

CHAPTER 15

Usability Evaluation

by Gilbert Cockton.

Put simply, usability evaluation assesses the extent to which an interactive system is easy and pleasant to use. Things aren't this simple at all though, but let's start by considering the following propositions about usability evaluation:

1. Usability is an inherent measurable property of all interactive digital technologies
2. Human-Computer Interaction researchers and Interaction Design professionals have developed evaluation methods that determine whether or not an interactive system or device is usable.
3. Where a system or device is usable, usability evaluation methods also

determine the extent of its usability, through the use of robust, objective and reliable metrics

4. Evaluation methods and metrics are thoroughly documented in the Human-Computer Interaction research and practitioner literature. People wishing to develop expertise in usability measurement and evaluation can read about these methods, learn how to apply them, and become proficient in determining whether or not an interactive system or device is usable, and if so, to what extent.

The above propositions represent an ideal. We need to understand where current research and practice fall short of this ideal, and to what extent. Where there are still gaps between ideals and realities, we need to understand how methods and metrics can be improved to close this gap. As with any intellectual endeavour, we should proceed with an open mind, and acknowledge that not only are some or all of the above propositions not true, but that they can never be so. We may have to close some doors here, but in doing so, we will be better equipped to open new ones, and even go through them.

15.1 FROM FIRST WORLD OPPRESSION TO THIRD WORLD EMPOWERMENT

Usability has been a fundamental concept for Interaction Design research and practice, since the dawn of Human-Computer Interaction (HCI) as an inter-disciplinary endeavour. For some, it was and remains HCI's core concept. For others, it remains important, but only as one of several key concerns for interaction design.

It would be good to start with a definition of usability, but we are in contested territory here. Definitions will be presented in relation to specific positions on usability. You must choose a one that fits your design philosophy. Three alternative definitions are offered below.

It would also be good to describe how usability is evaluated, but alternative understandings of usability result in different practices. Professional practice is very varied, and much does not generalise from one project to the next. Evaluators must choose how to evaluate. Evaluations have to be designed, and designing requires making choices.

15.1.1 The Origins of HCI and Usability

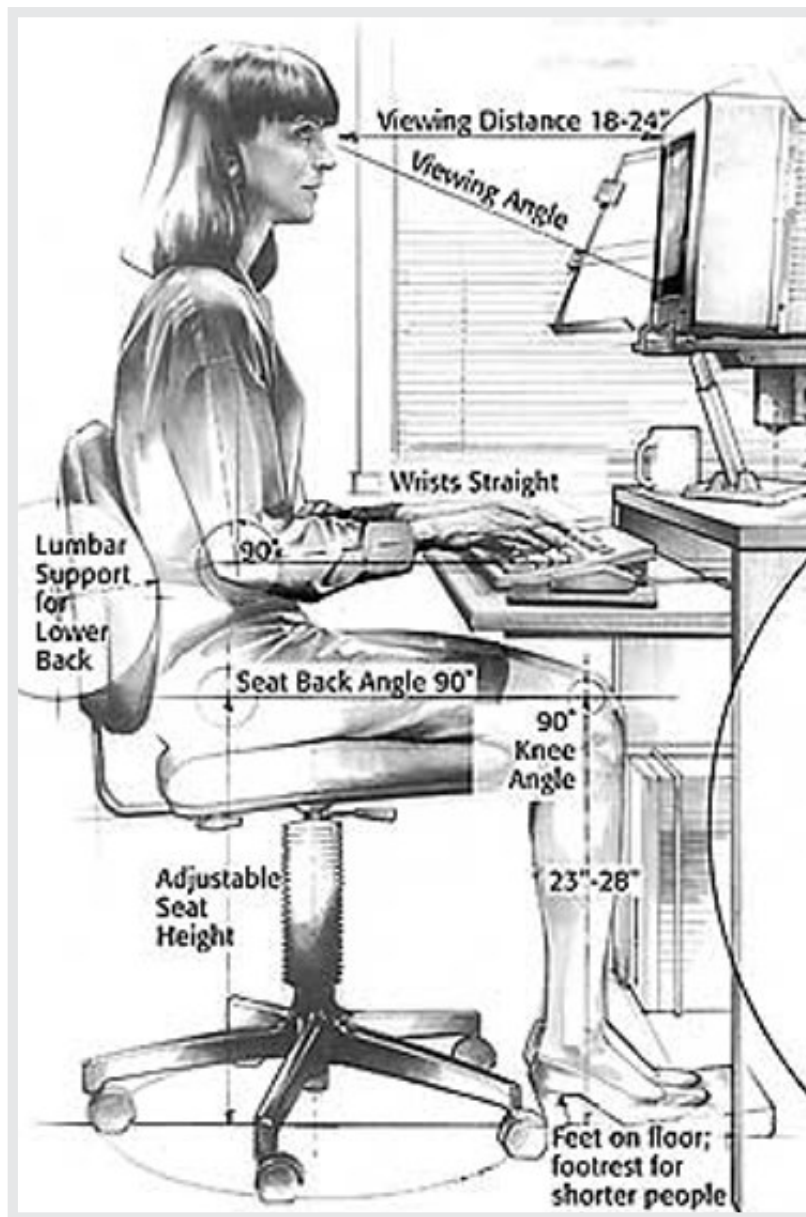
HCI and usability have their origins in the falling prices of computers in the 1980s, when for the first time, it was feasible for many employees to have their own *personal computer* (a.k.a PC). For their first three decades of computing, almost all users were highly trained specialists of expensive centralised equipment. A trend towards less well trained users began in the 1960s with the introduction of timesharing and minicomputers. With the use of PCs in the 1980s, computer users increasingly had no, or only basic, training on operating systems and applications software. However, software design practices continued to implicitly assume knowledgeable and competent users, who would be familiar with technical vocabularies and systems architectures, and also possess an aptitude for solving problems arising from computer usage. Such implicit assumptions rapidly became unacceptable. For the typical user, interactive computing became associated with constant frustrations and consequent anxieties. Computers were obviously too hard to use for most users, and often absolutely unusable. *Usability* thus became a key goal for the design of any interactive software that would not be used by trained technical computer specialists. Popular terms such as “user-friendly” entered everyday use. Both usability and user-friendliness were initially understood to be a property of interactive software. Software either was usable or not. Unusable software could be made usable through re-design.



Courtesy of Boffy b. Copyright: CC-Att-SA-3 (Creative Commons Attribution-ShareAlike 3.0).



Courtesy of Jeremy Banks. Copyright: CC-Att-2 (Creative Commons Attribution 2.0 Unported).



Courtesy of Berkeley Lab. Copyright: pd (Public Domain (information that is common property and contains no original authorship)).

FIGURE 15.1 A-B-C: The Home Personal Computer (PC) and Associated Peripherals is Now an Everyday Sight in Homes Worldwide. Usability became a critical issue with PC's introduction.

15.1.2 From Usability to User Experience via Quality in Use

During the 1990s, more sophisticated understandings of usability shifted from an all-or-nothing binary property to a continuum spanning different extents of usability. At the same time, the focus of HCI shifted to *contexts of use* (Cockton 2004). Usability ceased to be HCI's dominant concept, with research increasingly focused on the fit between interactive software and its surrounding usage contexts. Quality in use no longer appeared to be a simple issue of how inherently usable an interactive system was, but how well it fitted its context of use. *Quality in use* became a preferred alternative term to *usability* in international standards work, since it avoided implications of usability being an absolute context-free invariant property of an interactive system. Around the turn of the century, the rise of networked digital media (e.g., web, mobile, interactive TV, public installations) added novel emotional concerns for HCI, giving rise to yet another more attractive term than usability: *user experience*.

Current understandings of usability are thus different from those from the early days of HCI in the 1980s. Since then, ease of use has improved though both attention to interaction design and improved levels of IT literacy across much of the population in advanced economies. Familiarity with basic computer operations is now widespread, as evidenced by terms such as “digital natives” and “digital exclusion”, which would have had little traction in the 1980s. Usability is no longer automatically the dominant concern in interaction design. It remains important, with frustrating experiences of difficult to use digital technologies still commonplace. Poor usability is still with us, but we have moved on from Thomas Landauer's 1996 *Trouble with Computers* (Landauer 1996). When PCs, mobile phones and the internet are instrumental in major international upheavals such as the Arab Spring of 2011, the value of digital technologies can massively eclipse their shortcomings.

15.1.3 From Trouble with Computers to Trouble from Digital Technologies

Readers from developing countries can today experience Landauer's *Trouble with Computers* as the moans of oversensitive poorly motivated western users. On 26th January 1999, a "hole in the wall" was carved at the NIIT premises in New Delhi. Through the this hole, a freely accessible computer was made available for people in the adjoining slum of Kalkaji. It became an instant hit, especially with children who, with no prior experience, learnt to use the computer on their own. This prompted NIIT's Dr. Mitra to [propose](#) the following hypothesis:

.....

" The acquisition of basic computing skills by any set of children can be achieved through incidental learning provided the learners are given access to a suitable computing facility, with entertaining and motivating content and some minimal (human) guidance "

-- <http://www.hole-in-the-wall.com/Beginnings.html>

.....

There is a strong contrast here with the usability crisis of the 1980s. Computers in 1999 were easier to use than those from the 1980s, but they still presented usage challenges. Nevertheless, residual usability irritations have limited relevance for this century's slum children in Kalkaji.

The world is complex, what matters to people is complex, digital technologies are diverse. In the midst of this diverse complexity, there can no simple day of judgement when digital technologies are sent to usability heaven or unusable hell.

The story of usability is a perverse journey from simplicity to complexity. Digital technologies have evolved so rapidly that intellectual understandings of usability have never kept pace with the realities of computer usage. The pain of old and new world corporations struggling to secure returns on investment in IT in the 1980s has no rendezvous with the use of social media in the struggles for democracy in third world dictatorships. Yet we cannot simply discard the concept of usability and move on. Usage can still be frustrating, annoying, unnecessarily difficult and even impossible, even for the most skilled and experienced of users.



FIGURE 15.2: NIIT's "hole in the wall" Computer in New Delhi.

Copyright © . All Rights Reserved. Used without permission under the Fair Use Doctrine (as permission could not be obtained). See the "Exceptions" section (and subsection "allRightsReserved-UsedWithoutPermission") on the page copyright notice.



FIGURE 15.3: NIIT's "hole in the wall" Computer in New Delhi.

Copyright © . All Rights Reserved. Used without permission under the Fair Use Doctrine (as permission could not be obtained). See the "Exceptions" section (and subsection "allRightsReserved-UsedWithoutPermission") on the page copyright notice.

15.1.4 From HCI's sole concern to an enduring important factor in user experience

This encyclopaedia entry is not a requiem for usability. Although now buried under broader layers of *quality in use* and *user experience*, usability is not dead. For example, I provide some occasional IT support to my daughter via SMS. Once, I had to explain how to force the restart of a recalcitrant stalled laptop. Her last message to me on her problem was:

.....
"It's fixed now! I didn't know holding down the power button did something different to just pressing it."
.....

Given the hidden nature of this functionality (a short press hibernates many laptops), it is no wonder that my daughter was unaware of the existence of a longer ‘holding down’ action. Also, given the rare occurrences of a frozen laptop, my daughter would have had few chances to learn. She had to rely on my knowledge here. There is little she could have known herself without prior experience (e.g., of iPhone power down).



FIGURE 15.4: Holding or Pressing? Who's to Know?.

Courtesy of Rico Shen. Copyright: CC-Att-SA-3 (Creative Commons Attribution-ShareAlike 3.0).

The enduring realities of computer use that usability seeks to encompass remain real and no less potentially damaging to the success of designs today than over thirty years ago. As with all disciplinary histories, the new has not erased the old, but instead, like geological strata, the new overlies the old, with outcrops of usability still exposed within the wider evolving landscape of user experience. As in geology, we need to understand the present intellectual landscape in terms of its underlying historical processes and upheavals.

What follows is thus not a journey through a landscape, but a series of excavations that reveal what usability has been at different points in different places over the last three decades. With this in place, attention is refocused on current changes in the interaction design landscape that should give usability a stable place within a broader understanding of designing for human values (Harper et al. 2008). But for now, let us begin at the beginning, and from there take a whistle stop tour of HCI history to reveal unresolved tensions over the nature of usability and its relation to interaction design.

15.2 FROM USABILITY TO USER EXPERIENCE - TENSIONS AND METHODS

The need to design interactive software that could be used with a basic understanding of computer hardware and operating systems was first recognised in the 1970s, with pioneering work within software design by Fred Hansen from Carnegie Mellon University (CMU), Tony Wasserman from University of California, San Francisco (UCSF), Alan Kay from Xerox Palo Alto Research Center (PARC), Engel and Granda from IBM, and Pew and Rollins from BBN Technologies (for a review of early HCI work, see Pew 2002). This work took several approaches, from detailed design guidelines to high level principles for both software designs and their development processes. It brought together knowledge and capabilities from psychology and computer science. The pioneering group of individuals here was known as the *Software Psychology Society*, beginning in 1976 and based in the Washington DC area (Shneiderman 1986). This collaboration between academics and practitioners from cognitive psychology and computer science forged approaches to research and practice that remained the dominant paradigm in Interaction Design research for almost 20 years, and retained a strong hold for a further decade. However, this collaboration contained a tension on the nature of usability.

The initial focus was largely cognitive, focusing on causal relationships between user interface features and human performance, but with different views on how user interface features and human attributes would interact. If human cognitive attributes are fixed and universal, then user interface features can be inherently usable or unusable, making usability an *inherent binary property* of interactive software, i.e., an interactive system simply *is* or *is not* usable. Software could be inherently usable by conformance to guidelines and principles that could be discovered, formulated and validated by psychological experiments. However, if human cognitive attributes vary not only between individuals, but across different settings, then usability becomes an emergent property that depends, not only on features and qualities of an interactive system, but also on who was using it, and on what they were trying to do with it. The latter position was greatly strengthened in the 1990s by the “turn to the social” (Rogers et al. 1994). However, much of the intellectual tension here was defused as HCI research spread out across a range of specialist communities focused on the Association for Computing Machinery’s conferences such as the ACM Conference on Computer Supported Cooperative Work (CSCW) from 1986 or the ACM Symposium on User Interface Software and Technology (UIST) from 1988. Social understandings of usability became associated with CSCW, and technological ones with UIST.

Psychologically-based research on usability methods in major conferences remained strong into the early 1990s. However, usability practitioners became dissatisfied with academic research venues, and the first UPA (Usability Professionals Association) conference was organised in 1992. This practitioner schism happened only 10 years after the Software Psychology Society had co-ordinated a conference in Gaithersburg, from which the ACM CHI conference series emerged. This steadily removed much applied usability research from the view of mainstream HCI researchers. This separation has been overcome to some extent by the UPA’s open access *Journal of Usability Studies*, which was inaugurated in 2005.

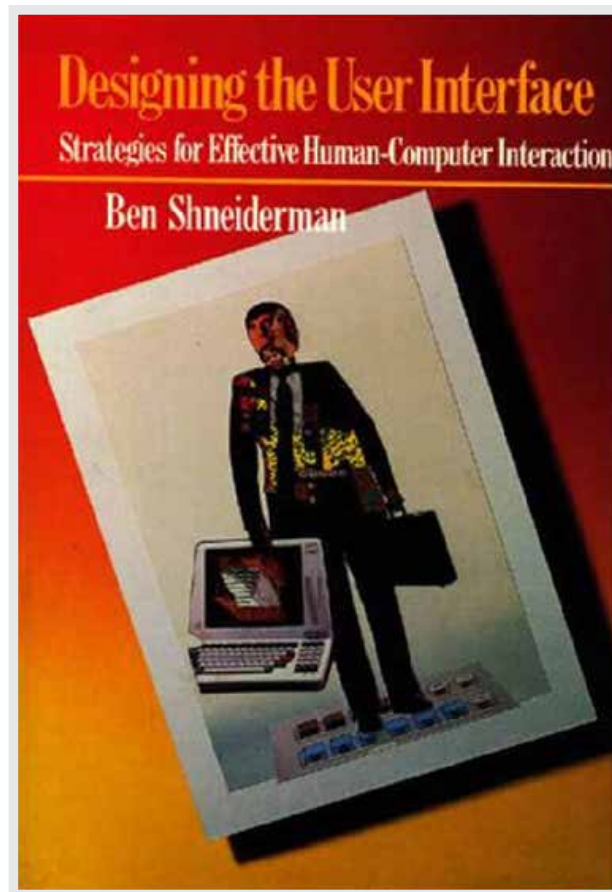


FIGURE 15.5: Ben Shneiderman, Software Psychology Pioneer, Authored the First HCI Textbook.

Copyright © Ben Shneiderman and Addison-Wesley. All Rights Reserved. Reproduced with permission. See section “Exceptions” in the copyright terms.

15.2.1 New Methods, Damaged Merchandise and a Chilling Fact

There is thus a dilemma at the heart of the concept of usability: is it a property of systems or a property of usage? A consequence of 1990s fragmentation within HCI research was such important conceptual issues were brushed aside in favour of pragmatism amongst those researchers and practitioners who retained a specialist interest in usability. By the early 1990s, a range of methods had been developed for evaluating usability. *User Testing* (Dumas and Redish 1993) was

well established by the late 1980s, as essentially a variant of psychology experiments with only dependent variables (the interactive system being tested became the independent *constant*). *Discount methods* included rapid low cost user testing, as well as inspection methods such as *Heuristic Evaluation* (Nielsen 1994). Research on *model-based methods* such as the *GOMS* model (Goals, Operators, Methods, and Selection rules - John and Kieras 1996) continued, but with mainstream publications becoming rarer by 2000.

With a choice of inspection, model-based and empirical (e.g., user testing) evaluation methods, questions arose as to which evaluation method was best and when and why. Experimental studies attempted to answer these questions by treating evaluation methods as independent variables in comparison studies that typically used problem counts and/or problem classifications as dependent variables. However, usability methods are too incompletely specified to be consistently applied, letting Wayne Gray and Marilyn Salzman invalidate several key studies in their *Damaged Merchandise* paper of 1998. Commentaries on their paper failed to undo the damage of the *Damaged Merchandise* charge, with further papers in the first decade of this century adding more concerns over not only method comparison, but the validity of usability methods themselves. Thus in 2001, Morten Hertzum and Niels Jacobsen published their “chilling fact” about use of usability methods: there are substantial evaluator effects. This should not have surprised anyone with a strong grasp of Gray and Salzman’s critique, since inconsistencies in usability method use make valid comparisons close to impossible in formal studies, and they are even more extensive in studies that attempt no control.

Critical analyses by Gray and Salzman, and by Hertzum and Jacobsen, made pragmatic research on usability even less attractive for leading HCI journals and conferences. The method focus of usability research shrunk, with critiques exposing not only the consequences of ambivalence over the causes of poor usability (system, user or both?), but also the lack of agreement over what was covered by the term usability.



FIGURE 15.6: 2020 Usability Evaluation Method Medal Winners.

Courtesy of kinnigurl. Copyright: CC-Att-SA-2 (Creative Commons Attribution-ShareAlike 2.0 Unported).

15.2.2 We Can Work it Out: Putting Evaluation Methods in their (Work) Place

Research on *usability* and *methods* has since the late 00s been superseded by research on *user experience* and *usability work*. *User experience* is a broader concept than usability, and moves beyond efficiency, task quality and vague user satisfaction to a wide consideration of cognitive, affective, social and physical aspects of interaction.

Usability work is the work carried out by usability specialists. Methods contribute to this work. Methods are not used in isolation, and should not be assessed in isolation. Assessing methods in isolation ignores the fact that usability work combines, configures and adapts multiple methods in specific proj-

ect or organisational contexts. Recognition of this fact is reflected in an expansion of research focus from *usability methods* to *usability work*, e.g., is in PhDs (Dominic Furniss, Tobias Uldall-Espersen, Mie Nørgaard) associated with the European MAUSE project (COST Action 294, 2004-2009). It is also demonstrated in the collaborative research of MAUSE Working Group 2 (Cockton and Woolrych 2009).

A focus on actual work allows realism about design and evaluation methods. Methods are only one aspect of usability work. They are not a separate component of usability work that has deterministic effects, i.e., effects that are guaranteed to occur and be identical across all project and organisational contexts. Instead, broad evaluator effects are to be expected, due to the varying extent and quality of design and evaluation resources in different development settings. This means that we cannot and should not assess usability evaluation methods in artificial isolated research settings. Instead, research should start with the concrete realities of usability work, and within that, research should explore the true nature of evaluation methods and their impact.

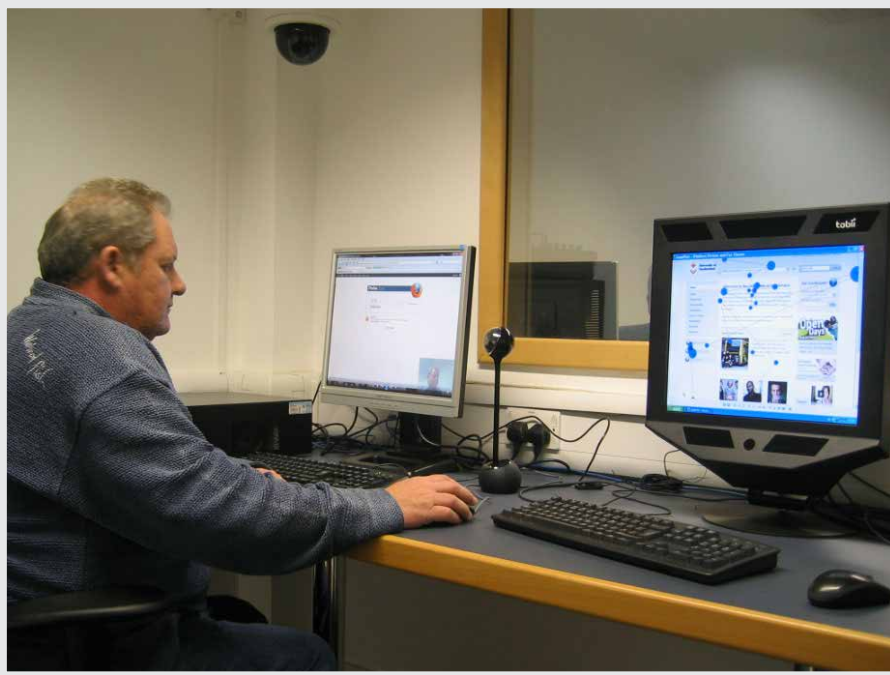


FIGURE 15.7: Usability Expert at Work: Alan Woolrych at the University of Sunderland using a minimal Mobile Usability Lab setup of webcam with audio recording plus recording of PC screen and sound, complemented by an eye tracker to his right.

Copyright status: Unknown (pending investigation). See section “Exceptions” in the copyright terms.

15.2.3 The Long and Winding Road: Usability’s Journey from Then to Now

Usability is now one aspect of user experience, and usability methods are now one loosely pre-configured area of user experience work. Even so, usability remains important. The value of the recent widening focus to user experience is that it places usability work in context. Usability work is no longer expected to establish its value in isolation, but is instead one of several complementary contributors to design quality.

Usability as a core focus within HCI has thus passed through phases of psychological theory, methodological pragmatism and intellectual disillusionment. More recent

foci on quality in use and user experience make it clear that Interaction Design cannot just focus on features and attributes of interactive software. Instead, we must focus on the interaction of users and software in specific settings. We cannot reason solely in terms of whether software is inherently usable or not, but instead have to consider what does or will happen when software is used, whether successfully, unsuccessfully, or some mix of both. Once we focus on interaction, a wider view is inevitable, favouring a broad range of concerns over a narrow focus on software and hardware features.



Copyright status: Unknown (pending investigation). See section “Exceptions” in the copyright terms.



Copyright status: Unknown (pending investigation). See section “Exceptions” in the copyright terms.

FIGURE 15.8 A-B: What’s Sailable: A Boat Alone or a Crewed Boat in Specific Sea Conditions? A Similar Question Arises for Usable Systems.

Many of the original concerns of 1980s usability work are as valid today as they were 30 years ago. What has changed is that we no longer expect usability to be the only, or even the dominant, human factor in the success of interactive systems. What has not changed is the potential confusion over what usability is, which has existed from the first days of HCI, i.e., whether software or usage is usable. While this may feel like some irrelevant philosophical hair-splitting, it has major consequences for usability evaluation. If software can be inherently usable, then usability can be evaluated solely through direct inspection. If usability can only be established by considering usage, then indirect inspection methods (walkthroughs) or empirical user testing methods must be used to evaluate.

15.2.4 Usability Futures: From Understanding Tensions to Resolving Them

The form of the word ‘usability’ implies a property that requires an *essentialist* position, i.e., one that sees properties and attributes as been inherent in objects, both natural and artificial (in Philosophy, this is called an essentialist or substantivist *ontology*). A literal understanding of usability requires interactive software to be inherently usable or unusable. Although a more realistic understanding sees usability as a property of interactive use and not of software alone, it makes no sense to talk of use as being usable, just as it makes no sense to talk of eating being edible. This is why the term *quality in use* is preferred for some international standards, because this opens up a space of possible qualities of interactive performance, both in terms of what is experienced, and in terms of what is achieved, for example, an interaction can be ‘successful’, ‘worthwhile’, ‘frustrating’, ‘unpleasant’, ‘challenging’ or ‘ineffective’.

Much of the story of usability reflects a tension between the tight software view and the broader sociotechnical view of system boundaries. More abstractly, this is a tension between substance (essence) and relation, i.e., between *inherent qualities* of interactive software and *emergent qualities* of interaction. In philosophy, the position that relations are more fundamental than things in themselves characterises a *relational ontology*.

Ontologies are theories of being, existence and reality. They lead to very different understandings of the world. Technical specialists and many psychologists within HCI are drawn to essentialist ontologies, and seek to achieve usability predominantly through consideration of user interface features. Specialists with a broader human-focus are mostly drawn to relational ontologies, and seek to understand how contextual factors interact with user interface features to shape experience and performance. Each ontology occupies ground within the HCI landscape. Both are now reviewed in turn. Usability evaluation methods are then

briefly reviewed. While tensions between these two positions have dominated the evolution of usability in principle and practice, we can escape the impasse. A strategy for escaping longstanding tensions within usability will be presented, and future directions for usability within user experience frameworks will be indicated in the closing section.

15.3 LOCATING USABILITY WITHIN SOFTWARE: GUIDELINES, HEURISTICS, PATTERNS AND ISO 9126

15.3.1 Guidelines for Usable User Interfaces

Much early guidance on usability came from computer scientists such as Fred Hansen from Carnegie Mellon University (CMU) and Tony Wasserman, then at University of California, San Francisco (UCSF). Computer science has been strongly influenced by mathematics, where entities such as similar or equilateral triangles have eternal absolute intrinsic properties. Computer scientists seek to establish similar inherent properties for computer programs, including ones that ensure usability for interactive software. Thus initial guidelines on user interface design incorporated a technocentric belief that usability could be ensured via software and hardware features alone. A user interface would be inherently usable if it conformed to guidelines on, for example, naming, ordering and grouping of menu options, prompting for input types, input formats and value ranges for data entry fields, error message structure, response time, and undoing capabilities. The following four example guidelines are taken from Smith and Mosier's 1986 collection commissioned by the US Air Force (Smith and Mosier 1986):

1.0/4 + Fast Response
Ensure that the computer will acknowledge data entry actions rapidly, so that users are not slowed or paced by delays in computer response; for normal operation, delays in displayed feedback should not exceed 0.2 seconds.
1.0/15 Keeping Data Items Short
For coded data, numbers, etc., keep data entries short, so that the length of an individual item will not exceed 5-7 characters.
1.0/16 + Partitioning Long Data Items
When a long data item must be entered, it should be partitioned into shorter symbol groups for both entry and display.
Example
A 10-digit telephone number can be entered as three groups, NNN-NNN-NNNN.
1.4/12 + Marking Required and Optional Data Fields
In designing form displays, distinguish clearly and consistently between required and optional entry fields.

FIGURE 15.9: Four example guidelines taken from Smith and Mosier's 1986 collection.

25 years after the publication of the above guidance, there are still many contemporary web site data entry forms whose users would benefit from adherence to these guidelines. Even so, while following guidelines can greatly improve software usability, it cannot guarantee it.

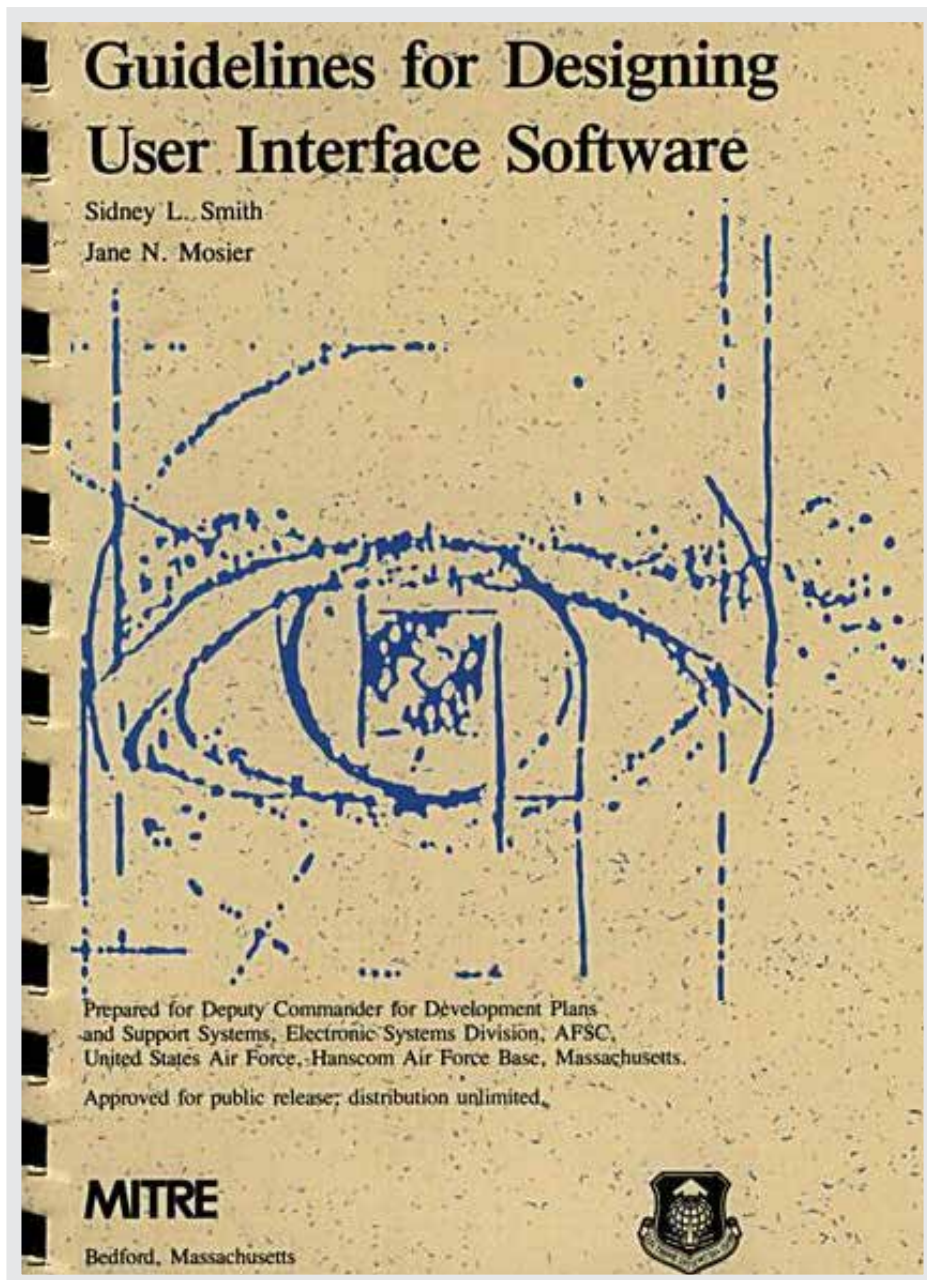


FIGURE 15.10: This Book Contains More Guidelines Than Anyone Could Imagine.

Copyright © Sidney L. Smith and Jane N. Mosier and The MITRE Corporation. All Rights Reserved. Reproduced with permission. See section "Exceptions" in the copyright terms.

15.3.2 Manageable Guidance: Design Heuristics for Usable User Interfaces

My original paper copy of Smith and Mosier's guidelines occupies 10cm of valuable shelf space. It is over 25 years old and I have never read all of it. I most probably never will. There are simply too many guidelines there to make this worthwhile (in contrast, I have read complete style guides for Windows and Apple user interfaces in the past).

The bloat of guidelines collections did not remove the appeal of technocentric views of usability. Instead, hundreds of guidelines were distilled into ten heuristics by Rolf Mohlich and Jakob Nielsen. These were further assessed and refined into the final version of in *Heuristic Evaluation* (Nielsen 1994), an inspection method that examines software features for potential causes of poor usability. Heuristics generalise more detailed guidelines from collections such as Smith and Mosier. Many have a technocentric focus, e.g.:

Visibility of system status

The system should always keep users informed about what is going on, through appropriate feedback within reasonable time.

User control and freedom

Users often choose system functions by mistake and will need a clearly marked “emergency exit” to leave the unwanted state without having to go through an extended dialogue. Support undo and redo.

Error prevention

Even better than good error messages is a careful design which prevents a problem from occurring in the first place. Either eliminate error-prone conditions or check for them and present users with a confirmation option before they commit to the action.

Recognition rather than recall

Minimize the user’s memory load by making objects, actions, and options visible. The user should not have to remember information from one part of the dialogue to another. Instructions for use of the system should be visible or easily retrievable whenever appropriate.

Flexibility and efficiency of use

Accelerators -- unseen by the novice user -- may often speed up the interaction for the expert user such that the system can cater to both inexperienced and experienced users. Allow users to tailor frequent actions.

FIGURE 15.11: Example heuristics originally developed in Molich and Nielsen 1990 and Nielsen and Molich 1990.

Heuristic Evaluation became the most popular user-centred design approach in the 1990s, but has become less prominent with the move away from desktop applications. Quick and dirty user testing soon overtook Heuristic Evaluation (compare the survey of Venturi et al. 2006 with Rosenbaum et al. 2000).

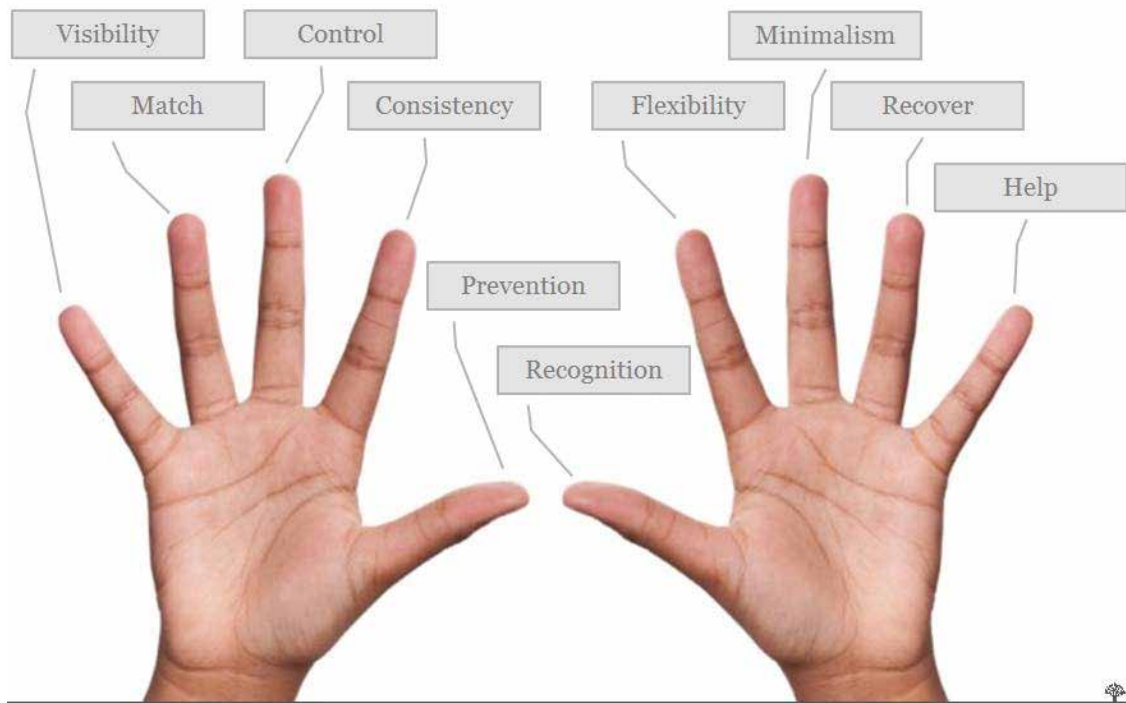


FIGURE 15.12: One Heuristic for Each Digit from Nielsen.

Copyright status: Unknown (pending investigation). See section “Exceptions” in the copyright terms.

15.3.3 Invincible Intrinsic: Patterns and Standards Keep Usability Essential

Moves away from system-centric approaches within user-centred design have not signalled the end of usability methods that focus solely on software artefacts, with little or no attention to usage. This may be due to the separation of the usability communities (now user experience) from the software engineering profession. System-centred usability remains common in user interface pattern languages. For example, a pattern from Jenifer Tidwell updates Smith and Mosier style guidance for contemporary web designers (designinginterfaces.com/Input_Prompt).

Pattern: Input Prompt

Prefill a text field or dropdown with a prompt that tells the user what to do or type.

Figure 15.13: An example pattern from Jenifer Tidwell's *Designing Interfaces*.

The 1991 ISO 9126 standard on *Software Engineering – Product Quality* was strongly influenced by the essentialist preferences of computer science, with usability defined as:

.....

“a set of [*product*] attributes that bear on the effort needed for use, and on the individual assessment of such use, by a stated or implied set of users.”

.....

This is the first of three definitions presented in this encyclopaedia entry. The attributes here are assumed to be software product attributes, rather than user interaction ones. However, the relational (contextual) view of usage favoured in HCI has gradually come to prevail. By 2001, ISO 9126 had been revised to define usability as:

.....

“(1’) the capability of the software product to be understood, learned, used and attractive to the user, when used under specified conditions”

.....

This revision remains product focused (essentialist), but the ‘when’ clause moved ISO 9126 away from a wholly essentialist position on usability by implicitly acknowledging the influence of a context of use (“specified conditions”) that extends beyond “a stated or implied set of users”.

In an attempt to align the technical standard ISO 9126 with the human factors standard ISO 9241 (see below), ISO 9126 was extended in 2004 by a fourth section on *quality in use*, resulting in an uneasy compromise between software engineers and human factors experts. This uneasy compromise persists, with the 2011 replacement standard for ISO 9126, ISO 25010 maintaining an essentialist view of usability. In ISO 25010, usability is both an intrinsic product quality characteristic *and* a subset of quality in use (comprising effectiveness, efficiency and satisfaction). As a product characteristic in ISO 25010, usability has the intrinsic subcharacteristics of:

- ▶ Appropriateness
- ▶ Recognisability
- ▶ Learnability
- ▶ Operability (degree to which a product or system has attributes that *make it* easy to operate and control - *emphasis added*)
- ▶ User error protection
- ▶ User interface aesthetics
- ▶ Accessibility

ISO 25010 thus had to include a note that exposed the internal conflict between software engineering and human factors world views:

.....

“ Usability can either be specified or measured as a product quality characteristic in terms of its subcharacteristics, or specified or measured directly by measures that are a subset of quality in use. ”

.....

A similar note appears for learnability and accessibility. Within the world of software engineering standards, a mathematical world view clings hard to an essentialist position on usability. In HCI, where context has reigned for decades, this could feel incredibly perverse. However, despite HCI's multi-factorial understanding of usability, which follows automatically from a contextual position, HCI evangelists' anger over poor usability always focuses on software products. Even though users, tasks and contexts are all known to influence usability, only hardware or software should be changed to improve usability, endorsing the software engineers' position within ISO 25010 (attributes *make software easy to operate and control*). Although HCI's world view typically rejects essentialist monocausal explanations of usability, when getting angry on the user's behalf, the software always gets the blame.

It should be clear that issues here are easy to state but harder to unravel. The stalemate in ISO 25010 indicates a need within HCI to give more weight to the influence of software design on usability. If users, tasks and contexts must not be changed, then the only thing that we can change is hardware and/or software. Despite the psychological marginalisation of designers' experience and expertise when expressed in guidelines, patterns and heuristics, these can be our most critical resource for achieving usability best practice. We should bear this in mind as we move to consider HCI's dominant contextual position on usability.

15.4 LOCATING USABILITY WITHIN INTERACTION: CONTEXTS OF USE AND ISO STANDARDS

The tensions within international standards could be seen within Nielsen's Heuristics, over a decade before the 2004 ISO 9126 compromise. While the five sample heuristics in the previous section focus on software attributes, one heuristic focuses on the relationship between a design and its context of use (Nielsen 1994):

.....

“ Match between system and the real world

The system should speak the users' language, with words, phrases and concepts familiar to the user, rather than system-oriented terms. Follow real-world conventions, making information appear in a natural and logical order. ”

.....

This relating of usability to the 'real world' was given more structure in the ISO 9241-11 Human Factors standard, which related usability to the usage context as the:

.....

“ Extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use ”

.....

This is the second of three definitions presented in this encyclopaedia entry. Unlike the initial and revised ISO 9126 definitions, it was not written by software

engineers, but by human factors experts with backgrounds in ergonomics, psychology and similar.

ISO 9241-11 distinguishes three component factors of *usability*: effectiveness, efficiency, satisfaction. These result from multi-factorial interactions between users, goals, contexts and a software product. Usability is not a characteristic, property or quality, but an *extent* within a multi-dimensional space. This extent is evidenced by what people can actually achieve with a software product and the costs of these achievements. In practical terms, any judgement of usability is a holistic assessment that combines multi-faceted qualities into a single judgement.

Such single judgements have limited use. For practical purposes, it is more useful to focus on separate specific qualities of user experience, i.e., the extent to which *thresholds* are met for different qualities. For example, a software product may not be deemed usable if key tasks cannot be performed in normal operating contexts within an acceptable time. Here, the focus would be on *efficiency* criteria. There are many usage contexts where time is limited. The bases for time limits vary considerably, and include physics (ballistics in military combat), physiology (medical trauma), chemistry (process control) or social contracts (newsroom print/broadcast deadlines).

Effectiveness criteria add to the complexity of quality thresholds. A military system may be efficient, but it is not effective if its use results in what is euphemistically called ‘collateral damage’, including ‘friendly fire’ errors. We can imagine trauma resuscitation software that enables timely responses, but leads to avoidable ‘complications’ (another domain euphemism) after a patient has been stabilised. A process control system may support timely interventions, but may result in waste or environmental damage that limits the effectiveness of operators’ responses. Similarly, a newsroom system may support rapid preparation of content, but could obstruct the delivery of high quality copy.

For *satisfaction*, usage could be both objectively efficient and effective, but cause uncomfortable user experiences that give rise to high rates of staff turnover (as common in call centres). Similarly, employees may thoroughly enjoy a fancy multimedia fire safety training title, but it could be far less effective (and thus potentially deadly) compared to the effectiveness of a boring instructional text-with-pictures version.

ISO 9241-11's three factors of usability have recently become five in by ISO 25010's quality in use factors:

- ▶ Effectiveness
- ▶ Efficiency
- ▶ Satisfaction
- ▶ Freedom from risk
- ▶ Context coverage

The two additional factors are interesting. *Context coverage* is a broader concept than the contextual fit of the *match between system and Nielsen's Match between System and Real World* heuristic (Nielsen 1994). It extends *specified users* and *specified goals* to potentially any aspect of a context of use. This should include all factors relevant to *freedom from risk*, so it is interesting to see this given special attention, rather than trusting *effectiveness* and *satisfaction* to do the work here. However, such piecemeal extensions within ISO 25010 open up the question of comprehensiveness and emphasis. For example, why are factors such as ease of learning either overlooked or hidden inside efficiency or effectiveness?



FIGURE 15.14: ISO Accessibility Standard Discussion.

Copyright © ISO and Lionel Egger. All Rights Reserved. Reproduced with permission. See section “Exceptions” in the copyright terms.

15.4.1 Contextual Coverage Brings Complex Design Agendas

Relational positions on usability are inherently more complex than essentialist ones. The latter let, interactive systems be *inspected* to assess their usability potential on the basis of their design features. Essentialist approaches remain attractive because evaluations can be fully resourced through guidelines, patterns and similar expressions of best practice for interaction design. Relational approaches require a more complex set of co-ordinated methods. As relational positions become more complex, as in the move from ISO 9241-11 to ISO 25010, a broader range of evaluation methods is required. Within the relational view, usability is the result of a set of complex interactions that manifests itself in a range of usage factors. It is very difficult to see how a single evaluation method could address all these factors. Whether or not this is possible, no such method currently exists.

Relational approaches to usability require a range of evaluation methods to establish its *extent*. Extent introduces further complexities, since all identified usability factors must be *measured*, then judgements must be made as to whether achieved extents are adequate. Here, usability evaluation is not a simple matter

of inspection, but instead it becomes a complex logistical operation focused on implementing a *design agenda*.

An *agenda* is list of things to be done. A *design agenda* is therefore a list of design tasks, which need to be managed within an embracing development process. There is an implicit design agenda in ISO 9241-11, which requires interaction designers to identify target beneficiaries, usage goals, and levels of efficiency, effectiveness and satisfaction for a specific project. Only then is detailed robust usability evaluation possible. Note that this holds for ISO 9241-11 and similar evaluation philosophies. It does not hold for some other design philosophies (e.g., Sengers and Gaver 2006) that give rise to different design agendas.

A key task on the ISO 9241-11 evaluation agenda is thus measuring the extent of usability through a co-ordinated set of metrics, which will typically mix quantitative and qualitative measures, often with a strong bias towards one or the other. However, measures only enable evaluation. To evaluate, measures need to be accompanied by targets. Setting such targets is another key task from the ISO 9241-11 evaluation agenda. This is often achieved through the use of generic *severity scales*. To use such resources, evaluators need to interpret them in specific project contexts. This indicates that re-usable evaluation resources are not complete re-usable solutions. Work is required to turn these resources into actionable evaluation tasks.

For example, the two most serious levels of Chauncey Wilson's problem severity scale (Wilson 1999) are:

- ▶ Level 1 — Catastrophic error causing irrevocable loss of data or damage to the hardware or software. The problem could result in large-scale failures that prevent many people from doing their work. Performance is so bad that the system cannot accomplish business goals.

- Level 2 — Severe problem causing possible loss of data. User has no workaround to the problem. Performance is so poor that ... universally regarded as ‘pitiful’.

Each severity level requires answers to questions about specific measures and contextual information, i.e., how should the following be interpreted in a specific project context: ‘many prevented from doing work’; ‘cannot accomplish business goals’; ‘performance regarded as pitiful’. These top two levels also require information about the software product: ‘loss of data’; ‘damage to hardware of software’; ‘no workaround’.

Wilson’s three further lower level scales add considerations such as: ‘wasted time’, ‘increased error or learning rates’, and ‘important feature not working as expected’. These all set a design agenda of questions that must be answered. Thus to know that performance is regarded as pitiful, we would need to choose to measure relevant subjective judgments. Other criteria are more challenging, e.g., how would we know whether time is wasted, or whether business goals cannot be accomplished? The first depends on values. The idea of ‘wasting’ time (like money) is specific to some cultural contexts, and also depends on how long tasks are expected to take with a new system, and how much time can be spent on learning and exploring. As for business goals, a business may seek, for example, to be seen as socially and environmentally responsible, but may not expect every feature of every corporate system to support these goals.

Once thresholds for severity criteria have been specified, it is not clear how designers can or should trade off factors such as efficiency, effectiveness and satisfaction against each other. For example, users may not be satisfied even when they exceed target efficiency and effectiveness, or conversely they could be satisfied even when their performance should not warrant that relative to design targets. Target levels thus guide rather than dictate the interpretation of results and how to respond to them.

Method requirements thus differ significantly between essentialist and relational approaches to usability. For high quality evaluation based on any relational position, not just ISO 9241-11's, evaluators must be able to modify and combine existing re-usable resources for specific project contexts. Ideally, the re-usable resources would do most of the work here, resulting in efficient, effective and satisfying usability evaluation. If this is not the case, then high quality usability evaluation will present complex logistical challenges that require extensive evaluator expertise and project specific resources.

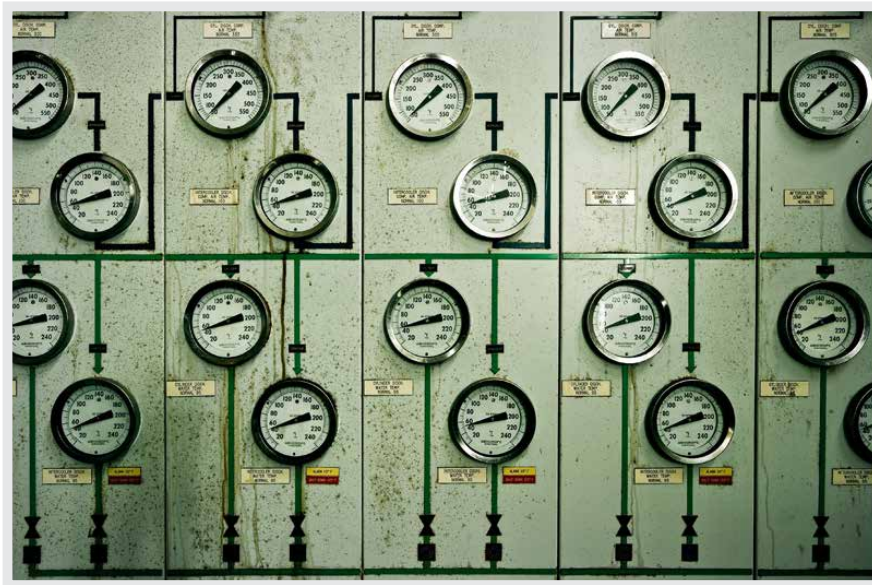


FIGURE 15.15: Relational Approaches to Usability Require Multiple Measures.

Copyright © Simon Christen - iseemooi. All Rights Reserved. Reproduced with permission. See section "Exceptions" in the copyright terms.

15.5 THE DEVELOPMENT OF USABILITY EVALUATION: TESTING, MODELLING AND INSPECTION

Usability is a contested historical term that is difficult to replace. User experience specialists have to refer to usability, since it is a strongly established concept with-

in the IT landscape. However, we need to exercise caution in our use of what is essentially a flawed concept. Software is not usable. Instead, *software gets used*, and the resulting user experiences are a composite of several qualities that are shaped by product attributes, user attributes and the wider context of use.

Now, squabbles over definitions will not necessarily impact practice in the ‘real world’. It is possible for common sense to prevail and find workarounds for what could well be semantic distractions with no practical import. However, when we examine usability evaluation methods, we do see that different conceptualisations of usability result in differences over the *causes* of good and poor usability.

Essentialist usability is, causally *homogeneous*. This means that all causes of user performance are of the same type, i.e., due to technology. System-centred inspection methods can identify such causes.

Contextual usability is causally *heterogeneous*. This means that causes of user performance are of different types, some due to technologies, others due to some aspect(s) of usage contexts, but most due to interactions between both. Several evaluation and other methods may be needed to identify and relate a nexus of causes.

Neither usability paradigm (i.e., essentialist or relational) has resolved the question of relevant *effects*, i.e., what counts as evidence of good or poor usability, and thus there are few adequate methods here. Essentialist usability can pay scant attention to effects (Lavery et al. 1997): who cares what poor design will do to users, it’s bad enough that it’s poor design! Contextual usability has more focus on effects, but there is limited consensus on the sort of effects that should count as evidence of poor usability. There are many examples of what could count as evidence, but what actually should is left to a design team’s judgement.

Some methods can predict effects. The *GOMS* model (Goals, Operators, Methods, and Selection rules) predicts effects on expert error free task completion time, which is useful in some project contexts (Card et al 1980, John and Kieras 1996). For example, external processes may require a task to be completed within

a maximum time period. If predicted expert error free task completion time exceeds this, then it is highly probable that non-expert error prone task completion take even longer. Where interactive devices such as in-car systems distract attention from the main task (e.g., driving), then time predictions are vital. Recent developments such as CogTool (Bellamy et al. 2011) have given a new lease of life to practical model-based evaluation in HCI. More powerful models than GOMS are now being integrated into evaluation tools (e.g., Salvucci 2009).



FIGURE 15.16: Model-Based methods can predict how long drivers could be distracted, and much more.

Courtesy of Ed Brown. Copyright: CC-Att-SA-2 (Creative Commons Attribution-ShareAlike 2.0 Unported).

Usability work can thus be expected to involve a mix of methods. The mix can be guided by high level distinctions between methods. Evaluation methods can be *analytical* (based on examination of an interactive system and/or potential interactions with it) or *empirical* (based on actual usage data). Some analytical methods require the construction of one or more models. For example, GOMS models

the relationships between software and human performance. Software attributes in GOMS all relate to user input methods at increasing levels of abstraction from the keystroke level up to abstract command constructs. System and user actions are interleaved in task models to predict users' methods (and execution times at a keystroke level of analysis).

15.5.1 Analytical and Empirical Evaluation Methods, and How to Mix Them

Analytical evaluation methods may be *system-centred* (e.g., Heuristic Evaluation) or *interaction-centred* (e.g., Cognitive Walkthrough). Design teams use the resources provided by a method (e.g., heuristics) to identify strong and weak elements of a design from a usability perspective. Inspection methods tend to focus on the *causes* of good or poor usability. *System-centred* inspection methods focus solely on software and hardware features for attributes that will promote or obstruct usability. *Interaction-centred* methods focus on two or more causal factors (i.e., software features, user characteristics, task demands, other contextual factors).

Empirical evaluation methods focus on evidence of good or poor usability, i.e., the positive or negative *effects* of attributes of software, hardware, user capabilities and usage environments. User testing is the main project-focused method. It uses project-specific resources such as test tasks, users, and also measuring instruments to expose usability problems that can arise in use. Also, essentialist usability can use empirical experiments to demonstrate superior usability arising from user interface components (e.g., text entry on mobile phones) or to optimise tuning parameters (e.g., timings of animations for windows opening and closing). Such experiments assume that the test tasks, test users and test contexts allow generalisation to other users, tasks and contexts. Such assumptions are readily broken, e.g., when users are very young or elderly, or have impaired movement or perception.

Analytical and empirical methods emerged in rapid succession, with empirical methods emerging first in the 1970s as simplified psychology experiments (for

examples, see early volumes of International Journal of Man-Machine Studies 1969-79). Model-based approaches followed in the 1980s, but the most practical ones are all variants of the initial GOMS method (John and Kieras 1996). Model-free inspection methods appeared at the end of the 1980s, with rapid evolution in the early 1990s. Such methods sought to reduce the cost of usability evaluation by *discounting* across a range of resources, especially users (none required, unlike user testing), expertise (transferred by heuristics/models to novices) or extensive models (none required, unlike GOMS).



FIGURE 15.17: Chicken Fajitas Kit: everything you need except chicken, onion, peppers, oil, knives, chopping board, frying pan, stove etc. Usability Evaluation Methods are very similar - everything is fine once you've worked to provide everything that's missing.

Copyright © Old El Paso. All Rights Reserved. Used without permission under the Fair Use Doctrine (as permission could not be obtained). See the "Exceptions" section (and subsection "allRightsReserved-Used-WithoutPermission") on the page copyright notice.

Achieving balance in a mix of evaluation methods is not straightforward, and requires more than simply combining analytical and empirical methods. This is because there is more to usability work than simply choosing and using methods. Evaluation methods are as complete as a Chicken Fajita Kit, which contains very little of what is actually needed to make Chicken Fajitas: no chicken, no onion, no peppers, no cooking oil, no knives for peeling/coring and slicing, no chopping board, no frying pan, no stoves etc. Similarly, user testing ‘methods’ as published miss out equally vital ingredients and project specific resources such as participant recruitment criteria, screening questionnaires, consent forms, test task selection criteria, test (de)briefing scripts, target thresholds, and even data collection instruments, evaluation measures, data collation formats, data analysis methods, or reporting formats. There is no complete published user testing method that novices can pick up and use ‘as is’. All user testing requires extensive project-specific planning and implementation. Instead, much usability work is about configuring and combining methods for project-specific use.

15.5.2 The Only Methods are the Ones that You Complete Yourself

When planning usability work, it is important to recognise that so-called ‘methods’ are more strictly loose collections of resources better understood as ‘approaches’. There is much work in getting usability work to work, and as with all knowledge-based work, methods cannot be copied from books and applied without a strong understanding of fundamental underlying concepts. One key consequence here is that only specific instances of methods can be compared in empirical studies, and thus credible research studies cannot be designed to collect evidence of systematic reliable differences between different usability evaluation methods. All methods have unique usage settings that require project-specific resources, e.g., for user testing, these include participant recruitment, test procedures and (de-)briefings. More generic resources such as problem extraction methods (Cockton and Lavery 1999) may also vary across user testing contexts. These inevitably obstruct reliable comparisons.



Copyright © George Eastman House Collection. All Rights Reserved. Reproduced with permission. See section "Exceptions" in the copyright terms.



FIGURE 15.18 A-B: Dogs or Cats: which is the better pet? It all depends on what sorts of cats and dogs you compare, and how you compare them. The same is true of evaluation methods.

Copyright © George Eastman House Collection. All Rights Reserved. Reproduced with permission. See section “Exceptions” in the copyright terms.

Consider a simple comparison of heuristic evaluation against user testing. Significant effort would be required to allow a fair comparison. For example, if the user testing asked test users to carry out fixed tasks, then heuristic evaluators would need to explore the same system using the same tasks. Any differences and similarities between evaluation results for the two methods would not generalise beyond these fixed tasks, and there are also likely to be extensive evaluation effects arising from individual differences in evaluator expertise and performance. If tasks are not specified for the evaluations, then it will not be clear whether differences and similarities between results are due to the approaches used or to the unrecorded tasks within for the evaluations. Given the range of resources that need to be configured for a specific user test, it is simply not possible to control all known potential confounds (still less all currently unknown ones). Without such controls, the main sources of differences between methods may be factors with no bearing on actual usability.

The tasks carried out by users (in user testing) or used by evaluators (in inspections or model specifications) are thus one possible confound when comparing evaluation approaches. So too are evaluation measures and target thresholds. Time on task is a convenient measure for usability, and for some usage contexts it is possible to specify worthwhile targets, e.g., for supermarket checkouts the target time to check out representative trolleys of purchases could be 30 minutes for 10 typical trolley loads of shopping). However, in many settings, there are no time thresholds for efficient use that can be used reliably (e.g., time to draft and print a one page business letter, as opposed to typing one in from a paper draft or a dictation).

Problems associated with setting thresholds are compounded by problems associated with choosing the measures for which thresholds are required. A wide range of potential measures can be chosen for user testing. For example, in 1988, usability specialists from Digital Equipment Corporation and IBM (Whiteside et al. 1988) published a long list of possible evaluation measures, including:

Measure without measure: there's so much scope for scoring

- ▶ Counts of:
 - ◆ commands used
 - ◆ repetitions of failed commands
 - ◆ runs of successes and of failures
 - ◆ good and bad features recalled by users
 - ◆ available commands not invoked/regressive behaviours
 - ◆ users preferring your system
- ▶ Percentage of tasks completed in time period
- ▶ Counts or percentages of:
 - ◆ errors
 - ◆ superior competitor products on a measure
- ▶ Ratios of
 - ◆ successes to failures
 - ◆ favourable to unfavourable comments
- ▶ Times
 - ◆ to complete a task
 - ◆ spent in errors
 - ◆ spent using help or documentation
- ▶ Frequencies
 - ◆ of help and documentation use
 - ◆ of interfaces misleading users

- ◆ users needing to work around a problem
- ◆ users disrupted from a work task
- ◆ users losing control of the system
- ◆ users expressing frustration or satisfaction

No claims were made for the comprehensiveness of the full list of measures that were known to have been used up to the point of publication within Digital Equipment Corporation or IBM. What was clear was a position that project teams must choose their own metrics and thresholds. No methods yet exist to reliably support such choices.

There are no universal measures of usability that are relevant to every software development project. Interestingly, Whiteside et al. (1988) was the publication that first introduced *contextual design* to the HCI community. Its main message was that existing user testing practices were delivering far less value for design than contextual research. A hope was expressed that once contexts of use were better understood, and contextual insights could be shown to inform successful design across a diverse range of projects, then new contextual measures would be found for more appropriate evaluation of user experiences. Two decades elapsed before bases for realising this hope emerged within HCI research and professional practice. The final sections of this encyclopaedia entry explore possible ways forward.

15.5.3 Sorry to Disappoint You But ...

To sum up the position so far:

1. There are fundamental differences on the nature of usability, i.e., it is either an inherent property of interactive systems, or an emergent property of usage. There is no single definitive answer to what usability 'is'.

Usability is only an inherent measurable property of all interactive digital technologies for those who refuse to think of it in any other way.

2. There are no universal measures of usability, and no fixed thresholds above or below which all interactive systems are or are not usable. There are no *universal*, robust, objective and reliable metrics. There are no evaluation methods that unequivocally determine whether or not an interactive system or device is usable, or to what extent. All positions here involve hard won expertise, judgement calls, and project-specific resources beyond what all documented evaluation methods provide.
3. Usability work is too complex and project-specific to admit generalisable methods. What are called ‘methods’ are more realistically ‘approaches’ that provide loose sets of resources that need to be adapted and configured on a project by project basis. There are no reliable pre-formed methods for assessing usability. Each method in use is unique, and relies heavily on the skills and knowledge of evaluators, as well as on project-specific resources. There are no off-the-shelf evaluation methods. Evaluation methods and metrics are not completely documented in any literature. Developing expertise in usability measurement and evaluation requires far more than reading about methods, learning how to apply them, and through this alone, becoming proficient in determining whether or not an interactive system or device is usable, and if so, to what extent. Even system-centred essentialist methods leave gaps for evaluators to fill (Cockton et al. 2004, Cockton et al. 2012).

The above should be compared with the four opening propositions, which together constitute an attractive ideology that promises certainties regardless of evaluator experience and competence. Each proposition is not wholly true, and can be mostly false. Evaluation can never be an add-on to software development projects. Instead, the scope of usability work, and the methods used, need to be

planned with other design and development activities. Usability evaluation requires supporting resources that are an integral part of every project, and must be developed there.

The tension between essentialist and relational conceptualisations of usability is only the tip of the iceberg of challenges for usability work. Not only is it not clear what usability is (although competing definitions are available), but it is also not clear specifically how usability should be assessed outside of the contexts of specific projects. What matters in one context may not matter in another. Project teams must decide what matters. The usability literature can indicate possible measure of usability, but none are universally applicable. The realities of usability work are that each project brings unique challenges that require experience and expertise to meet them. Novice evaluators cannot simply research, select and apply usability evaluation methods. Instead, actual methods in use are the critical *achievement* of all usability work.

Methods are *made* on the ground on a project by project basis. They are not archived ‘to go’ in the academic or professional literature. Instead there are two starting points. Firstly, there are literatures on a range of approaches that provide some re-usable resources for evaluators, but require additional information and judgement within project contexts before practical methods can be completed. Secondly, there are detailed case studies of usability work within specific projects. Here the challenge for evaluators is to identify resources and practices within the case study that would have a good fit with other project contexts, e.g., a participant recruitment procedure from a user testing case study may be re-usable in other projects, perhaps with some modifications.

Readers could reasonably draw the conclusion from the above that usability is an attractive idea in principle that has limited substance in reality. However, the reality is that we all continue to experience frustrations when using interactive digital technologies, and often we would say that we do find them difficult to use. Even so, frustrating user experiences may not be due to some single abstract con-

struct called ‘usability’, but instead be the result of unique complex interactions between people, technology and usage contexts. Interacting factors here must be considered together. It is not possible to form judgements on the severity of isolated usage difficulties, user discomfort or dissatisfaction. Overall judgements on the quality of interactive software must balance what can be achieved through using it against the costs of this use. There are no successful digital technologies without what could be usability flaws to some HCI experts (I can *always* find some!). Some technologies appear to have severe flaws, and are yet highly successful for many users. Understanding why this is the case provides insights that move us away from a primary focus on usability in interaction design.

15.6 WORTHWHILE USABILITY: WHEN AND WHY USABILITY MATTERS, AND HOW MUCH

While writing the previous section, I sought advice via Facebook on transferring contacts from my vintage Nokia N96 mobile phone to my new iPhone. One piece of advice turned out to be specific to Apple computers, but was still half-correct for a wintel PC. Eventually, I identified a possible data path that required installing the Nokia PC suite on my current laptop, backing up contacts from my old phone to my laptop, downloading a freeware program that would convert contacts from Nokia’s proprietary backup format into a text format for spreadsheets/databases (comma separated values - .csv), failing to import it into a cloud service, importing it into the Windows Address Book on my laptop (after spreadsheet editing), and then finally synchronising the contacts instead via iTunes with my new iPhone.

15.6.1 A Very Low Frequency Multi-device Everyday Usability Story

From start to finish, my phone number transfer task took two and a half hours. Less than half of my contacts were successfully transferred, and due to problems in the spreadsheet editing, I had to transfer contacts in a form that required further editing on my iPhone or in the Windows contacts folder.

Focusing on the ISO 9241-11 definition of usability, what can we say here about the usability of a complex ad hoc overarching product-service system involving social networks, cloud computing resources, web searches, two component product-service systems (Nokia 96 + Nokia PC Suite, iPhone + iTunes) and Windows laptop utilities?

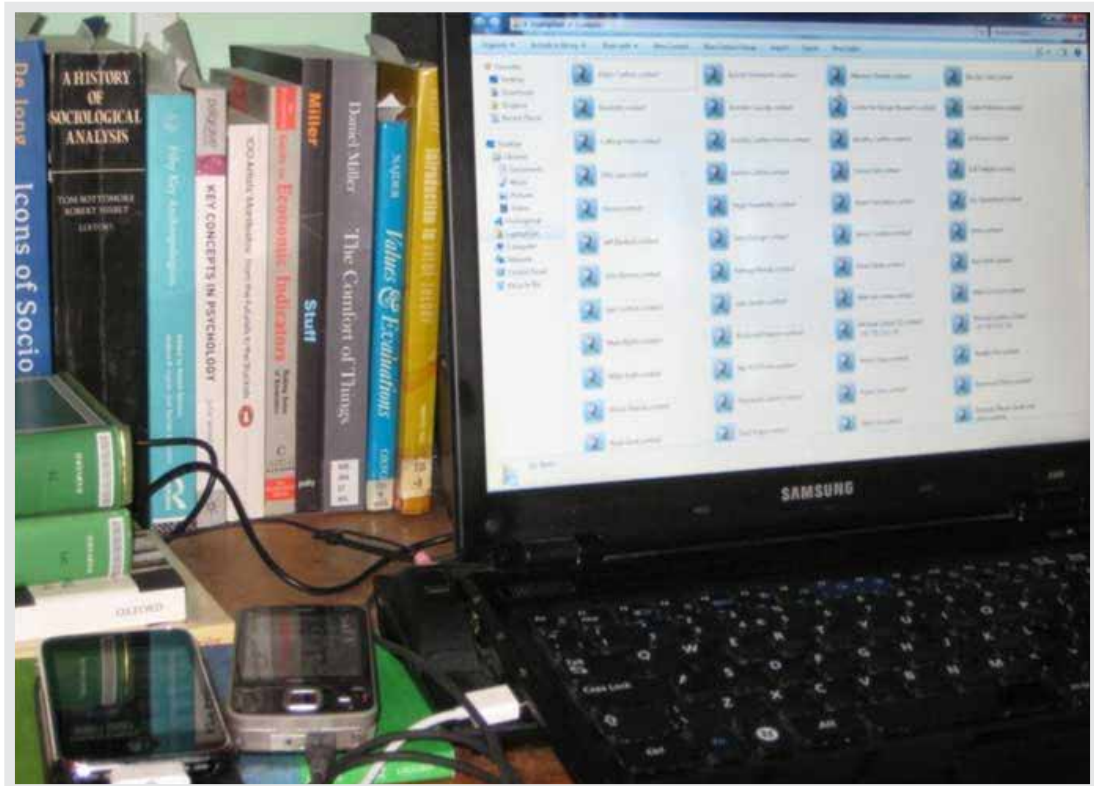


FIGURE 15.19: A Tale of Two Mobiles and Several Software Utilities.

Copyright status: Unknown (pending investigation). See section “Exceptions” in the copyright terms.

Was it efficient taking 2.5 hours over this? Around 30 minutes each were spent on:

1. web searches, reading about possible solutions, and a freeware download
2. installing mobile phone software (new laptop since I bought the Nokia), attempts to connect Nokia to laptop, laptop restart, successful backup, extraction to .csv format with freeware

3. exploring a cloud email contacts service, failing to upload to it.
4. test upload to Windows address book, edits to improve imports, failed edits of phone numbers, successful import
5. Synchronisation of iPhone and iTunes

To reach a judgement on efficiency, we need to first bear in mind that during periods of waiting (uploads, downloads, synchronisations, installations), I proof read my current draft of this entry and corrected it. This would have taken 30 minutes anyway. Secondly, I found useful information from the web searches that lead me to the final solution. Thirdly, I had to learn how to use the iTunes synchronisation capabilities for iPhones, which took around 10 minutes and was an essential investment for the future. However, I wasted at least 30 minutes on a cloud computing option suggested on Facebook (I had to add email to the cloud service before failing to upload from the .csv file). There were clear usability issues here, as the email service gave no feedback as to why it was failing to upload the contacts. There is no excuse for such poor interaction design in 2011, which forced me to try a few times before I realised that it would not work, at least with the data that I had. Also, the extracted phone numbers had text prefixes, but global search and replace in the spreadsheet resulted in data format problems that I could not overcome. I regard both of these as usability problems, one due to the format of the telephone numbers as extracted, and one due to the bizarre behaviour of a well known spreadsheet programme.

I have still not answered the question of whether this was efficient in ISO 9241-11 terms. I have to conclude that it was not, but this was partly due to my lack of knowledge on co-ordinating a complex combination of devices and software utilities. However, back when contacts were only held on mobile phone SIMs, the transfer would have taken a few minutes in a phone store. So, current usability problems here are due to the move to storing contacts in more comprehensive formats separately from a mobile phone's SIM. However, while there used to be

a more efficient option, most of us now make use of more comprehensive phone memory contacts, and thus the previous fast option was at the cost of the most primitive contact format imaginable. So while the activity was relatively inefficient, there are potentially compensating reasons for this.

The only genuine usability problems relate to the lack of feedback in the cloud-based email facility, the extracted phone number formats, and bizarre spreadsheet behaviour. However, I only explored the cloud email option following advice via Facebook. My experience of problems here was highly contextual. For the other two problems, if the second problem had not existed, then I would never have experienced the third.

There are clear issues of efficiency. At best this took twice as long as it should have once interleaved work and much valuable re-usable learning are discounted. However, the causes of this inefficiency are hard to pin-point within the complex socially shaped context within which I was working.

Effectiveness is easy to evaluate. I only transferred just under 50% of the contacts. Note how straightforward the analysis is here when compared to efficiency in relation to a complex product-service system.

On balance, you may be surprised to read that I was fairly satisfied. Almost 50% is better than nothing. I learned how to synchronise my iPhone via iTunes for the first time. I made good use of the waits in editing this encyclopaedia entry. I was not in any way *happy* though, and I remain dissatisfied over the phone number formats, inscrutable spreadsheet behaviour and mute import facility on a top three free email facility.

15.6.2 And the Moral of My Story Is: It was Worth It, on Balance

What overall judgement can we come to here? On a *binary* basis, the final data path that I chose *was usable*. An abandoned path was not, so I did encounter one unusable component during my attempt to transfer phone numbers. As regards a more realistic *extent* of usability (as opposed to binary usable vs. unusable), we must trade off factors such as efficiency, effectiveness and satisfaction against

each other. I could rate the final data path as 60% usable, with effective valuable learning counteracting the ineffective loss of over half of my contacts, which I had to subsequently enter manually. I could raise substantially this to 150% by adding the value of the resulting example for this encyclopaedia entry! It reveals the complexity of evaluating usability of interactions involving multiple devices and utilities. Describing usage is straightforward: judging its quality is not.

So, poor usability is still with us, but it tends to arise most often when we attempt to co-ordinate multiple digital devices across a composite ad-hoc product-service system. Forlizzi (2008) and others refer to these now typical usage contexts as *product ecologies*, although some (e.g., Harper et al. 2008) prefer the term *product ecosystems*, or product-service ecosystems (ecology is the discipline of ecosystems, not the systems themselves).

Components that are usable enough in isolation are less usable in combination. Essentialist positions on usability become totally untenable here, as the phone formats can blame the bizarre spreadsheet and vice-versa. The effects of poor usability are clear, but the causes are not. Ultimately, the extent of usability, and its causes in such settings, is a matter of interpretation based on judgements of the value achieved and the costs incurred.

Far from being an impasse, regarding usability as *a matter of interpretation* actually opens up a way forward for evaluating user experiences. It is possible to have robust interpretations of efficiency, effectiveness and satisfaction, and robust bases for overall assessments of how these trade-off against each other. To many, these bases will appear to be subjective, but this is not a problem, or at least it is far less of a problem than acting erroneously as if we have generic universal objective criteria for the existence or extent of usability in any interactive system. To continue any quest for such criteria is most definitely inefficient and ineffective, even if the associated loyalties to seventeenth century scientific values bring some measure of personal (subjective) satisfaction.

It is poor usability that focused HCI attention in the 1980s. There was no positive conception of *good* usability. Poor usability could degrade or even destroy the intended value of an interactive system. However, good usability can not *donate* value beyond that intended by a design team. Usability evaluation methods are focused on finding problems, not on finding successes (with the exception of Cognitive Walkthrough). Still, experienced usability practitioners know that an evaluation report should begin by commending the strong points of a design, but these are not what usability methods are optimised to detect.

Realistic relevant evaluations must assess *incurred costs* relative to *achieved benefits*. When transferring my contacts between phones, I experienced the following problems and associated costs:

1. Could not upload contacts into cloud email system, despite several attempts (cost: wasted 30 minutes)
2. Could not understand why I could not upload contacts into cloud email system (costs: prolonged frustration, annoyance, mild anger, abusing colleagues' company #1)
3. Could not initiate data transfer from Nokia phone first time, requiring experiments and laptop restart as advised by Nokia diagnostics (cost: wasted 15 minutes)
4. Over half of my contacts did not transfer (future cost: 30-60 further minutes entering numbers, depending on use of laptop or iPhone, in addition to 15 minutes already spent finding and noting missing contacts)
5. Deleting type prefixes (e.g., TEL CELL) from phone numbers in a spreadsheet resulted in an irreversible conversion to a scientific format number (cost: 10 wasted minutes, plus future cost of 30-60 further minutes editing numbers in my phone, bewilderment, annoyance, mild anger, abusing colleagues' company #2)

6. Had to set a wide range of synchronisation settings to restrict synchronisation to contacts (cost extra 10 minutes, initial disappointment and anxiety)
7. Being unable to blame Windows for anything (this time)!

By forming the list above, I have taken a position on what, in part, would count as poor usability. To form a judgement as to whether these costs were *worthwhile*, I also need to take a position on positive outcomes and experiences:

1. an opportunity to ask for, and receive, help from Facebook friends (realising some value of existing social capital)
2. a new email address gilbertcockton@... via an existing cloud computing account (future value unknown at time, but has since proved useful)
3. Discovered a semi-effective data path that transferred almost half of my contacts to my iPhone (saved: 30-60 minutes of manual entry, potential re-usable knowledge for future problem solving)
4. Learned about a nasty spreadsheet behaviour that could cause problems in the future unless I find out how to avoid it (future value potentially zero)
5. Learned about the Windows address book and how to upload new contacts as .csv files (very high future value – at the very least PC edits/updates are faster than iPhone, with very easy copy/paste from web and email)
6. Learned how to synchronise my new iPhone with my laptop via iTunes (extensive indubitable future value, repeatedly realised during the editing of this entry, including effortless extension to my recent new iPad)
7. Time to proof the previous draft of this entry and edit the next version (30 minutes of effective work during installs, restarts and uploads)

8. Sourced the main detailed example for this encyclopaedia entry (hopefully as valuable to you as a reader as to me as a writer: I've found it really helpful)

In many ways the question as to whether the combined devices and utilities were 'usable' has little value, as does any question about the extent of their combined usability. A more helpful question is whether the interaction was *worthwhile*, i.e., did the achieved resulting benefits justify the expended costs? *Worth* is a very useful English word that captures the relationship between costs and benefits: achieved *benefits* are (not) *worth* the incurred costs. Worth relates positive value to negative value, considering the balance of both, rather than, as in the case of poor usability, mostly or wholly focusing on negative factors.



FIGURE 15.20: Usability Verdict: Not Guilty.

Copyright status: Unknown (pending investigation). See section "Exceptions" in the copyright terms.

So, did my resulting benefits justify my expended costs? My answer is yes, which is why I was satisfied at the time, and am more satisfied now as frustrations fade and potential future value has been steadily realised. Given the two or three usability problems encountered, and their associated costs, it is quite clear that the interaction could have been more worthwhile (increased value at lower costs), but this position is more clear cut than having to decide on the extent and severity of usability problems in isolation. The interaction would have been more worthwhile in the absence of usability problems (but I would not have this example). It would have also been more worthwhile if I'd already known in advance how to extract contacts from a Nokia backup file in a format where they could have been uploaded into the Windows address book of contacts. Still better, the utility suite that came with my phone could have [had.csv](#) file import/export. Perhaps the best solution would be for phones to enable Windows to import contacts from them. Also, if I had used my previous laptop, the required phone utility suite was already installed and there should have been no initial connection problems. There were thus ways of reducing costs and increasing value that would not involve modifications to the software that I used, but would instead have replaced them all with *one simple purpose built tool*. None of the experienced usability problems would have been fixed. Once the complexity of the required data path is understood, it is clear that the best thing to do is to completely re-engineer it. Obliteration beats iteration here.

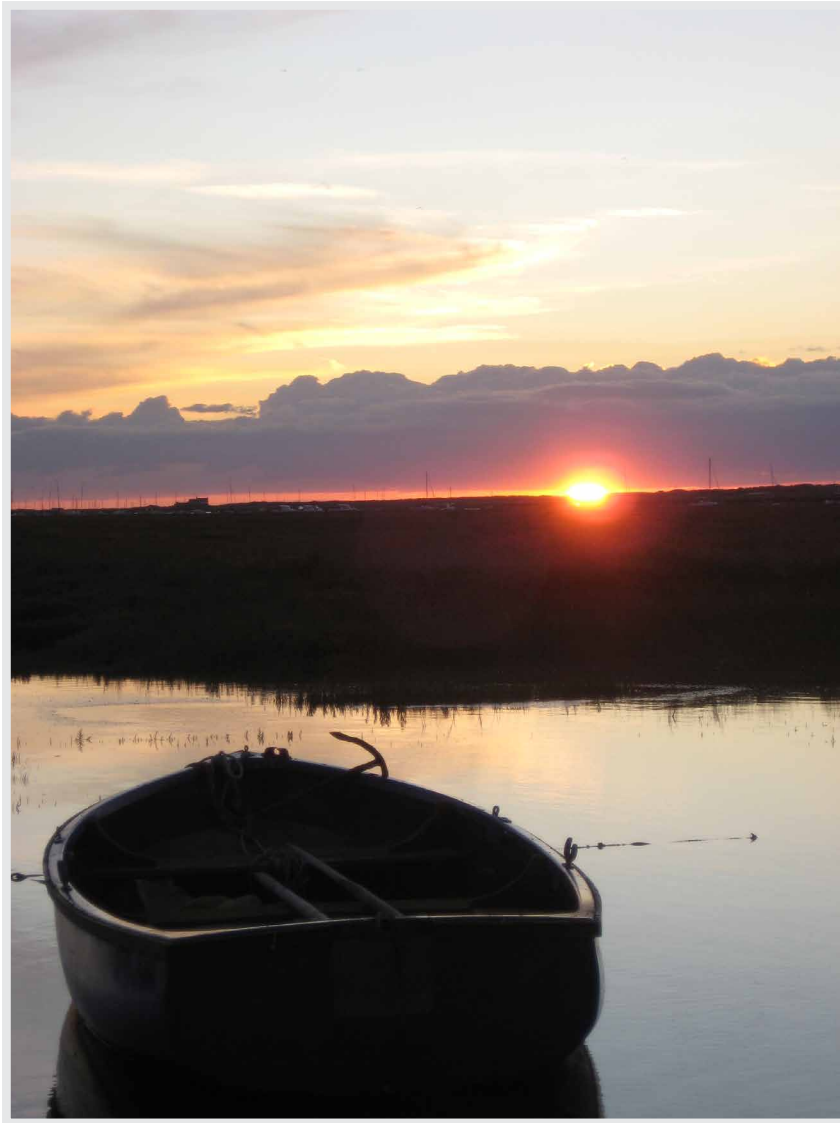


FIGURE 15.21: At the end of the day, you have to look at the big picture.

Copyright status: Unknown (pending investigation). See section “Exceptions” in the copyright terms.

15.6.3 Usability is Only One Part of a BIG Interaction Design Picture

By considering usability within the broader context of experience and outcomes, many dilemmas associated with usability in isolation disappear. This generalises

to human-centred design as a whole. In his book *Change by Design*, Tim Brown, CEO of IDEO, builds a compelling case for the human-centred practices of multi-disciplinary design teams. Even so, he acknowledges the lack of truly compelling stories that fully establish the importance of human-centred design to innovation, since these are undermined by examples of people regularly surmounting inconveniences (Brown 2009, pp.39-40), to which I have just added above. Through examples such as chaining bicycles to park benches, Brown illustrates *worth in action*: the benefit (security of bike) warrants the cost (finding a nearby suitable fixed structure to chain to). The problem with usability evaluations is that they typically focus on incurred costs without a balancing focus on achieved benefits. Brown returns to the issue of balance in his closing chapter, where design thinking is argued to achieve balance through its integrative nature (p.229).

Human-centred contributions to designs are just one set of inputs. Design success depends on effective integration of all its inputs. Outstanding design always overachieves, giving users/owners/sponsors far more than they were expecting. The best design is thus **B**alanced, **I**ntegrative and **G**enerous – or plain **BIG** for short. Usability needs to fit into the big picture here.

Usability evaluation finds usage problems. These need to be understood holistically in the full design context before possible solutions can be proposed. Usability evaluation cannot function in isolation, at least, not without isolating the usability function. Since the early 90s, usability specialists have had a range of approaches to draw on, which, once properly adapted, configured and combined can provide highly valuable inputs to the iterative development of interaction designs. Yet we continue to experience interaction design flaws, such as lack of instructive actionable feedback on errors and problems, which can and should be eliminated. However, appropriate use of usability expertise is only one part of the answer. A complete solution requires better integration of usage evaluation into other de-

sign activities. Without such integration, usability practices will continue to be met often with disappointment, distrust, scepticism and a lack of appreciation in some technology development settings (Iivari 2005).

This sets us up for a third alternative definition of usability that steers a middle course between essentialism and relationalism:

.....

“Usability is the extent of impact of negative user experiences and negative outcomes on the achievable worth of an interactive system. A usable system does not degrade or destroy achievable worth through excessive or unbearable usage costs.”

.....

Usability can thus be understood as a major facet of user experience that can reduce achieved worth through adverse usage costs, but can only add to achieved worth through the iterative removal of usability problems. Usability improvements reduce usage costs, but cannot increase the value of usage experiences or outcomes. In this sense, usability has the same structural position as Herzberg's (Herzberg 1966) *hygiene factors* in relation to his *motivator* factors.

15.6.4 From Hygiene Factors to Motivators



FIGURE 15.22: Usability is a negative hygiene factor, not a positive motivator.

Courtesy of Office for Emergency Management. U.S. Office of War Information. Domestic Operations Branch. Bureau of Special Services. Copyright: pd (Public Domain (information that is common property and contains no original authorship)).

Herzberg studied motivation at work, and distinguished positive *motivators* from negative *hygiene* factors in the workplace. Overt and sustained recognition at work is an example of a motivator factor, whereas inadequate salary is an example of a hygiene factor. Motivator factors can cause job satisfaction, whereas hygiene factors can cause dissatisfaction. Although referred to as Herzberg's two-factor theory (after the two groups of factors), it spans three valences: positive, neutral and negative. The absence of motivators does not result in dissatisfaction, but in the (neutral) absence of (dis)satisfaction. Similarly, the absence of negative hygiene factors does not result in satisfaction, but in the (neutral) absence of (dis)satisfaction. Loss of a positive motivator thus results in being unsatisfied, whereas loss of an adverse hygiene factor results in being undissatisfied! Usability can thus be thought of as an overarching term for hygiene factors in user experience. Attending to poor usability can remove adverse demotivating hygiene factors, but it cannot introduce positive motivators.

Positive motivators can be thought of as the opposite pole of user experience to poor usability. Poor usability demotivates, but good usability does not motivate, only positive experiences and outcomes do. The problem with usability as originally conceived in isolation from other design concerns is that it only supports the identification and correction of defects, and not the identification and creation of positive qualities. Commercially, poor usability can make a product or service uncompetitive, but usability can only make it competitive relative to products or services with equal value but worse usability. Strategically, increasing value is a better proposition than reducing usage costs in any market where overall usability is 'good enough' across competitor products or services.

15.7 FUTURE DIRECTION FOR USABILITY EVALUATION

Usage costs will always influence whether an interactive system is worthwhile or not. These costs will continue to be so high in some usage contexts that the

achieved worth of an interactive system is degraded or even destroyed. For the most part, such situations are avoidable, and will only persist when design teams lack critical human-centred competences. While originally encountered in systems developed by software engineers, poor usability is now also linked to design decisions imposed by some visual designers, media producers, marketing ‘suits’, interfering managers, inept committees, or in-house amateurs. Usability experts will continue to be needed to fix their design disasters.

15.7.1 Putting Usability in its Place

In well directed design teams, there will not be enough work for a pure usability specialist. This is evidenced by a trend within the last decade of a broadening from usability to user experience expertise. User experience work focuses on both positive and negative value, both during usage and after it. A sole focus on negative aspects of interactive experiences is becoming rarer. Useful measures of usage are extending beyond the mostly cognitive problem measures of 1980s usability to include positive and negative affect, attitudes and values, e.g., fun, trust, and self-affirmation. The coupling between evaluation and design is being improved by user experience specialists with design competences. We might also include interaction designers with user experience competences, but no interaction designer worthy of the name should lack these! Competences in high-fidelity prototyping, scripting and even programming are allowing user experience specialists firstly to communicate human context insights through working prototypes (Rosenbaum 2008), and secondly to communicate possible design responses to user experience issues revealed in evaluations.

Many user experience professionals have also developed specific competences in areas such as brand experience, trust markers, search experience/optimisation, usable security and privacy, game experience, self and identity, and human values. We can see two trends here. The first involves complementing human-

centred expertise with strong understandings of specific technologies such as search and security. The second involves a broadening of human-centred expertise to include business competences (e.g., branding) and humanistic psychological approaches (e.g., phenomenology, meaning and value). At the frontiers of user experience research, the potentials for exploiting insights from the humanities are being increasingly demonstrated (e.g., Bardzell 2011, Bardzell and Bardzell 2011, Blythe et al. 2011).

The extension of narrow usability expertise to broader user experience competences reduces the risk of inappropriate evaluation measures (Cockton 2007). However, each new user experience attribute introduces new measurement challenges, as do longer term measures associated with achieved value and persistent adverse consequences. A preference for psychometrically robust metrics must often be subordinated to the needs to measure specific value in the world, however and wherever it occurs. User experience work will thus increasingly require the development of custom evaluation instruments for experience attributes and worthwhile outcomes. Standard validated measures will continue to add value, but only if they are the right measures. There is however a strong trend towards custom instrumentation of digital technologies, above the level of server logs and low level system events (Rosenbaum 2008). Such custom instrumentation can extend beyond a single technology component to all critical user touch points in its embracing product-service ecosystem. For example, where problems arise with selecting, collecting, using and returning hired vans, it is essential to instrument the van hire depots, not the web site. Where measures relate directly to designed benefits and anticipated adverse interactions, this approach is known as *direct worth instrumentation* (Cockton 2008b).



FIGURE 15.23: The right instrumentation is crucial for worthwhile evaluation.

Courtesy of Michael Sandberg. Copyright: pd (Public Domain (information that is common property and contains no original authorship)).

Risks of inappropriate standard metrics arise when web site evaluations use the ready-to-hand measures of web server logs. What is easy to measure via a web server is rarely what is needed for meaningful relevant user experience evaluation. Thus researchers at Google (Rodden et al. 2010) have been developing a set of more relevant user experience ('HEART') measures to replace or complement existing log-friendly metrics ('PULSE' measures). The HEART measures are Happiness, Engagement, Adoption, Retention, and Task success. The PULSE measures are Page views, Uptime, Latency, Seven-day active users (i.e. count of unique users who used system at least once in last week), and Earnings. All PULSE measures are easy to make, but none are always relevant.

Earnings (sales) can of course be a simple and very effective measure for e-commerce as a measure of not one, but every, user interaction. As an example of the effectiveness of sales metrics, Sunderland University's Alan Woolrych (see Figure 15.7) has contributed his expertise to commercial usability and user experience projects that have increased sales by seven digits (in UK sterling), increasing sales in one case by at least 30%. Improved usability has been only one re-design input here, albeit a vital one. Alan's most successful collaborations involve marketing experts and lead business roles. Similar improvements have been recorded by collaborations involving user experience agencies and consultancies worldwide. However, the relative contributions of usability, positive user experience, business strategy and marketing expertise are not clear, and in some ways irrelevant. The key point is that successful e-commerce sites require *all* such inputs to be co-ordinated throughout projects.

There are no successful digital technologies without what might be regarded as usability flaws. Some appear to have severe flaws, and are yet highly successful for many users. Usability's poor reputation in some quarters could well be due to its focus on the negative at the expense of the positive. What matters is the resulting balance of worth as judged by all relevant stakeholders, i.e., not just users, but also, for example, projects' sponsors, service provision staff, service management, politicians, parents, business partners, and even the general public.



FIGURE 15.24: Evaluation of the Whereabouts Clock revealed unexpected benefits.

Copyright © Microsoft Research in Cambridge and Abigail Sellen, Shahram Izadi, Richard Harper, and Rachel Eardley. All Rights Reserved. Reproduced with permission. See section “Exceptions” in the copyright terms.



FIGURE 15.25: The Whereabouts Clock in its usage context.

Copyright © Microsoft Research in Cambridge and Abigail Sellen, Shahram Izadi, Richard Harper, and Rachel Eardley. All Rights Reserved. Reproduced with permission. See section “Exceptions” in the copyright terms.

Evaluation needs to focus on both positives and negatives. The latter need to be identified and assessed for their impact on achieved worth. Where there are unacceptable adverse impacts, re-design and further evaluation is needed to confirm that unintended negative experiences and/or outcomes have been ‘designed out’. However, evaluation misses endless opportunities when it fails to identify unintended positives experiences and/or outcomes. Probe studies have proved to be highly effective here, identifying positive appropriative use that was completely unanticipated by design teams (e.g., Brown et al. 2007, Gaver et al. 2008). It is refreshing to encounter evaluation approaches that identify unexpected successes as well as unwanted failures. For example, the evaluation of the Whereabouts

Clock (Brown et al. 2007) revealed one boy's comfort at seeing his separated family symbolically united on the clock's face.

Designers and developers are more likely to view evaluation positively if it is not overwhelmingly negative. Also, this spares evaluators from ritually starting their reports with a 'few really good points about the design' before switching into a main body of negative problems. There should always be genuine significant positive experiences and outcomes to report.

Evaluation becomes more complicated once positive and negative phenomena need to be balanced against each other across multiple stakeholders. *Worth* has been explored as an umbrella concept to cover all interactions between positive and negative phenomena (Cockton 2006). As well as requiring novel custom evaluation measures, this also requires ways to understand the achievement and loss of worth. There have been some promising results here with novel approaches such as *worth maps* (Cockton et al. 2009a, Cockton et al. 2009b, Otero and José 2009). Worth maps can give greater prominence to system attributes while simultaneously relating them to contextual factors of human experiences and outcomes. Evaluation can focus on worth map elements (system attributes, user experience attributes, usage outcomes) or on the connections between them, offering a practical resource for moving beyond tensions between essentialist and relational positions on software quality.

Worth-focused evaluation remains underdeveloped, but will focus predominantly on outcomes unless experiential values dominate design purpose (as in many games). Where experiential values are not to the fore, detailed evaluation of user interactions may not be worthwhile if products and services have been shown to deliver or generously donate value. Evaluation of usage could increasingly become a relatively infrequent diagnostic tool to pinpoint where and why worth is being degraded or destroyed. Such a strategic focus is essential now that we have new data collection instruments such as web logs and eye tracking that gather mas-

sive amounts of data. Such new weapons in the evaluation arsenal must be carefully aimed. A 12-bore shotgun scattershot approach cannot be worthwhile for any system of realistic complexity. This is particularly the case when, as in my personal example of phone contacts transfer, whole product ecologies (Forlizzi 2008) must be evaluated, and not component parts in isolation. When usage within such product ecologies here is mobile, intermittent and moves through diverse social contexts, it becomes even more unrealistic to evaluate every second of user interaction.

In the future, usability evaluation will be put in its place. User advocates will not be given free rein to berate and scold. They will become integral parts of design teams with **B**alanced, **I**ntegrated and **G**enerous (BIG!) design practices. It's time for all the stragglers in usability evaluation to catch up with the **BIG** boys and girls. Moaning on the margins about being ignored and undervalued is no longer an option. Usability must find its proper place within interaction design, as an essential part of the team, but rarely King of the Hill. The reward is that usability work could become much more rewarding and less fraught. That has got to be worthwhile for all concerned.



Courtesy of Tety. Copyright: CC-Att-2 (Creative Commons Attribution 2.0 Unported).



Courtesy of Joel Rogers. Copyright: pd (Public Domain (information that is common property and contains no original authorship)).

FIGURE 15.26 A-B: From Solo Specialist to Team Member: User Experience as an integrated part of design teams.

15.8 WHERE TO LEARN MORE

HCI Remixed (Erickson and McDonald 2008) is an excellent collection of short essays on classic HCI books and papers, plus other writing that has influenced leading HCI researchers. It contains a short essay (Cockton 2008a) on the Whiteside et al. (1988) classic, and many more of interest.

There is a short account of BIG Design in Cockton, G. Design: [BIG and Clever](#), *Interfaces Magazine*, 87, British Interaction Group, ISSN 1351-119X 2011, 5-7

Sears' and Jacko's HCI Handbook (Sears and Jacko 2007) is a very comprehensive collection of detailed chapters on key HCI topics. The 3rd edition will be published in 2012. There are chapters on user testing, inspection methods, model-based methods and other usability evaluation topics.

Darryn Lavery prepared a set of tutorial materials on inspection methods in the 1990s that are still available:

- ▶ Lavery, D., Cockton, G., and Atkinson, M. P. 1996. Heuristic Evaluation: Usability Evaluation Materials, Technical Report TR-1996-15, University of Glasgow. Accessed 15/9/11 at http://www.dcs.gla.ac.uk/asp/materials/HE_1.0/materials.pdf
- ▶ Lavery, D., Cockton, G., and Atkinson, M. P. 1996. Heuristic Evaluation for Software Visualisation: Usability Evaluation Materials, Technical Report TR-1996-16, University of Glasgow, 1996. Accessed 15/9/11 at http://www.dcs.gla.ac.uk/asp/materials/SVHE_1.0/materials.pdf
- ▶ Lavery, D., Cockton, G., and Atkinson, M. P. 1996. Cognitive Dimensions: Usability Evaluation Materials, Technical Report TR-1996-17, University of Glasgow. Accessed 15/9/11 at http://www.dcs.gla.ac.uk/asp/materials/CD_1.0/materials.rtf
- ▶ Lavery, D., and Cockton, G. 1997. Cognitive Walkthrough: Usability Evaluation Materials, Technical Report TR-1997-20, Department of Computing Science, University of Glasgow. Edited version available 15/9/11 as <http://www.dcs.gla.ac.uk/~pdg/teaching/hci3/cwk/cwk.html>

Europe's COST programme has funded two large research networks on evaluation and design methods. The MAUSE project (COST Action 294, 2004-2009) focused on maturing usability evaluation methods. The TwinTide project (COST Action

ICo904, 2009-2013) has a broader focus on design and evaluation methods for interactive software. There are several workshop proceedings on the MAUSE web site (www.cost294.org), including the final reports, as well as many publications by network members on the associated MAUSE digital library. The TwinTide web site (www.twintide.org) is adding new resources as this new project progresses.

The [Usability Professionals Association](#), UPA, have developed some excellent resources, especially their open access on-line [Journal of Usability Studies](#). Their [Body of Knowledge](#) project, BOK, also has created a collection of resources on evaluation methods that complement the method directory prepared by [MAUSE WG1](#). Practically minded readers may prefer BOK content over more academically oriented research publications.

Jakob Nielsen has developed and championed discount evaluation methods for over two decades. He co-developed Heuristic Evaluation with Rolf Mohlich. Jakob's www.useit.com web site contains many useful resources, but some need updating to reflect some major developments in usability evaluation and interaction design over the last decade. For example, in the [final version of his heuristics](#) some known issues with Heuristic Evaluation are not covered. Even so, the critical reader will find many valuable resources on www.useit.com. Hornbæk (2010) is a very good source of critical perspectives on usability engineering, and should ideally be read alongside browsing within www.useit.com.

The American Association for Computing Machinery (ACM) sponsors many key HCI conferences through its SIGCHI special interest group. The annual CHI (Computer-Human Interaction) conference is an excellent source for research papers. There is no specialist ACM conference with a focus on usability evaluation, but the SIGCHI DIS (Designing Interactive Systems) conference proceedings and the DUX (Designing for User Experiences) conference proceedings do contain some valuable research papers, as does the SIGCHI CSCW conference series. The SIGCHI UIST conference (Symposium on User Interface Software and Technol-

ogy) often includes papers with useful experimental evaluations of innovative interactive components and design parameters. All ACM conference proceedings can be accessed via the [ACM Digital Library](#). Relevant non ACM conferences include UPA (The Usability Professionals' Association international conference), ECCE (the European Conference on Cognitive Ergonomics), Ubicomp (International Conference on Ubiquitous Computing), INTERACT (the International Federation for Information Processing Conference on Human-Computer Interaction) and the British HCI Conference series. UPA has a specific practitioner focus on usability evaluation. Most HCI publications are indexed on www.hcibib.org. In November 2011, a search for *usability evaluation* found almost 1700 publications.

15.9 ACKNOWLEDGEMENTS

I have been immensely fortunate to have collaborated with some of the most innovative researchers and practitioners in usability evaluation, despite having no serious interest in usability in my first decade of work in Interaction Design and HCI!

One of my first PhD students at Glasgow University, Darryn Lavery, changed this through his struggle with what I had thought was going to be a straightforward PhD on innovative inspection methods. Darryn exposed a series of serious fundamental problems with initial HCI thinking on usability evaluation. He laid the foundations for over a decade of rigorous critical research through his development of conceptual critiques (Lavery et al. 1997), problem report formats (Lavery and Cockton 1997), and problem extraction methodologies (Cockton and Lavery 1999). From 1998, Alan Woolrych, Darryn Lavery (to 2000), myself and colleagues at Sunderland University built on these foundations in a series of studies that exposed the impact of specific resources on the quality of usability work (e.g., Cockton et al. 2004), as well as demonstrat-

ing the effectiveness of these new understandings in Alan's commercial and e-government consultancies. Research tactics from our studies were also used to good effect from 2005-2009 by members of WG2 of COST Action 294 (MAUSE - see *Where to learn more* above), resulting in a new understanding of evaluation methods as usability work that adapts, configures and combines methods (Cockton and Woolrych 2009). COST's support for MAUSE and the the follow on TwinTide Action (*Where to learn more*, below) has been invaluable for maintaining a strong focus on usability and human-centred design methods in Europe. Within Twintide, Alan Woolrych, Kasper Hornbæk, Erik Frøkjær and I have applied results from MAUSE to the analysis of Inspection Methods (Cockton et al. 2012), and more broadly within the broader context of usability work (Woolrych et al. 2011).

Nigel Bevan, a regular contributor to MAUSE and TwinTide activities has provided helpful advice, especially on international standards. Nigel is one of many distinguished practitioners who have generously shared their expertise and given me feedback on my research. At the risk of omission and in no particular order, I would particularly like to acknowledge generous sharing of knowledge of and insights on usability and emerging evaluation practices by Tom Hewett, Fred Hansen, Jonathan Earthy, Robin Jeffries, Jakob Nielsen, Terry Roberts, Bronwyn Taylor, Ian McClelland, Ken Dye, David Caulton, Wai On Lee, Mary Czerwinski, Dennis Wixon, Arnie Lund, Gaynor Williams, Lynne Coventry, Jared Spool, Carolyn Snyder, Will Schroeder, John Rieman, Giles Colborne, David Roberts, Paul Englefield, Amanda Prail, Rolf Mohlich, Elizabeth Dykstra-Erickson, Catriona Campbell, Manfred Tscheligi, Verena Giller, Regina Bernhaupt, Lucas Noldus, Bonnie John, Susan Dray, William Hudson, Stephanie Rosenbaum, Bill Buxton, Marc Hassenzahl, Carol Barnum, William Hudson, Bill Gaver, Abigail Sellen, Jofish Kaye, Tobias Uldall-Espersen, John Bowers and Elizabeth Buie. My apologies to anyone who I have left out!

15.10 COMMENTARY BY DAVID A. SIEGEL

How to [cite this commentary in your report](#)

David A. Siegel



© *David A. Siegel*

David is a well-known user experience researcher and consultant, and co-owner of Dray & Associates, Inc. He specializes in using field user studies and contextual research, and naturalistic usability evaluation to help guide product concept, strategy, and interaction design. He has published and taught on a variety of user-centered design topics, including many workshops and tutorials ...

David A. Siegel

David A. Siegel is a member of The Interaction Design Foundation

I appreciate the opportunity to comment on Gilbert Cockton's chapter on usability. My comments come from the perspective of someone who has practiced user experience (UX) research of many types as a consultant. Although I have done my

share of usability evaluations, almost all of my work currently consists of in vivo contextual research, with a focus on finding ways to increase value to the user. The product teams I work with often include internal usability specialists, and I well understand their roles within their teams and the challenges they face. Finally, my prior career as a psychologist has given me a very healthy respect for the difficulties of measuring and understanding human behavior in a meaningful way, and impatience with people who gloss over these challenges.

To begin with points of agreement, I applaud Gilbert's emphasis on the need to consider usability in the context of all other factors that influence the value people obtain from interactive products. I also agree with his critique of the methodological limitations of laboratory usability evaluation. I could not agree more that contextual research is usually much more powerful than laboratory usability evaluation as an approach to understanding the user experience holistically and to gaining insights that will drive UX design towards greater overall value. I also agree with Gilbert's call to usability professionals to focus on the larger issues.

With this said, however, I have a number of concerns about the chapter's portrayal and critique of usability as an inherently limited, marginal contributor to development of great products. In regard to practice, there are many gradations of skill and wisdom, and some unknown proportion of usability practitioners may deserve to be confronted with the criticisms Gilbert raises. However, I question the idea that these criticisms are true of usability practice in principle. I believe that most mature usability practitioners are aware of the issues he raises, would agree with many of his points, and work hard to address them in various ways. In the discussion that follows, I will present an alternate view of usability's role as a fundamental contributor to product value. This requires considering usability at two levels: as an abstract concept and as a field of practice.

First, one note on terminology: throughout this commentary I use the word "product" to refer to anything that is being designed for interactive use, be it software, website, system, or device, or any new features of these.

15.10.1 Usability and Value as Abstract Constructs

It has become commonplace to emphasize a distinction between usability and value, and also to claim that “experience” has superseded usability. This treats usability as though it is distinct from both of these other concepts. Even though usability is generally acknowledged to be important, it is portrayed as quite subordinate. In Gilbert’s chapter, this is reflected in the idea that usability is merely a “hygiene factor,” the absence of which can block the delivery of value or reduce it by adding to costs, but one which can never go beyond neutral as a contributor to value. In my view, this greatly understates the contribution of usability to value. The two concepts are far more intertwined than this. Attempts to abstract value from usability are just as flawed as the reverse.

The notion that ease of use is a separate issue from value, although one that affects it, has much face validity. It seems to make sense to think of value as a function of benefit somehow related inversely with costs, with usability problems counted in the costs column. Unfortunately, this is consistent with the notion of usability as “a feature,” something that makes usability professionals cringe, just as the idea of design as the “lipstick” applied to a product in the last stage makes designers cringe. In my view, usability divorced from value is as undefined as the sound of one hand clapping. Usability can only be defined in the context of benefit. By this I do not mean benefit in principle, but rather the benefit anticipated by or experienced by the user. At one level, this is because usability and experienced benefit interact in complex ways. But beyond this, there are many products where usability is itself the primary value proposition. In fact, the central value proposition of most technological tools is that they make something of value easier to achieve than it used to be. A mobile phone has value because its portability enables communication while mobile, and its portability matters because it makes it more usable when mobile.

In another example, a large medical organization I am familiar with recently adopted a new, integrated digital medical record system. Initially, there was a

great deal of grumbling about how complex and confusing it was. I saw the classic evidence of problems in the form of notes stuck on computer monitors warning people not to do seemingly intuitive things and reminding them of the convoluted workarounds. However, more recently, I have heard nurses make comments about the benefit of the system. Doctors' orders are entered electronically and made automatically available to the appropriate departments. As a result, patients now can come to the clinic for a follow up laboratory test without having to remember to bring a written copy of the lab order. "Usability" is not simply the issue of whether doctors can figure out how to enter the order in the system and direct it to the lab rather than the ophthalmology department, although that is part of it. The benefit has to do with its overall success in reducing the usability problems of an earlier process that used to be difficult to coordinate and error prone, and this increase in usability only matters because it is delivering a real benefit.

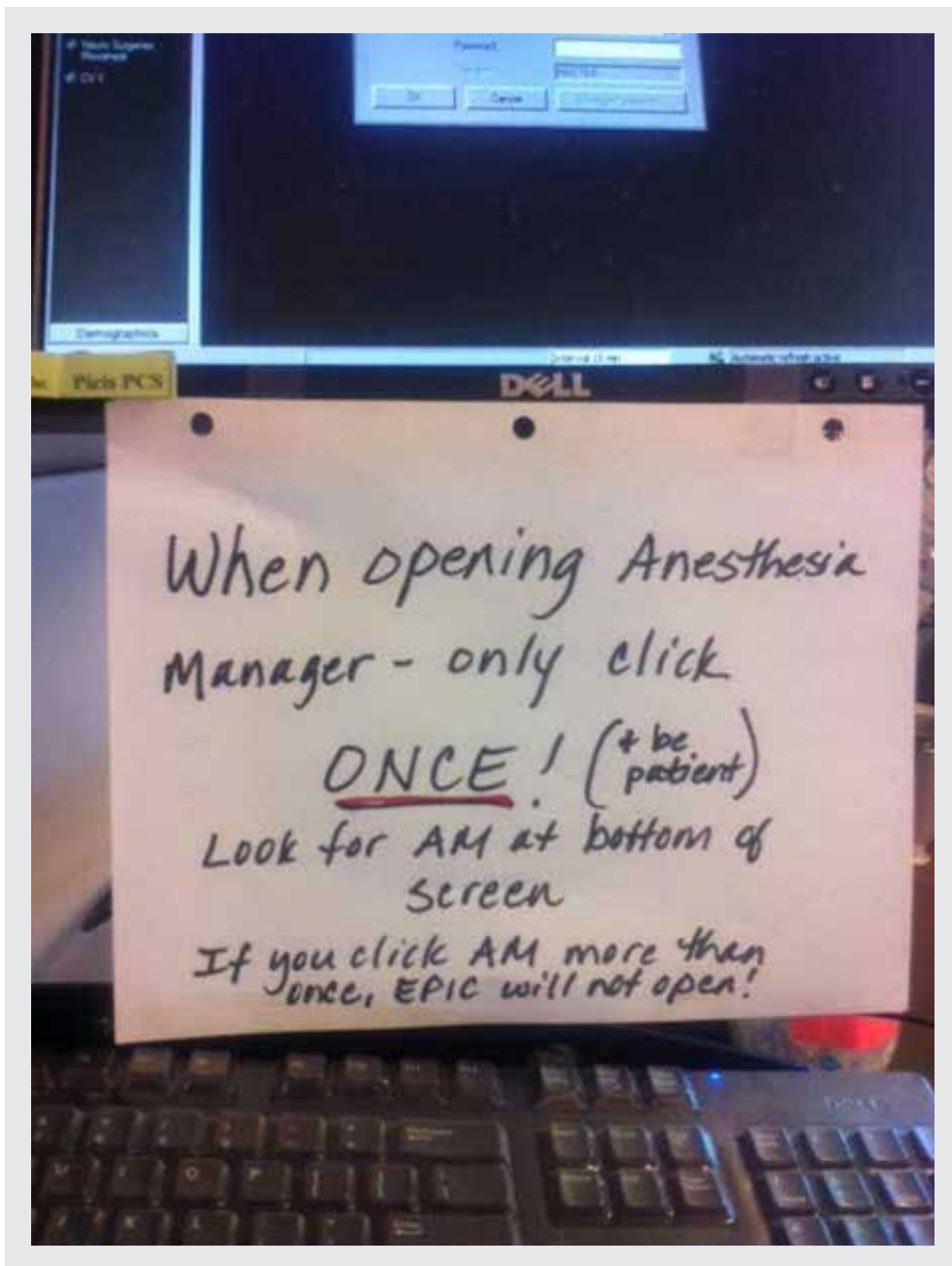


FIGURE 15.1

Copyright © David Siegel - Dray and Associates. All Rights Reserved. Reproduced with permission. See section "Exceptions" in the copyright terms.

Sometimes, usability seems detached from value when the goal is fulfilled at the end of a sequence of steps, but the steps along the way are confusing. However, it can be the separation from the experience of value that creates the usability problem. For example, if people trying to book an online hotel reservation get lost in preliminary screens where they first have to create an account, we might see usability as only relevant to the cognitive aspects of the sign up process, and as mere hygiene factors. But when users become disoriented because they do not understand what a preliminary process has to do with their goal, it can be precisely because they cannot see the value of the preliminary steps. That is, they can't see how the steps contribute to something they care about, and lead them towards their goal. If they did, the subparts of the process would both be more understandable and would acquire value of their own, just as a well-designed hammer gains value not simply in its own right, but because it is understood as a more effective tool for driving nails (which are valued because of the value of the carpentry tasks they enable, and so on.) This is simultaneously a usability problem and an "experience of value problem." For this reason, a common challenge of usability is to convey to users that they are making progress towards an outcome that they value.

For example, in one product that I worked on, users were offered the opportunity to enroll for health insurance benefits that claimed to be highly personalized. In addition to setting different benefit levels for different members of their families, users could compose their own preferred networks of medical specialists, for which they would receive the highest reimbursement levels. Unfortunately, the actual user experience did not appear to live up to this. As soon as the user entered identifying information, the system applied defaults to all the decisions that the user was supposedly able to personalize. It only fulfilled its value proposition of personalization by allowing the user to "edit" the final configuration—13 screens into the process. Along the way, the user experienced the sense that decisions were being imposed. There was not even an indication to the user that the

opportunity to make personal choices was coming eventually. Unfortunately, the system did not start by asking the user which choices mattered to them and what their preferences were, so it could factor these things in before presenting a result to the user.

How should we construe this? As a usability problem? As a problem in delivery of value? As a failure in the design of a user experience? It is all of these at the same time. The discrepancy from the expected perception of value is a primary cause of the confusion users felt. None of these constructs (usability, value, experience) can be defined without incorporating the others. If we parse and remove the meaning that we can attribute to any of them, we drain the meaning from the others. Disputes about which is the legitimate language to describe them are at best just ways to emphasize different faces of the same phenomenon, and at worst semantic quibbling. This means that usability is something more than just another item to add into the costs column when we weigh them against benefits to arrive at value. It also means we can't answer the question of whether something is usable without also answering the question, "What matters?"

15.10.2 Usability Practice in Product Development

While Gilbert and I may agree on the need for a more holistic focus on user experience, we may disagree about whether usability in practice actually takes this holistic view. Reducing the profession to a particular type of laboratory evaluation makes it seem limited and can raise questions about its relevance. While as I said, I agree with Gilbert's critique of the methodological limitations of this approach, the profession is far broader and more diverse than this. Furthermore, even despite its limitations, traditional usability evaluation often contributes significant value in the product development context, at least when practiced by reflective professionals. Below, I comment on some of the major issues Gilbert raises with regard to usability practice.

15.10.2.1 Is ‘Ease of Use’ still relevant?

Although some interaction design patterns have become established, and an increasing number of users have gained generalizable skills in learning a variety of new interaction patterns, this does not mean that ease of use as an issue has gone away or even declined in importance. For several reasons, it makes more sense to see the spectrum of usability issues to be addressed as having evolved. First, the spectrum of users remains very large and is constantly expanding, and there are always some at an entry level. Second, although with experience users may gain knowledge that is transferrable from one family of products to another, this can be both an asset and a source of confusion, because the analogies among product designs are never perfect. Third, as innovation continues to create new products with new capabilities, the leading edge of UX keeps moving forward. On that leading edge, there are always new sets of design challenges, approaches, and tradeoffs to consider. Finally, the world does not consist only of products intended to create experiences for their own sake as opposed to those that support tasks (a distinction that is not necessarily so clear). Products that are designed to facilitate and manage goal-oriented tasks and to support productivity continue to have a tremendous impact on human life, and we have certainly not learned to optimize ease of interaction with them. Finally, usability is continually driven forward by competition within a product domain.

Another claim in the chapter that suggests limited relevance for usability is that good product teams do not need a dedicated usability person. This is too simplistic. Of course, a designated usability person does not create usability single handedly. That is the cumulative result of everything that goes into the product. However, how much specialized work there is for a usability person depends on many factors. We need to take into account the variability among ways that product teams can be structured, the magnitude of the UX design challenges they face in their product space, the complexity of the product or family of inter-related products that the usability person supports, how incremental versus innovative

the products are, what the risk tolerance is for usability problems, how heterogeneous the user population and user contexts are, how much user persistence is needed for usage to be reinforced and sustained by experiences of value, etc. The simplistic statement certainly does not address the fact that some usability work takes more effort to carry out than others. To do realistic research with consumers is generally much easier than doing realistic research inside enterprises.

As a matter of fact, in actual practice teams often do not have usability professionals assigned to them full time, because these people often support multiple product teams, in a matrix organizational structure. There are benefits to this in terms of distributing a limited resource around the company. But there are also drawbacks. This structure often contributes to the usability person being inundated with requests to evaluate superficial aspects of design. It can also exclude the usability person from integrative discussions that lead to fundamental aspects of product definition and design and determine the core intended value of the product. Some usability people may accept this limited role complacently and passively respond to team requests, in the hopes of providing “good service,” but many others recognize the challenges of this role structure and work very hard to get involved with deeper issues of value, exactly as Gilbert urges them to.

15.10.2.2 Do usability professionals only focus on cognition?

Several points in Gilbert’s critique of practice are based on a limited view of what usability people do. It is true that laboratory usability evaluation typically does try to isolate cognitive factors by treating the users goals and motivation as givens, rather than attempting to discover them. Often, it is the fit of the assumed goal that is in question, and that makes the biggest difference in user experience.

But many usability professionals spend a great deal of time doing things other than laboratory tests, including, increasingly, fundamental in context user research. For many years, usability evaluation has served as a platform to promote systematic attention to deeper issues of value to the user. Many usability professionals deeply understand the complex, entangled relationship between

ease of use and value, and work to focus on broad questions of how technology can deliver experienced value. Some usability people have succeeded in getting involved earlier in the design process when they can contribute to deeper levels of decision-making. This has led to their involvement in answering questions about value, like “What will matter to the user?” or “What will influence whether people will really adopt it?” rather than only asking, “Could the user, in principle, figure out how to do it if they wanted to?” There are certainly people who are narrow specialists in a particular set of techniques focused on ease of use, but they do not own the definition of the field, and specialization per se is not bad.

15.10.2.3 What can usability people contribute?

Gilbert is correct that UX skills are increasingly distributed across roles. He lists a number of such skills, but missing from the list is the skill of doing disciplined research to evaluate evidence for the assumptions, claims, beliefs, or proposed designs of the product team, whether these are claims about what people need and will value, or whether a particular interface design will enable efficient performance.

Gilbert points out that there is no cookbook of infallible usability approaches. This is not a surprise, and indeed, we should never have expected such a thing. Such cookbooks do not exist for any complex field, and there is no way to guarantee that a practical measurement approach captures the core meaning of a complex construct. I do agree wholeheartedly with Gilbert when he points out the many factors that can complicate the process of interpreting usability findings due to this lack of a cookbook of infallible methods and the presence of many confounds. These issues argue for the need for greater professionalism among usability practitioners, not for the downgrading of the profession or marginalizing it on the periphery of the product development team. Professionalism requires that practitioners have expert understanding of the limitations of methods, expertise in modifying them to address different challenges, the dedication to continually advance their own processes, and the skill to help drive the evolution of practice over time. At a basic level, mature usability professionals recognize that results from a single evalua-

tion do not give an absolute measure of overall usability. They are careful about overgeneralizing. They at least attempt to construct tasks that they expect users will care about, and attempt to recruit users who feel will engage realistically with the tasks. They wrestle with how best to achieve these things given the constraints they work under. Those who do not recognize the challenges of validity, or who apply techniques uncritically are certainly open to criticism, or should be considered mere technicians, but, again, they do not represent the best of usability practice.

In the absence of scientific certainty, where is the value of usability practice? In the product development context, this should not be judged by how well usability meets criteria of scientific rigor. It is more relevant to ask how it compares to and compliments other types of evidence that are used as a basis for product definition, audience targeting, functional specification, and design decisions. This means we need to consider usability's role within the social and political processes of product development.

Membership in product teams often requires allegiance to the product concept and design approach. Sometimes, demonstrations of enthusiasm are a prerequisite for hiring. Often, it is risky for team members to challenge the particular compromises that have been made previously to adapt the product to various constraints or a design direction that has become established, since these all have vested interests behind them. In this context, the fact that usability methods (or approaches as Gilbert rightfully calls them) are scientifically flawed does not mean they are without value. It is not as though all the other streams of influence that affect product development are based on solid science while usability is voodoo. When you consider the forces that drive product development, it is clear that subjective factors dominate many of them, for example:

- ▶ Follow the leader design mentality
- ▶ Imperfect and sometimes cherry-picked market research data
- ▶ Internal politics
- ▶ Impressions derived from other faulty approaches to testing the product's appeal

- ▶ Beliefs about who future users will be

Product decisions are also deeply influenced by legitimate considerations that are difficult to evaluate objectively, much less to weigh against each other, such as:

- ▶ How to deal with legacy issues and the difficulty of innovating with an installed base
- ▶ How to weigh engineering feasibility and manufacturing cost against other considerations
- ▶ Market timing
- ▶ The business need to differentiate from the competition, even if the competition has a better approach
- ▶ The need to promote the inherent advantages of your technical strengths and downplay its limitations

In this context, a discipline that offers structured and transparent processes for introducing evidence-based critical thinking into the mix adds value, even though its methods are imperfect and its evidence open to interpretation. Sometimes, usability evaluation is a persuasive tool to get product teams to prioritize addressing serious problems that everyone knew existed, but that could not receive focus earlier. Sometimes this is needed to counterbalance the persuasive techniques of other disciplines, which may have less scientific basis than usability. Sometimes usability results provide a basis to resolve disputes that have no perfect answer and that have previously paralyzed teams. And sometimes they have the effect of triggering discussions about controversial things that would otherwise have been suppressed.

15.10.2.4 Does usability contribute to innovation?

Sometimes, usability in practice is portrayed as a mere quality assurance process, or as Gilbert says, a hygiene factor. It is often equated with evaluation as distinct from discovery and idea generation. In many ways, this is a false distinction.

Careful evaluation of what exists now can inspire invention and direct creativity towards things that will make the most difference. Practices like rapid iterative design reflect efforts to integrate evaluation and invention. Practices that are considered to be both discovery and invention processes, like contextual design, fall on a continuum with formative usability evaluation and naturalistic evaluation in the usage context. Of course, usability professionals differ in their skills for imagining new ways of meeting human needs, envisioning new forms of interactive experience, or even generating multiple alternative solutions to an information architecture problem or interface design problem. Some may lack these skills. However, the practice of usability is clearly enhanced by them. Those who can integrate evaluation and invention can add more value to the product development process and can help ensure usability/value in the ultimate product.

15.10.3 Conclusion

Certainly one can find examples of bad usability practice, and I cannot judge what other people may have encountered. Of course, there is also a lot of bad market research, bad design, bad business decision-making, bad engineering, and bad manufacturing. Let us not define the field based on its worst practice, or even on its lowest-common denominator practice. Failure to take into account the kinds of confounds Gilbert identifies is indeed bad practice because it will lead to misleading information. Handing over to a team narrow findings, minimally processed, excludes the usability practitioner from the integrative dialogue in which various inputs and courses of action are weighed against each other, and from the creative endeavor of proposing solutions. This will indeed limit usability practitioners to a tactical contributor role and will also result in products that are less likely to provide value for the users.

Finally, to any usability practitioners who think that usability is some kind of essence that resides in a product or design, and that can be objectively and accurately measured in the lab: Stop it. If you think that there is a simple definition of ease of use that can be assessed in an error-free way via a snapshot with an

imperfect sample of representative users and simulated tasks: Stop it. If you think usability does not evolve over time or interact with user motivation and expectations and experience of benefit: Stop it. If you think that ease of use abstracted from everything else is the sole criterion for product success or experienced value: Stop it! If you think you are entitled to unilaterally impose your recommendations on team decision-making: Stop it. You are embarrassing the profession!

15.11 COMMENTARY BY STEPHANIE ROSENBAUM

How to [cite this commentary in your report](#)

Stephanie Rosenbaum



© *Stephanie Rosenbaum*

Stephanie Rosenbaum is founder and president of TecEd, Inc., a 15-person consultancy specializing in usability research and information design. Now in its 41st year of business, TecEd maintains four offices: Ann Arbor, Michigan; Palo Alto, California; Rochester, New York; and Milwaukee, Wisconsin. Clients include Microsoft, Cisco Systems, Sun Microsystems, Yahoo!, eBay, ...

Stephanie Rosenbaum

Stephanie Rosenbaum is a member of The Interaction Design Foundation

Gilbert Cockton's chapter on Usability Evaluation includes a great deal of valuable, interesting, and well-reasoned information. However, its presentation and focus could be more helpful to the Interaction-Design.org audience. If some of this audience consists of practitioners—especially less-experienced practitioners—then Cockton is not speaking to their needs.

Who are you, the readers of this chapter? One of Interaction-Design.org's tag lines says "making research accessible," and its mission statement talks about producing top-grade learning materials to benefit industry and academia. It seems likely that many of you are practitioners in business, technology, healthcare, finance, government, and other applied fields.

As founder and CEO of a user experience consultancy, I find that most people—in both industry and academia—want to learn about usability evaluation as part of their goal to design better products, websites, applications, and services. Especially in industry, philosophical debates about points of definition take second place to the need to compete in the marketplace with usable, useful, and appealing products.

This is not a new observation. As early as 1993, Dumas and Redish [1] pointed out that we don't do usability testing as a theoretical exercise; we do it to improve products. Unfortunately, Cockton loses sight of this key objective and instead forces his readers to follow him as he presents, and then demolishes, an increasingly complex series of hypotheses about the meaning of usability. The danger of this approach is that a casual reader—especially one with a limited command of English—may learn from the chapter precisely the ideas Cockton eventually disproves.

For example, Cockton begins his chapter with several "ideal" propositions about usability as an inherent property of software that can be measured accurately by well-defined methods, regardless of the context of use. Yet as he states later in the chapter, the contextual nature of design—and thus usability—has long been known, not only in the 1988 Whiteside et al. publication Cockton mentions,

but also in the work of Gould and Lewis in the 1970s, published in their seminal 1985 article [2].

Throughout his chapter, Cockton continues to build and revise his definitions of usability. The evolution of these definitions is interesting to me personally because of my academic degrees in the philosophy of language. But reading this chapter gives my colleagues in industry only limited help in their role as user experience practitioners conducting usability evaluations of products under development.

In Section 15.1.1—and implicitly throughout the chapter—Cockton associates usability primarily with interactive software. The concept of usability has never applied only to software; ease of use is important to all aspects of our daily life. In 1988, Don Norman wrote about the affordances of door handles [3]. Giving a guest lecture on usability evaluation, I was surprised and impressed by an attendee’s comment describing how his company conducted usability testing of electric table saws.

In Section 15.1.2, Cockton describes “a dilemma at the heart of the concept of usability: is it a property of systems or a property of usage?” Why can’t it be both? Interactive systems are meaningless without users, and usage must be of something.

The discussion of damaged merchandise (invalid usability methods) in Section 15.2.1 misses the point that most usability work involves applied empirical methods rather than formal experiments. There are—and will always be—evaluator effects in any method which has not been described in enough detail to replicate it. The fact that evaluator effects exist underlines the importance of training skilled evaluators.

I am concerned that Cockton is emphasizing a false dichotomy when he says, “If software can be inherently usable, then usability can be evaluated solely through direct inspection. If usability can only be established by considering usage, then indirect inspection methods (walkthroughs) or empirical user testing

methods must be used to evaluate.”

There need not be a dichotomy between essentialist ontologies and relational ontologies of usability as described in Section 15.2.4—and it’s not clear that this classification adds to the reader’s understanding of usability evaluation. Rather, if enough people in enough different contexts have similar user experiences, then guidelines about how to improve those experiences can be created and applied effectively, without using empirical methods for every evaluation.

Also, from a practical standpoint, it is simply not realistic to usability test every element of every product in all of its contexts. A sensible model is to include both inspection/heuristics and empirical research in a product development program, and move back and forth among the methods in a star pattern similar to the star life cycle of interactive system development, with its alternating waves of creative and structuring activities [4]. See Figures 1 and 2.

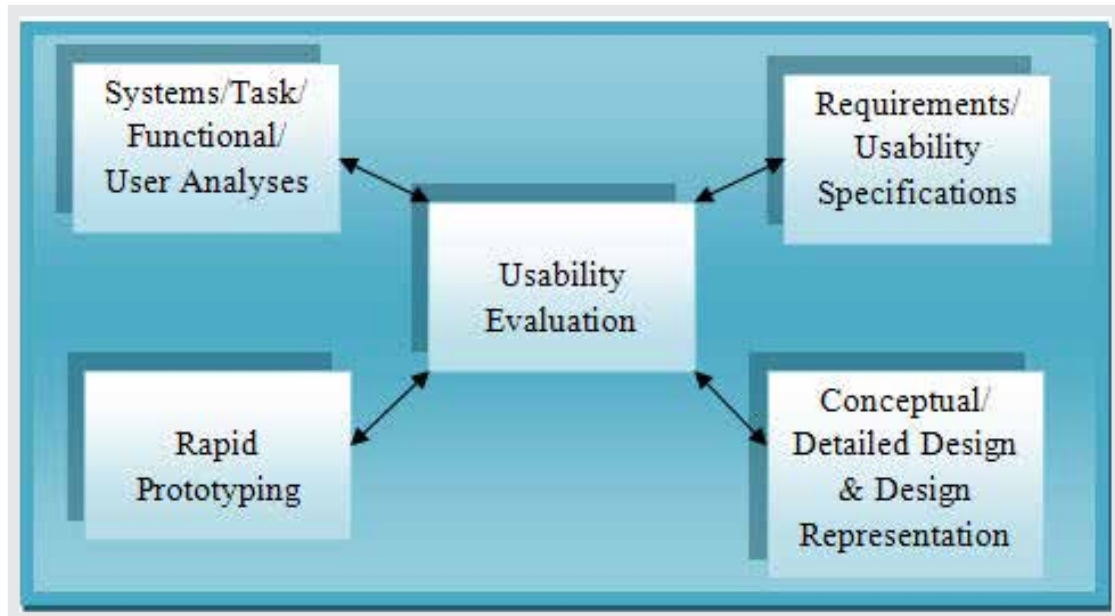


FIGURE 15.1: Star Life Cycle of Interactive System Development (adapted from Hix and Hartson, 1993).

Copyright © Stephanie Rosenbaum. All Rights Reserved. Reproduced with permission. See section “Exceptions” in the copyright terms.

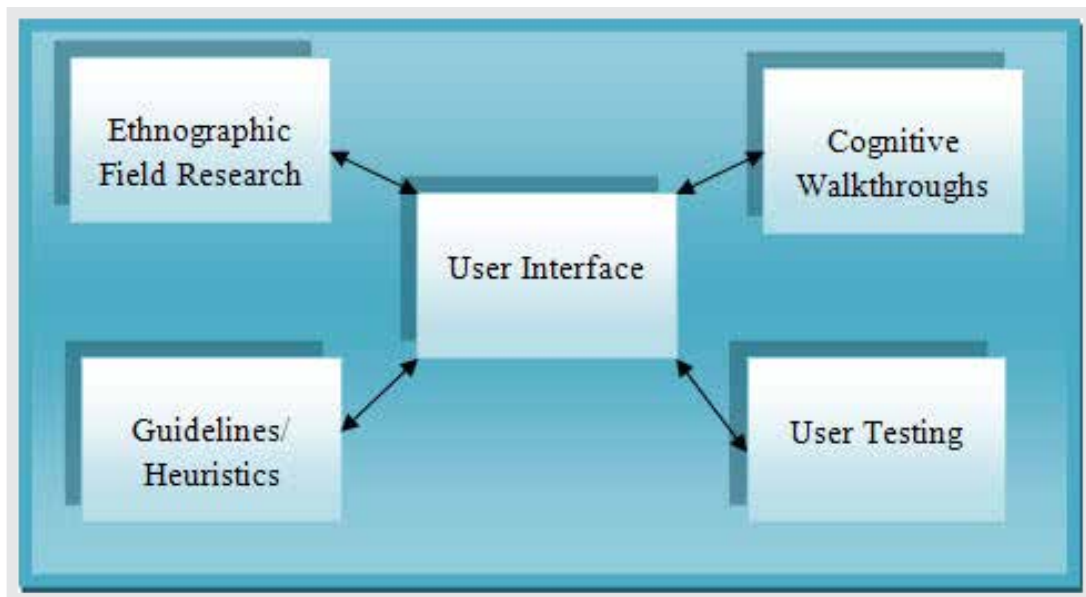


FIGURE 15.2: Usability Evaluation Star Model.

Copyright © Stephanie Rosenbaum. All Rights Reserved. Reproduced with permission. See section “Exceptions” in the copyright terms.

Thus a key element of usability evaluation is deciding when to employ guidelines and inspection (user-free methods) and when it’s critical to perform empirical research such as usability testing or contextual inquiry with the target audience. Planning the activities in a usability evaluation program—and the schedule and budget appropriate to each—is central to the responsibilities of an experienced and skilled usability practitioner. An encyclopedia chapter on usability evaluation should help readers understand this decision-making process.

By the time we get to sections 15.5.1 and 15.5.2, Cockton is accurately describing the situation usability practitioners face: “There is no complete published user testing method that novices can pick up and use ‘as is’. All user testing requires extensive project-specific planning and implementation. Instead, much usability work is about configuring and combining methods for project-specific use.”

It’s true that practitioners in industry, who perform most of today’s usability work, typically do not have time or resources to describe their methods in as much

detail as do academic researchers. I wish this chapter had provided more references about how to learn usability evaluation skills; adding such a focus would make it more valuable for readers. (I have included a selection of these at the end of my commentary.) Although Cockton correctly points out that such resources are not sufficiently complete to follow slavishly, they are still helpful learning tools.

From my own experience at TecEd, the selection and combination of methods in a usability initiative are the most challenging—and interesting—parts of our consulting practice. For example, our engagements have included the following sequences:

Comcast	Qualitative research over a four-week period to learn about customers' enjoyment of Comcast Video Instant Messaging and the ease of its use over time, as well as feature and guest-service preferences. The longitudinal study included three phases—we observed and interviewed pairs of Comcast customers, first in their own homes during in-home installation, then in the usability lab to collect more structured behavioral data in a controlled environment, and finally in focus groups to collect preference data after a month's experience with the new service.
Ford Motor Company	Ethnographic interviews at the homes of 19 vehicle owners throughout the United States. We observed vehicle records and photographed and analyzed artifacts (see Figure 3) to learn how Web technology could support the information needs of vehicle owners. Next we conducted interviews at the homes of 10 vehicle buyers to learn what information they need to make a purchase decision, where they find it, and what they do with it. We subsequently conducted another cycle of interviews at the homes of 13 truck buyers to learn similar information, as well as how truck buyers compare to other vehicle buyers.

Philips Medical Systems	Multi-phase qualitative research project with physicians and allied health personnel during the alpha test of a clinical information system at a major U.S. hospital. After initial “out of box” usability testing at the hospital, we coordinated audiotape diary recording and conducted weekly ethnographic interviews, then concluded the project with a second field usability test after six weeks.
A Major Consumer Electronics Company	Unmoderated card sorting, followed by an information architecture (IA) exploration to help define the user interface for a new product. We began with a two-hour workshop to brainstorm terms for the card sorting, then created and iterated lists of terms, and launched the sorting exercise. For the qualitative IA exploration, we emulated field research in the usability laboratory, a methodology for gaining some benefits of ethnography when it isn’t practical to visit users in the field. We used stage design techniques to create three “environments”: home, office, and restaurant (see Figure 4). In these environments, we learned some contextual information despite the lab setting.

**Cisco
Systems**

Early field research for the Cisco Unified Communications System, observing how people use a variety of communication methods and tools in large enterprise environments. We began each site visit with a focus group, then conducted contextual inquiry with other participants in their own work settings. Two teams of two researchers (one from TecEd, one from Cisco) met in parallel with participants, to complete each site visit in a day. After all the visits, Cisco conducted a full-day data compilation workshop with the research teams and stakeholders. Then TecEd prepared a project report (see Figure 5) with an executive summary that all participating companies received, which was their incentive to join the study.

15.12 COMMENTARY BY ANN BLANDFORD

How to [cite this commentary in your report](#)

Ann Blandford



© *Ann Blandford*

Ann Blandford is Professor of Human–Computer Interaction in the Department of Computer Science at University College London, and served as Director of UCL Interaction Centre (UCLIC) (2004-2011). Her teaching includes User-Centred Evaluation Methods on the MSc in HCI with Ergonomics at UCL. She started her career in industry, as a software engineer, but soon moved into academia, where she...

Ann Blandford

Ann Blandford is a member of The Interaction Design Foundation

Gilbert Cockton’s article on Usability Evaluation does a particularly good job of drawing out the history of “usability” and “user experience” (UX), and highlighting the limitations as well as the importance of a classical “usability” perspective. For several years, I taught a course called “Usability Evaluation Methods”, but I

changed the name to “User-centred Evaluation Methods” because “usability” had somehow come to mean “the absence of bad” rather than “the presence of good”. Cockton argues that “user experience” is the more positive term, and we should clearly be aiming to deliver systems that have greater value than being “not bad”.

However, there remains an implicit assumption that evaluation is summative rather than formative. For example, he discusses the HEART measures of Happiness, Engagement, Adoption, Retention and Task success, and contrasts these with the PULSE measures. Used effectively, these can give a measure of the quality (or even the worth) of a system, alone or in the product ecologies of which it is a part. However, they do not provide information for design improvement. A concern with the quantifiable, and with properties of evaluation methods such as reliability (e.g. Hertzum & Jacobsen, 2001), has limited our perspective in terms of what is valuable about evaluation methods. Wixon (2003) argues that the most important feature of any method is its downstream utility: does the evaluation method yield insights that will improve the design? To deliver downstream utility, the method has to deliver insights not just about *whether* a product improves (for example) user happiness, but also *why* it improves happiness, and *how* the design could be changed to improve happiness even further (or reduce frustration, or whatever). This demands evaluation methods that can inform the design of next-generation products.

Of course, no method stands alone: a method is simply a tool to be used by practitioners for a purpose. As Cockton notes, methods in practice are adopted and adapted by their users, so there is in a sense no such thing as a “method”, but a repertoire of resources that can be selected, adapted and applied, with more or less skill and insight, to yield findings that are more or less useful. To focus this selection and adaptation process, we have developed the Pret A Rapporter framework (Blandford et al, 2008a) for planning a study. The first important element of the framework is making explicit the obvious point that every study is conducted for a purpose, and that that purpose needs to be clear (whether it is formative or

summative, focused or exploratory). The second important element is that every study has to work with the available resources and constraints: every evaluation study is an exercise in the art of the possible.

Every evaluation approach has a potential scope — purposes for which it is and is not well suited. For example, an interview study is not going to yield reliable findings about the details of people's interactions with an interface (simply because people cannot generally recall such details), but might be a great way to find out people's attitudes to a new technology; a GOMS study (John and Kieras, 1996) can reveal important points about task structure, and deliver detailed timing predictions for well structured tasks, but is not going to reveal much about user attitudes to a system; and a transaction log analysis will reveal what people did, but not why they did it.

Cockton draws a distinction between analytical and empirical methods, where analytical methods involve inspection of a system and empirical methods are based on usage. This is a good first approximation, but hides some important differences between methods. Some analytical methods (such as Heuristic Evaluation or Expert Walkthrough) have no direct grounding in theory, but provide more or less support for the analyst (e.g. in the form of heuristics); others (including GOMS) have a particular theoretical basis which typically both constrains the analyst, in terms of what issues can be identified through the method, and provides more support, yielding greater insight into the underlying causes of any issues identified, and hence a stronger basis to inform redesign. In a study of several different analytical methods (Blandford et al, 2008c), we found that methods with a clear theoretical underpinning yielded rich insights about a narrow range of issues (concerning system design, likely user misconceptions, how well the system fits the way users think about their activities, the quality of physical fit between user and system, or how well the system fits its context of use); methods such as Heuristic Evaluation, which do not have theoretical underpinnings, tend to yield insights across a broader range of issues, but also tend to focus more on

the negative (what is wrong with a system) than the positive (what already works well, or how a system might be improved).

Cockton rightly emphasises the importance of context for assessing usability (or user experience); surprisingly little attention has been paid to developing methods that really assess how systems fit their users in their various contexts of use. In the context of e-commerce, such as his van hire example, it is widely recognised that the Total Customer Experience matters more than the UX of the website interface (e.g. Minocha *et al*, 2005): the website is one component of a broader system, and what matters is that the whole system works well for the customers (and also for the staff who have to work within it). The same is true in most contexts: the system has to perform well, it has to be usable and provide a positive user experience, but it also has to fit well into the context of use.

In different contexts, different criteria become prominent. For example, for a banking system, security is at least as important as usability, and having confidence in the security of the system is an important aspect of user experience. A few days ago, I was trying to set up a new standing order (i.e. regular payment from my bank account to a named payee) to pay annually at the beginning of the year ... but the online banking system would only allow me to set up a new standing order to make a payment in the next four months, even though it would permit payment to be annual. This was irritating, and a waste of time (as I tried to work out whether there was a way to force the system to accept a later date for first payment), but it did not undermine my confidence in the system, so I will continue to use it because in many other situations it provides a level of convenience that old-fashioned banking did not.

Cockton points out that there are many values that a system may offer other than usability. We have recently been conducting a study of home haemodialysis. We had expected basic usability to feature significantly in the study, but it does not: not because the systems are easy to use (they are not), but because the users

have to be very well trained before they are able to dialyse at home, their lives depend on dialysis (so they are grateful to have access to such machines), and being able to dialyse at home improves their quality of life compared to having to travel to a dialysis centre several times a week. The value to users of usability is much lower than the values of quality of life and safety.

Particularly when evaluating use in context, there doesn't have to be an either-or between analytical and empirical methods. In our experience, combining empirical studies (involving interviews and observations) with some form of theory-based analysis provides a way of generalising findings beyond the particular context that is being studied, while also grounding the evaluation in user data. If you do a situated study of (for example) a digital library in a hospital setting (Adams et al, 2005), it is difficult to assess how, or whether, the findings generalise to even a different hospital setting, never mind other contexts of use. Being able to apply a relevant theoretical lens (in this case, Communities of Practice) to the data gives at least some idea of what generalises and what doesn't. In this case, the theory did not contribute to an understanding of usability *per se*, but to an understanding of how the deployment of the technology influenced its acceptance and take-up in practice. Similarly, in a study of an ambulance dispatch system (Blandford and Wong, 2004), a theory of situation awareness enabled us to reason about which aspects of the system design, and the way it was used in context, supported or hindered the situation awareness of control room staff. It was possible to apply an alternative theoretical perspective (Distributed Cognition) to the same context of use (ambulance dispatch) (Furniss and Blandford, 2006) to get a better understanding of how the technology design and workspace design contribute to the work of control room staff, including the ways that they coordinate their activity. By providing a semi-structured method (DiCoT) for conducting Distributed Cognition analyses of systems (Blandford and Furniss, 2006), we are encoding key aspects of the theory to make it easier for others to apply it (e.g. McKnight and

Doherty, 2008), and we are also applying it ourselves to new contexts, such as an intensive care unit (Rajkomar and Blandford, in press). Even though particular devices are typically at the centre of these studies, they do not focus on classical usability of the device, or even on user experience as defined by Cockton, but on how the design of the device supports work in its context of use.

Another important aspect of use in context is how people think about their activities and how a device requires them to think about those activities. Green (1989) and others (Green et al, 2006) developed Cognitive Dimensions as a vocabulary for talking about the mismatch between the way that people conceptualise an activity and the way they can achieve their goals with a particular device; for example, Green proposes the term “viscosity” to capture the idea that something that is conceptually simple (e.g. inserting a new figure in a document) is practically difficult (requiring each subsequent figure to be renumbered systematically in many word processors). We went on to develop CASSM (Blandford et al, 2008b) as a method for systematically evaluating the quality of the conceptual fit between a system and its users. Where there are different classes of users of the same system, which you might regard as different personas, you are likely to find different qualities of fit (Blandford et al, 2002). CASSM contrasts with most established evaluation methods in being formative rather than summative; in focusing on concepts rather than procedures; in being a hybrid empirical-analytical approach; and in focusing on use in context rather than either usability or user experience as Cockton describes them. It is a method for evaluating how existing systems support their users in context, which is a basis for identifying future design opportunities to either improve those systems or deliver novel systems that address currently unmet needs. Evaluation should not be the end of the story: as Carroll and Rosson (1992) argue, systems and uses evolve over time, and evaluation of the current generation of products can be a basis for designing the next generation.

This commentary has strayed some way from the classical definitions of usability as encapsulated in many of the standards, and cited by Cockton, to focus more

on how to evaluate “quality in use”, or the “extent to which a product can be used by specified users to achieve specified goals” *within their situated context of use*. Cockton argues that “several evaluation and other methods may be needed to identify and relate a nexus of causes”. I would argue that CASSM and DiCoT are examples of formative methods that address this need, focusing on how products are used in context, and how an understanding of situated use can inform the design of future products. Neither is a silver bullet, but each contributes to the agenda Cockton outlines.

15.12.1 References

- ▶ Adams, A., Blandford, A. & Lunt, P. (2005) Social empowerment and exclusion: A case study on digital libraries. In *ACM Transactions on CHI*. 12.2. 174-200. DOI <http://doi.acm.org/10.1145/1067860.1067863>
- ▶ Blandford, A., Adams, A., Attfield, S., Buchanan, G., Gow, J., Makri, S., Rimmer, J. & Warwick, C. (2008a) The PRET A Reporter framework: Evaluating digital libraries from the perspective of information work. *Information Processing & Management*. 44. 4-21. DOI <http://dx.doi.org/10.1016/j.ipm.2007.01.021>
- ▶ Blandford, A. & Furniss, D. (2006) DiCoT: a methodology for applying Distributed Cognition to the design of team working systems. In S. Gilroy & M. Harrison (Eds.) *Proc. DSVIS 2005*. Springer: LNCS 3941. 26-38. DOI http://dx.doi.org/10.1007/11752707_3
- ▶ Blandford, A., Green, T. R. G., Furniss, D. & Makri, S. (2008b) Evaluating system utility and conceptual fit using CASSM. *International Journal of Human—Computer Studies*. 66. 393-409. DOI <http://dx.doi.org/10.1016/j.ijhcs.2007.11.005>
- ▶ Blandford, A., Hyde, J. K., Green, T. R. G. & Connell, I. (2008c) Scoping Analytical Usability Evaluation Methods: A Case Study.

Human Computer Interaction Journal. 23.3. 278 — 327. DOI 10.1080/07370020802278254

- ▶ Blandford, A. & Wong, W. (2004) Situation Awareness in Emergency Medical Dispatch. *International Journal of Human—Computer Studies*. 61(4). 421-452. DOI <http://dx.doi.org/10.1016/j.ijhcs.2003.12.012>
- ▶ Blandford, A. E., Wong, B. L. W., Connell, I. & Green, T. R. G. (2002) Multiple Viewpoints On Computer Supported Team Work: A Case Study On Ambulance Dispatch. In X. Faulkner, J. Finlay & F. Détienne (Eds.) *People and Computers XVI: Proc. HCI'02*. 139-156. Springer./li>
- ▶ Carroll, J. M. & Rosson, M.B. (1992) Getting around the task-artifact cycle: how to make claims and design by scenario. *ACM Transactions on Information Systems*, 10(2), 181-212.
- ▶ Furniss, D. & Blandford, A. (2006), Understanding Emergency Medical Dispatch in terms of Distributed Cognition: a case study. *Ergonomics Journal*. 49. 12/13. 1174-1203. DOI <http://dx.doi.org/10.1080/00140130600612663>
- ▶ Green, T.R.G. (1989) Cognitive dimensions of notations. In R. Winder and A. Sutcliffe (Eds), *People and Computers V*. Cambridge University Press< 443-460.
- ▶ Green, T.R.G., Blandford, A.E., Church, L., Roast, C.R. & Clarke, S. (2006) Cognitive dimensions: Achievements, new directions, and open questions, *Journal of Visual Languages & Computing*, Volume 17, Issue 4, Ten Years of Cognitive Dimensions, August 2006, Pages 328-365.
- ▶ Hertzum, M., and Jacobsen, N.E. (2001) [The Evaluator Effect: A Chilling Fact about Usability Evaluation Methods](#). *International Journal of Human-Computer Interaction*, 13(4), 421-443.

- ▶ John, B. & Kieras, D.E. (1996) Using GOMS for user interface design and evaluation: which technique? *ACM ToCHI* 3.4. 287-319.
- ▶ McKnight, J. and Doherty, G. (2008) Distributed cognition and mobile healthcare work. In *Proceedings of the 22nd British HCI Group Annual Conference on People and Computers: Culture, Creativity, Interaction - Volume 2* (BCS-HCI '08), Vol. 2. British Computer Society, Swindon, UK, 35-38.
- ▶ Minocha, S., Dawson, L., Blandford, A. & Millard, N. (2005) Providing value to customers in e-commerce environments: the customer's perspective. In *Contemporary Research in E-Marketing Vol 2*. Ideas Group publishing. 119-146.
- ▶ Rajkomar, A. & Blandford, A. (in press) Understanding Infusion Administration in the ICU through Distributed Cognition. *Journal of Biomedical Informatics*. To appear. DOI <http://dx.doi.org/10.1016/j.jbi.2012.02.003>
- ▶ Wixon, D. (2003) Evaluating Usability Methods; Why the Current Literature Fails the Practitioner. *Interactions* July and August 2003. 28-34.

15.13 COMMENTARY BY THOMAS VISBY SNITKER

How to [cite this commentary in your report](#)

Thomas Visby Snitker



© *Thomas Visby Snitker*

Thomas is one of Denmark's leading usability specialists. He has extensive experience in user research and usability studies for Danish as well as for international companies. Thomas is a frequent speaker at Danish and international conferences, such as The UX Masterclass, and blogs for the Danish edition of Computerworld. In addition, he serves as external reviewer at the IT Universi...

Thomas Visby Snitker

Thomas Visby Snitker is a member of The Interaction Design Foundation

15.13.1 Making usability simpler - the way forward?

I work with usability on a daily basis and my clients - most annoyingly - do not really take much interest in what I do for them. Unless of course I break something in the process. Usually they just want to know how they can improve their user

interfaces (UI). Well, that's acceptable for me - as a usability specialist that is what I am concerned with.

My customers may ask when to do what and why, but they only listen for as long as it takes to make up their minds — they look for the immediate UI tweaks and solutions, not for insight into the complex intricacies and interactions between users, contexts, media and services. They request my complex research but would rather get a quick fix.

As a consequence, my company has launched a new service, UsabilityForce, to take the complexity out of usability research from the perspective of the customer. UsabilityForce allows producers, designers, developers and others to simply order videos of users thinking aloud while using the clients' service or product at their own leisure, following a test script with test tasks provided by the client. The client can watch the videos and sum up the findings himself or we can provide that through one of our consultants.

The testers install a bit of code on their computer that allows them to hit Record, Pause and Stop. They also use a microphone to capture their audio and an internet connection to upload their video. In the standard test setup with five users, it usually takes only 3-4 hours to collect the five videos.

The contrast between simple and complex research is strong. Quicker versus slower, cheaper versus more expensive, simple versus complex. The ramifications of simpler research are far-reaching and include:

- ▶ Less data takes less time to analyse, to report and react to.
- ▶ Less time spent in research means quicker reaction time.
- ▶ You go into detail with a smaller scope or a specific part of a larger scope.
- ▶ Those who find it difficult to see the gains in complex usability because of the size of the investment will have less reason to hesitate. Maybe no reason at all.

- ▶ Those who find that complex research takes too much time will find that the simple research is fast.
- ▶ Simple testing may make new research topics relevant and feasible; things that were previously too small to test (in terms of costs versus benefits), like a Facebook page, a newsletter, a search engine result page, etc.
- ▶ Simple testing can make continuous evaluations feasible - by making simple tests at a regular interval. This could feed into a structured measurement process using Key Performance Indicators (KPIs) all through a development process and ongoing after launch.

I imagine that the simple usability testing will provide a useful supplement to complex testing. As long as complex products and services are conceived and developed they of course need complex research. Furthermore, I speculate that these simpler research technologies will not only have an impact on how usability specialists conduct research, but also on how my ‘annoyingly usability ignorant clients’ will change. I imagine that some of them will understand better how to benefit from a user research project. They will do the math, they will build the business case, they will include their stakeholders and they will persuade reluctant gate keepers.

I also imagine, and hope, that simpler research will allow our community to grow. Those clients who consume our services will grow more committed to the usability of their products and services; they will be more demanding and assertive in the field of usability and perhaps user experience (UX) as well. They will start to question our expertise and they will link their success to the usability of their user interfaces. They will set up a strategy for the user experience (UX targets) and measure the performance of their UX. Simpler research is a strong force and can change how we work and how well usability is adopted by those who need it.

15.14 COMMENTARY BY TILDE BEKKER

How to [cite this commentary in your report](#)

Tilde Bekker



Tilde Bekker is an associate professor in the Industrial Design department at the Eindhoven University of Technology. Her research interests include designing for playful interaction, and designing products for children and older adults. She leads and participates in research projects on playful interactions that examine how to persuade people to a healthier lifestyle. She has over 75 pu...

Tilde Bekker

Tilde Bekker is a member of The Interaction Design Foundation

The area of usability evaluation is on the move, as Gilbert Cockton describes. The chapter provides a thorough description of the historical development of usability evaluation methods and provides a good starting point for considering what needs to be done next.

In my commentary I expand on one aspect of evaluation methods: eliciting information from users. I describe how, in the area of *Interaction Design and Children*, evaluation methods have been adapted to increase the output of children participants in evaluation sessions. Two approaches have been applied: by providing different strategies for supporting verbalizations and by providing non-verbal ways to children for expressing their opinion.

For more than 10 years I have been teaching HCI and Industrial Design students how to apply a wide variety of evaluation approaches to various kinds of products and interfaces. Applying evaluation methods to the design of technologies for children can provide a new perspective because it forces us to re-examine some of the assumptions we make about usability evaluation methods.

15.14.1 Adapting evaluation methods to participants' skills

What is an interesting challenge in designing and evaluating interactive products for children is to find a good match between the skills and qualities of the participants and the properties of the design and evaluation activity. This approach, which has been widespread in the research area of Interaction Design and Children, has led to some interesting adaptations to existing usability evaluation methods and also to the development of new usability evaluation methods.

In the past 10 to 15 years various studies have examined whether children have the skills and qualities required for a variety of evaluation methods. We can of course argue that when a participant has trouble participating in an evaluation session, we have to train the participant. Another or complementary option is to adjust or redesign the evaluation method to make it easier and possibly more fun to participate in an evaluation session.

15.14.2 Verbalization techniques

An important skill required for many evaluation methods is the ability to verbalize one's thoughts. Such verbalizations can be used as a basis for interpreting what

usability problems are embedded in the user interface. Different techniques are applied for eliciting verbal output.

One common approach for eliciting verbal output is the *think-aloud method*. Participants are asked to verbalize their thoughts while they are interacting with the product. The evaluation facilitator may prompt the participant to keep talking during the session. However, can children think aloud during usability evaluation sessions? Initially it was suggested that children of 13 years and older can think aloud (Hanna et al., 1997). More recent research showed that children of 7 years and older can think aloud when the protocol for facilitating the verbalizations is adjusted to a more relaxed dialogue (Donker and Markopoulos, 2002).

Evaluation methods may also incorporate other strategies to support participants in verbalizing their thoughts than being prompted by a facilitator. Examples of other strategies are participating in an evaluation session together with peers, tutoring another child, or being prompted by a social robot as a proxy for the facilitator. However, the success of these strategies may depend on children having other skills required for these set-ups, such as the ability to collaborate.

An evaluation method called *co-discovery or constructive interaction* applies a technique where two participants collaborate in performing tasks in an evaluation setting. Supporting verbalizations by talking to a peer may be a more natural setting than holding a monologue or talking to a test facilitator. However, children do need to collaborate for the evaluation sessions to be effective. Some research has shown that younger children of 6 to 7 years old participating may still lack sufficient social skills to be effective participants. They may forget to collaborate and work on a task on their own, thus not providing many verbal utterances. They may sometimes actually compete when doing a task (Markopoulos and Bekker, 2003; Van Kesteren et al., 2003). Older children (between 13 and 14) have been shown to collaborate quite well in co-discovery sessions (Als et al, 2005). Other factors that may influence the quality of the collaboration and the outcome of the session are whether the pairs are friends or not and gender.

Another method, called *peer tutoring*, is based on the idea that one child explains to another child how a product works (Höysniemi et al, 2003). At the beginning one child will try out using a product. Then the first child will become the tutor of a second child. The tutor will help the second child to interact with the product. From the dialogue between the two children their understanding about the product and usability problems can be distilled. The success of this approach depends on whether the tutor is able to fulfill his tutor role effectively, and whether the tutee is open to being taught by another child. Evidence from peer tutoring indicates that when the tutor forgets to play the tutor role the pairs of children take on roles more similar to those in co-discovery sessions. Furthermore, tutors may have trouble only explaining the interaction to the other child without taking over doing a task (Van Kesteren et al, 2003).

A more recently developed method, in which a child is being prompted by a facilitator through a robot interface, is called the *robotic intervention method* (Fransen and Markopoulos, 2010). Providing a context in which children can talk to a playful and toy-like robot is expected to be less inhibiting than talking to an adult. So far, no increase in problems uncovered using this method compared to an active intervention method was found. Children did seem more at ease when participating in the sessions. A slight drawback of the methods was that children perceived the questions asked by the robot to be more difficult than those asked by a human facilitator.

15.14.3 Complementing verbal with non-verbal approaches

A different strategy than facilitating verbal output is to provide alternative non-verbal ways to indicate positive and negative aspects in an interface. This was applied in the PhD work by Wolmet Barendregt who developed the *picture cards method* (Barendregt et al., 2008). The method was developed to find problems with children's computer games. It includes cards that children can pick up to indicate both positive and negative aspects in an interface. They can place a card in a box every time they experience a particular emotion shown on one of the cards. They pick up a picture card to express their feelings when interacting with a game or product.

The categories of the cards correspond to various types of problems and fun issues. In a study with children of 5 and 6 years old children expressed more problems explicitly with the picture cards method than in a think-aloud session.

15.14.4 A rich usability evaluation context

I agree with Cockton that there are no generalizable evaluation methods. Learning how to conduct usability evaluations requires developing an understanding of the complete evaluation context. This context includes many factors, such as who applies the method in what type of development process. And it also includes, as I illustrated earlier, specific requirements of the user group.

Evaluation methods can be further improved by adapting them to the skills and qualities of all the stakeholders involved, by providing diverse ways to provide input, addressing positive and negative experiences, and possibly even making the activity more fun and enjoyable.

Developing evaluation approaches is like developing products and systems: for every improvement we try to incorporate in an evaluation method, we run the risk of adding new challenges for the participants.



FIGURE 15.1: Usability Evaluation with children.

Copyright © Bart van Overbeeke Fotografie. All Rights Reserved. Reproduced with permission. See section "Exceptions" in the copyright terms.

15.14.5 References

- ▶ Als, B.S., Jensen, J.J. & Skov, M.B. (2005) Exploring verbalization and collaboration of constructive interaction with children. In *Proceedings of the 2005 IFIP TC13 international conference on Human-Computer Interaction (INTERACT'05)*, Springer-Verlag, Berlin, Heidelberg, 443-456.
- ▶ [Barendregt, W., Bekker, M.M., & Baauw, E. \(2008\) Development and evaluation of the problem identification picture cards method, *Cognition, Technology and Work*, 10 \(2\), 95-105.](#)

- Donker, A., & Markopoulos, P. (2002) A comparison of think-aloud, questionnaires and interviews for testing usability with children. In *People and Computers XVI - Memorable Yet Invisible, Proceedings of HCI 2002*, London: Springer, 305-316.
- ▶ Fransen, S. & Markopoulos, P. (2010). Let robots do the talking. In *Proceedings of the 9th International Conference on Interaction Design and Children (IDC '10)*. ACM, New York, NY, USA, 59-68.
 - Hanna ,J., Ridsen , K. & Alexander, K. Guidelines for usability testing with children, (1997) *interactions*, 4(5), 9-14.
 - ▶ Höysniemi,J., Hämäläinen, P. &Turkki, L. (2003) Using peer tutoring in evaluating the usability of a physically interactive computer game with children. *Interacting with Computers* 15(2): 203-225
 - ▶ Markopoulos, P. & Bekker, M. (2003) On the assessment of usability testing methods for children. *Interacting with Computers*, 15(3), 227-243.
 - ▶ Van Kesteren, I., Bekker, M.M., Vermeeren, A.P.O.S. & Lloyd, P. (2003) Assessing usability evaluation methods on their effectiveness to elicit verbal information from children subjects, In *Proceedings of the Conference on Interaction Design and Children*, Preston, 1-3 July 2003, 41-49.

15.15 REFERENCES

Bardzell, Jeffrey (2011): *Interaction criticism: An introduction to the practice*. In *Interacting with Computers*, 23 (6) pp. 604-621

Bardzell, Jeffrey (2009): *Interaction criticism and aesthetics*. In: *Proceedings of ACM CHI 2009 Conference on Human Factors in Computing Systems* 2009. pp. 2357-2366

Bardzell, Shaowen and Bardzell, Jeffrey (2011): *Towards a feminist HCI methodology: social science, feminism, and HCI*. In: *Proceedings of ACM CHI 2011 Conference on Human Factors in Computing Systems* 2011. pp. 675-684

[Bellamy](#), Rachel, [John](#), Bonnie E. and [Kogan](#), Sandra (2011): Deploying CogTool: integrating quantitative usability assessment into real-world software development. In: [Proceeding of the 33rd international conference on Software engineering](#) 2011. pp. 691-700

[Blythe](#), Mark, [Petrie](#), Helen and [Clark](#), John A. (2011): F for fake: four studies on how we fall for phish. In: [Proceedings of ACM CHI 2011 Conference on Human Factors in Computing Systems](#) 2011. pp. 3469-3478

[Brown](#), Tim (2009): *Change by Design: How Design Thinking Transforms Organizations and Inspires Innovation*. HarperBusiness

[Brown](#), Barry A. T., [Taylor](#), Alex S., [Izadi](#), Shahram, [Sellen](#), Abigail, [Kaye](#), Joseph Jofish and [Eardley](#), Rachel (2007): Locating Family Values: A Field Trial of the Whereabouts Clock. In: [Krumm](#), John, [Abowd](#), Gregory D., [Seneviratne](#), Aruna and [Strang](#), Thomas (eds.) [Ubi-Comp 2007 Ubiquitous Computing - 9th International Conference](#) September 16-19, 2007, Innsbruck, Austria. pp. 354-371

[Card](#), Stuart K., [Moran](#), Thomas P. and [Newell](#), Allen (1980): The keystroke-level model for user performance with interactive systems. In [Communications of the ACM](#), 23 pp. 396-410

[Cockton](#), Gilbert (2006): Designing worth is worth designing. In: [Proceedings of the Fourth Nordic Conference on Human-Computer Interaction](#) 2006. pp. 165-174

[Cockton](#), Gilbert (2007): Some Experience! Some Evolution. In: [Erickson](#), Thomas and [McDonald](#), David W. (eds.). "HCI Remixed: Reflections on Works That Have Influenced the HCI Community". The MIT Presspp. 215-219

[Cockton](#), Gilbert (2004): Value-centred HCI. In: [Proceedings of the Third Nordic Conference on Human-Computer Interaction](#) October 23-27, 2004, Tampere, Finland. pp. 149-160

[Cockton](#), Gilbert (2008): Load while Aiming; Hit? (invited keynote). In: [Brau](#), Henning (ed.). "Usability Professionals 2008 Conference". pp. 17-22

[Cockton](#), Gilbert and [Lavery](#), Darryn (1999): A Framework for Usability Problem Extraction. In: [Angela.Sasse](#), M. and [Johnson](#), Chris (eds.) [Proceedings of Interact 99](#) 1999, Edinburgh.

[Cockton](#), Gilbert and [Woolrych](#), Alan (2009): A. Comparing UEMs: Strategies and Implementation, Final Report of COST 294 Working Group 2. In: [Law](#), Effie Lai-Chong, [Scapin](#), Dominique, [Cockton](#), Gilbert, [Springett](#), Mark, [Stary](#), C. and [Winckler](#), Marco (eds.). "Maturation of Usability Evaluation Methods: Retrospect and Prospect: Final Reports of COST294-MAUSE Working Groups".

[Cockton](#), Gilbert, [Woolrych](#), Alan, [Hornbæk](#), Kasper and [Frøkjær](#), Erik (2012): Inspection-based methods. In: [Sears](#), Andrew and [Jacko](#), Julie A. (eds.). "The Human-Computer Inter-

action Handbook: Fundamentals, Evolving Technologies and Emerging Applications, Third Edition". pp. 1275-1293

[Cockton](#), Gilbert, [Woolrych](#), Alan and [Hindmarch](#), Mark (2004): Reconditioned merchandise: extended structured report formats in usability inspection. In: [Dykstra-Erickson](#), Elizabeth and [Tscheligi](#), Manfred (eds.) [Extended abstracts of the 2004 Conference on Human Factors in Computing Systems, CHI 2004, Vienna, Austria, April 24 - 29, 2004](#) 2004. pp. 1433-1436

[Cockton](#), Gilbert, [Kujala](#), Sari, [Nurkka](#), Piia and [Hölttä](#), Taneli (2009): Supporting Worth Mapping with Sentence Completion. In: [Gross](#), Tom, [Gulliksen](#), Jan, [Kotze](#), Paula, [Oestreicher](#), Lars, [Palanque](#), Philippe A., [Prates](#), Raquel Oliveira and [Winckler](#), Marco (eds.) [Human-Computer Interaction - INTERACT 2009, 12th IFIP TC 13 International Conference, Uppsala, Sweden, August 24-28, 2009, Proceedings, Part II](#) 2009. pp. 566-581

[Cockton](#), Gilbert, [Kirk](#), Dave, [Sellen](#), Abigail and [Banks](#), Richard (2009): Evolving and augmenting worth mapping for family archives. In: [Proceedings of the HCI09 Conference on People and Computers XXIII](#) 2009. pp. 329-338

[Dumas](#), Joseph S. and [Fox](#), Jean E. (2007): Usability Testing: Current Practice and Future Directions. In: [Sears](#), Andrew and [Jacko](#), Julie A. (eds.). "The Human-Computer Interaction Handbook: Fundamentals, Evolving Technologies and Emerging Applications, Second Edition (Human Factors and Ergonomics)". CRC Presspp. 1129-1150

[Dumas](#), Joseph S. and [Redish](#), Janice C. (1993): A Practical Guide to Usability Testing. Norwood, NJ, Intellect

[Erickson](#), Thomas and [McDonald](#), David W. (2008): HCI Remixed: Reflections on Works that Have Influenced the HCI Community. Cambridge, Massachusetts, MIT Press

[Forlizzi](#), Jodi (2008): The product ecology: Understanding social product use and supporting design culture. In [International Journal of Design](#), 2 (1) pp. 11-20

[Gaver](#), William W., [Boucher](#), Andy, [Law](#), Andy, [Pennington](#), Sarah, [Bowers](#), John, [Beaver](#), Jacob, [Humble](#), Jan, [Kerridge](#), Tobie, [Villar](#), Nicholas and [Wilkie](#), Alex (2008): Threshold devices: looking out from the home. In: [Proceedings of ACM CHI 2008 Conference on Human Factors in Computing Systems](#) April 5-10, 2008. pp. 1429-1438

[Gray](#), Wayne D. and [Salzman](#), Marilyn C. (1998): Damaged Merchandise? A Review of Experiments That Compare Usability Evaluation Methods. In [Human-Computer Interaction](#), 13 (3) pp. 203-261

[Harper](#), Richard, [Rodden](#), Tom, [Rogers](#), Yvonne and [Sellen](#), Abigail (2008): Being Human: Human Computer Interaction in 2020. Microsoft Research Ltd

Hertzum, Morten and Jacobsen, Niels Ebbe (2001): *The Evaluator Effect: A Chilling Fact About Usability Evaluation Methods*. In International Journal of Human-Computer Interaction, 13 (4) pp. 421-443

Herzberg, Frederick (1966): *Work and the Nature of Man*.

Herzberg, Frederick (1973): *Work and the Nature of Man*. Signet

Hornbæk, Kasper (2010): *Dogmas in the assessment of usability evaluation methods*. In Behaviour and Information Technology, 29 (1) pp. 97-111

Iivari, N. (2005): *Usability Specialists - 'Mommy Mob', 'Realistic Humanists' or 'Staid Researchers'? An Analysis of Usability Work in the Software Product Development*. In: Proceedings of IFIP INTERACT05: Human-Computer Interaction 2005. pp. 418-430

International Standards Association (1991). *ISO/IEC 9126: Software engineering - Product quality*. Retrieved 1 December 2011 from International Standards Association: http://en.wikipedia.org/wiki/ISO/IEC_9126

International Standards Association (1998). *ISO 9241-11: Ergonomic requirements for office work with visual display terminals (VDTs) -- Part 11: Guidance on usability*. Retrieved 1 December 2011 from International Standards Association: http://en.wikipedia.org/wiki/ISO/IEC_9126

International Standards Association (2001). *ISO/IEC 9126-1:2001 Software engineering - Product quality - Part 1: Quality model*. Retrieved 1 December 2011 from International Standards Association: http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=22749

International Standards Association (2004). *ISO/IEC TR 9126-4:2004 Software engineering -- Product quality -- Part 4: Quality in use metrics*. Retrieved 1 December 2011 from International Standards Association: http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=39752

International Standards Association (2011). *ISO/IEC 25010 Systems and software engineering -- Systems and software Quality Requirements and Evaluation (SQuaRE) -- System and software quality models*. Retrieved 1 December 2011 from International Standards Association: http://www.iso.org/iso/catalogue_detail.htm?csnumber=35733

John, Bonnie E. and Kieras, David E. (1996): *Using GOMS for User Interface Design and Evaluation: Which Technique?*. In ACM Transactions on Computer-Human Interaction, 3 (4) pp. 287-319

[Landauer](#), Thomas K. (1996): *The Trouble with Computers: Usefulness, Usability, and Productivity*. The MIT Press

[Lavery](#), Darryn, [Cockton](#), Gilbert and [Atkinson](#), Malcolm P. (1997): Comparison of Evaluation Methods Using Structured Usability Problem Reports. In [Behaviour and Information Technology](#), 16 (4) pp. 246-266

[Law](#), Effie Lai-Chong, [Hvanberg](#), Ebba, [Cockton](#), Gilbert, [Palanque](#), Philippe A., [Scapin](#), Dominique L., [Springett](#), Mark, [Stary](#), Christian and [Vanderdonckt](#), Jean M. (2005): Towards the maturation of IT usability evaluation (MAUSE). In: [Costabile](#), Maria Francesca and [Paterno](#), Fabio (eds.). "Human-computer interaction: INTERACT 2005: IFIP TC13 international conference: proceedings. Lecture Notes in Computer Science. (3585)". pp. 1134-1137

[Nielsen](#), Jakob (1994): Enhancing the explanatory power of usability heuristics. In: [Plaisant](#), Catherine (ed.) [Conference on Human Factors in Computing Systems, CHI 1994, Boston, Massachusetts, USA, April 24-28, 1994, Conference Companion](#) 1994. p. 210

[Otero](#), Nuno and [José](#), Rui (2009): Worth and Human Values at the Centre of Designing Situated Digital Public Displays. In [IIAPUC](#), 1 (4) pp. 1-13

[Pew](#), Richard W. (2002): Evolution of human-computer interaction: from Memex to Bluetooth and beyond. In: [Sears](#), Andrew and [Jacko](#), Julie A. (eds.). "The Human-Computer Interaction Handbook: Fundamentals, Evolving Technologies and Emerging Applications (Human Factors and Ergonomics)". Lawrence Erlbaumpp. 1-17

[Rodden](#), Kerry, [Hutchinson](#), Hilary and [Fu](#), Xin (2010): Measuring the user experience on a large scale: user-centered metrics for web applications. In: [Proceedings of ACM CHI 2010 Conference on Human Factors in Computing Systems](#) 2010. pp. 2395-2398

[Rogers](#), Yvonne, [Bannon](#), Liam and [Button](#), Graham (1994): Rethinking theoretical frameworks for HCI: A Review. In [ACM SIGCHI Bulletin](#), 26 (1) pp. 28-30

[Rosenbaum](#), Stephanie (2008): The Future of Usability Evaluation : Increasing Impact on Value. In: [Law](#), Effie Lai-Chong, [Hvannberg](#), Ebba Thora and [Cockton](#), Gilbert (eds.). "Maturing Usability". Springerpp. 344-378

[Rosenbaum](#), Stephanie, [Rohn](#), Janice Anne and [Humburg](#), Judee (2000): A Toolkit for Strategic Usability: Results from Workshops, Panels, and Surveys. In: [Turner](#), Thea, [Szwillus](#), Gerd, [Czerwinski](#), Mary, [Paterno](#), Fabio and [Pemberton](#), Steven (eds.) [Proceedings of the ACM CHI 2000 Human Factors in Computing Systems Conference](#) April 1-6, 2000, The Hague, The Netherlands. pp. 337-344

Salvucci, Dario D. (2009): Rapid prototyping and evaluation of in-vehicle interfaces. In ACM Transactions on Computer-Human Interaction, 16 (2) p. 9

Sears, Andrew and Jacko, Julie A. (2007): *The Human-Computer Interaction Handbook: Fundamentals, Evolving Technologies and Emerging Applications* (2nd Edition). Lawrence Erlbaum Associates

Sengers, Phoebe and Gaver, William (2006): Staying open to interpretation: engaging multiple meanings in design and evaluation. In: Proceedings of DIS06: Designing Interactive Systems: Processes, Practices, Methods, & Techniques 2006. pp. 99-108

Shneiderman, Ben (1986): No members, no officers, no dues: A ten year history of the software psychology society. In ACM SIGCHI Bulletin, 18 (2) pp. 14-16

Smith, Sidney L. and Mosier, Jane N. (1986). GUIDELINES FOR DESIGNING USER INTERFACE SOFTWARE, The MITRE Corporation, Bedford, Massachusetts, USA, Prepared for Deputy Commander for Development Plans, and Support Systems, Electronic Systems Division, AFSC, United States Air Force, Hanscom Air Force B.

Venturi, Giorgio, Troost, Jimmy and Jokela, Timo (2006): People, Organizations, and Processes: An Inquiry into the Adoption of User-Centered Design in Industry. In International Journal of Human-Computer Interaction, 21 (2) pp. 219-238

Wasserman, Anthony I. (1973): The design of 'idiot-proof' interactive programs. In: Proceedings of the June 4-8, 1973, national computer conference and exposition 1973. pp. m34-m38

Whiteside, John, Bennett, John and Holtzblatt, Karen (1988): Usability Engineering: Our experience and Evolution. In: Helander, Martin and Prabhu, Prasad V. (eds.). "Handbook of human-computer interaction". pp. 791-817

Wilson, Chauncey (1999). Severity Scale for Classifying Usability Problems. Retrieved 1 December 2011 from The Usability SIG Newsletter: <http://www.stcsig.org/Usability/newsletter/9904-severity-scale.html>

Woolrych, Alan, Hornbæk, Kasper, Frøkjær, Erik and Cockton, Gilbert (2011): Ingredients and Meals Rather Than Recipes: a Proposal for Research That Does Not Treat Usability Evaluation Methods As Indivisible Wholes. In International Journal of Human-Computer Interaction, 99999 (1) pp. 1-1

ABOUT THE AUTHOR

Gilbert Cockton



HAS ALSO PUBLISHED UNDER THE NAME OF:

“G. Cockton”

PERSONAL HOMEPAGE:

designsituations.blogspot.com/

CURRENT PLACE OF EMPLOYMENT:

[Northumbria University](http://www.northumbria.ac.uk)

I’m Professor of Design Theory in the School of Design at Northumbria University, which has roots back to 1844 as one of the original British Government Schools of Design. I’m an Interaction Designer who occasionally dabbles in product and service design. I have a multi-disciplinary background in humanities (History), applied human sciences (Education), and engineering design (Computer Science, old school HCI). In 2005 I was awarded a UK NESTA fellowship to work on value-centred approaches to design. As a first result, I moved beyond value (to worth) and beyond centredness (to multiple design foci). As a second result, my research

has developed a more general design focus, and I moved from Computing (Sunderland) to Design (Northumbria) in September 2009. I now work alongside very talented design educators and researchers with a broad range of craft skills and design philosophies. I increasingly find the user-centred positions of 1980s HCI naive and uninformed. When I'm not supporting an amazing group of colleagues in my role as Associate Dean for Research and Innovation, my research focuses on balanced, integrated and generous (BIG!) fusions of the main design paradigms (Applied Arts, Engineering, User-Centred). Examples of such fusions include leveraging crafted forms within user research, focusing evaluation practices on achieved worth, combining engineering specification with humane design purposes, and blending tacit creative and explicit systematic design work. I have developed the Working to Choose (W2C) framework to provide an overarching structure for co-ordinating research on re-usable resources in design practice.

YOUR NOTES AND THOUGHTS ON CHAPTER 15

Record your notes and thoughts on this chapter. If you want to share these thoughts with others online, go to the bottom of the page at:

http://www.interaction-design.org/encyclopedia/usability_evaluation.html

NOTES:
