USABILITY EVALUATION: AN EMPIRICAL VALIDATION OF DIFFERENT MEASURES TO QUANTIFY INTERFACE ATTRIBUTES

Matthias Rauterberg

Work and Organisational Psychology Unit
Swiss Federal Institute of Technology (ETH)
Nelkenstrasse 11, CH–8092 Zuerich
Tel: +41-1-63-27082, Email: rauterberg@rzvax.ethz.ch

Abstract: One of the main problems of standards (e.g., DIN 66234, ISO 9241) in the context of usability of software quality is, that they can not be measured in product features. We present a new approach to measure user interface quality in a quantitative way. First, we developed a concept to describe user interfaces on a granularity level, that is detailed enough to preserve important interface characteristics, and is general enough to cover most of known interface types. We distinguish between different types of 'interaction points'. With these kinds of interaction points we can describe several types of interfaces (CUI: command, menu, form-fill-in; GUI: desktop, direct manipulation, multimedia, etc.). We carried out two different comparative usability studies to validate our quantitative measures. The results of one other published comparative usability study can be predicted. Results of six different interfaces are presented and discussed.

Keywords: user interfaces, utility functions, testability, quantization.

1. INTRODUCTION

One of the main problems of standards (e.g., DIN 66234, ISO 9241) to quantify software quality of usability is, that they can not be measured in product features (Kirakowski and Corbett, 1990). Four different views on human computer interaction to measure interactive qualities currently exists (see also Bevan, et al., 1991, p. 651; Rengger, 1991).

(1) The *interaction-oriented view:* usability quality is measured in terms of how the user interacts with the product ("usability testing"). This view is the most common one. All kinds of usability testing with "real" users are subsumed in this category (IFIP, 1981).

(2) The *user-oriented view:* usability quality is measured in terms of the mental effort and attitude of the user ("questionnaires" and "interviews").

(3) The *product-oriented view:* usability quality is measured in terms of the ergonomic attributes of the product itself (quantitative measures).

(4) The *formal view:* usability is formalised and simulated in terms of mental models (formal concepts). Karat (1988) describes formal methods in the context of "theory-based" evaluation.

The interactive qualities of user interfaces currently are quantified in the context of *interaction-oriented view* and *user-oriented view*, but these both approaches are time consuming and more or less expensive (Jeffries and Desurvire, 1992).

2. A DESCRIPTIVE CONCEPT OF INTERACTION POINTS

We present a new approach to measure user interface quality in a quantitative way. First, we developed a concept to describe user interfaces on a granularity level, that is detailed enough to preserve important interface characteristics, and is general enough to cover most of known interface types (command language, CUI, GUI, multimedia, etc.). Different types of user interfaces can be quantified and distinguished by the general concept of "interaction points". Regarding to the interactive semantic of "interaction points" (IPs), different types of IPs must be discriminated (see also Denert, 1977).

An interactive system can be distinguished in a dialog and an application manager. So, we distinguish between dialog objects (DO, e.g. "window") and application objects (AO, e.g. "text document"), and dialog functions (DF, e.g. "open window") and application functions (AF, e.g. "insert section mark"). Each function $f \in FS$, that changes the state of an application object, is an application function. All other functions are dialog functions (e.g., window operations like move, resize, close). The complete set of all description terms is defined in Table 1.

Table 1 The interactive space (IS) consists of the object space (OS) and the function space (FS); FS can be distinguished in perceptible and hidden interactive functions (PF and HF).

$IS := OS \cup FS$	[interaction space]
$DC \in IS$	[dialog context]
$OS := PO \cup HO$	[object space]
$FS := PF \cup HF$	[function space]
$PO := PDO \cup PAO$	[(perceptible) representations of objects]
$HO := HDO \cup HAO$	[hidden objects]
$PF := PDFIP \cup PAFIP$	[(perceptible) representations of functions]
$HF := HDFIP \cup HAFIP$	[hidden functions]
$PDFIP := \{(df,pf) \in HDFIP \times PF: pf = \delta(df)\}$	[(perceptible) represented DFIP]
$PAFIP := \{(af,pf) \in HAFIP \times PF: pf = \alpha(af)\}$	[(perceptible) represented AFIP]
$IP := DFIP \cup AFIP$	[interaction points]
$DFIP := PDFIP \cup HDFIP$	[IPs of dialog functions]
$AFIP := PAFIP \cup HAFIP$	[IPs of application functions]
δ := mapping function of a $df \in HDFIP$ to an appropriate $pf \in PF$.	
α := mapping function of an $af \in HAFIP$ to an appropriate $pf \in PF$.	
$PDO := \{(do,po) \in HDO \times PO: po = \mu(do)\}$	[(perceptible) represented DO]
$PAO := \{(ao,po) \in HAO \times PO: po = \nu(ao)\}$	[(perceptible) represented AO]
μ := mapping function of a dialog object $do \in DO$ to an appropriate $po \in PO$.	
ν := mapping function of an application object $ao \in AO$ to an appropriate $po \in PO$.	

A dialog context (DC) is defined by all available objects and functions in the actual system state. If in the actual DC the set of available functions changes, then the system changes from one DC to another. All dialog objects (functions, resp.) in the actual DC are perceptible (PO, PF) or hidden (HO, HF). Four different mapping functions relate perceptible structures to hidden objects or functions.

Each interaction point (IP) is related to at least one interactive function. If both mapping function's δ and α are of the type 1:m(any), then the user interface is a command interface. If both mapping function's δ and α are of the type 1:1, then the user interface is a menu or direct manipulative interface where each $f \in FS$ is related to a perceptible structure PF (see Figure 2). The perceptual structure (visible, audible, or tactile) of a function (PF) can be, e.g., an icon, earcon, menu option, command prompt, or other mouse sensitive areas.

The intersection of PF and PO is sometimes not empty: $PF \cap PO \neq \emptyset$. In the context of graphical interfaces icons are elements of this intersection, e.g., PDFIP "copy" $\equiv$ PDO "clipboard", PAFIP "delete" $\equiv$ PAO "trash". Each interaction point (IP) is related to at least one interactive function (see Figure 1).

One important difference between a menu and a direct manipulative interface can be the "interactive directness". A user interface is 100% interactively direct, if the user has fully access in the actual dialog context to all $f \in FS$ (Laverson, et al., 1987). Good interface design is characterised by optimising the multitude of DFIPs (e.g. "flatten" the menu tree; see Paap and Roske-Hofstrand, 1988) and by allocating an appropriate PDFIP to the remaining HDFIPs.

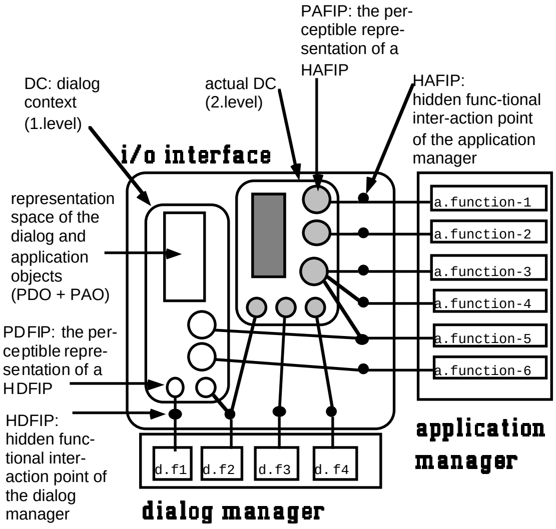
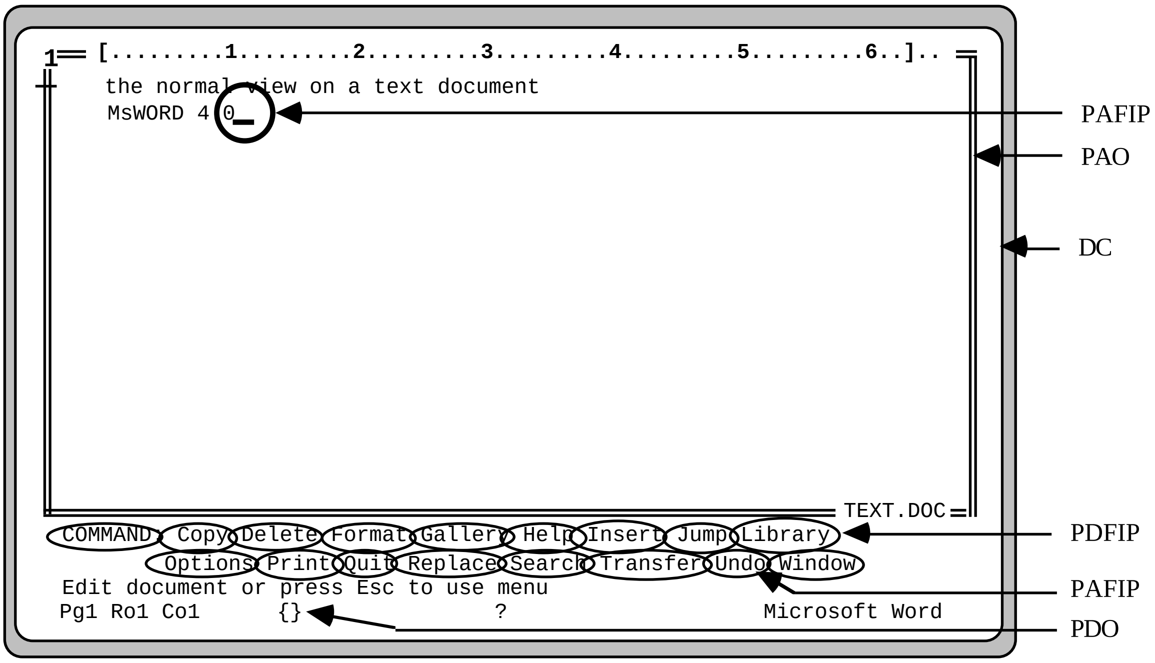

PAFIP: the perceptible representation of a HAFIP
DC: dialog context (1.level)
actual DC (2.level)
HAFIP: hidden func-tional inter-action point of the application manager
i/o interface
representation space of the dialog and application objects (PDO + PAO)
PDFIP: the perceptible representation of a HDFIP
HDFIP: hidden functional interaction point of the dialog manager
a.function-1
a.function-2
a.function-3
a.function-4
a.function-5
a.function-6
application manager
d.f1
d.f2
d.f3
d.f4
dialog manager

Fig. 1. A schematic presentation of the I/O interface, the dialog and the application manager of an interactive system with a menu tree of two levels.



```
1═ [........1.........2.........3.........4.........5.........6..].. ═
     the normal view on a text document
     MsWORD 4 0_
                                                          TEXT.DOC
 Edit document or press Esc to use menu
 Pg1 Ro1 Co1      {}              ?                  Microsoft Word
```

Fig. 2. An actual dialog context (DC) of the text processing program MsWord with the representation space of the interactive object (PAO: text document; PDO: clipboard), and the representation space (PF: marked by circles) of the interactive functions (PAFIP: text entry point, undo; PDFIP: menu options).

3. FOUR QUANTITATIVE MEASURES OF INTERFACE ATTRIBUTES

To estimate the amount of "feedback" of an interface a ratio is calculated: "number of PFs" (#PF = #PDFIP + #PAFIP) divided by the "number of HFs" (#HF = #HDFIP + #HAFIP) per dialog context. This ratio quantifies the average "amount of functional feedback" of the function space (FB; see Formula 1). We abbreviate the number of all different dialog contexts with D. A GUI has often a very large number of DCs. To handle this problem we take only all task related DCs into account. Doing this, our measures will give us only a lower estimation for GUIs.

The average length of all possible sequences of interactive operations (PATH) from the top level dialog context (DC, e.g., 'start context') down to DCs with the desired HAFIP or HDFIP can be used as a possible quantitative measure of "interactive directness" (ID, see Formula 2). The measure ID delivers two indices: one for HAFIPs and one for HDFIPs. A PATH has no cycles and has not more than two additional dialog operations compared with the shortest sequence. An interface with the maximum ID of 100% has only one DC with path lengths of one dialog step. We abbreviate the number of all different dialog paths with P.

$$\text{Functional feedback:} \quad \text{FB} = 1/\text{D} \sum_{d=1}^{D} (\#\text{PF}_d \,/\, \#\text{HF}_d) \quad * \; 100\% \qquad (1)$$

$$\text{Interactive directness:} \quad \text{ID} = \left\{ 1/\text{P} \sum_{p=1}^{P} \text{lng}(\text{PATHp}) \right\}^{-1} * \; 100\% \qquad (2)$$

$$\text{Application flexibility:} \quad \text{DFA} = 1/\text{D} \sum_{d=1}^{D} (\#\text{HAFIP}_d) \qquad (3)$$

$$\text{Dialog flexibility:} \quad \text{DFD} = 1/\text{D} \sum_{d=1}^{D} (\#\text{HDFIP}_d) \qquad (4)$$

To quantify the flexibility of the application manager we calculate the average number of HAFIPs per dialog context (DFA; see Formula 3). To quantify the flexibility of the dialog manager we calculate the average number of HDFIPs per dialog context (DFD; see Formula 4). A modeless dialog state has maximal flexibility (e.g., "command" interfaces, or Oberon; Wirth and Gutknecht, 1992).

Let us apply the five measures to our example in Figure 1. The average amount of functional feedback is:
FB = (4/4 + 6/8) /2 * 100% = 87.5%.
The average amount of interactive directness is:
$ID_{HAFIP} = ((2*1 + 5*2) / 7)^{-1} * 100\% = 58.3\%$;
$ID_{HDFIP} = ((2*1 + 3*2) / 5)^{-1} * 100\% = 62.5\%$.
The average amount of flexibility is:
DFA = (2 + 5) / 2 = 3.5 and
DFD = (2 + 3) / 2 = 2.5.
To interpret the results of our measures appropriately, we need empirical studies.

4. RESULTS AND DISCUSSION

We carried out two different comparative usability studies to validate our measures (Rauterberg, 1992; Brunner and Rauterberg, 1993). A third external comparative study (Grützmacher, 1988) was used for a cross validation. All three investigated software products have the same application manager, but two different dialog managers each.

4.1 Results of experiment-I

We compared an old CUI-interface of a relational database management system with a new GUI-interface (Rauterberg, 1992). The main result of this empirical investigation was, that the mean task solving time with the GUI is significantly shorter than with the CUI interface. How can we explain this difference? Our first interpretation of this outcome was the supposed different amount of 'transparency' (Ulich, et al., 1991). One aspect of 'transparency' is 'feedback' (see Dix, et al., 1993, pp. 318-321).

Interesting is the fact, that the GUI supports the user with less "visual feedback" (FB = 66%, see Table 3) on average than the CUI (FB = 73%). This amount of FB of the CUI is caused by 22 small DCs with FB = 100%; the GUI has only 14 DCs with FB = 100%. The amount of functional feedback seems not to be related to the advantage of GUIs. There must be another reason.

The "interactive directness" is not quite different between both interfaces
(CUI: ID = 24.7% for AFIPs and 23.2% for DFIPs versus
GUI: ID = 22.5% for AFIPs and 25.5% for DFIPs, see Table 3).

Only the two measures of "flexibility" show an important difference between both interfaces
(CUI: DFA = 12.1 and DFD = 10.1 versus
GUI: DFA = 19.5 and DFD = 20.4, see Table 3).
We interpret this result to the effect that flexibility must exceed a threshold to be effective (DFD, DFA > 15).

4.2 Results and discussion of experiment-II

If our interpretation of the outcome of experiment-I is correct then we can not find a significant performance difference for dialog structures that remain under the assumed threshold of 15. To control the factor of feedback we carried out this second experiment with a multimedia information system that has 100% functional feedback for both interfaces (Brunner and Rauterberg, 1993).

We picked out a multimedia information system with a hierarchical dialog structure where DFA and DFD are clearly under 15. We implemented a comparable system with a net-shaped dialog structure where DFA and DFD have nearly the same ratio of flexibility as in experiment-I:
$DFA_{GUI} / DFA_{CUI} = 1.6$ and
$DFA_{MMnet} / DFA_{MMhier} = 1.2$;
$DFD_{GUI} / DFD_{CUI} = 2.0$ and
$DFD_{MMnet} / DFD_{MMhier} = 2.6$.

Table 2 Overview of our three empirical validation studies. DV means 'dependent variable' in the analysis of variance. The alpha-error is abbreviated with p.

Experiment	Interface type and dialog structure	Application type	Source of the empirical comparison study	Result of the empirical comparison study
I	CUI-hierarchical	Relational database	Rauterberg (1992)	DV: Task solving time
I	GUI-hierarchical	Relational database	Rauterberg (1992)	GUI << CUI (p≤.002)
II	Multimedia-hierarchical	Information system	Brunner & Rauterberg (1993)	DV: Task solving time
II	Multimedia-net shaped	Information system	Brunner & Rauterberg (1993)	MM_{hier} <= MM_{net} (p≤.085)
III	CUI-hierarchical	Simulation tool	Grützmacher (1988)	DV: Target discrepancy
III	CUI-net shaped	Simulation tool	Grützmacher (1988)	CUI_{hier} == CUI_{net} (p≤.784)

Table 3 Comparison our three empirical validation studies relating to the quantitative measures ID, FB, DFA, and DFD. P is the number of all different dialog PATHs for an AFIP or a DFIP; D is the number of all different DCs.

Expe-riment	Interface type and dialog structure	$P_{(AFIP)}$	$ID_{(AFIP)}$ %	$P_{(DFIP)}$	$ID_{(DFIP)}$ %	D	FB %	DFA	DFD
I	CUI-hierarchical	434	24.7	362	23.2	36	73	12.1	10.1
I	GUI-hierarchical	547	22.5	570	25.5	28	66	19.5	20.4
II	Multimedia-hierarchical	241	25.1	34	28.1	68	100	3.6	0.5
II	Multimedia-net shaped	276	40.7	87	46.3	65	100	4.2	1.3
III	CUI-hierarchical	720	20.9	693	23.9	363	86	2.0	1.9
III	CUI-net shaped	490	15.8	1053	21.9	389	90	1.3	2.7

As we predicted, we can not find a significant performance difference between both types of dialog structures (see Table 2). To make sure that our results are not biased by our own expectations, we carried out a cross validation study. To do this, (1) we need the outcomes of an external independent comparison study between two different interfaces and (2) the possibility to apply our quantitative measures to all DCs of both interfaces. The empirical investigation of Grützmacher (1988) fulfils both conditions.

4.3 Results and discussion of experiment-III

The study of Grützmacher (1988) was carried out to investigate research questions in the context of how to control a complex domain with a simulation tool. One independent factor was varied: the dialog structure (hierarchical versus net-shaped). This simulation tool was implemented on a mainframe computer system with character oriented terminals (IBM 3270).

The dependent variable was not 'task solving time' but 'target discrepancy' as a performance measure. The sample consists of 20 users with the hierarchical dialog structure and 15 users with the net-shaped structure. The main result was that the factor 'dialog structure' was not significant. Given our interpretation of the last two experiments we expected a value for DFA and DFD under 15.

With the generous support of Grützmacher we were able to analyse all 752 dialog contexts for both interfaces. For the hierarchical CUI we get the following results: DFA = 2.0 and DFD = 1.9; for the net-shaped CUI: DFA = 1.3 and DFD = 2.7 (see last two rows in Table 3). These results for DFA and DFD of both CUI interfaces give us a sufficient evidence that the following assumptions are correct:

(1) We can measure the dialog flexibility in a task independent and quantitative way, and
(2) the values of DFA and DFD must exceed the threshold of 15.

5. CONCLUSION

Using the four quantitative measures for "feedback", "interactive directness" and "flexibility" to measure the interactive quality of user interfaces, we are able to classify the most common types: command, menu, desktop (see Rauterberg, 1993). The command interface is characterised by high interactive directness, but this interface type has a very low amount of visual feedback. Especially graphical interfaces (e.g., multimedia) can support users with sufficient interactive directness. GUIs are characterised by high dialog flexibility.

The presented approach to quantify usability attributes and the interactive quality of user interfaces is a first step in the right direction. The next step is a more detailed analysis of the relevant characteristics and validation of these characteristics in further empirical investigations. In the context of standardisation we can use our criteria to test user interfaces for conformity with standards.

ACKNOWLEDGEMENTS

We have to thank the following persons for their generous support: Dr. Karl Schlagenhauf and Raimund Mollenhauer at ADI Software Inc., Karlsruhe (D), Andreas Grützmacher and Prof. Eberhard Ulich at the ETH, Zürich (CH).

REFERENCES

Bevan, N., J. Kirakowski and J. Maissel (1991). What is Usability? In: *Human Aspects in Computing: Design and Use of Interactive Systems with Terminals* (H-J. Bullinger, (Ed.)), 651-655. Elsevier, Amsterdam.

Brunner, M. and M. Rauterberg (1993). *Hierarchische oder netzartige Dialogstruktur bei multimedialen Informationsystemen: eine experimentelle Ver-*

gleichsstudie. Technical Report MM-2-93. Institut für Arbeitspsychologie, Eidgenössische Technische Hochschule, Zürich.

Denert, E. (1977). Specification and design of dialogue systems with state diagrams. In: *International Computing Symposium 1977* (E. Morlet and D. Ribbens (Eds.)), 417-424. North-Holland, Amsterdam.

Dix, A., J. Finlay, G. Abowd and R. Beale (1993). *Human-Computer Interaction*. Prentice Hall, New York.

Grützmacher, B. (1988). *Datenpräsentation und Lösungsverhalten in einer komplexen, simulierten Problemsituation*. Unpublished Master Thesis. (Philosophische Fakultät I, Psychologisches Institut, Abteilung Angewandte Psychologie). Universität Zürich, Zürich.

IFIP (1981). *Report of the 1st Meeting of the European User Environment Subgroup of IFIP WF 6.5*. German National Center for Computer Science (GMD), P.O. 1316, D-5202 Sankt Augustin (Germany).

Jeffries, R. and H. Desurvire (1992). Usability testing vs. heuristic evaluation: was there a contest? *SIGCHI Bulletin* **24**(4), 39-41.

Karat, J. (1988). Software Evaluation Methodologies. In: *Handbook of Human-Computer Interaction* (M. Helander, (Ed.)), 891-903. Elsevier, Amsterdam.

Kirakowski, J. and M. Corbett (1990). Effective Methodology for the Study of HCI. In: *Human Factors in Information Technology Vol. 5* (H. Bullinger and P. Polson, (Eds.)). North-Holland, Amsterdam.

Laverson, A., K. Norman and B. Shneiderman (1987). An evaluation of jump-ahead technique in menu selection. *Behaviour and Information Technology* **6**(2), 97-108.

Paap, K. and R. Roske-Hofstrand (1988). Design of menus. In: *Handbook of Human-Computer Interaction* (M. Helander, (Ed.)), 205-235. Elsevier, Amsterdam.

Rauterberg, M. (1992). An empirical comparison of menu-selection (CUI) and desktop (GUI) computer programs carried out by beginners and experts. *Behaviour and Information Technology* **11**(4), 227-236.

Rauterberg, M. (1993). Quantitative Measures to Evaluate Human-Computer Interfaces. In: *Human-Computer Interaction: Applications and Case Studies* (M. Smith and G. Salvendy, (Eds.), Advances in Human Factors/Ergonomics Vol. 19A), 612-617. Elsevier, Amsterdam.

Rengger, R. (1991). Indicators of usability based on performance. In: *Human Aspects in Computing: Design and Use of Interactive Systems with Terminals* (H-J. Bullinger, (Ed.)), 656-660. Elsevier, Amsterdam.

Ulich, E., M. Rauterberg, T. Moll, T. Greutmann and O. Strohm (1991). Task orientation and user-oriented dialog design. *International Journal of Human-Computer Interaction* **3**(2), 117-144.

Wirth, N. and J. Gutknecht (1992). *Project Oberon - The design of an operating system and compiler*. Addison-Wesley, Reading (MA).