Beyond Usability: Process, Outcome and Affect in human computer interactions

Andrew Dillon

This paper was presented as the Lazerow Lecture 2001, at the Faculty of Information Studies, University of Toronto, March 2001.

Abstract

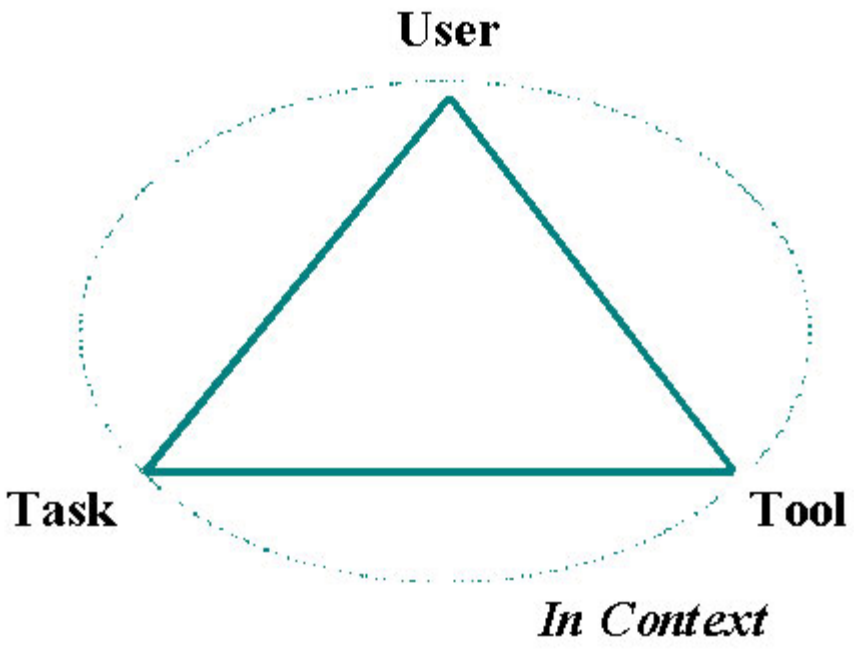
The present paper reviews the general usability framework that has dominated discussion in the field of human-computer interaction (HCI) and finds it wanting. An alternative view of the important determinants of user experience of interactive devices is presented with examples.

Systematizing Interaction Evaluation

Evaluations of information technology for human use may take many forms, from consumer reactions to screens (on a small scale) to studies of widespread adoption of technology (on a large scale). The classic human-computer interaction (HCI) approach to evaluation has focused on usability and this term is become routinely used to describe a quality of information technology that must be taken seriously in design. Most new applications are tested for usability at some stage in their design and it is common for usability to be trumpeted in marketing campaigns and advertising as a distinguishing index of product quality. While such prominence represents a victory for the user-centered design movement, it is clear that usability alone, at least as it is currently conceived, is insufficient for ensuring high quality user experiences with a new technology.

Usability defined

Historically, usability (of an application) has moved from a concern with features of an interface to address aspects of the interaction expressed in terms of human action. While multiple definitions still exist, the nearest to an agreed standard is the ISO 9241-derived definition of usability as the effectiveness, efficiency, and satisfaction with which specified users can achieve specified goals in particular environments.



This is an extremely useful definition that even now is overlooked by many people in favor of simpler conceptions of usability as a combination of features or an interface style to be copied. This definition places emphasis on measurable criteria of performance (effectiveness, efficiency and satisfaction) that are context-bound by the type of user, the type of task and situation of use.

Evaluation of usability is thus conducted by having representative users interact with the design in a suitably task-oriented fashion while evaluators (also known as usability engineers) record such data as time per task, errors, and user attitude. While this operational approach is highly grounded and suitable for quantification, there are several problems that have never been fully resolved by usability engineers.

Effectiveness as Quality?
Some tasks do not have a definitive correct answer, and the scoring of effectiveness in such contexts is problematic. Consider the tasks listed in Table 1.

Creative production (writing, design)
Information retrieval
Reading
Data analysis
Management
Making a purchase…..

Table 1. Tasks supported by technology that may prove difficult to measure

Effectiveness certainly can be deduced here. After all, we assume that managers, analysts, writers and others are subject to some performance assessment in their work.

However, in designing technology to support these tasks, the classic usability engineering approach calls for explicit measures of task performance. Determining what these measures may be in typical test scenarios and their relationship to meaningful task outcomes for users in context is not immediately clear.

Efficiency

When usability professionals talk of efficiency they invoke measures of resources used to perform a task such as time, effort, cost. For example, considering a user interaction with a web page we may assess efficiency through such indices as:

Time taken to complete task
Number of steps taken
Number of deviations from ideal path

Such variables are frequently highly positively correlated - but not always! Raw counts of steps taken can hide time on specific tasks, e.g., where the user is either spending a lot or almost no time at a specific step.

We should also ask if efficiency is really what most users want. The push to efficiency is symptomatic of an engineering-oriented approach, and we need to ask who determines efficiency? Are path deviations always inefficient? Is time equally weighted by user, designer or owner? Almost certainly it is not, yet speed measures are continually invoked as a primary quality of interaction.

Socio-technical theorists such as Eason (1988) make a strong case for negotiation of such variables among stakeholders, but such negotiations are often beyond typical usability tests. This is not necessarily the fault of usability professionals but it is a hazard of any operationalization of usability that emphasizes speed and efficiency (see Dillon 2000 for more on this).

Satisfaction- extending performance to affect

The third component of the ISO-92411 definition of usability is satisfaction. The need for this third measure is made clear by recent data showing that performance and affect are clearly correlated (Frojkaer et al 2001). Many users express preferences for tools that they do not use to maximum effectiveness or efficiency, and a complication for designers is determining the relative weights to place on each measure.

What determines a user's satisfaction? There are possibly more factors than we yet know but it is likely that satisfaction is influenced by such factors as personal experience with other technologies, preferred working style, the manner of introduction, and the aesthetics of the product.

Is usability enough?

Obviously I believe usability is important, but I have come to the conclusion that it is insufficient. There are two major reasons for this. First, for some of the reasons outlined above, I consider the set of metrics that falls out of the effectiveness, efficiency and satisfaction model can place undue emphasis on speed and accuracy. For many contemporary interactions, and more importantly, for the type of technology I envisage us using in the coming century, I suspect user experience will prove more complicated.

Second, the nature of interaction with many discretionary technologies is really more about enhancement of work and leisure. The type of user responses that we will need to capture and analyze to help us design more innovative software are not clearly task-based in the classical performance sense. Consider for example, the processes underlying creative thinking or stimulation, where the cause-effect relationships between input and output are not so clearly delineated.

Beyond usability: Process, Outcome and Affect

In extending the classic ISO approach to usability of effectiveness, efficiency and satisfaction I recognize the value operational measures taken in context. However, over the last 10 years of teaching usability evaluation and watching the results being interpreted by design teams I began to feel that existing measures were not capturing all that was of interest. Let me propose alternatives.

User experience can be thought of as existing at three levels:
· Process
· Outcome
· Affect

Process refers to the actions and responses involved in interacting with a device. Classically we have measured the screen-by-screen transactions of users and with good reason. But I am also interested in comparing such transactions with ideal paths, deviations from ideal, recovery from problems, situated cognition, and shifting motivations and attention across interaction sequences.

Outcome covers the range of variables that measure or refer to what the user attains from the interaction, both in the short and the long term. Classically we have measured task completion (an effectiveness score) but we might also consider learning, purchases made, details submitted, information located, or anything that brings closure to the user's interaction. Increasingly, different outcomes will be enabled by interactive technologies and our evaluations need to reflect these.

Affect covers the host of attitudinal, emotional and mood-related elements of experience. These exist in all human endeavor yet have been seriously overlooked in studies of usability. I envisage us paying much more attention to user choice, preference, perception of aesthetics, frustration and sense of enhancement or accomplishment.

Classic usability does a good job of covering some outcomes (e.g., task completion) but it misses many others. Usability measures satisfaction as if it were the only affective component worthy of consideration, and traditional process measures are hard to locate in most usability work beyond simple metrics of navigation. I am arguing for casting the net much further, to get at the type of data that will truly inform us of what users are experiencing with technologies.

Experiencing IT at 3 levels:

Put simply, the POA (Process, outcome, affect) approach emphasizes three key issues:

· What user does
· What user attains
· How user feels

The type of measures we may take for each of these are summarized briefly in Table 2:

Process: what user does	Outcome: What user attains	Affect: What user feels
· Navigation paths taken · Use of back button or links · Use of menus, help, etc. · Focus of attention	· What constitutes the end of the interaction? · Purchase made? · Details submitted? · Information located? · Comprehension attained?	Beyond satisfaction, we need to know if user feels: · Empowered? · Annoyed, frustrated? · Enriched? · Unsure or wary? · Confident? · Willing to come back?
Aim: Understanding user's moves and attention through the information space	Aim: Observe what it means for user to feel accomplishment or closure	Aim: Identify what the interaction means for the user in their world

Table 2: Measures and motives of the POA approach to evaluation

User experience = actions +result +emotion

Another, simplified, way of looking at this is to say that user experience is made up of actions, results and emotions, and we have not, as a field, been particularly strong in measuring these over the course of the last 20 years of software design. Indeed, cognitive science seems to have chosen to ignore emotion from the outset until it solved the problems of understanding rational thought and memory. However, even the most rigid rationalist must agree that much of what we do as humans is driven by emotional and affective factors.

This approach naturally scales up beyond traditional usability measures. First, by taking us into the terrain of user emotions we move beyond ability to use and into willingness or desire to use. Research in the business world has shown that ease of use is insufficient to predict intention to use (Davis et al 1989) and to understand more fully these dynamics we must move beyond the narrow focus of usability. I have written more on this in Dillon and Morris (1996) but the notion of emotion in use and adoption remains to be studied more fully.

Second, at the organizational level, the concept of usability is not powerful enough to allow us to understand what drives a group or team to exploit technology. Socio-technical theorists have long argued that enhancement of one's is a crucial determinant of technology adoption and diffusion (for an excellent overview of the socio-technical approach see Eason, 1988). I have written more on this in Dillon (2000).

'New' measures of user experience

In studying users in our lab, my students and I are keen to identify what makes interactive experience more interesting or appealing to a user. Further, we are trying to get at the dynamics of use that are more typical of discretionary users e.g., living with an application over time and learning to use it over repeated sessions (rather than training to a criterion and then giving responses to an experimenter, as in the traditional usability test). Below is a list of the major issues we are examining, but it should be noted that these represent ideas and research in progress, not completed programs. As such, what I describe is impressionistic rather than formal.

· Aesthetics,
· Perceived usability,
· Learning over time
· Cognitive effort,
· Perception of information shapes,
· Intention to use
· Self-efficacy

In the space provided here it is impossible to outline all of these or even any one of them in complete detail but in this section I will provide an example of how extending the analysis of interaction beyond usability has opened up new lines of enquiry that promise to yield exciting insights.

Aesthetics as a driver of user perception

I have been studying aesthetics and their impact on perceptions of usability since 1999, mostly with Maria Black (a former student of mine). Our approach has been to expose users to screen shots of designs and to have them rate the design both on aesthetics ("is it attractive, appealing or beautiful?") and likely or perceived usability ("do you think you would find it easy to use?").

In our first study we took 7 interface designs with known user performance data. These were presented originally in Tullis and Kodimer's (1993) study of alternative designs for a spreadsheet function. We asked 15 users to rate "aesthetics" and "likely usability" of each alternative design, and then compared our results with the original study's ranking and performance data. The results can be seen in Table 3 below:

Interface	Performance	Preference	Rating of Aesthetics	Perceived usability
A	2	4	1	3
B	7	5	1	1
C	6	6	4	2
D	1	1	3	4
E	4	2	6	5
F	3	3	5	3
G	5	7	7	7

We have conducted a follow up study with30 users who rated the aesthetics, the likely usability, and then used 4 web search interfaces
After use we asked them to rate aesthetics and usability again. We had anticipated that use would lead to a greater calibration of performance and perception ratings but once again we found that rankings, even after use showed no correlation with performance.

So what?

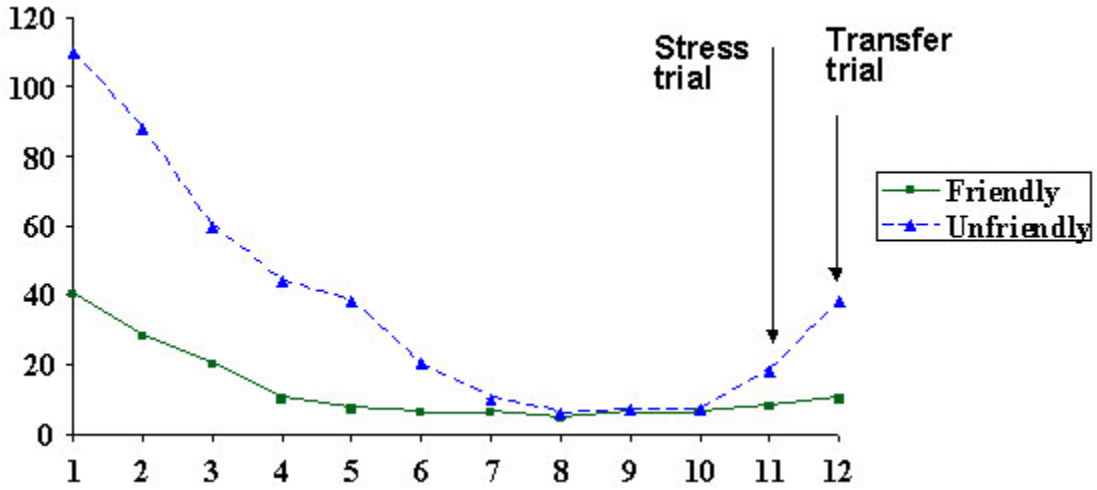
Users respond to interface 'beauty', and tend to relate usability to aesthetics in a fairly predictable fashion. Furthermore, users do not predict their own performance (process and outcome) accurately, especially at the early stages of use with a new interface. It is clear therefore from our studies of users that evaluations based on single experiences with technology do not provide stable estimates of long-term (or even medium-term) usability.

Use over time

Most usability tests are based on quick time slices of interaction, frequently captured as the user is learning to interact with the application for the first time. More accurately, such data provides a test of learnability, not usability, and these are not the same. Even where users may be trained to criterion before conducting the formal evaluation, there is still the sense that such interactions are occurring at the earliest stages of use and may not

be the best predictors of stable interaction. Yet most users are fairly stable in their choice of applications, investing the time to learn only rarely, sticking with applications for months or even years before changing or updating.

In a study tackling the pattern of interactive behavior over time, I examined 20 users attempting to master one of two versions of a bibliographic database, designer either in accordance with or in violation of design guidelines. The pattern of errors, for example, over 12 days of trials was highly informative and can be seen in the following graph:



What is seen in this graph is a telling indication of how interface designs matter both in the short term and under specific longer-term conditions. Here, users in the better-designed interface mastered the technology faster, and made significantly fewer errors from the outset. However, with practice, users of the poorly designed interface managed to overcome their difficulties and perform at a level where their error scores were equivalent to the users of the better interface. So, one may have thought from this that practice is sufficient.

However, what is even more telling are the results from days 11 and 12. Here, users were required to perform under stress (day 11) and to transfer to the alternative interface (day 12). That is, for the stress condition, users were told to work faster than they were typically performing. For the transfer task, users who learned on the bad interface now got to use the better one, and vice-versa. In both cases, the differences that had disappeared in the basic error scores on previous days reappeared at significant levels. In other words, while the regular error data suggested both sets of users were performing at an equivalent level, stress and transfer caused the users who learned on the poor interface to suffer a decrement in performance.

It would seem from these data that the quality of the design one learns on has both immediate and longer-term benefits. That users of the poor interface suffered a decrement in performance when they were offered a demonstrably better design runs counter to our naïve assumptions about usability, though it may have many parallels in our own lives.

It should be noted that the data in Figure 3 represents a behavioral source only - errors. If one listens to the verbal protocols that emerged from users in these studies it is clear that large differences could still be identified between user groups, even when their error scores indicated equivalence in performance. In other words, behavior alone is not a strong measure of what people are thinking.

Genres in information space

One final area I wish to talk about now is our research into information as space. With many new information resources being created, and the web giving rise to forms of information that have no obvious parallels in the paper domain, it is of some importance to examine how users are handling such information.

It is clear that any experienced user learns how information is laid out and the regularity of form that characterizes many established paper resources (such as journals, newspapers, fiction, manuals etc) can be used by both producer and consumer of information to aid communication. The term 'genre' is often used in this context, and while the definition can be fuzzy, genre really reflects cognitive expectations of the form of information under discussion. I coined the term 'shape' to help operationalize how this may work for digital information users and have been working on explicating this into measurable form via experimental studies.

We analyzed a sample of 100 home pages for features and produced a ranked list of features one finds on such pages. We then created 8 test pages that manipulated the presence or absence of the four most common and four least common features. New users were asked to rate the pages they thought were the 'best' home page designs (where 'best' was a subjective assessment of the users alone).

There was a significant positive correlation between the ratings of 'best' and the number of common features presented on a page. In other words - expectations for digital information spaces are forming quickly (there were no 'home pages' before 1995) and violation of expectancy would seem to impact initial user ratings (Dillon and Gushrowski, 2000).

What can we learn from this for user-centered design?

There are several lessons to learn from these data and they emphasize the importance of extending the classic usability approach to evaluation to include a more holistic set of user experience measures.

First, to be valid and reliable, user data in evaluation must reflect all aspects of the user experience. By this I mean that, as outlined here, we must learn to look at processes, outcomes and affect beyond simple task measures of efficiency, effectiveness and satisfaction.

Second, user experience is dynamic - most evaluations miss this. It is almost certain that a user's initial reaction to an interface is determined by multiple factors, chief among them being the perception of aesthetics, experience with equivalent designs, and immediate feedback. Vital as it is to understand these responses, a predictable shift occurs in users as they spend more time with an interface. What was initially seen as unusable often comes to be liked. What starts out appearing attractive and usable may later be disliked and rejected. Basing one's evaluation of a design solely on the first reactions of users can be extremely misleading, and evaluators need to build in a greater sense of the role of time when planning and conducting evaluations.

User data is the best indicator of interaction quality. Currently, our best theories are limited in terms of their applicability to design. However, we cannot retreat into the easy empiricism of current usability perspectives where everything is measured in terms of effectiveness, efficiency and satisfaction. Theory building must occur if we are to have long term impact and the diversity of experiences users can have with technology are not simply reduced to these operational criteria. We need to stretch our conception of interaction beyond performance and simple likes/dislikes. I argue for a richer sense of user experience, one that allows for aesthetics as much as efficiency and the creation of community discourse forms over time as much as the measurement of effectiveness in a single task. There is much work ahead but unless we embrace these issues as part of our research agenda, then the study of HCI will forever be piecemeal and weak, and its results will find little positive reception among the many designers and consumers who could most benefit from them.

Refs:

Davis, F. D. (1989). "Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology." MIS Quarterly 13(3): 319-340.

Dillon, A. (2000) Group dynamics meet cognition: applying socio-technical concepts in the design of information systems. In Coakes, E., Willis, D. and Lloyd-Jones, R. (eds) The New SocioTech: Graffiti on the Long Wall Springer Verlag Series on CSCW,London: Springer, 119-125.

Dillon, A. and Gushrowski, B. (2000) Genres and the Web - is the home page the first digital genre? Journal of the American Society for Information Science, 51,2,202-205

Dillon, A. and Morris, M. (1996) User acceptance of new information technology - theories and models. In M.Williams (ed.) Annual Review of Information Science and Technology, Vol 31, 3-32 Medford NJ: Information Today.

Eason, K. (1988) Information Technology and Organizational change. London: Taylor and Francis.

Frøkjær, E. Hertzum M. and Hornbæk, K. (2000) Measuring usability: are effectiveness, efficiency, and satisfaction really correlated? Proc. Of the ACM SIGCHI 2000 Conference. New York: ACM Press.

Tullis, T. and Kodimer. M (1992) A Comparison of Direct-Manipulation, Selection, and Data-Entry Techniques for Reordering Fields in a Table Proceedings of the Human Factors Society 36th Annual Meeting, 1992, pp. 298-302