confusion Matrix for LSA on restricted terms

\[
\begin{array}{cccccc}
\text{Stim.} & \text{Learn.} & \text{LT Usab.} & \text{Usef.} & \text{Social} \\
\hline
C1. & 24 & 0 & 0 & 0 & 0 \\
C2. & 41 & 67 & 3 & 6 & 4 \\
C3. & 0 & 0 & 89 & 0 & 0 \\
C4. & 2 & 8 & 11 & 90 & 0 \\
C5. & 0 & 0 & 0 & 5 & 19 \\
\end{array}
\]

Table 6.7: Confusion Matrix for Concept Analysis (proposed procedure)

into five overall categories).
6.5. DISCUSSION

Researcher is assisted by the iterative process of coding and visualization where the impact of newly defined concepts is directly visible on the full set of data.

This, however, also introduces some complications for the coding process. When defining a concept through the interpretation of a certain term or phrase, the researcher is also forced to assess the generalizability of the relation between term and concept under different contexts. In principle, the researcher needs to avoid coding lengthy phrases as the likelihood of occurrence of the same phrase in other narratives decreases as its length increases. On the other hand, a narrow interpretation of a single term in a given context entails the risk of false attribution of meaning of the term in other contexts. Thus the researcher needs to find a balance between, on the one hand, effective characterization of narratives, and on the other hand, omitted risks of false interpretation in other contexts. We argue that this process minimizes the risks of over-interpretation of qualitative data as the researcher is forced towards a systematic comparison of the use of a given term or phrase across different contexts.

While the proposed approach argues for the aggregation of experiences into generalized knowledge, it does not intend to downplay the role of idiosyncratic insights in the design process. Instead, it argues for an interplay between generalized knowledge and idiosyncratic insights. The latter ones may provide a rich understanding on how the product is experienced in a given physical and social context. They may lead to an understanding of users' needs and generate ideas for design. But, design needs prioritization. Generalized knowledge may point at the important issues. For instance, how frequent is a given experience? What experiences dominate over prolonged use? Next, not all experiences are equally interesting to the designers. An information-theoretic approach would suggest that interestingness relates to the rarity of an experience. Thus, by visualizing the interrelation between experience narratives (see fig. 6.2) we enable the researcher to identify experiences that bear limited similarity to others and are thus more likely to contain new information for design.

The proposed approach is only a first step towards more automated approaches for visualizing and interacting with qualitative data. A number of limitations and future directions may be noted. First, coding is now restricted to matching terms or phrases whereas more advanced coding procedures might also include logical operations. For instance, the researcher might desire to identify a concept when two terms appear (at any place) within the narrative or when a term appears while another term is absent (e.g. the term beautiful is likely to relate to the concept of visual aesthetics only when terms like simple and interactive are absent). Similarly, axial coding is now restricted to establishing a hierarchical structure between the concepts while a wider spectrum of relations would enable a richer exploration of data. For instance, causal relations between concepts might be explicitly identified by the researcher or automatically established by the relative position of concepts within the unit of analysis.
6.6 Conclusion

This chapter highlighted two problems in the qualitative analysis of experience narratives. First, qualitative analysis is a labor intensive activity which becomes increasingly a concern when qualitative data may be elicited from a large amount of participants as in the case of iScale. Second, qualitative analysis has been shown to be prone to researcher bias as humans often rely on heuristics in forming judgments about the relevance or similarity of two or more data instances (Kahneman et al., 1982).

This chapter proposed a semi-automated approach that aims at supporting the researcher in the content analysis of experience narratives. This approach relies on a combination of traditional qualitative coding procedures (Strauss and Corbin, 1998) with computational approaches to the assessment of semantic similarity of documents (Salton et al., 1975). The approach shares a number of advantages over traditional content analysis procedures as the coding scheme derived in the analysis of a small set of data is used to characterize all remaining data. Thus, through an iterative process of coding and visualization of insights, the approach enables the researcher in moving between highly idiosyncratic insights and generalized knowledged. Secondly, as the researcher is forced to examine the use of language under different contexts, it minimizes the risk of over-interpretation which is common in traditional qualitative analysis practices.

Using data from chapter 4, the performance of the proposed approach was compared to the one of a fully automated semantic analysis procedure, the Latent-Semantic Analysis (Deerwester et al., 1990). The proposed approach was found to display a substantially closer fit to the results of manual clustering of narratives in comparison to Latent-Semantic Analysis.
Part IV

Reflections
Conclusions

The focus of Human-Computer Interaction has evolved from the study of the usability of interactive products towards a more holistic understanding of the psychological and social impact of products in people’s lives. Firstly, this has resulted in a wealth of new concepts such as pleasure (Jordan, 2000), fun (Blythe et al., 2003), aesthetics (Tractinsky et al., 2000), and hedonic qualities in the use of personal interactive products (Hassenzahl, 2004), but also aspects of trust in online transactions (Egger, 2003), and the increased social connectedness that awareness systems bring among family members (IJsselsteijn et al., 2009; Markopoulos et al., 2004; Van Bel et al., 2008). Secondly, it has lead to an increased emphasis on methods for assessing the subjective quality and psychological consequences of product use. While a wealth of methods and techniques are available for assessing the usability of interactive products, research on methods for the subjective assessment of users’ experiences is only at its infancy (e.g. Van Schaik and Ling, 2007; Fenko et al., 2009; Ben-Bassat et al., 2006; Zimmerman et al., 2009; Väänänen-Vainio-Mattila et al., 2008).

7.1 Contributions of this Thesis

This thesis identified the notion of diversity in subjective judgments of the quality of interactive products. It identified two distinct schools of thought in the field of user experience and highlighted the methodological issues they pose when one is concerned about capturing the diversity in users’ responses to interactive products. More specifically, the focus of this thesis has been threefold:
• to conceptualize the notion of diversity in subjective judgments of users’ experiences with interactive products
• to establish empirical evidence for the prevalence of diversity, and
• to provide a number of methodological tools for the study of diversity in the context of product development.

7.1.1 Conceptualizing Diversity in User Experience

A long-standing assumption in the field of psychophysics is that different individuals will more or less agree on perceptual judgments such as how much noise, or blur, an image contains, or how much friction, or inertia, one may find in a haptic control. This thesis highlighted that this assumption, often referred to as the principle of homogeneity of perception (Martens, 2003), does not necessarily hold in the context of cognitive judgments of the quality of interactive products.

Using a framework from (Hassenzahl, 2005), we conceptualized diversity as existing at two different stages in the formation of an overall evaluative judgment about an interactive product. Perceptual diversity lies in the process of forming product quality perceptions (e.g. novel, easy to use) on the basis of product features. For instance, different individuals may infer different levels on a given quality of the same product, e.g. disagree on its novelty. Evaluative diversity lies in the process of forming overall evaluations of the product (e.g. good-bad) on the basis of product quality perceptions. For instance, different individuals may form different evaluative judgments even while having no disagreement on the perceived quality of the product, e.g. both might think of it as a novel and hard-to-use product, but they disagree on the relative importance of each quality.

7.1.2 Establishing empirical evidence for the prevalence of diversity in user experience

We identified two critical sources for diversity in the context of users’ experiences with interactive products: (a) interpersonal diversity in users’ responses to early conceptual designs, and (b) dynamics of users’ experiences over time.

Interpersonal diversity in users’ responses to early conceptual designs

Den Ouden (2006) highlighted that design decisions made early in the conceptual phase and particularly the ones relating to product definition are responsible for the majority of soft reliability problems. This leads to questioning the degree to which such design decisions are grounded on empirical insights about users’ preferences. Based on a number of informal communications with stakeholders in concept design
practices, we observed that it is often difficult to trace back the reasons that motivated certain design decisions. The questions raised then were on what basis are design decisions made? Can designers really foresee users’ preferences? Chapter 2 presented a study that inquired into the differences between designers’ and users’ views on a set of early conceptual designs. The study highlighted that designers’ views can be substantially different from the ones of designers. Even in a case where designers and users preferred the same product, users were found to base their preference on substantially different reasons, i.e. they valued different product qualities than the ones valued by the designers. It was further highlighted that both designers and users, in judging the quality of these conceptual designs, elicited a wealth of personal attributes that could not be captured by standardized questionnaires. Chapter 2 highlighted that diversity also exists within a single individual, in the sense that different attribute judgments of a participant may reveal different, complementary, views. This diversity in perspective cannot be represented in a single averaged view but requires multiple diverse views.

**Dynamics of users’ experiences over time**

Product evaluation practices have traditionally been focusing on early interactions and product qualities that are salient in these interactions. This thesis highlighted that diversity exists not only across different individuals but also across time, in the sense that different product qualities are valued at different phases in the adoption of the product. Chapter 4 presented two studies that inquired into the dynamics of users’ experiences with interactive products. In the first study, judgments of the overall *goodness* of a novel pointing device were found to shift from a use-based evaluation dominated by the pragmatic quality of the product, i.e. usefulness and ease-of-use, to an ownership-based evaluation dominated by aspects of identification, i.e. what the product expressed about their self-identify in social contexts. Judgments of *beauty* seemed to relate to perceptions of novelty during initial interactions, but this effect disappeared after four weeks of use. The second study followed six individuals through an actual purchase of an Apple iPhone and inquired into how their expectations and experiences developed one week before and four weeks after the purchase of the product. The study revealed that the product qualities that provided positive initial experiences were less crucial for motivating prolonged use. A tentative framework was proposed that identifies three distinct phases in the adoption of a product: *orientation, incorporation* and *identification*, each referring to a different relation between user and product, with different qualities being most salient.
7.1.3 Proposing methodological tools for the study of diversity

Lastly, we proposed a number of methodological tools for the study of the two different sources of diversity: (a) understanding interpersonal diversity through personal attribute judgments, and (b) understanding the dynamics of experience through experience narratives.

Understanding interpersonal diversity through personal attribute judgments

Traditional approaches to measuring users’ responses to artifacts derived through parallel design (Nielsen and Desurvire, 1993) lie in the use of validated psychometric scales where measures are being defined a-priori by the researchers. This thesis identified two limitations of such practices when one is concerned about inquiring into the diverse ways with which users form evaluative judgments of interactive products. First, it was argued that the a-priori definition of relevant measures is inherently limited as researchers might fail to consider a given dimension as relevant for the given product and context, or they might simply lack validated measurement scales for a relevant dimension. Secondly, such approaches assume that participants are able to interpret correctly and position a given statement, that is defined by the researcher, in their context. Recent literature has challenged this assumption, suggesting that in certain cases participants are unable to interpret the personal relevance of the statement in their own context; instead, they employ shallow processing (Sanford et al., 2006), i.e. responding to surface features of the language rather than attaching personal relevance to the question.

This thesis argued for an alternative approach that lies in a combination of structured interviewing, that aims at eliciting the attributes that are personally meaningful for each individual, with a subsequent rating process performed on the attributes that were elicited during the interview. The Repertory Grid Technique was proposed as a promising attribute elicitation technique as it aligns well with parallel design practices that are typically met in the concept design phase (Nielsen and Desurvire, 1993; Hassenzahl and Wessler, 2000).

Despite the promising nature of the Repertory Grid Technique, it was argued that current analysis procedures are not suited for interpreting such data. Two dominant approaches in the analysis of repertory grid data were identified: a) a qualitative, idiosyncratic approach with a narrative summarization, and b) a quantitative approach that employs averaging procedures using exploratory multivariate techniques. We argued that these two approaches are limited in two respects. Qualitative approaches do not take properly into account the elicited attribute ratings and thus do not fully exploit the true value of the Repertory Grid Technique, which is to quantify rich qualitative insights. Averaging procedures, on the other hand, treat diversity
among participants as error and thereby contradict the basic idea of The Repertory Grid and relevant personal attribute elicitation techniques.

Thus the lack of appropriate statistical analysis techniques was established and two Multi-Dimensional Scaling procedures that aim at quantifying the diversity of repertory grid data were proposed. The inadequacy of traditional averaging practices was established as in two case studies they were found to account for 15% and 43% of the available attributes only; the remaining attributes had minimal contribution to the configuration space and thus could not be adequately modeled by the average model. The technique proposed in chapter 3 was found to account for more than double of the attributes accounted for by the average model, to provide a better model fit even for the attributes that were adequately predicted by the average model, and to result in semantically richer insights, since the diverse models can account for more semantically different attributes.

Understanding the dynamics of experience through experience narratives

Traditional approaches to measuring the dynamics of experience over time typically employ validated measurement and structural models across different phases in the adoption of a system (e.g. Venkatesh and Davis 2000, Venkatesh and Johnson 2002, Kim and Malhotra 2005). This thesis highlighted a number of limitations of such approaches. First, they rely on an assumption that the relevant latent constructs remain constant, but their perceived value and relative dominance change over time. As it was shown in chapter 4, however, prolonged use might relate to a different set of user experiences such as daily rituals and personalization that do not become apparent in users’ initial interactions and that may thus be not captured by the measurement model. Reversely, other constructs that are salient in initial use situations may become irrelevant over prolonged use. This may lead to distorted data, for instance when participants are not able to interpret the personal relevance of a given measure (e.g. learnability) in the current usage situation. Last, such approaches provide rather limited insight into the exact reasons for changes in users’ experiences. They may, for instance, reveal a shift in the dominance of perceived ease-of-use and perceived usefulness on intention to use a product (e.g. Venkatesh and Davis 2000), but provide limited insight to the exact experiences that contributed to such changes. Insights into these rich and contextualized data is what designers need in designing for a given context.

This thesis argued for an alternative approach to the measurement of the dynamics of experience over time, that relies on a) the elicitation of idiosyncratic self-reports of one’s experiences with a product, the so-called experience narratives, and b) the extraction of generalized knowledge from a pool of experience narratives through
content analytical procedures. In this sense, each narrative may provide rich insights into a given experience and the context in which it takes place. However, generalized knowledge may also be gained from these experience narratives. Such generalized knowledge may be reflected in questions like: how frequent is a certain kind of experience, what is the ratio of positive versus negative experiences and how does this compare to competitive products, how does the dominance of different product qualities fluctuate over time and what should we improve to motivate prolonged use?

This leads to two research problems. Firstly, how can we elicit experience narratives efficiently? Chapter 5 reviewed existing methodological paradigms for inquiring into users’ experiences over time. Longitudinal designs such as the one employed in chapter 4 were identified as the gold standard in the study of long-term effects of product use, but their labor-intensive nature was highlighted as a barrier towards their adoption but also as an antecedent of restricted samples in terms of the user population, the product population and studied time. An alternative methodological approach was proposed that relies on the elicitation of user’ experiences with a product from memory. The chapter presented iScale, a tool that was designed with the aim of increasing participants’ effectiveness in recalling their experiences with a product. Two different versions of iScale, the Constructive and the Value-Account were motivated by two distinct theoretical approaches in the reconstruction of one’s emotional experiences. These two versions were tested in two separate studies against a control condition, free-recall employing no process of enhancing users’ recall. Overall, iScale was found to result in a) an increase in the number of experience reports that participants provided, b) an increase in the amount of contextual information for the reported experiences, and c) an increase in participants’ accuracy in recalling concrete details of the experienced events, thus suggesting that iScale is able to minimize retrospection biases when recalling one’s past experiences with a product.

Secondly, how can we aggregate the idiosyncratic experiences into generalized knowledge? Chapter 4 presented a case where content-analysis was employed in deriving key themes in the data, classifying narratives into a set of main categories and identifying the distinct distributions over time across the different categories. Two limitations were identified in this procedure. First, it is a labor intensive activity which becomes increasingly a concern when qualitative data may be elicited from a large amount of participants as in the case of iScale. Second, it is prone to researcher bias as humans often rely on heuristics in forming judgments about the relevance or similarity of two or more data instances (Kahneman et al., 1982). Chapter 6 proposed a novel technique for the semi-automated analysis of experience narratives that combines traditional qualitative coding procedures (Strauss and Corbin, 1998) with computational approaches for assessing the semantic similarity between documents (Salton et al., 1975). It was argued that the proposed approach supports
7.2 IMPLICATIONS FOR THE PRODUCT CREATION PROCESS

7.2.1 Integrating subjective and behavioral data

This thesis focused on methods for the subjective evaluation of interactive products. How can such methods be combined with objective information regarding users’ behavior? Behavioral data may not only provide insights into the actual usage of a product but may also inform the elicitation of users’ experience. In Funk et al. (2010)
we provided a case study where usage information was used to augment subjective feedback in two ways (figure 7.2). First, users were able to provide subjective feedback at the times they liked as in traditional event-based diaries (Bolger et al., 2003). Through process mining techniques (van der Aalst et al., 2007) we were able to inquire into users’ interactions that preceded the users’ reports. Second, through an observation specification language (Funk et al., 2008) we were able to identify interaction patterns that we were interested in and probe for subjective feedback as in the experience sampling method (Hektner et al., 2007). In this way, we were able to identify two distinct usage modes of a media recommender system, i.e. goal-oriented interaction as in the case of searching for specific information and action-oriented interaction as in the case of browsing through media content (Hassenzahl, 2005).

This approach enabled us in capturing momentary experiences that pertained to interactions that we were interested in. We observed however that experience sampling is often intrusive to users’ interactions and can lead to loss of information as participants often cannot interrupt the activity they are engaged in to provide subjective feedback regarding their experience (see also Hsieh et al., 2008). One could thus integrate such in-situ field feedback with methods for retrospective feedback as in the case of iScale, or the Day Reconstruction Method (Khan et al., 2008 for instance, see). These methods will be attempted in UXSuite (Koca and Funk, 2009), a start-up company supported by a STW valorization grant.

### 7.2.2 The end of specifications?

The starting point of this research project was the observation that about half of returned products are attributed to so-called soft reliability problems, i.e. cases where the user complaints despite the product meeting its specifications. The question raised then is, can quality models rely on specification compliance as a metric for the quality of a product? Brombacher et al. (2005) criticized traditional quality and reliability models in that they assume a) the existence of a full set of specifications that provide a complete description of the functional behavior of a product, and b) that these specifications provide complete coverage of all requirements of the prod-
uct. This thesis further supported this view by illustrating that user requirements may substantially vary across different individuals and that these requirements may develop over time as the users appropriate the product in their own context.

7.3 Avenues for future research

This thesis proposed a number of methods for quantifying the diversity in users’ experiences with interactive products. A number of limitations may however be noted, leading to interesting avenues for future research.

7.3.1 Leveraging insights across different exploratory studies

One of the advantages of the Repertory Grid Technique is that it provides rich insights into individuals’ idiosyncratic views on the set of stimuli. This thesis argued for exploratory quantitative analysis techniques that can adequately model the different views that one or more individuals have. A limitation of the Repertory Grid Technique however lies in its lack of scalability across different empirical studies. Each study yields a wealth of relations between stimuli, perceptions (i.e. personal constructs) and evaluations (e.g. preference judgments). Leveraging the insights of different empirical studies is cumbersome, if not impossible.

Structured interview techniques, however, such as triading [Kelly 1955], ladder-ing [Reynolds and Gutman 1988] and pyramiding [Fransella et al. 2003], that are typically employed in repertory grid interviews have the advantage of imposing a set of relations between the constructs that are being elicited, and are thus computational friendly. Empirical insights could thus be modeled through a graph \( G=(V,E) \) where the vertices \( V \) reflect the identified constructs and the edges \( E \) reflect the empirically identified relations between the constructs, and each edge, i.e. a relation between two constructs, could be weighted by its frequency of appearance in empirical studies. This might lead to a new set of tools that could have a substantial impact on science and design.

7.3.2 Computational tools for making survey research scalable

Psychometric scales can be characterized as one of the most ubiquitous measurement tools in the social sciences. However, they are not free of problems. The development of a set of scales is often described as a three-step process: item generation, scale development, and scale evaluation [Hinkin 1995]. The first step aims at enhancing the content validity of the questionnaire (i.e. that a complete coverage of the domain of interest is obtained through the proposed items); the latter two steps aim at enhancing the convergent and discriminant validity of the questionnaire (i.e. that each item
correlates highly with other items that attempt to measure the same latent construct, and weakly with items that attempt to measure different latent constructs).

While the later two phases are supported by a wealth of statistical techniques, the item generation phase (related to the content validity of the questionnaire) is still regarded as a largely subjective procedure (Scandura and Williams 2000; Hinkin 1995; Larsen et al. 2008b). Questionnaire items are typically generated through brainstorming with domain experts, or through empirical studies with participants from the targeted audience, most often involving structured interviewing techniques (c.f. Hinkin 1995). A number of limitations can be identified in item generation practices.

First, brainstorming practices require a firm understanding of the problem domain and clear definitions of the constructs to be measured. However, as Haynes et al. (1995) noted, questionnaires are often grounded in contemporaneous theories that evolve over time and that are supported by limited empirical data, thus, it is inevitable that early questionnaires fail to capture all possible facets of a construct and consecutive iterations are required for an adequate measurement of the construct.

Second, domain experts in brainstorming item generation practices often resort to lexical similarity (i.e. synonymy) when deriving new items in an effort to assure high convergent validity of the proposed constructs. This may have substantial implications to the rating process. Larsen et al. (2008b) found for the majority of constructs in a sample of questionnaires, the semantic similarity between items to be a significant predictor of participants’ ratings ($0.00 < R^2 < 0.63$). In such cases, participants are more likely to have employed shallow processing (Sanford et al. 2006), i.e. responding to surface features of the language rather than attaching personal relevance to the question.

Ideally, questionnaire items should be grounded on a large pool of empirical studies that are likely to have identified multiple facets of a given construct. Constructs could then evolve as new knowledge is being added to the corpus. Such a practice is however currently infeasible due to a lack of scalability across different empirical studies, as the exact procedure of item generation, the full-list of generated items, and their relationships are frequently not properly reported (Hinkin 1995; Scandura and Williams 2000). A computational infrastructure such as the one sketched above would enable social scientists in leveraging insights across different exploratory empirical studies which would in turn lead to scalable scale development practices within and across different research groups.

### 7.3.3 Empirical knowledge bases for forming design goals

Designers form beliefs. In the absence of empirical insights, designers base these beliefs on prior experiences and their own intuition. Empirical knowledge derived
from similar artifacts in similar contexts may not only make these beliefs grounded on actual evidence, but may also inspire the process of forming beliefs. Computational infrastructures such as the ones described above may provide rich insights into the relationships between design parameters, contextual parameters, and their psychological consequences (e.g. product quality perceptions).

7.3.4 A new basis for user insights?

The computational infrastructure proposed above assumes the use of structured interview techniques such triading, laddering and pyramiding; these share the benefit of imposing a set of relations between the constructs being elicited and thus are more "computational-friendly". The iScale tool lead to a new challenge rooted in providing access to immense, unstructured qualitative data. Chapter 6 proposed a computational approach for the content analysis of unstructured qualitative data such as in the case of experience narratives. One may however note that content analysis is limited in its scope as it looks only for one type of relation, that of similarity/dissimilarity between different data instances (e.g. narratives). The question then becomes: how can computational tools assists researchers in identifying a richer set of relationships such as causal effects in unstructured qualitative data? Can this procedure be partly automated, e.g. uncovering causal relationships through the relative position of concepts within the unit of analysis? If not, how can we assist the researcher in transforming qualitative data to a computational format (i.e. a graph network) that would enable him in rapid hypothesis testing, e.g. counterfactual analysis (King et al. 1994) within her own dataset as well as across different datasets?

Next, given the access to computational tools that partly automate the analysis of rich qualitative data and support the identification of emerging patterns, we foresee a new basis for user insights, one derived from internet forums, complaint centers, emails and other sources of public statements. These may lead to user insights relating to product success in the market but also to emerging trends and unidentified needs in the marketplace. This thesis, through the development of the iScale tool and the Computational Content Analysis method, made a first step in inquiring into richer sources of user insight data.
Part V

Bibliography
Bibliography


Bartlett, F. (1932) *Remembering*. Cambridge, MA.


ISO (1996) ISO 9241 - *Ergonomic requirements for office work with visual display terminals (VDTs) - Part 11: Guidance on usability*.


Likert, R. (1932) A technique for the measurement of attitudes. No. 140 in Archives of Psychology.


BIBLIOGRAPHY


BIBLIOGRAPHY


Zimmerman, J. (2009) Designing for the self: making products that help people become the person they desire to be. In: CHI ’09:


A.1 Descriptions to participants - Day Reconstruction & Experience Narration

A.1.1 Day Reconstruction

Hereby, we would like to ask you to write a diary describing your daily activities that had something to do with an iPhone, like sending an SMS or showing your iPhone to your friends. Think of those activities as a continuous series of scenes or episodes in a film. Give each episode a brief name which clearly describes the activity that

Figure A.1: A tentative framework of the design process proposed by Martens (2009b, personal communication).
you were engaged in. Exemplary activities could range from: making a phone call, to playing around with iPhone features, to looking for information on the internet relating to the iPhone, to talking about it with friends, etc. What matters is that you provide an as accurate as possible picture of types of activities that you were engaged in and an estimation of the time you spent on each activity. When you had several adjacent and/or similar activities (e.g. making several phone calls) you might want to group them as one activity and estimate the total time you spent on that (e.g. calling my friends 1 hour might mean that you have made several phone calls that day and the estimated time of those phone calls was 1 hour).

### A.1.2 Experience Narration

We ask you to consider all your experiences of today that somehow relate to iPhone. Think back about those that were most essential and important for you. Now think of the three experiences that were for you personally most satisfying or unsatisfying experiences of today. Please, use your own feeling or a definition of what “satisfying” and “unsatisfying experience” means. Take a couple of minutes to be sure to come up with three most crucial experiences; you may also want to write them down for yourself. We want you to be open as to which experiences to report. Bear in mind our aim is NOT to evaluate either iPhone or you. Our goal is to understand what kinds of experiences you have when using iPhone during this month and how these experiences change over time. Example experiences could range from a frustrating episode while trying to find certain functionality, to positive experiential encounters when you discover exciting features, to moments of pride when you show your new product to your friends etc. What matters is that the experience is truly personal and impactful.

For each of those three experiences, we would like you to capture an as rich picture as possible and report it to us. We want you to tell a story that describes in great detail the situation that occurred (i.e. what happened), what were your impressions regarding that situation (i.e. how did you feel about it), what were your motivations to act in a certain way in that situa-

---

**Table A.1: An example reconstructed day**

<table>
<thead>
<tr>
<th>Today’s activities with iPhone</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1. phoning</td>
<td>3 mins</td>
</tr>
<tr>
<td>2. checking web</td>
<td>15 mins</td>
</tr>
<tr>
<td>3. installing games</td>
<td>5 mins</td>
</tr>
<tr>
<td>4. playing games</td>
<td>90 mins</td>
</tr>
<tr>
<td>5. editing phone book</td>
<td>2 mins</td>
</tr>
<tr>
<td>6. checking mail</td>
<td>5 mins</td>
</tr>
</tbody>
</table>

---
A.2. PARTICIPANTS’ SKETCHES IN ISCALE ACROSS THE TWO SESSIONS

...tion (i.e. why did you do what you did) and also how would you reflect on what has happened (i.e. what was your interpretation of that situation).

Example experience narrative

I needed to call fedex services to send a package. The day before I said - it would be so cool if I checked the phone number with iPhone and would be immediately able to clock on that number and call the required number. Yeah, but how possible would it be... It was! In the morning I’ve browsed the web to find the number and saw it to be a hyperlink. Once I’ve pressed on it, iPhone immediately started to dial the number. It felt exactly like I would have dreamed it. Really nice feeling - to have the phone do what you hoped for but not really believed it to be possible...

A.2 Participants’ sketches in iScale across the two sessions

Participants’ sketches using iScale, across the two sessions, in the Constructive and in the Value-Account condition. Time (perceived) reflects the x-axis of the iScale tool. The area between the two sketches, the one elicited in the first session and the second in the second session, may be used as an index of (lack of) test-retest reliability of participants recall process.
Figure A.2: Participants’ sketches resulting from the Constructive iScale tool. Part A.
Figure A.3: Participants' sketches resulting from the Constructive iScale tool. Part B.
Figure A.4: Participants’ sketches resulting from the Constructive iScale tool. Part C.
Figure A.5: Participants’ sketches resulting from the Constructive iScale tool. Part D.
<table>
<thead>
<tr>
<th>Time</th>
<th>Perceived value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.00</td>
<td>0.50</td>
</tr>
<tr>
<td>0.80</td>
<td>0.00</td>
</tr>
<tr>
<td>0.60</td>
<td>-0.50</td>
</tr>
<tr>
<td>0.40</td>
<td>-1.00</td>
</tr>
<tr>
<td>0.20</td>
<td>-1.50</td>
</tr>
<tr>
<td>0.00</td>
<td>-2.00</td>
</tr>
</tbody>
</table>

Session 2

Session 1

Session

Figure A.6: Participants’ sketches resulting from the Value-Account iScale tool. Part A.
A.2. PARTICIPANTS’ SKETCHES IN iSCALE ACROSS THE TWO SESSIONS

Figure A.7: Participants’ sketches resulting from the Value-Account iScale tool. Part B.
Figure A.8: Participants’ sketches resulting from the Value-Account iScale tool. Part C.
A.2. PARTICIPANTS’ SKETCHES IN iSCALE ACROSS THE TWO SESSIONS

Figure A.9: Participants’ sketches resulting from the Value-Account iScale tool. Part D.
# List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>A modification of Hassenzahl’s 2005 framework, highlighting diversity at two stages in forming overall judgments about products.</td>
<td>20</td>
</tr>
<tr>
<td>1.2</td>
<td>Four different sources of diversity in user experience. These are assumed to modulate the dominance of different product qualities on users’ experience and overall evaluative judgments.</td>
<td>22</td>
</tr>
<tr>
<td>1.3</td>
<td>A simplified framework of the design process, derived from Martens’ 2009b personal communication) depicted in Appendix A, figure A.1. According to this framework, the designer forms beliefs about the potential consequences of use of a given product in a given context, grounded on empirical knowledge. These beliefs are in turn externalized to artifacts and the artifacts are validated on a number of criteria that have been derived from prior beliefs, or empirical insights from the previous design iteration.</td>
<td>23</td>
</tr>
<tr>
<td>2.1</td>
<td>Multi-Dimensional Scaling on dissimilarity ratings for three products with latent constructs fitted as vectors. Note that the latent constructs employed in this study fail to explain the dissimilarity between products 1 and 3.</td>
<td>34</td>
</tr>
<tr>
<td>2.2</td>
<td>Consistency between positive and negative poles for negation and opposition attributes. Negative scores were inverted.</td>
<td>39</td>
</tr>
<tr>
<td>2.3</td>
<td>A two-dimensional MDS configuration of three websites using ratings from two attributes. Website $j$ is perceived as more legible and colorful than websites $i$ and $k$.</td>
<td>41</td>
</tr>
<tr>
<td>2.4</td>
<td>An example storyboard describing one of the concepts used in the study.</td>
<td>43</td>
</tr>
<tr>
<td>2.5</td>
<td>User segmentation map based on correlations between dissimilarity scores.</td>
<td>45</td>
</tr>
<tr>
<td>2.6</td>
<td>Perceptual maps for groups 1 to 5. The light vectors represent individual preferences, while the other vectors represent the most significant attributes.</td>
<td>49</td>
</tr>
<tr>
<td>Figure</td>
<td>Description</td>
<td>Page</td>
</tr>
<tr>
<td>--------</td>
<td>-------------</td>
<td>------</td>
</tr>
<tr>
<td>2.7</td>
<td>Attributes (a) positively and (b) negatively ranked when TouchTo-Print is preferred (along with 95% exact confidence intervals)</td>
<td>50</td>
</tr>
<tr>
<td>2.8</td>
<td>Importance rankings for reliability attributes for designers and users. Only 2 designers reported reliability concerns in contrast to 7 users. Five users ranked reliability attributes within their three most important attributes.</td>
<td>51</td>
</tr>
<tr>
<td>2.9</td>
<td>Two dimensional visualization of the dissimilarity of the individual views derived from the reanalysis of the data using the procedure proposed in chapter 3. Views connected through lines belong to the same cluster. Abbreviations used: Doc=Documentation Expert, CD=Concept Developer, UE=Usability Expert, PM=Product Manager, SOFTW=Software Expert, DES=Visual Interface Designer, MKT=Market Expert, U=User</td>
<td>54</td>
</tr>
<tr>
<td>3.1</td>
<td>The eight stimuli used in the study</td>
<td>59</td>
</tr>
<tr>
<td>3.2</td>
<td>Two diverse views for one participant</td>
<td>63</td>
</tr>
<tr>
<td>3.3</td>
<td>A 2-D perspective of the 3-D visualization of distances between individual's different views</td>
<td>66</td>
</tr>
<tr>
<td>3.4</td>
<td>The average model and the three diverse models corresponding to clusters 1 to 3 in figure 3.3</td>
<td>69</td>
</tr>
<tr>
<td>3.5</td>
<td>(left) Scatter plot of the accounted variance by the three diverse models of the heuristic technique versus the variance in the original attribute ratings, and (right) scatter plot of the range versus the estimated noise in predicted attribute ratings</td>
<td>71</td>
</tr>
<tr>
<td>3.6</td>
<td>$R^2$ (top) and $R_k$ (bottom) values for the 38 attributes that were adequately modeled by the three models of the heuristic technique proposed in this chapter (left) and the remaining attributes (right), for the heuristic technique and the algorithmic technique proposed by Martens (2009a)</td>
<td>72</td>
</tr>
<tr>
<td>4.1</td>
<td>Hassenzahl’s (2004) framework distinguishing between two overall evaluative judgments, i.e. goodness and beauty, and three quality perceptions, i.e. pragmatic quality, hedonic quality - stimulation, and hedonic quality - stimulation.</td>
<td>80</td>
</tr>
<tr>
<td>4.2</td>
<td>A novel pointing device for Interactive TV context, used in Study 1</td>
<td>82</td>
</tr>
<tr>
<td>4.3</td>
<td>2D view on the 3-dimensional visualization of distances between quality attributes, beauty and goodness. Users' perceptions during the 1st week of use (left) and after 4 weeks of use (right)</td>
<td>84</td>
</tr>
<tr>
<td>4.4</td>
<td>Interface used for rating the product after reporting an experience</td>
<td>89</td>
</tr>
</tbody>
</table>
4.5 Temporality of experience, consisting of three main forces, an increasing familiarity, functional dependency and emotional attachment, all responsible for shifting users’ experience across three phases: orientation, incorporation and identification. In each phase, different product qualities are appreciated. 92

4.6 Number of reported satisfying & dissatisfying experiential episodes over the four weeks relating to the three phases of adoption. 95

5.1 (a) Constructive iScale, (b) Value-Account iScale. 115

5.2 (a) interface for reporting experiences, and (b) overall questions asked in the beginning of the survey. 116

5.3 Examples of free-hand sketching. 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 sketch (C: Constant, L: Linear, NL: Non-Linear, D: Discontinuous). 121

5.4 Average number of experience reports elicited (a) in the presence or absence of sketching in the constructive and in the Value Account modes of reconstruction, and (b) for the two respective product qualities in relation to mode of reconstruction. 131

5.5 Relation between actual time (reported time for a sketched node) and position of node along the iScale’s x-axis: (a) linear relation, (b) power-law relation. Actual time (days) is adjusted to the full time of ownership of the product for each participant. 134

5.6 Temporal inconsistency in the sketching and the no-sketching condition for the two respective qualities of reporting. 137

5.7 Example sketches elicited in the constructive (top) and the value-account (bottom) conditions during the two sessions. 138

5.8 Number of narratives in a certain period relating to learnability, stimulation, long-term usability and usefulness as a function of time of ownership. 140

5.9 Averaged across participants sketches for ease-of-use and innovativeness. 143

6.1 Multi-Dimensional Scaling on concept dissimilarity (eq. 6.7). Concepts connected through lines belong to the same cluster. A-priori defined concepts are depicted in italics, while superordinate concepts are denoted in bold. 156
6.2 Interactive scatter plot depicting the similarity between narratives derived from the co-occurrence of coded concepts and processed by means of Multi-Dimensional Scaling. The researcher may browse through different data instances, or identify the least coded narratives. By selecting a narrative the full text appears with the coded part highlighted and annotated by the respective concept. The analyst can further code the data, which will in turn influence the visualizations. 157

6.3 Interactive scatter plot depicting the number of codes within each narrative. The researcher may browse through different data instances, or identify the least coded narratives. By selecting a narrative the full text appears with the coded part highlighted and annotated by the respective concept. The analyst can further code the data, which will in turn influence the visualizations. 158

6.4 Four possible states of a given concept appearing in none, one, or both narratives. 159

6.5 Visualizing frequency of experiences (y axis) and average impact (size of circle) for four types of experiences colored differently over the four weeks of the study. 162

6.6 Visualizing relations between product qualities and product features. Each column depicts the distribution of experiences over the five product qualities for a given product feature. 162

6.7 Histogram on the length (number of words) of experience narratives. 5% of the narratives, i.e. the ones containing less than 24 words, were excluded from further analysis. 165

6.8 Percentage of narratives being correctly classified, i.e. assigned to the same category as in the hand-coded analysis of the initial study Karapanos et al. (2009), both for the optimal approach and for the approach that employs human raters in assigning the 9 clusters to five categories. The 95% confidence intervals depict the uncertainty derived from the human classification of the nine clusters into the five overall categories. 167
7.1 A simplified framework of the design process, derived from (Martens, 2009b, personal communication) depicted in Appendix A, figure A.1. According to this framework, the designer forms beliefs about the potential consequences of use of a given product in a given context, grounded on empirical knowledge. These beliefs are in turn externalized to artifacts and the artifacts are validated on a number of criteria that have been derived from prior beliefs, or empirical insights from the previous design iteration. 179

7.2 Information about preceding user actions can (a) provide the context for user-initiated feedback, and (b) trigger event-based experience sampling. 180

A.1 A tentative framework of the design process proposed by (Martens, 2009b, personal communication). 203
A.2 Participants’ sketches resulting from the Constructive iScale tool. Part A. 206
A.3 Participants’ sketches resulting from the Constructive iScale tool. Part B. 207
A.4 Participants’ sketches resulting from the Constructive iScale tool. Part C. 208
A.5 Participants’ sketches resulting from the Constructive iScale tool. Part D. 209
A.6 Participants’ sketches resulting from the Value-Account iScale tool. Part A. 210
A.7 Participants’ sketches resulting from the Value-Account iScale tool. Part B. 211
A.8 Participants’ sketches resulting from the Value-Account iScale tool. Part C. 212
A.9 Participants’ sketches resulting from the Value-Account iScale tool. Part D. 213
List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Percentages of attributes types from Karapanos and Martens (2007) (Study 1) and Hassenzahl and Wessler (2000) (Study 2) studies.</td>
<td>37</td>
</tr>
<tr>
<td>2.2</td>
<td>Amount of variance accounted for by the latent one-dimensional construct. Original attributes were in Dutch.</td>
<td>39</td>
</tr>
<tr>
<td>2.3</td>
<td>Classification of attributes elicited in the study into 16 categories and three broad themes</td>
<td>46</td>
</tr>
<tr>
<td>2.4</td>
<td>Number (and percentage) of attributes being adequately modeled by the resulting views of the averaging analysis and the two proposed procedures.</td>
<td>52</td>
</tr>
<tr>
<td>2.5</td>
<td>Number of attributes that are adequately modeled by the resulting one or two views for each participant.</td>
<td>53</td>
</tr>
<tr>
<td>3.1</td>
<td>Attribute categories and examples</td>
<td>58</td>
</tr>
<tr>
<td>3.2</td>
<td>Goodness of fit Criteria. Attributes that are adequately predicted are employed in model 1. A second model is attempted only on attributes that display the least fit, to ensure diversity between the two models.</td>
<td>61</td>
</tr>
<tr>
<td>3.3</td>
<td>Goodness of fit statistics for the two diverse models of participant one. Attributes (2,4,6,8,12,13) were adequately predicted (R^2&gt;.5 &amp; R_k&gt;6) by model 1. Attributes (1,5,7,9,10,11) displayed the least fit (R_k&lt;4) and were used to derive a second model. Attributes (5,7,9,10) were adequately predicted by model 2.</td>
<td>62</td>
</tr>
<tr>
<td>3.4</td>
<td>Number of attributes explained by the two views for the ten participants of the study.</td>
<td>65</td>
</tr>
<tr>
<td>3.5</td>
<td>Goodness of fit of the average and the three diverse models for the 38 in total attributes resulting from models 1 to 3</td>
<td>68</td>
</tr>
<tr>
<td>4.1</td>
<td>Bipolar attributes measuring pragmatic quality, stimulation and identification (Hassenzahl, 2004)</td>
<td>83</td>
</tr>
<tr>
<td>4.2</td>
<td>Latent constructs measured through a single scale</td>
<td>90</td>
</tr>
<tr>
<td>4.3</td>
<td>The coding scheme of the analysis</td>
<td>91</td>
</tr>
</tbody>
</table>
4.4 Multiple Regression analysis with usefulness, ease-of-use, stimulation and identification as predictors and Goodness as predicted ($\beta$ values and significances * $p<.05$, ** $p<.001$) for both satisfying and dissatisfying experiences. 98

4.5 Multiple Regression analysis with usefulness, ease-of-use, stimulation and identification as predictors and Beauty as predicted ($\beta$ values and significances * $p<.05$, ** $p<.001$) for both satisfying and dissatisfying experiences. 98

5.1 The three product qualities that participants reported on, along with definitions and word items. 118

5.2 Relationship between sketching and reporting in free-hand sketching. Types of sketching: C=Constant, L=Linear, NL=Non-Linear and D=Discontinuous. 122

5.3 Design qualities elicited in the interview sessions. Number of participants using a given quality to differentiate the three sketching techniques, and number of participants mentioning the given quality as affecting the sketching or recalling process. 124

5.4 Experimental design: manipulated variables (rows) and process (columns). Each participant joined two sessions, each consisting of two tasks. Participants in group A, for instance, which belong in the Constructive condition, in the first session, they used the sketching tool to report on ease-of-use followed by the non-sketching tool to report on innovativeness. In the second session the order was counterbalanced but the two coupling of the two variables, sketching and quality, remained the same. 129

5.5 Number of experience reports judged for the three dimensions of richness and for the four levels of design relevance, for the four conditions resulting from presence or absence of sketching and constructive or value-account mode of reconstruction. 133

6.1 Domain-specific knowledge employed for the analysis of the 329 narratives. The domain-specific concepts were derived from Hassenzahl (2004) and Lavie and Tractinsky (2004b) psychometric scales. 154

6.2 Concepts that were derived from the analysis of the 329 narratives along with examples of coded data. Next to existing domain-specific knowledge, the researcher annotates data by relating individual terms or phrases to a set of latent concepts (linked concepts are denoted in brackets). 155
<table>
<thead>
<tr>
<th>Table</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.3</td>
<td>Three most dominant concepts in each cluster</td>
<td>160</td>
</tr>
<tr>
<td>6.4</td>
<td>Concepts relating to product features derived from the analysis of the 329 narratives along with examples of coded data. Inclusion of subordinate concepts is denoted through brackets.</td>
<td>163</td>
</tr>
<tr>
<td>6.5</td>
<td>Confusion Matrix for traditional LSA</td>
<td>168</td>
</tr>
<tr>
<td>6.6</td>
<td>Confusion Matrix for LSA on restricted terms</td>
<td>168</td>
</tr>
<tr>
<td>6.7</td>
<td>Confusion Matrix for Concept Analysis (proposed procedure)</td>
<td>168</td>
</tr>
<tr>
<td>A.1</td>
<td>An example reconstructed day</td>
<td>204</td>
</tr>
</tbody>
</table>
Summary

Quantifying Diversity in User Experience

Evaluation should be integral to any design activity. Evaluation in innovative product development practices however is highly complicated. It often needs to be applied to immature prototypes, while at the same time users’ responses may greatly vary across different individuals and situations.

This thesis has focused on methods and tools for inquiring into users’ experiences with interactive products. More specifically, it had three objectives: a) to conceptualize the notion of diversity in subjective judgments of users’ experiences with interactive products, b) to establish empirical evidence for the prevalence of diversity, and c) to provide a number of methodological tools for the study of diversity in the context of product development.

Two critical sources of diversity in the context of users’ experiences with interactive products were identified and respective methodological solutions were proposed: a) understanding interpersonal diversity through personal attribute judgments, and b) understanding the dynamics of experience through experience narrati ves.

Personal Attribute Judgments, and in particular, the Repertory Grid Technique, is proposed as an alternative to standardized psychometric scales, in measuring users’ responses to artifacts in the context of parallel design. It is argued that traditional approaches that rely on the a-priori definition of the measures by the researchers have at least two limitations. First, such approaches are inherently limited as researchers might fail to consider a given dimension as relevant for the given product and context, or they might simply lack validated measurement scales for a relevant dimension. Secondly, such approaches assume that participants are able to interpret and position a given statement that is defined by the researcher to their own context. Recent literature has challenged this assumption, suggesting that in certain cases participants are unable to interpret the personal relevance of the statement in their own context, and might instead employ shallow processing, that is respond to surface features of the language rather than attaching personal relevance to the ques-
tion. In contrast, personal attributes are elicited from each individual respondent, instead of being a-priori imposed by the experimenter, and thus are supposed to be highly relevant to the individual. However, personal attributes require substantially more complex quantitative analysis procedures. It is illustrated that traditional analysis procedures fail to bring out the richness of the personal attribute judgments and two new Multi-Dimensional Scaling procedures that extract multiple complementary views from such datasets are proposed.

An alternative approach for the measurement of the dynamics of experience over time is proposed that relies on a) the retrospective elicitation of idiosyncratic self-reports of one’s experiences with a product, the so-called experience narratives, and b) the extraction of generalized knowledge from these narratives through computational content analysis techniques. iScale, a tool that aims at increasing users’ accuracy and effectiveness in recalling their experiences with a product is proposed. iScale uses sketching in imposing a structured process in the reconstruction of one’s experiences from memory. Two different versions of iScale, each grounded in a distinct theory of how people reconstruct emotional experiences from memory, were developed and empirically tested. A computational approach for the extraction of generalized knowledge from experience narratives, that combines traditional coding procedures with computational approaches for assessing the semantic similarity between documents, is proposed and compared with traditional content analysis.

Through these two methodological contributions, this thesis argues against averaging in the subjective evaluation of interactive products. It proposes the development of interactive tools that can assist designers in moving across multiple levels of abstractions of empirical data, as design-relevant knowledge might be found on all these levels.
Acknowledgements

Four years seem like an enormous time in the beginning, somehow they seem too little throughout the process, but retrospectively, they seem to have been full of learning, doing, and meeting wonderful people. I am grateful to a number of people that contributed to this period of my life.

First, and foremost, the people who helped me grow, my advisors, Jean-Bernard Martens and Aarnout Brombacher. Jean-Bernard, doing a PhD under your supervision is above all a learning trip. Your knowledge in statistics and your rigor in the scientific progress contributed most to my development in these four years. I find no way to appropriately thank you for that. I will also not forget your willingness to challenge my thoughts in our recent discussions that helped me define my new line of research. Aarnout, you have challenged my view on the need for separating research from practice. I warmly thank you for this and for providing the infrastructure to perform industrial research.

Next, the people who inspired me most, Marc Hassenzahl, John Zimmerman, and Jodi Forlizzi. Marc, your work is what attracted me to the field of User Experience and what kept me motivated throughout the process. You have an amazing ability to sustain rigor while meeting practical relevance. Thank you for giving me the option to work with and learn from you. Besides being a greater researcher, you are a great person. John and Jodi, I only realized how different you are when I came to Pittsburgh. I also realized how amazingly you work together. Theorizing is always much trickier than what it initially seems. I am glad to have done this work close to you. John, you have a never-ending will to question things. Thank you for helping see further and for introducing me to the complexities of Silverstone’s framework. Jodi, you have an amazing sense of where to stop a qualitative analysis process or where to look for. Your help was critical for not getting lost in the qualitative data analysis. I would like to thank you both for welcoming me to Pittsburgh and for helping me on this project. Besides that, your work on designing for the self and user experience frameworks have largely influenced my thoughts and goals in my research. Marc and John, I am grateful and honored to have you here today.

My thesis committee, Prof. Paul Hekkert, Assoc. Prof. Wijnand IJsselsteijn, Prof.
Berry Eggen and Assoc. Prof. Panos Markopoulos, for reading my thesis and helping
improving its quality through their constructive feedback.

Several parts of this work have been conducted in collaboration with industrial
partners. Some people proved to be indispensable in this process. Especially, Fred de
Jong from Océ and Mariëlle Meuffels and Nico van den Gaarden from Philips.

The Soft Reliability team, Anne Rozinat, Aylin Koca, and Mathias Funk. You guys
rock! Next to working closely together, you are great friends. I really enjoyed work-
ing and spending time with you.

My closest colleagues and friends, Aga Matysiak-Szóstek, Wouter Sluis, Maurits
Kaptein, Javed Khan, and Georgios Metaxas. Aga, you are the most proactive and
fast-paced people I have ever worked with. Somehow, you have an ability to never
get stuck! Besides that, you are always "wearing" a smile and discussions are never
boring with you. I really enjoyed doing things together over these years and I hope
we can continue working together in the future. Wouter, you were one of the most
visible components of our office. Not only due to your height but also due to your
personality. I enjoyed sharing the same desk "island", I admire your positive stance
in life, and I wish you personal and professional happiness. Maurits, thank you for
our smoking breaks (although I - almost - never smoked; that also says something
about how much I enjoyed discussing with you). You are a great person and a great
researcher. I look forward to seeing how your work will shape over time. Javed and
Georgie, even here I prefer to address you as a team. Throughout these years, we
have shared lots of discussions and time together. Thank you for being here with me
today.

The User-Centered Engineering group and my colleagues at 2.44 who provided a
enjoyable and stimulating environment. Especially, Abdullah, Alex, Berke, Evelien,
Jeroen, Joep, Iris, Karin, Natalia, Nora, Kostas, Oscar, Saskia and Zeljko.

I would like to thank a number of people that influenced my education path before
I start working on this project. Prof. Athanassios Skodras for guiding and motivating
me throughout my undergraduate studies, Prof. Nikolaos Avouris and Dr. Nikolaos
Tselios for introducing me to the field of Human-Computer Interaction and provid-
ing endless support through my undergraduate dissertation and Prof. Ann Bland-
ford for being there for me during my MSc thesis despite the London-Eindhoven
distance while being at Philips Research.

Last, but most important, my family. My parents Thomas and Kostoula. You have
provided endless support throughout all my studies. I know when you funded my
MSc studies in the UK it wasn’t in your plans to stay abroad for that long, but things
don’t always turn in life as we plan them :). I thank you for supporting all my choices
irrespective of whether you agreed with them or not. The older I get the more I see
you in me and I thank you for that. My big brother, Kostis. You taught me how to
program on GW-Basic when we were kids, you were also the one to make me aware of the HCI field. As far as I can remember, I have been trying to reach your level of knowledge, maturity, and ethics. Thank you for motivating me to pursue things in life and for looking after me. My partner in life, Olga. You have been close to me for all the ups and downs in this process. From giving the first feedback to all my ideas, to proofreading my papers, and above, for being there for me no matter what. I know you since we were kids and ever since we were together, our needs, wishes and goals have always been growing together. I cannot express by words what you mean to me. I would like to devote this book to you.
Biography

Evangelos Karapanos was born on the 8th of April 1981 in Athens, Greece. He obtained a BSc degree in Physics with a specialization in electronics, computing and signal processing. During his studies he worked on a part time basis as a web programmer which brought him to the field of Human-Computer Interaction (HCI). He conducted his BSc dissertation at the HCI group of the department of Electrical and Computer Engineering of the same university, on model-based design and evaluation of walk-up-and-use interfaces. He subsequently pursued a MSc in HCI with Ergonomics at University College London. His MSc dissertation, which was part of an internship at Philips Research Eindhoven, focused on the user acceptance of nomadic user interfaces. On November 15th 2005 he joined the TU/e project Soft Reliability as a PhD candidate. He has been a visiting researcher at Philips Consumer Lifestyle from 2006 to 2008 and a visiting scholar of the Human-Computer Interaction Institute at Carnegie Mellon University in 2008.
Publications related to this research

Journals

Proceedings


Informally reviewed