Lecture 5-2: Usability Methods II

- Heuristic Analysis
  - Heuristics versus Testing Debate
  - Some Common Heuristics

Expert Reviews (1)

- *Heuristic Evaluation*
  - Nielsen & Molich (1990) CHI Proceedings
  - Based upon empirical article Molich & Nielsen (1990) (in readings)
  - Inspection of a prototype or finished system to identify all changes necessary to optimize human performance and preference
  - Evaluators use a set of guidelines or general principle
    - hence term: “heuristics”
- Distinctions not always made clear in studies and criticisms of heuristic evaluation:
  - Use of heuristics (guidelines) or not
  - Experience level of reviewers
    - experts vs. non-experts using just heuristics)
  - Review by lone individual or joint review by group
    - Research shows it makes a difference
  - Use of prescribed tasks versus self-guided evaluation

Expert Reviews (2)

- *Cognitive Walkthrough*
  - Distinct and more formal technique than heuristic evaluation
  - Proceed step-by-step through system using task scenarios
    - use context of several core tasks user must accomplish
    - operation and feedback of the system are compared to users' goals and expectations
  - Contrast with simple inspection by individual
  - Often these techniques define this as a group review
  - Analogy to software walkthrough
  - Several techniques defined in literature
    - Articles appearing same time as Nielsen and Molich:
    - Lewis et al (1990), Wharton et al. 1992, Jeffries et al

Usability Inspection Methods

- Nielsen & Mack (Eds.) (1994) Usability Inspection Methods.
- Nielsen – Methods
  http://www.useit.com/papers/heuristic/inspection_summary.html
  - Heuristic Evaluation
  - Heuristic Estimation
  - Cognitive Walkthrough
  - Pluralistic Walkthrough
  - Feature Inspection
  - Consistency Inspection
  - Standards Inspection
  - Formal Usability Inspection

The Bias Against Expert Reviews

- Interview with John Karlin *(*Klemmer, 1989, *Ergonomics)*
  - Karlin “founder of human factors in industry” circa 1945 (Bell Labs)
  - Q: “Next, let’s consider where human factors people get their answers. I’ll name four general sources of answers and ask your opinion and ranking of each. First: expert opinion; second: human factor or psychological principles; third: prior data; fourth: new laboratory data.”
  - A: “New laboratory data is far and away the most important. I would rank principles second, but more as a foundation for obtaining new data than a source of answers in themselves. Prior data is third, but it was probably most useful when first done and seldom applies to the present situation. *Regarding expert opinion, I wouldn’t give it the time of day.* [...]”

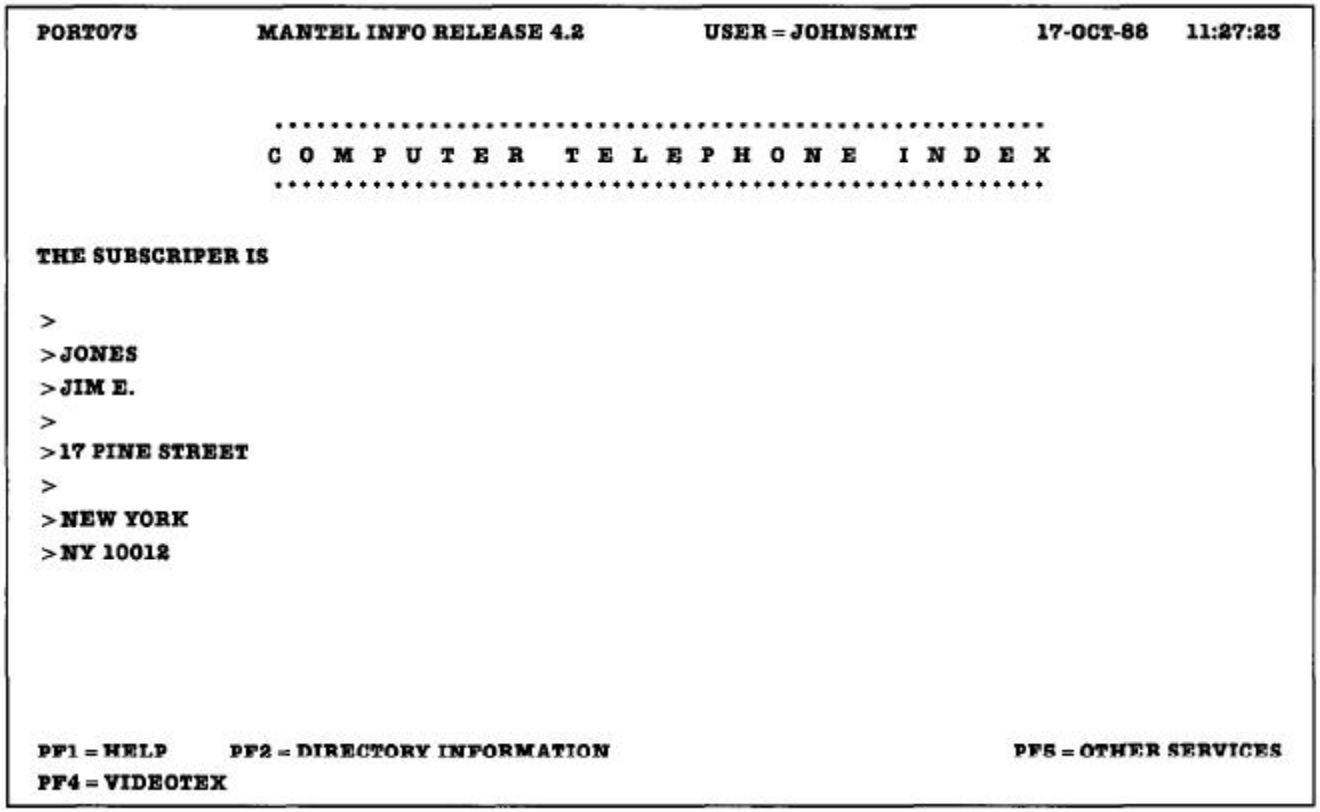
(Italics are mine)

What is the Reality of the Situation?

- User testing is extremely expensive in time and money
  - This despite the good arguments for its value in the long term
- Practitioners are often faced with the reality of providing an expert review or *no design input whatsoever*
  - What is preferable?
- Can there really be *no* generalizability from prior human factors data or psychological research?
  - Are guidelines and principles useless
- The answer: Research shows that expert reviews find problems that later show up in user testing
  - But the literature displays an ongoing debate about the *validity* and *effectiveness* of heuristic evaluation and cognitive walkthroughs

Molich and Nielsen (1990)

- Used computer professionals
- Fictitious system
  - Two-screen character user interface computer system
  - Look up telephone numbers from customers' bill
- Molich and Nielsen identified 30 problems with system
  - This was an "expert evaluation" of sorts
- Evaluators provided with set of heuristics to use
- Evaluators found it difficult to identify all 30 problems
  - Range: 0-18, average: 11 problems identified
- Note their original points:
  - HCI design is neither common knowledge nor intuitive
  - Knowledge of a few design principles is useful
  - The more people look at a design, more problems identified


PORT075 MANTEL INFO RELEASE 4.2 USER = JOHNSMIT 17-OCT-86 11:27:23
COMPUTER TELEPHONE INDEX
THE SUBSCRIPER IS
>
>JONES
>JIM E.
>
>17 PINE STREET
>
>NEW YORK
>NY 10012
PF1 = HELP
PF2 = DIRECTORY INFORMATION
PF5 = OTHER SERVICES
PF4 = VIDEOTEX

Mentioned by %	Problem number		Description
95	15	Serious	Re-display input (telephone number) with subscriber information
92	9		Avoid the use of English terms if a Danish term exists
92	10		Use the Danish national characters wherever possible
77	4		Remove unnecessary information
74	18	Serious	Inform the user if it may take 30 seconds before a reply appears
73	5		Avoid mysterious characters (>); consider using field labels
64	8		The function keys should be listed in some natural order
64	24	Serious	The error messages are too vague
62	19	Serious	The options available to the user should be displayed
58	3		Avoid spelling errors
52	7		The first name should be written before the last name
42	26	Serious	The error messages should be more constructive
38	11		Do not distort information (username) entered by the user
32	12		Clarify or remove information that is difficult to understand
29	23		The word ILLEGAL may intimidate the user
27	30		"Enter number and RETURN" may be taken literally
18	31		Show an example of a telephone number in the initial prompt
17	6		Interspersed blank lines reduce the readability of an address
16	14		Questions must be expressed from the user's point of view
14	25		The system should tell how it has interpreted the user's input
13	16		3 different terms are used for "Telephone number"
12	13		The meaning of the notation PF1=HELP is not clear to novices
12	17		Coordinate placement of error messages with the rest of the system
12	27		The request "Try again" in an error message is meaningless
9	2		Avoid the use of abbreviations
9	28		Allow lower case L and the letter O instead of digits 1 and 0
8	29	Serious	Accept parentheses, spaces and hyphen in telephone number
4	20	Serious	There may be no emergency exit from the initial prompt
4	21		There is no emergency exit during a long retrieval
1	22	Serious	It may not be possible to edit input in the initial prompt

TABLE 1

Summary of 77 entries submitted in a contest for evaluating a human-computer dialogue. For each problem the table shows the percentage of the entries that identified the problem. Problem 1 does not appear, since it was described in an example in the text of the exercise. The problem numbers refer to the detailed solution in Appendix 2. Some of the problems may prevent some users from using the system in a meaningful way. These problems are marked "Serious" in the table.

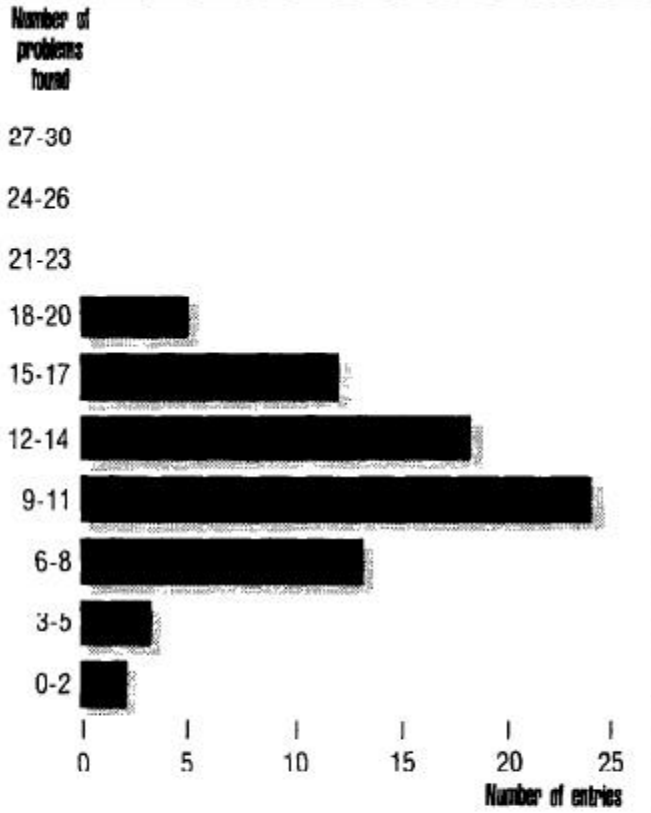
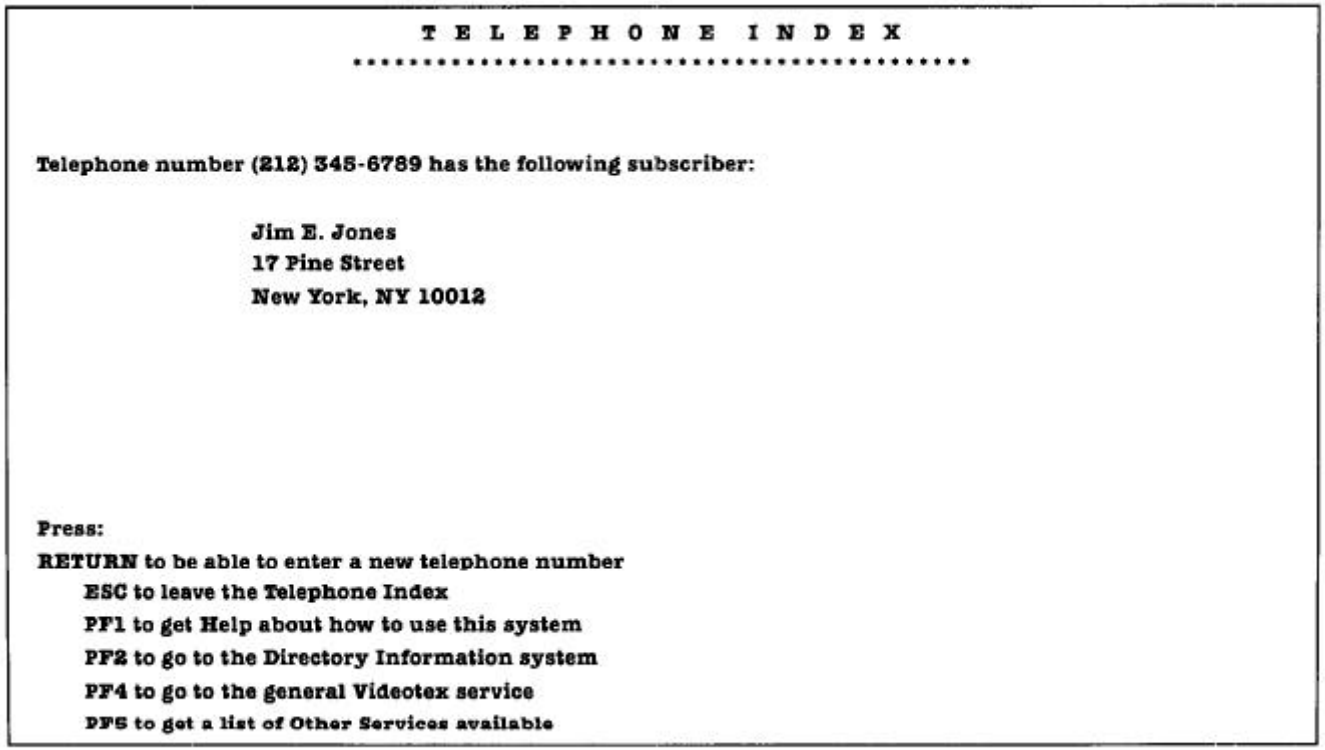


FIGURE 1

This diagram shows the distribution of the number of problems identified per entry in the contest referred to in the article. The median number of problems mentioned was 11 while the average was 11.2 out of 30 problems. The winner mentioned 18 problems. Problem 1 was described in an example in the text of the exercise; therefore, any mentioning of problem 1 is not included in this figure.


T E L E P H O N E I N D E X
Telephone number (212) 345-6789 has the following subscriber:
Jim E. Jones
17 Pine Street
New York, NY 10012
Press:
RETURN to be able to enter a new telephone number
ESC to leave the Telephone Index
PF1 to get Help about how to use this system
PF2 to go to the Directory Information system
PF4 to go to the general Videotex service
PF5 to get a list of Other Services available

Bailey, Allan, & Raiello (1992)

- Used Molich and Nielsen's (1990) task
  - Claimed that many of the 29 problems would not have a real effect on actual users' performance on, or preference of, the system
- Did a usability test on simulation of M&N system
  - Collected performance data and preference ratings
    - time to complete task and errors
  - One group used the original M&N system
  - Three more groups used system modifed by one change (each time) based upon previous testing results
  - Fifth group used M&N's ideal system, with the 29 problems fixed
- Results
  - Significant difference found between first and second group with one improvement
  - No reliable differences found between other successive improvements of system, including ideal system

Bailey, et al. Conclusions

- Only two problems out of the 29 made a difference in performance and preference (one change per screen)
- Conclude: Heuristic evaluation identifies many problems that are not related to performance or preference when tested on real users
  - Heuristic evaluation produces many “false positives”
- This is wasteful: will go through the expense of fixing many problems that are not real problems

- What problems might there be with Bailey et al.’s conclusions?

Bailey, R. W., Allan, R. W. & Raiello, P. (1992) Usability testing vs. Heuristic evaluation: A head-to-head comparison. *Human Factors Society Proceedings,* p. 409.

Jeffries, Miller, Wharton, & Uyeda (1991)

- Software user interface evaluated by four groups using four different techniques
  - Heuristic evaluation
  - Software guidelines
  - Cognitive walkthroughs
  - Usability testing
- User interface specialists (“experts”) did the heuristic evaluation
- Non-experts (software developers) did guidelines and walkthrough methods
- User interface expert conducted study on six users
  - Evaluated HP-VUE, GUI for Unix system (prior to Motif)

Jeffries et al. Results

- Three times more usability problems were identified by experts using heuristic evaluation
- Severity of problems rated and number of severe problem found by each method evaluated
- Heuristic evaluation produced the best results
  - Found the most problems
  - Found more of the most serious problems
  - Lowest cost
- Usability testing was second at finding serious problems
  - Good at finding recurring and general problems
  - Good at avoiding low-priority problems
- Analysis of time to conduct review versus problems found makes heuristic evaluation by experts the most cost-effective

Jeffries et al. (1991)

Number of error found by groups

	Heuristic	Testing	Guidelines	Walkthrough
Total	152	38	38	40
Severe	28	18	12	9

Jeffries, R., Miller, J. R., Wharton, C., & Uyeda, K. M. (1991) User interface evaluation in the real world: A comparison of four techniques. *CHI Proceedings*, p. 119.

Why Might an "Expert" be Different?

- What might an expert bring to an heuristic analysis?
  - Technical background
    - Knowledge of design guidelines
    - Greater and more detailed knowledge of more guidelines and principles (than simple Nielsen heuristics, for example)
    - General knowledge of cognitive psychology, behavioral science, human factors literature and concepts
  - Practical experience in user interface design
    - Experience with the results of user testing on systems
    - Experience with released products in the same domain as the product in development -- knows the problems in the field
    - Knowledge of mistakes made in the past on similar systems
    - Knowledge of the user population

Response to Jeffries et al. (1991)

- Bailey et al. (1992): Heuristic evaluation and usability testing find different types of problems
  - Ideal is to use both
  - Must identify high from low priority problems in heuristic evaluation
- What is the “true” measure of what is a “problem”?
- Karat et al (1992) compared usability testing to walkthroughs conducted by groups and individuals
  - Walkthroughs conducted by non-experts
  - Testing found 2x the number of problems found by groups and 3x number of problems found by individual
- Day & Boyce (1993):
  - Difference between explained by user of experts or not
  - Both methods valuable and should be used at different stages in the design process

Karat, C. M., Campbell, R. & Fiegel, T. (1992). Comparison of empirical testing and walkthrough methods in user interface evaluation. *CHI Proceedings*, p. 397.

Desurvire, Lawrence & Atwood (1991)

- Interactive telephone-based user interface
- Compared violations of UI against Smith & Mosier guidelines
- Four groups
  - User method, nine tasks on prototype
  - Heuristic analysis with experts, based on requirements
  - Heuristic analysis with non-experts, requirements
  - Usability testing
- Ratings collected from all groups on 10 selected guidelines
- Experts predicted percentage of users completing task and completing task without errors

Desurvire, H., Lawrence, D., & Atwood, M. (1991) Empiricismversus judgement: Comparing user interface evaluation methods on a new telephone-based interface. *SIGCHI Bulletin, 23*(4), p. 58-59.

Desurvire, et al. Results

- Ratings from user method and experts predicted observed test performance
- Best guess predictions correlated highly with actual task completions: $R^2 = .61$
- Supports the value of heuristic evaluation
- Note that evaluation was done on paper specification!

Rooden, Green, & Kanis (1999)

- Exisiting programmable coffeemaker
  - Actual difficulties observed in field
- Compared with “practitioners” evaluations done while inspecting design *models* and viewing videotapes of user testing
  - Models were drawings and computer simulations
- Results:
  - Identified 7-23 problems, total of 86 as a group
  - 42 of those problems were actually observed in use of real products
- Characteristics of model played a role
- Problems did not appear in model or user testing
  - e.g. Lights not visible in sunny kitchen
  - Events happen in field which escape all evaluation methods
- Appears to support Bailey, but ...
  - User testing was done, and it suffered same consequence
  - Severity of problems not assessed

Rooden, M. J., Green, W. S., & Kanis, H. (1999). Difficulties in usage of coffeemaker predicted on the basis of design models. *HFES Proceedings*, p. 476.

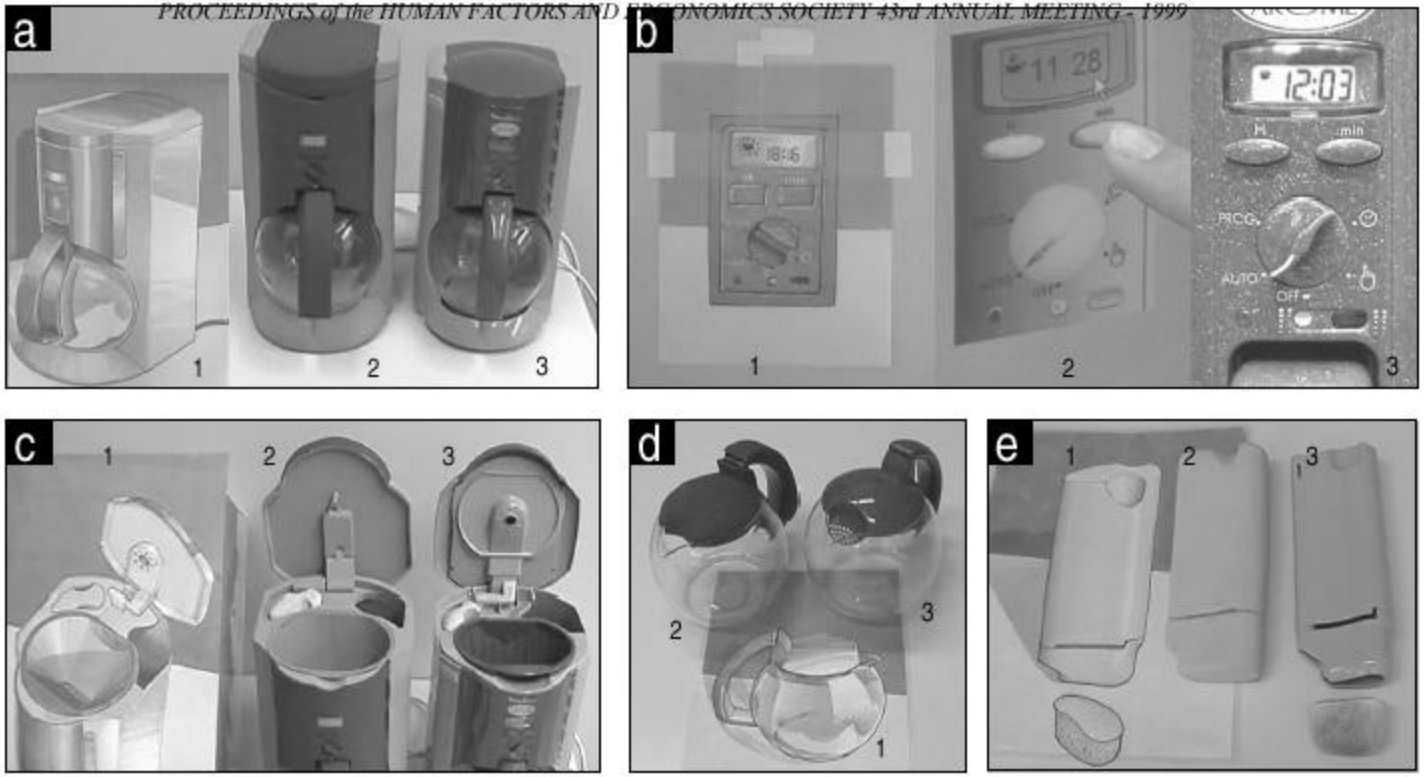
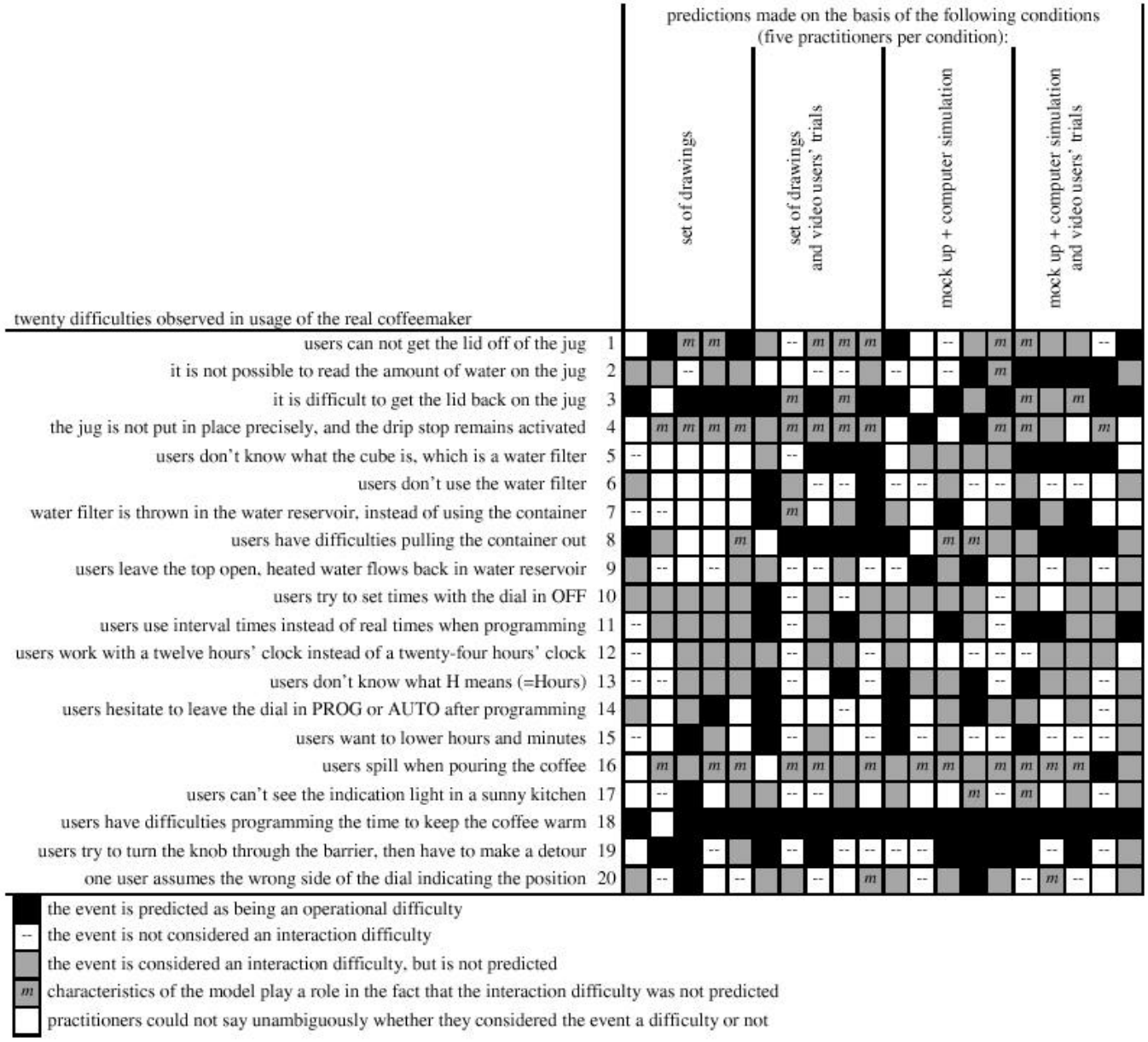


Figure 2. The coffeemaker and its derived design models. Each labeled picture (a, b, c, d and e) shows a part of the coffeemaker (a: whole product, b: control panel, c: opened top, d: jug, e: water filter and container) in each of the two design models (1: set of drawings, 2: foam mock-up and computer simulation) and the real product (3).


predictions made on the basis of the following conditions (five practitioners per condition):
set of drawings
set of drawings and video users' trials
mock up + computer simulation
mock up + computer simulation and video users' trials
twenty difficulties observed in usage of the real coffeemaker
users can not get the lid off of the jug 1
it is not possible to read the amount of water on the jug 2
it is difficult to get the lid back on the jug 3
the jug is not put in place precisely, and the drip stop remains activated 4
users don't know what the cube is, which is a water filter 5
users don't use the water filter 6
water filter is thrown in the water reservoir, instead of using the container 7
users have difficulties pulling the container out 8
users leave the top open, heated water flows back in water reservoir 9
users try to set times with the dial in OFF 10
users use interval times instead of real times when programming 11
users work with a twelve hours' clock instead of a twenty-four hours' clock 12
users don't know what H means (=Hours) 13
users hesitate to leave the dial in PROG or AUTO after programming 14
users want to lower hours and minutes 15
users spill when pouring the coffee 16
users can't see the indication light in a sunny kitchen 17
users have difficulties programming the time to keep the coffee warm 18
users try to turn the knob through the barrier, then have to make a detour 19
one user assumes the wrong side of the dial indicating the position 20
the event is predicted as being an operational difficulty
-- the event is not considered an interaction difficulty
the event is considered an interaction difficulty, but is not predicted
m characteristics of the model play a role in the fact that the interaction difficulty was not predicted
practitioners could not say unambiguously whether they considered the event a difficulty or not

Catani & Biers (1998)

- MS Windows library search software
- Compared effect of high versus low fidelity prototype (paper versus Visual Basic)
  - Found no effect of prototype with user testing
- 5 “usability professionals” identified problems on high fidelity prototype, 3-9 years experience
  - Not clear whether any formal heuristic analysis method used
- Total of 99 usability problems
  - 66 identified by professionals, 16 unique
  - 83 identified in usability testing, 33 unique
  - 50 problems idenitified both by professionals and testing
  - Most frequent problems found in testing were not the most frequent problems identified by the experts
  - *But note: Test users had defined tasks, experts were free to explore*
- Severity of problems rated by professionals, could not get good agreement on severity

Catani, M. B., & Biers, D. W. (1998). Usability evaluation and prototype fidelity; Users and usability professionals. *HFES Proceedings*, p. 1331.

Fu, Salvendy, & Turley (1998)

- Literature review: experts in heuristic evaluation and typical user testing subject in usability testing find different, *distinct* sets of usability problems
- Classes of problems:
  - Skill-based
    - perceptual and motor difficulties with signals and displays
  - Rule-based
    - consistency problems, can't detect system states, apply wrong rules
  - Knowledge-based
    - mental models
- Predict experts are effective in identifying skill-based and rule-based usability problems and usability testing with users will be effective in identifying knowledge-based problems

Fu, L., Salvendy, G., & Turley, L. (1998). Who finds what in usability evaluation. *HFES Proceedings*, p. 1341.

Fu, et al. Experiment and Results

- Internet multi-media training application
- Usability test, eight tasks
- Heuristic evaluation, eight tasks, used guidelines, were experts
- Total of 39 distinct problems
  - Only considered problems which were replicated
  - User testing: 21 problems identified
  - Heuristic evaluation: 34 identified
  - 41% overlap
- Predictions verified
  - Users found more knowledge-based problems
  - Experts more skill- and rule-based problems
- Explanations:
  - Mental models of users and experts are different
  - Users have best access to their own mental models
  - Expertise and experience is effective in identifying the skill and rule-based problems

Some Conclusions?

- Fu et al: It is best to do both testing and heuristic evaluation
  - Best at finding different sorts of problems
  - Use in the context of an iterative design process
- What of practical considerations?
  - Schedules and budgets
  - User interface professionals called in on limited basis
  - Expert evaluation is very cost effective
- What is the “true” measure of “real” problems?
  - Usability tests?
  - Problems found in field after product introduction?
  - Priority or significance of problems found is an important issue
  - How can this be assessed, from either of these various methods
- Expertise is important
  - Nielsen was wrong: Heuristics given alone to non-experts is *not* as effective
  - Evaluation by groups in better, and groups of experts even better
- Cognitive walkthrough methods, with designated task scenarios, may have advantages

What about Heuristics?

- Molich and Nielsen (1990)
  - Use simple and natural dialog
  - Speak the user's language
  - Minimize the user's memory load
  - Be consistent
  - Provide feedback
  - Provide clearly marked exits
  - Provide shortcuts
  - Provide good error messages
  - Prevent errors
  - Provide help and documentation

- Nielsen (1994) Improved heuristics:
  - Visibility of system status
  - Match between system and the real world
  - User control and freedom
  - Consistency and standards
  - Error prevention
  - Recognition rather than recall memory
  - Flexibility and efficiency of use
  - Aesthetic and minimalist design
  - Helping users recognize, diagnose, and recover from errors

http://www.useit.com/papers/heuristic/heuristic_list.html

"Research-based" Heuristics (1)

Gerhardt-Powals (1996)

1. Automate unwanted workload
   - Free cognitive resources for high-level tasks
   - Eliminate mental calculations, estimations, comparisons, and unnecessary thinking

2. Reduce uncertainty
   - Display data in a manner that is clear and obvious

3. Fuse data
   - Reduce cognitive load by bringing together lower level data into a higher level summation

4. Present new information with meaningful aids to interpretation
   - Use a familiar framework, making it easier to absorb
   - Use everyday terms, metaphors, etc.

5. Use names that are conceptually related to function
   - Context-dependent
   - Attempt to improve recall and recognition

"Research-based" Heuristics (2)

Gerhardt-Powals (1996)

6. Group data in consistently meaningful ways to decrease search time

7. Limit data-driven tasks
   - Reduce the time spent assimilating raw data
   - Make appropriate use of color and graphics

8. Include in the displays only that information needed by the user at a given time
   - Allow users to remain focused on critical data
   - Exclude extraneous information that is not relevant to current tasks

9. Provide multiple coding of data when appropriate

10. Practice judicious redundancy (to resolve the possible conflict between heuristics 6 and 8)

References

Nielsen, J. (1994) Enhancing the explanatory power of usability heuristics. *CHI Proceedings*.

Gerhardt-Powals, J. (1996) Cognitive engineering prinicples for enhancing human-computer performance. *Human-Computer Interaction, 8*(2), 189-211.

Bailey, R. (1999). http://www.humanfactors.com/library/may992.htm

Straub, K. (2003). http://www.humanfactors.com/downloads/sep032.html
→ Recent review concurs with opinion of lecture

A Possible Ideal?