Clarifying the Metrics of Usability

Version 1.0 April 11, 2001

James F. Zdralek
Department of Psychology; Carleton University
Ottawa, Canada

Abstract

Classification of the metrics used for the measurement of usability in the iterative design process as well as through out the product lifecycle will be clarified. The use of metrics for specifying usability will be compared with measuring usability for diagnostics. A different approach will be suggested combining the ISO definition of Usability, the theory of Innovation Diffusion, and incorporating the measures of Inherent and Apparent Usability.

Usability Metrics

Why use a usability metric? Europe has started to look at introducing legislation regarding usability in the work place. Usable software has shown value both in the short and long terms. Usability can create a marketing edge that differentiates a product in the market place.

Usability is an abused term and as such has needed a fair amount of clout from a large organization to define it.

ISO

The ISO 9245 specification uses a three-prong attack. It defines usability in terms of three measures and stabilizes the measures by requiring a specified context.

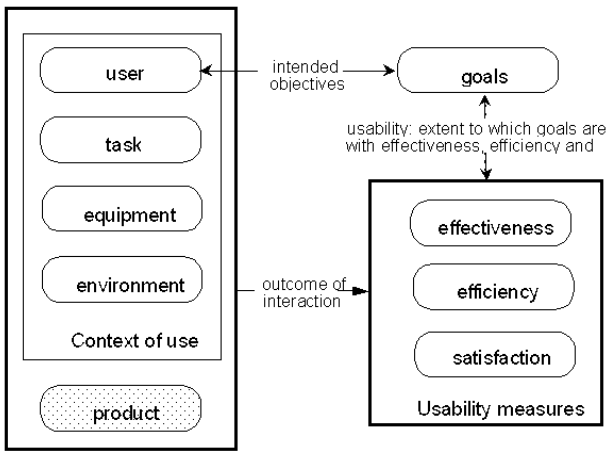

Figure 1 - ISO Diagram of Usability

> 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.
> (ISO DIS 9241-11)

For the purposes of simplicity and to ease combination with Innovation Diffusion the ISO diagram above will be modified. The product and equipment concepts are merged into a "tool" as well the goal is integrated into the task.

The Context

What might effect the measures of Usability? To illustrate, a usability metric for a piece of software would be different for a naïve user than a trained user. To properly measure usability the context should be fixed to enable repetition of the results. The Task, the User, the Tool and the Environment are the context of use. The attributes of these components to context can be discovered in qualitative rounds of research.

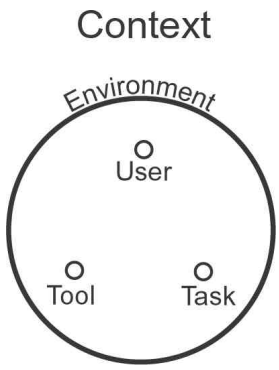

Illustration 1 - Context

These attributes are needed prior to measuring any Usability metric so that any measures detected as being low or high can be related back to the causes. It will also highlight any confounding situations where a testing situation may not reflect the real world.

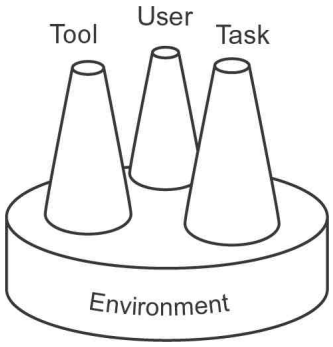

Illustration 2 - Attribute Distinction

The attributes that are components of a context are some times hard to categorize within the scheme. Is a light fixture an attribute of the environment or is it a tool? Adjusting the light might be a task. Where would such an attribute of the context be categorized within the scheme? As the illustration above shows, the categorization of attributes is a logical construct and does not require black and white answers. All things are connected and so an attribute can be part of two or more of the components that make up a context.

Two Terms For the Same Thing

In the ISO spec "Usability" as a term refers to metrics measured when the changes are made to the product being designed. A separate term, "Quality of a work system in use", is used to describe the same usability metric but only when the possibly to change other things exists. Although the majority of usability tests are done to achieve changes to the equipment, designs can also be recommendations made to change any of the elements that make up the context.

Change the Task.
Lengthen the time required to finish.
Change the order in which tasks are done.

Change the User.
Improve on the job training.
Require downtime to rest.

Change the Environment.
Improve the organization.
Improve the lighting.

Most often the improvements are on the equipment which in the HCI specific case is the software of the Graphical User Interface.

Plethora of Triads

Referring back to the 1st century BC, Kiana Matthews reiterates a concept that many have seen, the concept of the three facets of the quality of an experience that is defined in the ISO specification.

> The Roman architect Vitruvius said buildings should have firmness, commodity and delight. By this principle, human-computer interfaces should be robust, useful and pleasurable. Their design should be as much an art as a science.
> (Matthews 1999)

This short table shows the many authors that have conjured up this same concept or wrote of metrics that nicely fit under these three factors.

ISO	Satisfaction	Efficiency	Effectiveness
Dillon & Morris	User Attitude Perception	Usability Performance	Utility Power
Vitruvius	Delight	Firmness	Commodity
Matthews	Pleasurable	Robust	Useful
Neilson	Satisfaction	Memorable Learnable Errors Efficiency	
Klienonen	Affect Presentation	Ease Of Use Logic	Useful Functional
Alben	Aesthetic	Learnable-Usable Manageable	Mutable Appropriate Needed
Rogers	Observable Try-Able	Complexity	Compatibility Relative Advantage

Table 1 - Three Factors

Definition of Terms

Standards are great because there are so many to choose from. Descriptive terms are redefined in this paper so as to distinguish this approach from others and further develop a philosophical approach to "the quality of a work system in use".

Interaction is used to describe the overall measurement of the ISO definition of "Usability" because the term usability has been so abused that it is now synonymous with Ease of Use.

Satisfaction is used as described in the ISO definition.

The other two measures of Interaction use Dillon and Morris's terminology.

Utility is used to describe the ISO definition of Effectiveness**.**

Usability is used to describe the ISO definition of Efficiency. The main reason for not using the term efficiency is that types of efficiencies can be categorized differently depending on if it was a measure of Perceived Effort, Resource Efficiency or Time Efficiency.

	Interaction Quality		
	Satisfaction	**Usability**	**Utility**
ISO	Satisfaction	Efficiency	Effectiveness
Dillon & Morris	User Attitude Perception	Usability Performance	Utility Power
Vitruvius	Delight	Firmness	Commodity
Matthews	Pleasurable	Robust	Useful
Efficiency	Effort	Time	Resources

Table 2 - Multiple Triplets and Types of Efficiency

The table above shows how the new terms relate to the other triplets and to the efficiency measures. The illustration below associates the measures of interaction Quality with particular components of the context. This is to show focus and proximity of association. It does not imply that the Interaction Quality Factor of "Utility" is only associated with the "Task" component of context. Rather, the entire Context contributes to form the Interaction Quality when focus is placed on one of the components. Similar to how, while appreciating a painting, the entire picture is still visible but the field of view can be centered on a particular item.

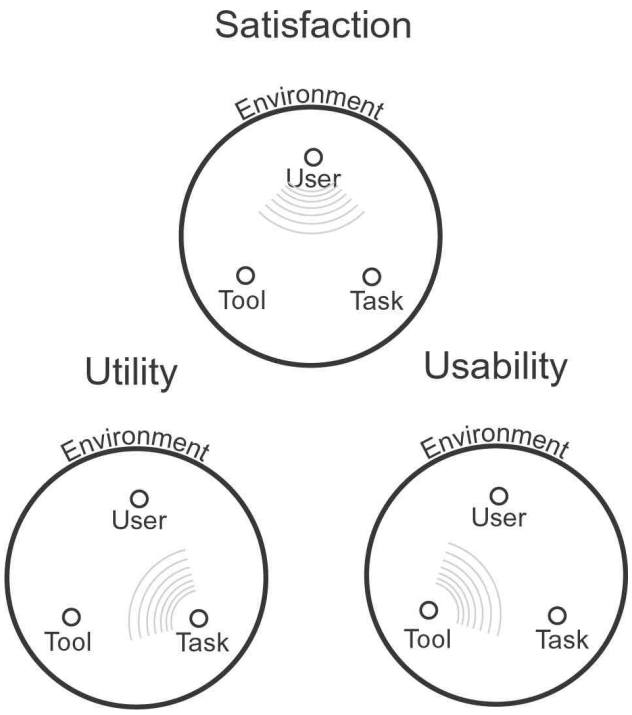

Illustration 3 - Factors of Interaction Quality

Interaction Quality

All of these factors of Interaction Quality can be thought of as waves that can create interference patterns.

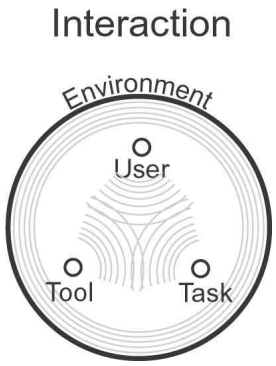

Illustration 4 - Interaction Quality

The metrics, repetitively ending in -ability or -able, can be thought of as some of the disturbances created by the interference patterns. The lack of a distinct association with only one of the factors of Interaction Quality shows how the categorization of metrics is a logical construct. This model does not require black and white categorizations of metrics into the factors composing Interaction Quality.

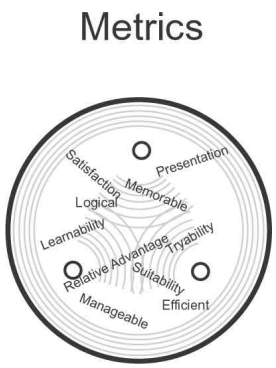

Illustration 5 - Metrics

Interaction Quality		
Satisfaction	Usability	Utility
Effort Efficient	Time Efficient	Resource Efficient
Satisfaction	Self Descriptive	Functional
Affect	Learnable	Portable
Appearance	Try-able	Suitable
Aesthetics	Observable	Compatible
	Individualize	Advantageous
	Memorable	Needed
	Errors	Appropriate
	Ease	
	Logic	
	Manageable	

Table 3A - Aggregate of Categorized Metrics

The table above shows the aggregate of metrics from the various usability definitions. The categorization is done associate the metric with the most logical factor. Using the alternate table below can solve arguments over whether a metric belongs to one factor or the other.

Retentive or expulsive, Type A or Type B, the choice of table is irrelevant but the key concept should be clear. The metric can be influenced by all the Context categories and be classified into more than one Interaction Quality factor.

Interaction Quality

Satisfaction	Usability	Utility
Effort Efficient	Time Efficient	Resource Efficient
Satisfaction	Self Descriptive	Functional
Affect	Learnable	Portable
Appearance	Try-able	Suitable
Aesthetics	Observable	Compatible
Individualize	Memorable	Needed
Ease		Appropriate
	Errors	Advantageous
	Logic	Manageable

Table 3B - Aggregate of Metrics

The scope of this paper is not to itemize the list of metrics available but to position usability within a broad conceptual picture involving the end result and adoption of the design by a population.

Quick and Dirty

In an industry rampant with the "quick and dirty methods", metrics proponents are left with little ammunition. Trying to convince people to do the research accurately with standard metrics is difficult. Research has been done as to whether one encompassing metric could be measured and the value used to predict the rest. (Frøkjær, Hertzum, Hornbæk, 2000) Research has also been done into the intercorrelations between metrics. (Keinonen 1997) Other questions have also been raised. Do the Metrics measure what they are categorized under. Should the face value of the measure be taken? How do we relate cause and effect? Do we look at mental artifacts or find clues to illogical interface elements. These are questions that may eventually get answered but much more must be done in the way of research. At present quick and dirty seems to get the funding.

Use in Development

What about the iterative design process? Interaction Quality metrics cannot be used as a guideline or a checklist. It is a measuring device.

> When guidelines are used for evaluation the result is a checklist showing which parts of a product conform to which recommendations. While this can be used to identify potential problems with the interface, it cannot be turned into measures, as there is no way to accurately weight the importance of the different recommendations.
> (Bevan & Macleod 1994)

Blood pressure is related to the state of the body. If a patient's blood pressure were extremely low a doctor would not immediately recommend medication. If the blood pressure is extremely low then relate it to the state of the body and situation to determine a solution. Is there a hole somewhere? Are any parts missing?

Interaction Quality metrics are clues to help designers focus on the attributes of Context that will inspire them to improve the design.

Pigeonholed

Typically in the industry people have been pigeonholed with their job function. As usability research evolves these job functions will cross-pollinate. Eventually the same lack of black and white categorizations that is exhibited in the metrics of Interaction Quality factors and attributes of Context will enter into the interface design field. As in the table below, User needs Assessment Research is at a distance from Satisfaction and Look and Feel while Visual Design is at a distance from Utility and Requirements. Designers and researchers should both be involved in the definition and gathering of Context attributes and Interaction Quality metrics.

Interaction Quality

Satisfaction	Usability	Utility
Visual Designer	Usability Engineer	User Needs Assessment
Look and Feel	3 Clicks -7 Items	Requirements

Table 4 - Pigeonholed Relationships to Job Function

Comparison or Spec

How should we use the measures? Should it be used to determine whether a delivered product meets a required level of usability? Or should it be used to provide feedback during the design process to help identify potential usability pitfalls.

Designs are improved upon by finding a Metric of Interaction Quality with a value that is questionable. Relating it to the context and generating improvements to the design.

To determine the Metrics of Interaction that have the most impact, the context must be looked at.

	Interaction	**Context**	
	Satisfaction Usability Utility	Users Tool Task Environment	
Spec	← ¾¾¾¾¾¾		
	¾¾¾¾¾¾→		Design

Table 5 - Means to an end

At the start of an iterative design process the context is often looked at to generate the design. It should also be looked at to determine which Metrics of Interaction should be selected to act as guidance or as a spec to adhere.

The first abuse is generating it by not looking at the Users, Task and Environment. Without that grounding in the context the Metric may be of little consequence while a valuable metric goes unmeasured.

The second abuse comes in the assigning of a number. This is a matter for debate. What does the number mean? Will improving one Metric have dangerous effects on other Metrics?

Subjective and Objective

Metrics can come in two distinct flavors. Subjective measures will give you a stated preference, a feeling, a human answer. Object measures will give you the scientific facts.

Currently there are standard tests to determine some subjective measures of Usability such as NASA's TLX, Task Load Index. MUSiC has a SMEQ, Subjective Mental Effort Questionnaire, as well as a SUMI, Software Usability Measurement Inventory, to determine through a questionnaire the user's perception of software quality.

There are also some objective measures of mental effort such as Heart Rate Variability and objective measures of performance such as task time and errors.

Measuring Before, During And After

When measuring an objective metric it is usually during the use of a product. Utility can be measured. Did it do the Job? Usability can be measured. How many mistakes were there? And theoretically you can measure Satisfaction through dopamine levels in the blood.

Measuring Interaction Quality after use, as is normally done, is a measure of the Experienced Interaction Quality.

Measuring Interaction Quality before use is a measure of the Expected Interaction Quality.

Expected:
User is asked to predict the different values.

Inherent:
The users' actions are measured.

Experienced:
User is asked to report the different values.

Expected values can have an impact on whether the product is tried. Bad results in these areas may predict a problem with the interface functionality not being "Self Evident"

The questions a user might ask themselves about each of the factors composing the Expected Interaction Quality is shown in the table below.

Interaction Quality

Satisfaction	Usability	Utility
Do I like it?	Will it be easy?	Can it do it?

Table 6 - Questions

Inherent values can have an impact on whether the product fulfills the contract. Bad results in these areas may predict a problem with the interface functionality not being "Within Spec"

Experienced values can have an impact on whether the product is accepted. Bad results in these areas may predict a problem with the interface functionality not being "Rewarding"

> The four major factors that influence the diffusion process are the innovation itself, how information about the innovation is communicated, time, and the nature of the social system into which the innovation is being introduced.
> (Rogers 1995)

Innovation Diffusion

Five distinct phases of The Innovation Decision Process theory (Rogers, 1995) are:

Knowledge:	Learn about the innovation.
Persuasion:	Be persuaded as to the merits.
Decision:	Decide to adopt.
Implementation:	Implement the innovation.
Confirmation:	Reaffirm or reject the decision.

The Metrics associated with Expected, Inherent and Experienced Usability fit into these phases through a second associated theory put forth by Rogers called The Theory of Perceived Attributes which states that potential adopters judge an innovation based on their perceptions in regard to five attributes of the innovation. The perceived attributes of Trial-ability, Observe-ability, Relative Advantage, Complexity, and Compatibility all fit under the three measures of Usability. The theory applies them only to "Expected Values" and ignores Inherent or Experienced Values. Those that have "Experienced Values of Usability" have an influence in the first two stages of Knowledge and Persuasion. Observe-ability can be thought of as a relation of Learn-ability in that features or methods of the product should be easily demonstrated. Trial-ability is the

Knowledge: Learn about the innovation.

Persuasion: Be persuaded as to the merits.

Decision: The decision phase of Innovation diffusion interacts with the "Expected Usability Metrics". It is at this point in the Diffusion of a product that a person decides to try the product. This is an investment in time and resources and is influenced by what they have heard about the product.

Implementation: Implement the innovation.

Confirmation: The confirmation phase of Innovation diffusion interacts with the "Experienced Usability Metrics". It is at this point in the Diffusion of a product that a person decides to continue use and also forms opinions that they will express to others.

It is not possible to measure the Expected, Inherent and Experienced Interaction Quality at all phases of development: Specification, Mock up, Early prototype and Final product. As the prototype or spec of a product develops different phases of the Innovation Decision process can be explored.

In the specification phase of product development, the knowledge and persuasion stages of innovation diffusion can be measured for usability purposes. What effects the users' perception of the product when they first hear about it?

These are very abstract usability goals.

Measuring the "Describable" usability prior to use. What will people be able to say about their experience with a product?

Measuring the "Observable" usability prior to use. While watching the product in use will the user be able to see the effect on the goal.

Measuring the "Apparent" usability prior to use. What judgements will a user make by just seeing the product?

In the Mock up phase of product development, the "decision" stage of innovation diffusion can be measured for usability purposes. What effects the users' perception of the product when they first see it?

Measuring subjective Usability of the product and taking into account how Expected, Inherent and Experienced Usability Metrics will interact with the innovation diffusion process will give designers a Measure of the attitudes towards the product once it is "in the wild". These measures reflect something more than Usability. These measures can give clues as to whether a product will suffer from a failure to thrive through no fault of Usability or Quality of Interaction.

Potency

Impact is the change that occurs during use. It relates to spontaneous order and chaos that any complex system will exhibit.

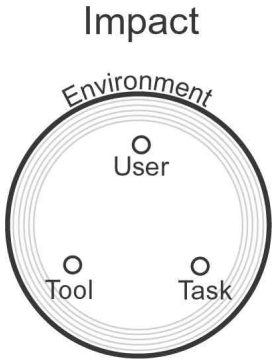

Illustration 3 - Impact

Potency describes Dillon and Morris's "User Attitude" which is also related to Kurosu and Kashimuras' "Apparent Usability".

Potency measures the attributes the user is aware of and how the users feeling toward the product effect implementation in the real world. This can be viewed as a longitudinal measure of Usability but incorporates aspects of marketing. Not the advertising aspect of marketing, but the matching of sizzle to steak.

The language of metrics, used for product evaluation during the formative and iterative design phases as well as performance "in the wild", needs to be clarified.

The Expected, Inherent and Experienced Interaction Quality of a product describe the Potency.

The Satisfaction, Usability and Utility of a product describe the factors that compose Interaction Quality.

The numerous Metrics describe the factors of Satisfaction, Usability and Utility.

The value of the Metrics indicates the match between the Tool, Users, Task and Environment for that particular component of the Interaction.

The Metric can be used as spec or an indicator to find and improve the Potency and Interaction Quality of a product.

References

Alben, L., 2000. Quality of experience: Defining the Criteria for Effective Interaction Design

B.E.R. de Ruyter & J.H.M. de Vet, Usability evaluations in user-centered design

Bevan, N., Macleod, M., 1994, Usability measurement in context, Behaviour and Information Technology, 13, pg 132-145

Dillon, A., and Morris, M., 1997, Power, Perception and Performance: From Usability Engineering to Technology Acceptance with the P3 Model of User Response, Indiana University and Air Force Institute of Technology

Dillon, A., Coakes, D. (et al) (Eds.), Group dynamics meet cognition: combining socio-technical concepts and usability engineering in the design of information systems, The Socio-Technical Revival. London: Springer-Verlag

Frøkjær, E., Hertzum, M., and Hornbæk, K., 2000. Measuring usability: Are effectiveness, efficiency, and satisfaction really correlated? ; Proceedings of the CHI 2000 conference on Human factors in computing systems, Pages 345 - 352

Gleiss, N., Håvard Skjetne, J., Richardson, S., Heim, J., Usability of Telecommunications - Equipment and Services, Telecommunications for all, Chapter 2.4

International ISO DIS 9241-11 Standard, 1994. Ergonomic requirements for office work with visual display terminals (VDTs): -Part 11: Guidance on Usability

Keinonen, T., 1997. Expected usability and product preference, Proceedings of the conference on designing interactive systems: processes, practices, methods, and techniques, Pages 197 - 204

Kurosu, M., Kashimura, K., 1995. Apparent Usability vs. Inherent Usability: Experimental Analysis on the Determinants of the Apparent Usability, Short Papers: Usability Evaluation, Proceedings of ACM CHI'95 Conference on Human Factors in Computing Systems v.2 p.292-293

Lindgaard, G., 2000. Aesthetics, User Experience, and HCI - do we have to bother? http://www.buildingsmallbusiness.com/pages/aesthetics.htm

Lindgaard, G., 2000. Designing Communication Systems for People: Making Sense Between the Lines, http://www.buildingsmallbusiness.com/pages/micon.htm

Matthews, K.K., 1999. Aesthetics and Usability; Paper: Advanced Topics in User Interface Design; University of Colorado at Boulder

Nielsen, J., 1999. How to Measure Usability, http://www.zdnet.com/devhead/stories/articles/0,4413,2321381,00.html

Nielsen, J., and Levy, J., 1994. Measuring usability: preference vs. performance. Communications ACM 37, 4, Pages 66 - 75

Rogers, E. M., 1995. Diffusion of innovations (4th ed.), New York, The Free Press

Surry, D.W., 1997. Diffusion Theory and Instructional Technology, University of Southern Mississippi, Annual Conference of the Association for Educational Communications and Technology (AECT), Albuquerque, New Mexico.

Tractinsky, N., 1997. Aesthetics and Apparent Usability: Empirically Assessing Cultural and Methodological Issues, PAPERS: Beauty and The Beat; Proceedings of ACM CHI 97 Conference on Human Factors in Computing Systems v.1 p.115-122