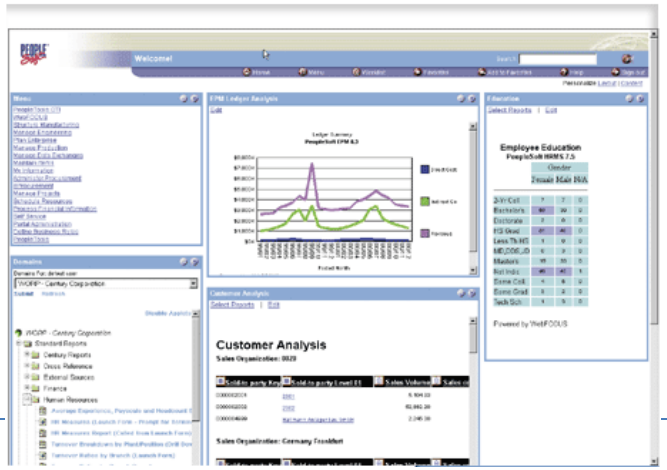
BENTLEY UNIVERSITY

Evaluating the Collaborative Critique Method

Tamara Babaian, Wendy Lucas, Mari-Klara Oja
CIS Department
Bentley University

Human-computer collaboration

- Human-Computer Collaboration paradigm of human-computer interaction requires the system to act as a user's intelligent partner.
- No user can be expected to know all of the system's features, so the system should assume the role of a collaborative partner.



Collaborative Critique (CC)

- A usability *inspection* method
  - Performed by trained analysts directly interacting with the artifact.

- Based on theory of collaboration

Tenets of collaboration

- Parties share a common *goal* and *plans* for accomplishing it.
- The goal and plans may be *incompletely* specified in the beginning and are refined or revised in the process of collaboration.
- Partners must *communicate* in order to refine goals and plans, maintain a *shared context*.
- Commitment to success of joint action:
  - implies helpful behaviors, e.g. learning and adapting.
  - implies helping a partner who is having a problem performing their part.

(Terveen'95, Grosz & Kraus'96, Bratman'92)

Why invent another usability inspection method?

Current methods (e.g. Cognitive Walkthrough and Heuristic Walkthrough)

- Limited to specific user difficulties
- Limited consideration of the task context of the system-user interaction
- Tend to discover a lot of insignificant problems, miss more severe ones.
- Difficult to generate design fixes
- Not well-suited for *complex* systems

(Cockton et al. '08)

Overview of CC: input to evaluators

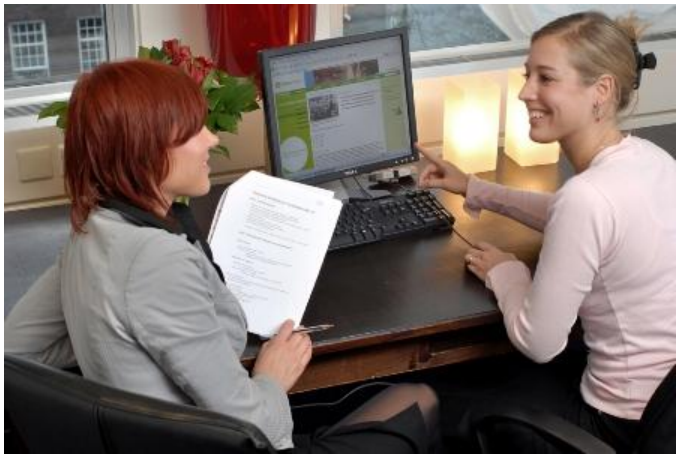
Input to evaluator:

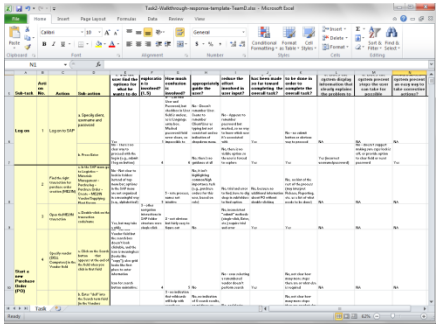
- The artifact for evaluation
- Description of typical user
- Task specification
- Correct step-sequence
- Spreadsheet to record the answers to CC questions

CC Process

Given the task description, and the artifact,

- See if you can discern the next action to be performed based on task description + exploration
- Compare your discerned action to the next step as specified in the correct step sequence
- Perform the correct action
- Record answers to **CC questions**
  - Yes/No(with explanation)/NA





Collaborative Critique Question 1

1. **Will the user find the options for what s/he wants to do in the current screen?**

YES ⇔ the option is available and easily identifiable

NO ⇔ user is confused about what to do, or

knows what s/he wants to do but can't find the option

Collaborative Critique Question 2

2. For the user to figure out what to do now:

Considering users with a range of experiences, answer with a number 1..5:

(a) how much exploration is involved?

1 -- 5

(b)how much confusion is involved?

1 -- 5

Exploration - the user's interaction with the system in order to find the right option or value, or to find how to do something within the system.

Assessing the **mental state** of the user with respect to understanding what is going on in the current interaction with the system, what needs to be done now, and how to do it.

Collaborative Critique Question 3

3. Is the system using knowledge of the task in general, the current user, and the context of the current action to the fullest extent in order to:

(a) appropriately guide the user?

YES ⇔ sufficient guidance for the user to be **effective and efficient**.
NO ⇔ there is a need for helping the user in the current situation, yet the system support is not adequate.

(b) reduce the effort involved in user input?

YES ⇔ system leveraging all contextual and general knowledge of the task and user to the fullest extent in minimizing the user effort involved in providing input to the system
NO ⇔ there is a way to further minimize user effort.

Example: task and context-aware guidance and effort reduction

Google Chrome browser –
displays a page preview automatically on print command.

Shows options controlling the printing process.

Collaborative Critique Question 4

4. After execution of the current action, will the user understand

(a) what progress has been made so far toward completing the overall task?

YES ⇔ the user will understand what has already been accomplished toward completing the overall task.
NO ⇔ user will be unsure of what has been accomplished.

(b) what remains to be done in order to complete the overall tasks?

YES ⇔ the user will know what is left to do.
NO ⇔ unsure of what remains to be done.

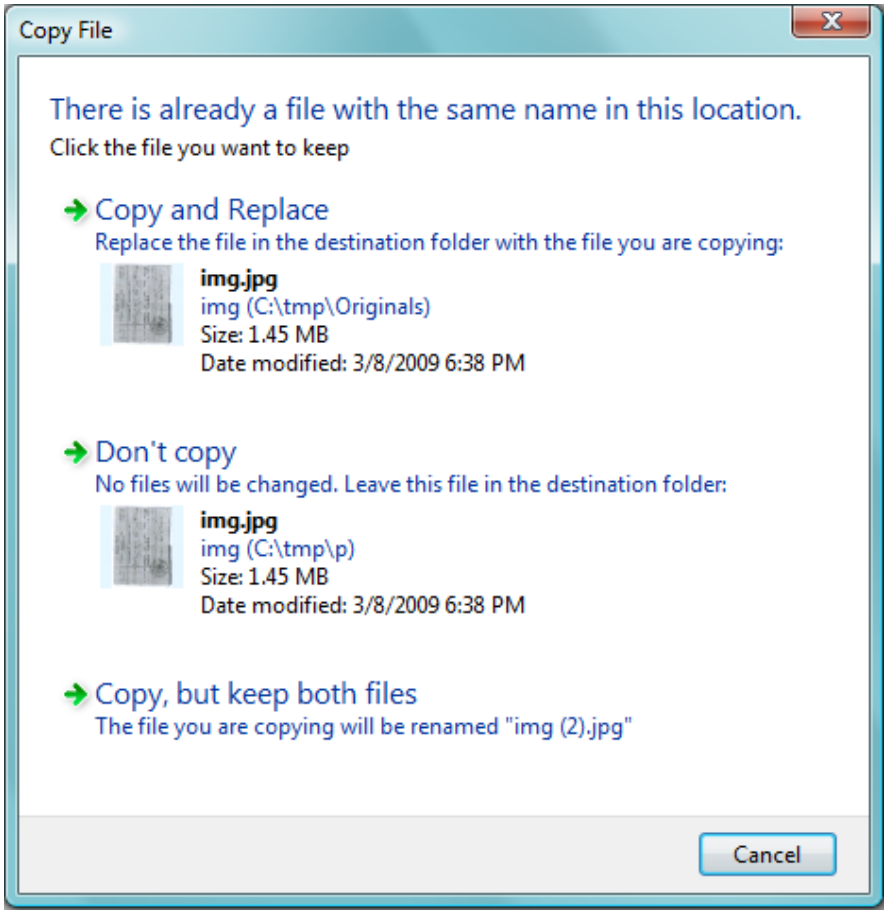
Error Situations: Questions 5 – 7

The following questions must be answered only in case an error condition is reported.

5. Does the system display information that clearly explains the problem to the user?

6. Does the system present steps the user can take for possible corrective actions?

7. Does the system present an easy way to take corrective actions?



Evaluating inspection methods

- Goal: assess how well the predictions of the CC reflect the reality

  Assuming,

  P – set of predicted usability problems

  A – set of actual problems, define

$$thoroughness = \frac{|P \cap A|}{|A|}$$

$$validity = \frac{|P \cap A|}{|P|}$$

- *A* – assessed using a form of user-testing.

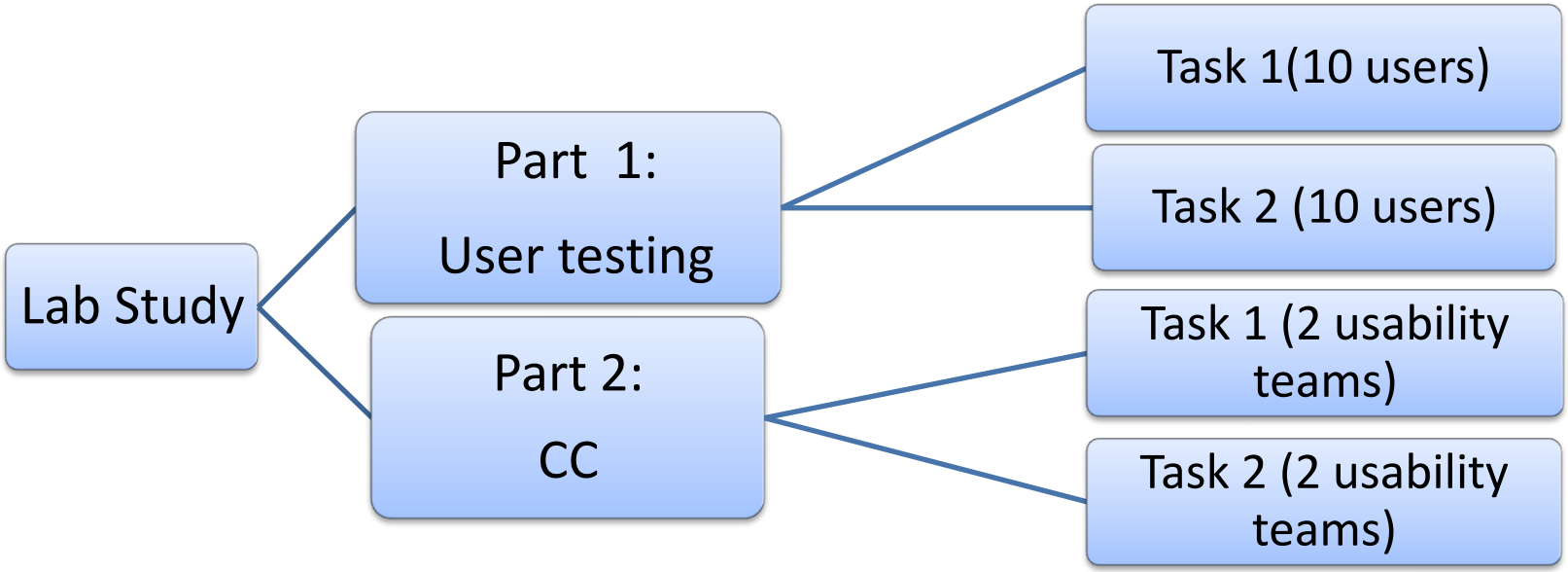
  (Sears'97; Hartson, et al. '03)

Evaluation problems

- No accepted methodology of usability problem categorization, therefore
  - *P (predicted usability problems)* is a subjective measure
  - *A (actual usability problems)* can only be approximated

(Hartson, et al. '03)

Two part study



- Evaluating two different tasks within a major ERP
  - Task 1 (create and test an Authorization profile) -52 steps
  - Task 2 (create a Purchase Order) included *seeded error situations* user needed to resolve - 66 steps

USER TESTING

Collecting the set of actual problems

Users

Task	Age (average)	ERP Experience		
		Novice	Intermediate	Expert
1	21 - 34 (26)	6	3	1
2	22 - 29 (26)	6	3	1

User-testing: critical incidents technique

1. User training:
   - Watch /practice task video tutorial (30 min)
     - Users were allowed to take notes
   - Watch video tutorial on reporting usability issues (10 min)

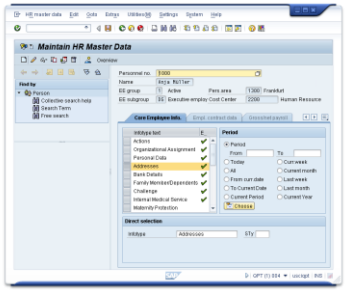
2. Perform the task, self-reporting usability issues (45 min)
   - Screen-capture the entire interaction

3. Retrospective reporting (30 min)

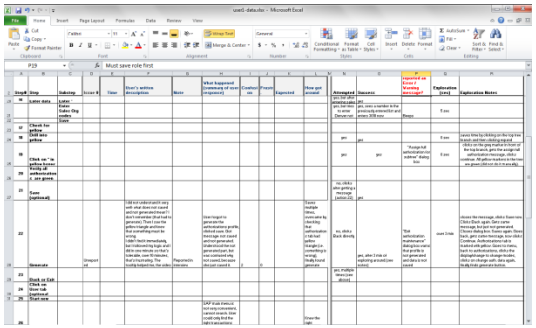
   For each self-reported problem, user was asked for:
   - Problem description
   - Confusion level from 1(not confused) -4 (extremely confused)
   - Frustration level from 1 -4
   - What s/he expected to happen
   - How s/he resolved the issue

   (Castillo, et al. 1997)

User-testing data transcription and analysis

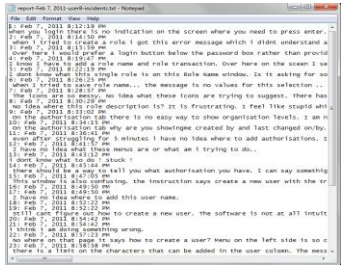


Screen-capture video of user's work on the task



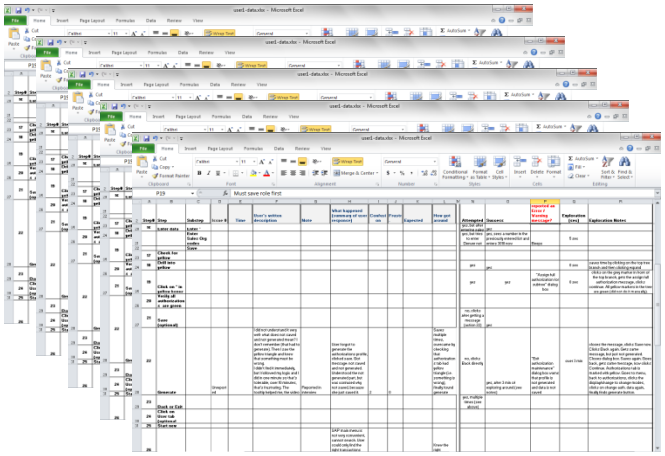
Stepwise transcript, including

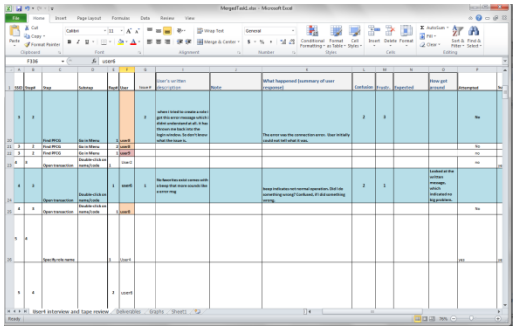
- If step was attempted (yes/no)
- If step was successfully completed (yes/no)
  - if system error occurred at the step (description, if yes)
- exploration time (if > 15 sec)
- exploration notes
- user-reported issue data



User-reported issues

Identify usability issues in user data



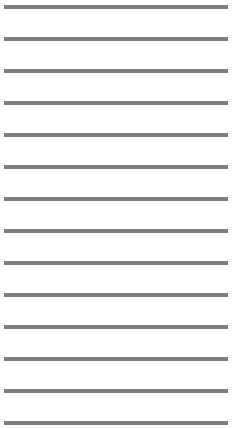
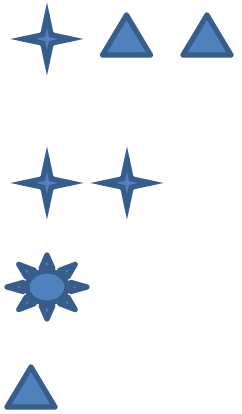
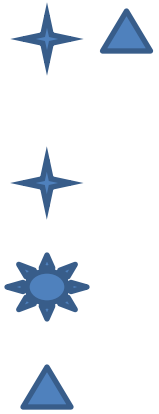


Merge the individual transcripts.
Identify usability problem instances
at each step of the sequence.
2-3 researchers working together.

Classify and describe usability problem instances in each step, including

- problem type, id
- if problem had a negative effect on the final outcome
- if user pursued a wrong path, but was able to recover
- if user asked a researcher for help
- if user consulted their notes

Actual problem data

Step Sequence	Problem Instances	Problem Type Per Step (PTPS)	Problem Types
			

Task/Steps	P. Instances [%User-Reported]	PTPS	P Types
1/52	115 [88%]	93	52
2/66	162 [72%]	109	62
Both	277 [79%]	202	98

COLLABORATIVE CRITIQUE WALKTHROUGHS

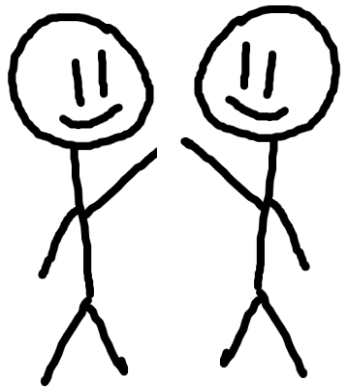
Collecting the set of predictions

Usability Analysts

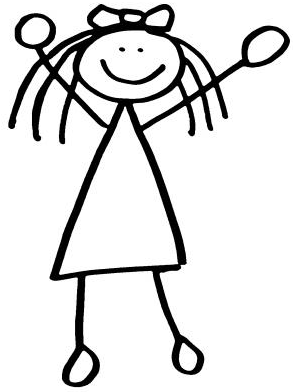
Team	Task	ERP Experience		Usability Evaluation		Total years in field
		person1	person 2	person1	person 2	
A	1	novice	inter.	expert	inter.	1
B	1	novice	novice	accomp.	inter.	4
C	2	inter.	novice	accomp.	accomp.	15
D	2	novice	novice	expert	inter.	11

CC walkthrough set up

- Prepared materials:
  - Task tutorial document, video tutorial
  - CC method tutorial document
  - Correct action sequence for the task
  - Spreadsheet for recording answers
- 2 computers: 1 for running the system, 1 for recording the answers.
- 2 analysts: 1 operating the system, 1 entering the answers in a spreadsheet. Both work together to compose the answers.
- Analysts speech and interaction with the system was recorded using screen capture software.
- Analysts were compensated for their participation.



Analysts

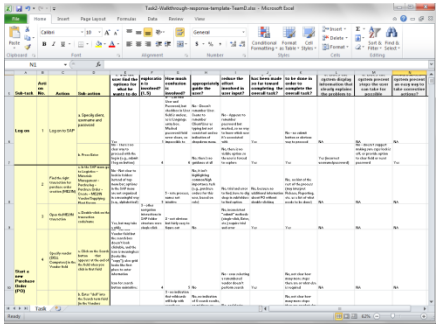


Researcher

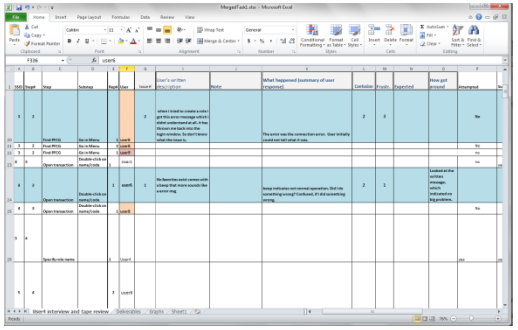
RESULTS

Comparing CC and User-testing data

Predicted actual problems (P∩A)



Collaborative Critique Answers



Actual problem set from user testing

- Mark a Problem Instance as **Predicted**, if an answer to the CC questions at some step described the Problem Instance directly, or identified its cause.
- Mark other Problem Instances as **Missed**.

Thoroughness data

Team	P Instances		PTPS		P Types	
	predicted (predictable)	%	predicted (predictable)	%	predicted (predictable)	%
Team A	41(98)	42	32(77)	42	22(45)	49
Team B	38(97)	39	28(78)	36	21(45)	47
Task 1 (A & B)	**56(104)**	**54**	**44(83)**	**53**	**28(48)**	**58**
Team C	102(135)	76	65(92)	71	38(56)	68
Team D	89 (133)	67	58(89)	65	36(56)	64
Task 2 (C & D)	**117(144)**	**81**	**78(98)**	**80**	**45(58)**	**78**
All	**173(248)**	**70**	**122(181)**	**67**	**66(93)**	**71**

Thoroughness data by severity

Lowest severity Highest severity

Team	Severity 1 predicted (predictable)	%	Severity 2 predicted (predictable)	%	Severity 3 predicted (predictable)	%
Team A	26(61)	43	6(12)	50	9(25)	36
Team B	19(59)	32	6(15)	40	13(23)	57
Task 1 (A & B)	**33(63)**	**52**	**9(15)**	**60**	**14(26)**	**54**
Team C	48(66)	73	21(27)	78	33(42)	79
Team D	40(65)	62	20(26)	77	29(42)	69
Task 2 (C & D)	**54(68)**	**79**	**25(28)**	**89**	**38(48)**	**79**
All	**87(131)**	**66**	**34(43)**	**79**	**52(74)**	**70**

Thoroughness in error handling situations

	all critiqued errors						seeded errors					
	P Instances		PTPS		P Types		P Instances		PTPS		P Types	
Team	predicted (predictable)	%	predicted (predictable)	%	predicted (predictable)	%	predicted (predictable)	%	predicted (predictable)	%	predicted (predictable)	%
Team C	37(39)	**95**	23(24)	**96**	10(11)	**91**	20(23)	**87**	12(14)	**86**	6(7)	**86**
Team D	25(32)	**78**	15(21)	**71**	6(10)	**60**	19(23)	**83**	11(14)	**79**	4(7)	**57**
together	43(45)	**96**	29(31)	**94**	13(14)	**93**	22(23)	**96**	14(14)	**100**	7(7)	**100**

- Task 2 included 3 *seeded* error situations
- Users experienced 9 different error situations
- Analysts encountered and evaluated 12 different errors

Mispredictions

- Computed false positives at the step level, i.e. those steps, where evaluators identified issues *not found* in the user data:
  - Task 1 – 4 steps out of 52(7.7%)
  - Task 2 – 3 steps out of 66 (4.5%)

- On steps 2a (exploration) – evaluators commonly over-estimated how many people would be unable to figure out what to do.

Conclusions

- CC is a promising new method.
- In the initial evaluation:
  - Good thoroughness measures in evaluation of complex tasks; ***great*** for problems related to error situations.
  - Uniform coverage of problems of different severity levels
  - Questions 2(a,b) useful, because it required the evaluators to explore and consider the user's mental and physical effort, but answers may not be accurate.
- Limitations of this study
  - Small number of evaluators
  - Did not evaluate method's effectiveness in driving design changes

Why another usability inspection method?

Current inspection methods

- Limited to specific user difficulties
- Limited consideration of the task context of the system-user interaction
- Tend to discover a lot of insignificant problems, miss more severe ones.
- Not well-suited for *complex* systems
- Difficulties in generating fixes

Collaborative Critique

- Systematically assesses total effect of the system behavior on user's cognitive and physical efforts
- Focus on system *usefulness* within task context
- Explicitly addresses error situations
- Takes into account the task and user context.
- CC questions directly point to the system's role in the task.

Future work

- Evaluate the CC method in the field
  - Fine tune the method
  - How valuable is the CC method as a driver for design changes?
- Evaluate the CC as an educational tool – teaching about collaborative user interaction design.

We are looking for opportunities to collaborate with companies, usability professionals, other researchers

Acknowledgement

This material is based in part upon work supported by the National Science Foundation under Grant No. 0819333. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.