SINTEF

SINTEF REPORT

SINTEF ICT

Address: NO-7465 Trondheim,
NORWAY
Location: Forskningsveien 1
Telephone: +47 22 06 73 00
Fax: +47 22 06 73 50

Enterprise No.: NO 948 007 029 MVA

TITLE

Usability evaluation of four functional identical versions of DES (Database of empirical studies)

AUTHOR(S)

Asbjørn Følstad and Jan Heim

CLIENT(S)

Simula research laboratory, dept. of software engineering

REPORT NO.	CLASSIFICATION	CLIENTS REF.	
SINTEF_A309	Open		
CLASS. THIS PAGE	ISBN	PROJECT NO.	NO. OF PAGES/APPENDICES
Open	82-14-04037-X	90B10114	17 + 2
ELECTRONIC FILE CODE		PROJECT MANAGER (NAME, SIGN.)	CHECKED BY (NAME, SIGN.)
NA		Asbjørn Følstad	Erik G. Nilsson
FILE CODE	DATE	APPROVED BY (NAME, POSITION, SIGN.)	
NA	2006-09-26	Bjørn Skjellaug, Research director	

ABSTRACT

Simula research laboratory, dept. of software engineering, has hired four different contractors to develop one version each of the department's new Database of Empirical Studies (DES). The four versions have been developed according to the same requirements specification.

Researchers at SINTEF ICT have compared the usability of the four DES-versions through expert evaluations (pilot) and user tests with 18 users. The users were highly representative of the DES user population.

The same general finding was made in both the expert evaluations and user tests. The systems differed with regard to usability. DES-versions A and D received the best overall usability scores. B and C received the lowest. The differences were statistically significant.

It should be noted that DES-version B was best for two of the three tasks. However, it's login (Task 1) had sufficiently low usability for this DES-version to receive the lowest global usability score. It is expected that a redesign of the B login (assumed to be a quick fix) would have made this DES-version the most usable.

Qualitative usability problems associated with the four DES versions are also presented.

KEYWORDS	ENGLISH	NORWEGIAN
GROUP 1	ICT	IKT
GROUP 2	HCI	Menneske-maskin interaksjon
SELECTED BY AUTHOR	Usability evaluation	Evaluering av brukskvalitet
	User test	Brukertest
	Expert evaluation	Ekspertevaluering

Executive summary

The usability of four functionally equal versions of DES (Database of empirical studies) has been investigated through formal expert walkthroughs and user tests; the expert walkthroughs serving as a pilot study.

Expert walkthroughs

The expert walkthrough was conducted as a simplification of the cognitive walkthrough as described by Lewis and Wharton (1997). A similar simplification was used by Følstad (in press). The walkthrough was conducted as a stepwise presentation of three tasks conducted with each of the four DES-versions. The tasks were developed on basis of the DES specification documentation. For each step and task four usability professionals serving as evaluators identified and classified usability problems with each DES-version. Aggregated usability scores were developed for each DES-version. The scores indicated potential differences between the DES-versions. A and D received scores indicating better usability than B and C.

User tests

User tests were conducted for the two tasks of the expert evaluations that showed greatest differences between the four DES-versions. It was judged necessary to split the first task of the expert walkthrough in two parts. Thus, also the user tests consisted of three tasks. The user tests were conducted according to Dumas and Redish (1993).

18 users were involved. The users were highly representative of the DES user population, being 17 Ph.D. students and 1 Scientific programmer working at Simula. Mean age was 29 years (min=22; max=36). All used PC and Internet daily. 13 had no previous knowledge of DES; 5 had superficial knowledge without having tried any of the DES-versions.

ISO 9241-11 states that a system is usable providing it enables *"specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use"* (ISO, 1998 p. 2). Task-specific measures of effectiveness, efficiency and satisfaction were defined as four dependent variables:

a) **Task completion** (yes/no).	Effectiveness
b) **User error rate**. Number of user errors made in consequence of usability problems.	Efficiency
c) **Task time.** Time from onset to completion of task.	
d) **Satisfaction.** Participants' response on a satisfaction questionnaire item following each task.	Satisfaction

To control for possible Order effects the participants were strategically assigned to one of four groups, each with different DES-version presentation order. The four orders were: ABCD, BDAC, CADB, DCBA.

The participants conducted three tasks on all four DES versions. The task sets were similar for all four DES versions:
1. Login.
2. Define new study. Including entry of study data.
3. Access existing study for editing.

Comparison of the four DES-versions was conducted with the compound measures Task usability and Global usability, in which the four dependent variables were weighted approximately equal. The construction of compound usability measures approximated Sauro (2006). The measure for each dependent variable included in the compound measures was converted to an index reflecting a percentage of a maximum:

a) Task completion: Task fully completed = 100. Task not fully completed = 0
b) User error: 100*(Error on task/Maximum error on task)
c) Task time: 100*((Task time-Min. task time)/(Max. task time-Min. task time))
d) Satisfaction: 100*((Task satisfaction score-1)/4)

The four dependent variables included in the measures had an inter-item reliability of Cronbach's $\alpha = 0.69$, indicating sufficient correlation. The compound usability measures were calculated as:

Task usability = Task completion – User error rate – Task time + Satisfaction

Global usability = (Task usability 1 + Task usability 2 + Task usability 3)/3

Mean Global usability and Task usability scores for the four DES-versions are presented in Figures I and II. Higher scores indicate better usability.

Figure I: Global usability scores for DES-versions A, B, C, and D

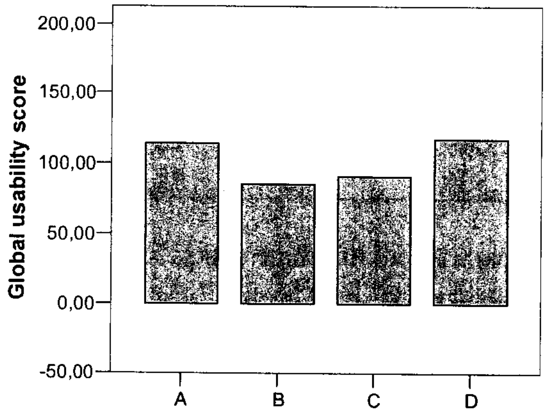
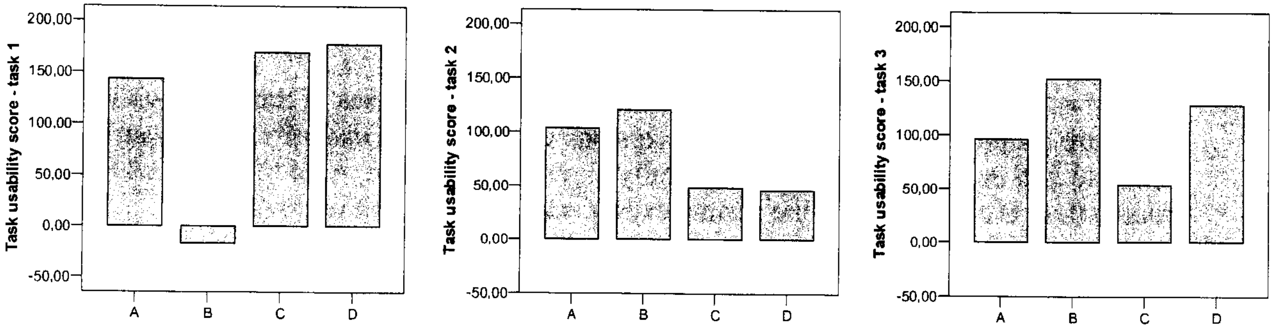


Figure II: Task usability scores on Task 1, 2 and 3. For DES-versions A, B, C, and D



Differences in Task usability between the four DES-versions were investigated as a within subjects effect in a General Linear Model (GLM) Repeated Measures analysis. Possible Order effects were investigated as a between subjects effect, on basis of the participants being assigned to one of four groups with different order of presentation for the four DES-versions. Possible interaction effects between DES-version and Order was investigated as a within subjects effect.

A test of the effects of DES-version, Order and DES-version*Order across all three tasks was conducted as GLM Repeated Measures multivariate test. Pillai's trace was used, because it is more robust with respect to small sample sizes and other violations of assumptions. Results are presented in Table I.

Table I: Multivariate Tests (Pillai's Trace)

Effect		Value	F	Hypothesis df	Error df	Sig.
Between Subjects:	Order	0,509	0,954	9	42	0,490
Within Subjects:	DES-version	0,948	12,135	9	6	0,003
	DES-version * Order	1,561	0,965	27	24	0,539

The between subject factor of Order was not significant, while the within subject factor DES-version was significant. The interaction between DES-version and Order was not significant, meaning that any position effect had the same impact on all interfaces.

A test of the effects of DES-version and DES-version*Order was conducted as univariate tests for each of the three Task usability measures. Results are presented in Table II.

Table II: Univariate Tests (Greenhouse-Geisser correction, Type III Sum of Squares)

Source	Measure	F	Hypothesis df	Error df	Sig.
Interface version	Task usability 1	60,617	1,54	21,54	0,000
	Task usability 2	14,569	1,91	26,77	0,000
	Task usability 3	13,049	2,14	29,96	0,000
DES-version * Order	Task usability 1	0,801	4,62	21,54	0,553
	Task usability 2	2,009	5,74	26,77	0,102
	Task usability 3	1,336	6,42	29,96	0,270

The within subject factor of DES-version is significant for all the three tasks. The interactions are not significant; any position effect had the same impact on all interfaces also at the task level.

The results of the GLM Repeated Measures analyses above were validated with post hoc and non-parametric methods. The validating analyses served to control for possible problems related to violations of GLM assumptions, and was conducted due to the low number of participants and occurrence of non-normal variable distributions.

Conclusions

Both the formal expert walkthrough and the quantitative analyses of the user tests indicate that there are differences between the DES-versions with regard to usability. The quantitative analyses indicate that the differences are statistically significant. In total, A and D received better usability scores than B and C in both the formal expert walkthroughs and user tests.

However, DES-version B received the highest Task usability scores for Tasks 2 and 3 and an extremely low Task usability score for Task 1. The cause for this low score on Task 1 was B's most unusable login procedure. Changing this login procedure could probably be done quite simply and at a low cost. If this had been done, B would have received the highest Global usability score, not the lowest.

It may also be noted that D, even though with one of the two best Global usability scores received poor Task usability scores on Task 2. Since all three tasks weighed the same, this was not sufficient to spoil D's high Global usability score. Whether or not the three tasks should weigh the same, when Task 2 is clearly more complex than the other two tasks, is debatable. If this is judged as necessary, additional analyses may be conducted where the three tasks are weighted differently.

TABLE OF CONTENTS

1. Introduction

Simula research laboratory, dept. of software engineering, has hired four different contractors to develop four different versions of the department's new Database of Empirical Studies (DES).

The four versions have been developed according to the same requirements specification. The contractors had no interaction or knowledge of each other apart from knowing that they were one of four companies doing the same job. The contractors received the same responses to questions raised during development.

The relative usability of the four systems has been evaluated by an independent research organisation, SINTEF ICT. The following evaluation activities have been conducted:

1. Introduction: Core user tasks established on basis of the DES requirements specification and acceptance test procedure
2. Expert evaluations: Informal expert evaluation and formal expert walkthrough conducted to identify possible usability problems, decide number of users to be included in user test, decide specific usability quality criteria, and define final task set for user test
3. User tests: Conducted with 18 users. All users conducted three equal tasks on all four systems. The systems order of presentation was changed to control for order effects.

The null hypothesis of the study was:

> *The four systems are not different with regard to usability*

2. Core user tasks

On basis of the DES specification document and the plan for DES acceptance test the following set of core tasks were established:

- Define new study
- Edit study
- Delete study
- Generate study overview report by search
- Access single study report
- Access graphical report

3. Expert evaluations

3.1 Informal expert evaluation: Establish tasks for formal expert walkthrough

An informal expert evaluation with one usability expert was conducted for all established core tasks. On basis of the informal evaluation, three tasks were chosen for a formal expert walkthrough with four usability specialists. The three tasks were chosen because they were judged to be the most challenging from a usability perspective. The three tasks were:

I. **Define new study.** Starting in logged out position. Data entry conducted with bogus data and project name – similar for all four versions of DES. Data entry included adding file and linking to publication.

II. **Access existing study for editing.** Study to be accessed in task: "DES bidding study by Magne Jørgensen"

III. **Generate study overview and access single study report.** Task conducted with user logged off DES. Study overview to be generated in task: Studies associated with "estimation" (search for "estimation"). Single study report to be accessed in task: "Improving estimation practices at Mogul by Bente Anda".

3.2 Formal expert walkthrough

A formal expert walkthrough, involving four usability experts, was conducted for all four DES-versions. One of the experts conducted the walkthrough alone, before serving as presenter for the three other. The three other evaluators conducted the walkthrough in a group (for purposes of efficiency), but without communicating their individual evaluation results to each other. The expert walkthrough is a simplification of the cognitive walkthrough as described by Lewis and Wharton (1997). A similar version of the expert walkthrough is used by Følstad (in press).

The walkthrough was conducted as a stepwise presentation of how the three established tasks should be conducted in each of the DES-versions. Task *n* was presented on each of the DES-versions before the evaluators moved on to Task *n+1*. The procedure for Task *n* on DES-version *m* was as follows:

- Test leader presents each step necessary to complete the task in the particular version of DES
- For each step the evaluators takes notes on usability problems and classifies these according to severity
- When all steps have been presented the evaluators gives an overall severity rating for the ability of the particular version of DES to solve the task.

Severity classification scale:

0. No usability problems identified
1. Cosmetic usability problems identified. Will probably make typical users experience minor delays in solving task
2. Serious usability problems identified. Will probably make typical users experience major delays in solving task
3. Critical usability problems identified. Will probably stop typical users from solving task

The resulting severity ratings from all four evaluators were averaged for each DES-version. Lower scores indicate better usability. The scores for the four versions of DES[1] were as presented in Table 1:

Table 1: Scores from formal expert walkthrough. Scores represent average severity classifications across tasks. Lower scores are better. Possible scores range from 0 to 3.

Scores	System A	System, B	System C	System D
Task I - average	1,5	2,25	2	2
Task II - average	1,5	1,75	3	1,5
Task III - average	1,25	1,75	1,25	1,25
All tasks - average	1,42	1,92	2,08	1,58

The scores on Task III did not vary as much as the scores on Task I and II. Thus, Task I and II may serve as better discriminators between the DES-versions as Task III, and was therefore recommended as tasks for the user tests. Due to the complexity of Task I it was recommended to split this in two, and thereby make "login" a separate task.

On basis of the expert walkthrough results, System A and D was expected outperform System B and C on usability, at least for the three tasks included in the evaluation.

[1] For internal use at Simula research laboratory: System A = des. System B = DES. System C = des1. System D = Des.

4. User tests

The results from the expert walkthroughs indicated that the null hypothesis (similar usability across all four versions of DES) may not be correct. To follow up the tentative conclusions based on results from the formal expert walkthrough, empirical usability data based on real users were collected through user tests for all four versions of DES. The user tests were conducted to enable both quantitative and qualitative analyses.

4.1 Participants and user groups

18 users participated. All participants used all four versions of DES. Consequently, all DES-versions were tested with 18 users. Nielsen (2006a) recommends testing with 20 users to achieve sufficiently tight confidence intervals when conducting quantitative user tests.

The participants were highly representative for the user group. All were employed at Simula as either Ph.D. students (17) or Scientific programmer (1). Mean age were 29 years (min=22; max=36). All used PC and Internet daily. 13 had no previous knowledge of DES. 5 had superficial knowledge of DES without having tried any of the DES-versions. Participation was conducted within work hours.

To control for Order effects (effects of the order in which the participants tested the four DES-versions) the participants were assigned to one of the following four groups: ABCD, BDAC, CADB, DCBA; the group names signifying the order in which the four DES-versions were tested. The subjects were strategically assigned to the groups, to avoid that no groups had participants from only one particular time of day or one particular day of the data collection. Two of the groups contained four participants each, the other two contained five.

4.2 Dependent variables

ISO 9241-11 states that a system is usable providing it enables *"specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use"* (ISO, 1998 p. 2).

Task-specific measures of effectiveness, efficiency and satisfaction were defined as four dependent variables:

a) **Task completion** (yes/no).	Effectiveness
b) **User error rate**. Number of user errors made in consequence of usability problems. c) **Task time.** Time from onset to completion of task.	Efficiency
d) **Satisfaction.** Participants' response on the following question, given after each task:	Satisfaction

How satisfied are you with using this application to complete this task?
Very unsatisfied 1 2 3 4 5 Very satisfied

4.3 Tasks

The tasks of the user tests were derived from Task I and II of the expert walkthrough. Task I was split in two, and became Task 1 and 2 of the user tests. Task II of the expert walkthrough became Task 3 of the user tests[2].

The user tests consisted of the following three tasks:

1. **Login:** Start in logged out position. Username and password provided on sheet of paper by test leader at task onset.
2. **Define new study.** Start in logged in position. Data entry conducted with bogus data and project name that were similar for all four versions of DES. Data entry included adding file and linking to publication. Bogus project name and data provided on sheet of paper by test leader at task onset.
3. **Access existing study for editing.** Start in the position following completion of Task 2. Information about study to be accessed given on sheet of paper provided by test leader at task onset.

For Task 2 and 3, the provided project information on names of persons and projects already existing in the Simula databases was strategically selected. Names that would show up either early or late in an alphabetically sorted list were avoided. (Selected person names: Jørgensen, Lines, Mardal, Skeie. Selected project names: History based …, Learning …, Optimism, Prediction interval)

Following completion of each task (of failure to do so) the participant answered the question on subjective satisfaction. Following all three tasks belonging to a DES-version, the participant also answered a similarly worded question on total satisfaction with the DES-version.

4.4 Procedure and set-up

The user tests were conducted with one test leader at a meeting room at Simula research laboratory. Procedure and setup corresponds to the recommendations Dumas and Redish (1993). One participant was tested at a time. One session consisted of pre-interview, introduction, testing and debriefing and lasted approximately 1 hour.

Pre-interviews were conducted by a separate form. Texts for introduction and task instructions were given in a written procedure description used by the test leader throughout all test sessions. The procedure descriptions also included step-descriptions for all four DES-versions at all three tasks.

The user interaction was conducted with Internet Explorer 6.0, broadband Internet connection, standard mouse and keyboard, and 17' screen. Screen resolution was 1024*768, as recommended by Nielsen (2006b). On-screen action and audio/video from the test room were recorded through MORAE Recorder 1.3.

4.5 Data registration

Registration of data from the user test recordings were conducted in MORAE Manager 1.3 by one SINTEF researcher. The following were registered:

- Task start
- Task stop
- User error caused by cosmetic usability problem (user spends some time recovering)

[2] Task III of the expert walkthrough was not included in the user test for reasons of efficiency. The expert walkthrough results indicated that this task represented the least severe usability problems, and was not predicted to contribute to differentiation between the four DES-versions.

- User error caused by serious usability problem (user spends long time and tries different possibilities before recovering)
- User error requiring help from test leader
- User error causing the user to achieve only partial completion
- User error causing user to abandon task

For all user errors a short contextual description was given, to enable qualitative structuring of the data. The description included information on: DES-version, on-screen element, description of usability problem causing the user error.

4.6. Results - quantitative analyses

Descriptive analyses

In the user tests 18 users conducted 3 tasks on 4 DES-versions. Totally 216 task instances (18*3*4) were conducted; 27 of these task instances were not fully completed. Totally 203 instances of user error were registered. Descriptive data for the four dependent variables are presented for each task in Table 2. (Descriptive data across each of the four DES-versions are presented in Table 6, Appendix 1. Graphical comparisons of the four DES-versions with regard to the four dependent variables are presented in Figure 3, Appendix 2)

Table 2: Descriptive data for the four dependent variables presented by task. All min. and max. values were observed.

		Min	Max	Mean	SD
Task 1	Task completion	0	1	0,2	0,4
	Task time (minutes)	0,3	6,5	1,1	1,1
	User error	0	2	0,3	0,6
	Satisfaction	1	5	3,6	1,3
Task 2	Task completion	0	1	0,2	0,4
	Task time (minutes)	2,1	11,5	4,6	1,7
	User error	0	7	1,8	1,5
	Satisfaction	1	5	3,0	0,9
Task 3	Task completion	0	1	0,0	0,2
	Task time (minutes)	0,2	3,8	1,3	0,8
	User error	0	3	0,7	0,7
	Satisfaction	1	5	3,3	1,1

To allow for high-level comparison of the four DES-versions, the compound measures Task usability and Global usability were constructed. In the compound measures, the four dependent variables were weighted equally. The construction of a compound usability measure followed Sauro (2006), with the exception that Sauro's standardization of the dependent variables was judged as not necessary in the present study since the DES-versions to be compared were functionally equal.

In order to construct the compound measures, the measure for each dependent variable was converted to an index reflecting a percentage of a maximum:

a)	Task completion:	Task fully completed = 100. Task not fully completed = 0
b)	User error:	100*(Error on task/Maximum error on task)
c)	Task time:	100*((Task time-Min. task time)/(Max. task time-Min. task time))
d)	Satisfaction:	100*((Task satisfaction score-1)/4)

The adequacy of the compound measures depend on the inter-item reliability of the four dependent variables included in the measures (e.g. whether the dependent variables correlate to a sufficient degree). A reliability analysis including the four dependent variables was conducted. In this analysis each of the 216 task instances was represented as rows in the data set. Cronbach's $\alpha = 0.69$, indicating sufficient correlation between the four dependent variables.

The compound usability measures were calculated as:

Task usability = Task completion – User error rate – Task time + Satisfaction

Global usability = (Task usability 1 + Task usability 2 + Task usability 3)/3

Mean Global usability and Task usability scores for the four DES-versions are presented in Figures 1 and 2. Higher scores indicate better usability.

Figure 1: Global usability scores for DES-versions A, B, C, and D

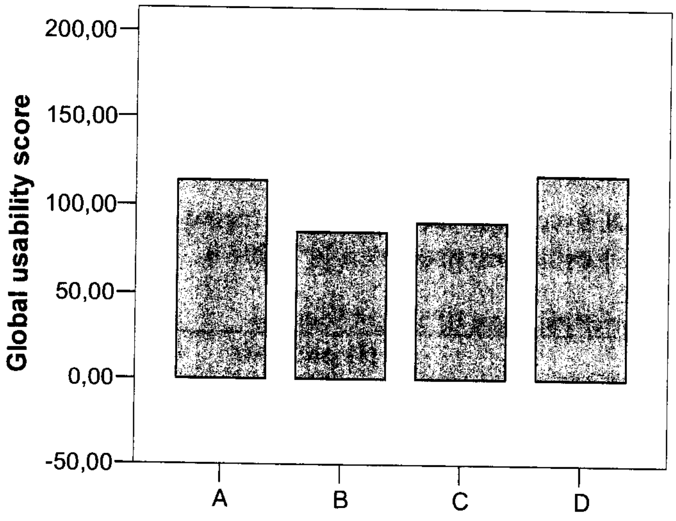
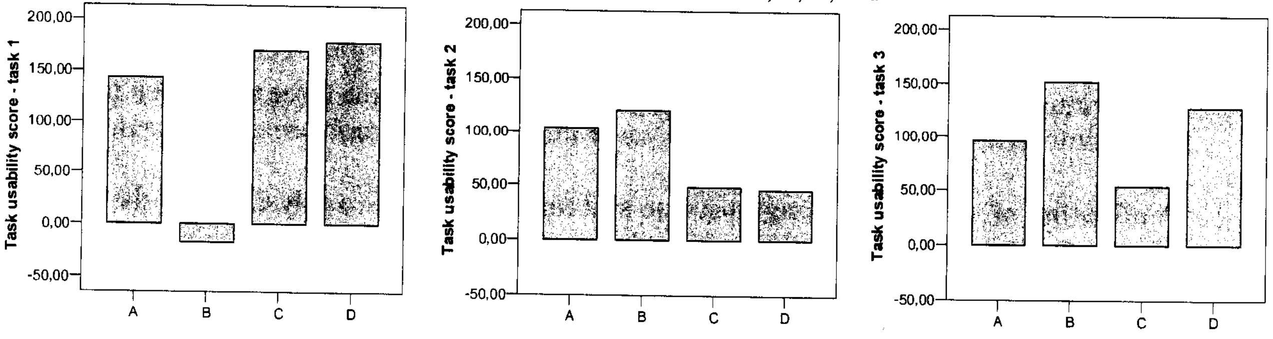


Figure 2: Task usability scores on Task 1, 2 and 3. For DES-versions A, B, C, and D



We see that the D and A received the higher Global usability scores, and B and C the lower. However, there are great differences between the three tasks. For example did DES-version B receive the lowest Task usability score on Task 1 and the highest Task usability score on Task 2 and 3.

Statistical testing of the differences between the four DES versions

A General Linear Model (GLM) Repeated Measures analysis was conducted. Differences in Task usability between the four DES-versions were investigated as a within subjects effect. Possible Order effects were investigated as a between subjects effect, on basis of the participants being assigned to one of four groups with different order of presentation for the four DES-versions.

Possible interaction effects between DES-version and Order was investigated as a within subjects effect.

A test of the effects of DES-version, Order and (DES-version*Order) across all three tasks was conducted as GLM Repeated Measures multivariate test. Pillai's trace was used, because it is more robust with respect to small sample sizes and other violations of assumptions. Results are presented in Table 3.

Table 3: Multivariate Tests (Pillai's Trace)

Effect		Value	F	Hypothesis df	Error df	Sig.
Between Subjects:	Order	0,509	0,954	9	42	0,490
Within Subjects:	DES-version	0,948	12,135	9	6	0,003
	DES-version * Order	1,561	0,965	27	24	0,539

The between subject factor of Order was not significant, meaning that the four Order groups to which the participants were assigned was not found to have any significant effect. The within subject factor DES-version was significant, meaning that it had significant effect which DES-version the individual participant used. The interaction between DES-version and Order was not significant, meaning that the any position effect had the same impact on all interfaces.

A test of the effects of DES-version and DES-version*Order was conducted as univariate tests for each of the three Task usability measures. Results are presented in Table 4.

Table 4: Univariate Tests (Greenhouse-Geisser correction, Type III Sum of Squares)

Source	Measure	F	Hypothesis df	Error df	Sig.
Interface version	Task usability 1	60,617	1,54	21,54	0,000
	Task usability 2	14,569	1,91	26,77	0,000
	Task usability 3	13,049	2,14	29,96	0,000
DES-version * Order	Task usability 1	0,801	4,62	21,54	0,553
	Task usability 2	2,009	5,74	26,77	0,102
	Task usability 3	1,336	6,42	29,96	0,270

The within subject factor of DES-version is significant for all the three tasks. The interactions are not significant; any Order effect had the same impact on all interfaces also at the task level.

Validation of the results using post hoc and non-parametric analysis methods

Due to the low number of participants and occurrence of non-normal variable distributions, the results of the GLM Repeated Measures analyses above were validated with post hoc and non-parametric methods.

A post hoc analysis of differences between means using the Bonferroni adjustment to compensate for multiple comparisons showed the following:

- Task 1: DES-version B was significantly lower (at the 0,05 level) on Task usability compared to each of the three others.
- Task 2: DES-versions A and B were significantly different from B and C.
- Task 3: For the Task usability measure, A and B was significantly different as were B and C, and C and D.

To allow for post hoc analyses of global differences in usability, the compound variable Global usability was entered as the dependent variable into a GLM Repeated Measures with DES-version as the repeated measure and Order as the between subject measure. For the Global usability measure, A and B were significantly different as were B and D, and C and D.

All post hoc findings were confirmed in non-parametric tests. Details are presented in Table 5.

Table 5: Wilcoxon Signed Ranks Test for each pair of interface for each task and all tasks together. Z values and Asymptotic 2-tailed Significance values.

		A		B		C	
		Z	**p**	**Z**	**p**	**Z**	**P**
	B	-3,59	0,000				
Task 1	C	-1,46	0,145	-3,72	0,000		
	D	-2,72	0,006	-3,72	0,000	-2,24	0,025
	B	-1,76	0,078				
Task 2	C	-3,20	0,001	-3,07	0,002		
	D	-2,94	0,003	-3,46	0,001	-0,33	0,744
	B	-3,11	0,002				
Task 3	C	-1,72	0,085	-3,51	0,000		
	D	-1,68	0,094	-2,29	0,022	-2,72	0,006
	B	-2,29	0,022				
All tasks	C	-1,98	0,048	-0,68	0,500		
	D	-0,07	0,948	-3,03	0,002	-2,68	0,007

4.7. Results from qualitative analysis

The four DES-versions had different usability problems associated with them. The usability problems were identified through analysis of user errors. Some usability problems were associated with only one user error. Other problems were associated with as many as 16 user errors (meaning that 16 of the 18 users experienced the particular usability problem).

In Table 6, usability problems associated with each of the four DES-versions is presented. In each cell the total number of usability problems associated with the particular DES-version is reported. DES-version B had as little as 10 usability problems, whereas C had as many as 19. It should be noted that even though B had the lowest number of usability problems, it also had the lowest global usability score. The reason for this is that one of the usability problems associated with B generated a large number of severe user errors.

Table 6 also describe all usability problems associated with each DES-version. Usability problems leading to (1) many and/or (2) severe user errors are presented first. "Many user errors" is understood as 4 or more. "Severe user errors" are understood as user errors classified as serious or catastrophic. Bold numbers in Table 6 indicate the occurrence of at least one severe user error in association with the particular usability problem.

Table 6: Overview of usability problems for DES-versions A, B, C, and D. References to the screens with the usability problem are provided in italics. The number following the usability problem indicates how many users that made errors because this problem. Numbers in bold indicate that at least one user made a severe user error.

DES-version A

12 usability problems were identified in this DES-version

4 usability problems caused many **and** severe user errors:

- *New study, add study material:* User clicks *submit* before providing *description* **(12)**
    - User does not see subsequent error message **(2)**
- *Studies (list overview)*: User clicks study name to open for edit **(7)**
- *New study, end date:* User does not provide value for day in date field **(4)**
- *Logged out:* User experiences difficulties in understanding/finding Login **(4)**

1 usability problem caused **either** many **or** severe user errors:

- *Receipt page:* User experiences difficulties returning to main menu **(3)**

The 7 remaining usability problems with DES-version A were:

- *New study, select publication:* User spends long time finding publication in list (3)
- *Studies (list overview)*: User spends long time finding adequate way to identify study for editing (2)
- *New study, end date:* User uses wrong date format (1)
- *New study, add study material:* User starts writing path to file in URL-field (1)
- *New study, add study material:* User does not close dialogue box with *Close*, and do not get subsequent confirmation of file added (1)
- *DES admin page:* User is confused with regard to difference between *Study search* and *Study overview* (1)
- *Studies (list overview)*: User clicks on link to Front page to find study for editing (1)

DES-version B.

10 usability problems were identified in this DES-version

2 usability problems caused many **and** severe user errors:

- *Simula Intranet frontpage (green):* User experiences difficulties in finding DES startpage **(16)**
- *Create study, end date:* User is confused by the date format **(5)**
    - User does not see subsequent error message **(1)**

4 usability problems caused **either** many **or** severe user errors:

- *Create study, related publications:* User experiences difficulties associated with not seeing whole publication name (5)
- *DES start page (logged in)*: User does not understand that that this is DES startpage **(3)**
- *Create study, upload files:* User clicks *Finish* before clicking *Add file* **(1)**
- *Create study, responsible persons:* User enters himself as responsible person for study, and does not find any place to enter the real study responsible **(1)**

The remaining 4 usability problems were:

- *DES start page (logged in)*: User enters data belonging to new study in search field instead of clicking *Create new study* (2)
- *Search results:* User spends long time finding Edit-link (2)
- *DES start page (logged in)*: User clicks repeatedly on *Study responsible* in search field to start search (1)
- *Create study, end date:* User does not provide value for day in date field (1)

DES-version C.
19 usability problems were identified in this DES-version

3 usability problems caused many **and** severe user errors were:

- *Manage studies:* User clicks *Search studies* to identify study for editing (only leads to studies in non-editable mode) **(15)**
- *Register new study, attach publications:* User experiences difficulties finding publication in list divided between several pages **(7)**
- *Register new study, attach study material:* User clicks *Finished* without clicking *Save and attach* **(6)**

4 usability problems caused **either** many **or** severe user errors were:

- *Register new study, attach study material:* User clicks *Save and attach* without providing *Description* (8)
- *Manage studies:* User checks checkbox instead of clicking study name to open for entering (5)
- *Register new study, attach publications:* User tries to use IE-search to find publications in list. Probably do not see that list is divided between several pages (4)
- *Register new study, attach publications*: User clicks *Search studies* to find publication in list **(3)**

The remaining 12 usability problems were:

- *Register new study, responsible:* User spends time to understand mechanism for choosing study responsible (2)
- *Register new study, add publications or files:* User experiences difficulties understanding which buttons that are associated with adding *Publications* and *Study material* (2)
- *Register new study:* User clicks Save changes before all obligatory data are entered (2)
- *Register new study, end date:* User does not provide value for day (2)
- *Admin home:* User clicks *Manage studies* instead of *Create new study* to enter new study (1)
- *Register new study, responsible:* User spends time to understand where to enter study responsible (1)
- *Register new study, add publications or files:* User experiences difficulties finding where to add *Publications* and *Study material* (1)
- *Register new study, notes:* User provides unnecessary information about added files (1)
- *Register new study, attach publications:* User clicks *Edit study* to return to editing, and do not get confirmation of publication added (1)
- *Register new study, attach publications:* User clicks Back in IE to return to editing, and do not get confirmation of publication added (1)
- *Register new study, attach study material:* User looks for study material to be uploaded in list of existing study material (1)
- *Manage studies:* User clicks on *Manage study material* to find study for editing (1)

DES-version D.
17 usability problems were identified in this DES-version

3 usability problems caused many **and** severe user errors:

- *New study, select study material:* User clicks *Add* without providing *Description* **(9)**
    - User does not see subsequent error message **(1)**
- *New study, select publications:* User tries to search by pushing Enter instead of clicking *Search.* Receive error message **(7)**
- *New study:* User experiences difficulties in finding where to register study responsible **(5)**

5 usability problems caused **either** many **or** severe user errors:

- *New study, select publications:* User clicks publication name to add, instead of marking Checkbox and clicking *Add* (8)
- *List studies:* User tries to conduct search with free text and other search criteria at the same time (5)
- *New study, select study material:* User does not use radio button to choose *File* instead of (default) *Url* **(3)**

- User does not see subsequent error message **(1)**

- *New study, select publications:* User clicks *Ok* without clicking *Add* **(3)**
- *New study, select study material:* User clicks *Ok* without clicking *Add* **(3)**

The remaining 9 usability problems were:

- *New study, select publications:* User is confused about publication identity because authors are not listed (3)
- *New study:* User clicks *Save study* without having entered all obligatory data (2)
- *New study, select study responsible:* User is confused by mechanism for selecting study responsible (2)
- *Study overview report:* User tries to identify study for editing by clicking *Edit study type* (2)
- *New study, end date:* User confuses date format (1)
- *New study, end date:* User tries to enter end date in field for *Duration of study* (1)
- *New study, add study material:* User spends time finding link to *Study material* (1)
- *Study overview report:* User tries to identify study in list by IE-search (1)
- *List studies:* User enters wrong date format in search field (1)

5 Discussion

The four DES-versions have been investigated with regard to differences in usability. There has been conducted a formal expert walkthrough and user tests. The results of the user tests has been made subject of quantitative and qualitative analyses.

Both the formal expert walkthrough and the quantitative analyses of the user tests indicate that there are differences between the DES-versions with regard to usability. A null hypothesis of no difference in usability between the four DES-versions cannot be sustained.

In total, A and D received better usability scores than B and C in both the formal expert walkthroughs and user tests. This difference in usability is also statistically significant according to the multivariate test of GLM Repeated Measures.

However, this picture needs some refinement. In particular with regard to DES-version B. B received the highest Task usability scores for Tasks 2 and 3 and an extremely low Task usability score for Task 1. The cause for this low score on Task 1 was B's most unusable login procedure. Changing this login procedure could probably be done quite simply and at a low cost. If that had been done, B would have received the highest Global usability score, not the lowest.

It may also be noted that D, even though with one of the two best Global usability scores received poor Task usability scores on Task 2. Since all three tasks weighed the same, this was not sufficient to spoil D's high Global usability score. Whether or not the three tasks should weigh the same, when Task 2 is clearly more complex than the other two tasks, is debatable. If this is judged as necessary, additional analyses may be conducted where the three tasks are weighted differently.

In cases of discrepancy between the usability experts and the user tests, results from the user tests are viewed as more valid. It may be noted that the usability experts in the formal expert walkthrough underestimated B's usability with regard to Task 2 and 3 of the user tests. The experts also overestimated D's usability with regard to Task 2 of the user tests[3].

[3] As explained earlier, the three user test tasks were derived from Task I and II of the expert walkthrough where Task I was split in two in the user test. Thus, Task 2 of the user tests was similar to parts of Task I of the expert walkthrough. Task 3 of the user test was similar to Task II of the expert walkthrough.

It is interesting to note that the qualitative analysis identified more usability problems in DES-version D (17) than in A (12) and B (10), even though D scored higher on Global usability than B, and received a similar Global usability score to A. This shows the importance of investigating usability quantitatively instead of qualitatively when the purpose is comparison between solutions, and not just redesign of a single solution.

The qualitative and quantitative results together indicate that the usability problems of C are not as severe and frequently experienced as the usability problems of B. Similarly, the few usability problems of B (in particular the login problem) have high impact on B's usability - even though their numbers are low. It may be noted that, given similar severity, usability problems associated with most user errors will be those usability problems that, if corrected, will do most to improve the usability of the four DES-versions.

References

Dumas, J. S., Redish, J. C. (1993). A practical guide to usability testing. Ablex publishing corporation. Norwood, New Jersey

Følstad, A. (in press). Work-Domain Experts as Evaluators: Usability Inspection of Domain-Specific Work Support Systems. Scheduled for publishing in 2007 in *International Journal of Human-Computer Interaction*, 22(1).

ISO/IEC 9241-11. Ergonomic requirements for office work with visual display terminals (VTDs). Part 11: Guidance on Usability. *ISO/IEC TR 9241-11* (1998)

Lewis, C., Wharton, C. (1997). Cognitive walkthroughs. In M. Helander, T. K. Landauer, P. Prabhu (eds.) *Handbook of Human-Computer Interaction. Second, completely revised edition.* Elsevier Science B. V.

Nielsen 2006a Quantitative Studies: How Many Users to Test? *Jakob Nielsen's Alertbox, June 26, 2006* Avaliable at: http://www.useit.com/alertbox/screen_resolution.html (Downloaded August 31, 2006).

Nielsen 2006b Screen Resolution and Page Layout, *Jakob Nielsen's Alertbox, July 31, 2006* Avaliable at: http://www.useit.com/alertbox/screen_resolution.html (Downloaded August 31, 2006).

Sauro, J., Kindlund, E. (2005). A method to standardize usability metrics into a single score. *Proceedings of the SIGCHI conference on Human factors in computing systems.* ACM Press, pp. 401-409.

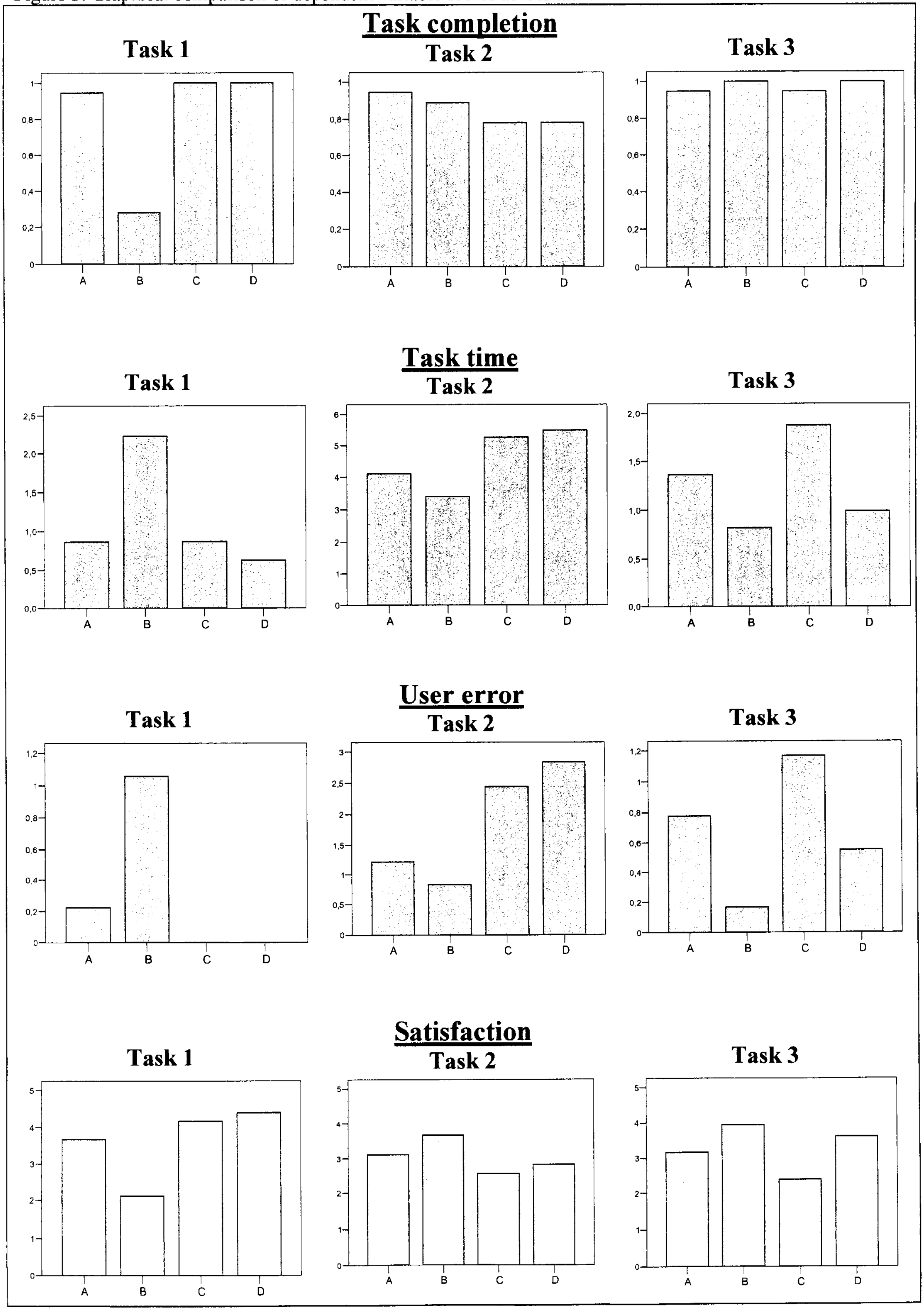
Appendix 1

Table 6: Data for the four dependent variables Task completion, Task time, User error and Satisfaction across tasks and DES-versions

		Total				A				B				C				D			
		Min	Max	Mean	SD	Min	Max	Mean	SD	Min	Max	Mean	SD	Min	Max	Mean	SD	Min	Max	Mean	SD
Task 1	Task completion	0	1	0,2	0,4	0	1	0,9	0,2	0	1	0,3	0,5	1	1	1	0	1	1	1	0
	Task time (minutes)	0,3	6,5	1,1	1,1	0,4	1,8	0,9	0,4	0,5	6,5	2,2	1,6	0,5	2,2	0,9	0,4	0,3	1,2	0,6	0,2
	User error	0	2	0,3	0,6	0	1	0,2	0,4	0	2	1,1	0,5	0	0	0	0	0	0	0	0
	Satisfaction	1	5	3,6	1,3	2	5	3,7	1,0	1	4	2,1	1,0	3	5	4,2	0,8	3	5	4,4	0,8
Task 2	Task completion	0	1	0,2	0,4	0	1	0,9	0,2	0	1	0,9	0,3	0	1	0,8	0,4	0	1	0,8	0,4
	Task time (minutes)	2,1	11,5	4,6	1,7	2,7	5,4	4,1	1,0	2,1	7,2	3,4	1,2	2,7	11,5	5,3	2,1	3,1	8,3	5,5	1,4
	User error	0	7	1,8	1,5	0	3	1,2	0,8	0	4	0,8	1,1	0	7	2,4	1,7	1	5	2,8	1,1
	Satisfaction	1	5	3,0	0,9	2	5	3,1	1,0	2	5	3,7	0,8	1	4	2,6	1,1	2	4	2,8	0,5
Task 3	Task completion	0	1	0,0	0,2	0	1	0,9	0,2	1	1	1	0	0	1	0,9	0,2	1	1	1	0
	Task time (minutes)	0,2	3,8	1,3	0,8	0,6	3,3	1,4	0,7	0,3	1,5	0,8	0,3	0,5	3,8	1,9	1,0	0,2	2,1	1,0	0,5
	User error	0	3	0,7	0,7	0	2	0,8	0,5	0	1	0,2	0,4	0	3	1,2	0,9	0	2	0,6	0,6
	Satisfaction	1	5	3,3	1,1	2	5	3,2	1,0	2	5	3,9	0,7	1	4	2,4	1,1	2	5	3,6	0,8

Appendix 2

Figure 3: Graphical comparison of dependent variable scores across tasks and DES-versions



Task completion

Task 1

Task 2

Task 3

Task time

Task 1

Task 2

Task 3

User error

Task 1

Task 2

Task 3

Satisfaction

Task 1

Task 2

Task 3

sirib

sirib

90F235 - Bestilling-SMODEM PPP.pdf
05/10/06 11:37

SIEMENS

90F235/YS

Bestilling

Side: 1 av 1

NR.* 4500870207

*Må oppgis i forsendelsesdokumenter og faktura

SINTEF IKT
O S Bragstads plass 2A
NO-7465 Trondheim
Norge

Attn.: Yngve Steinheim

Sted/Dato*
Oslo, 29.09.2006
Vår ref.
Vidar Steigen T&I / 1441
Deres ref.
90F235

90F235
4/10 06
Y. Steinheim
FS, PS. OK

Forsendelsesadresse Siemens AS Postboks 1, Alnabru N-0613 Oslo Att. Vidar Steigen K7	Faktureres til Siemens AS Regnskapsservice K2 Postboks 1, Alnabru N-0613 Oslo
Kunde- eller prosjektnavn + ordrenr.* (Merk også sendingen) SISOG DPM Modem	Forsendelsesmåte

For bestillingen gjelder "Standard Purchasing Terms for Siemens AS".

Betalingsbetingelser Netto + 30 dager		Leveringsbetingelser DDP Oslo
KS-standard ISO 9001	Leveringsdato Utført delarbeid	Siemens AS

Pos. nr./Varenr.	Antall	Spesifikasjon	Enhetspris	Pris NOK, ekskl. MVA

DPM POWERLINE MODEM

Specification for powerline modem with Ethetnet support

The Order is related to your e-mail dated 248h of September 2006 encompassing accumulated work amounting to 55 hours.

This order covers initial work complying with your project offer from October 15th.

The Order is limited to the number of man-hours stated below.

Total man-hours:	55
Total cost:	NOK 52 250,00

Siemens AS Oil & Gas

Foretaksregisteret
NO 915 826 946 MVA
Hovedkontor
Østre Aker vei 90
Postboks 1
N-0613 OSLO

Technology & Innovation

Postboks 1, Alnabru
N-0613 OSLO

Bearbeider: Vidar Steigen

Østre Aker vei 90
OSLO

Telefon: 22633937
Mobil: 92283378
Telefax: 22633863
vidar.steigen@siemens.com