Software for Context-Aware Multi-User Systems

Session 4
Quality Evaluation

João Pedro Sousa
CS 895 / SWE 821
George Mason University

not enough to evaluate quality
quality is built in

- in the 70's Japan's auto industry
  had trouble exporting because of low quality
- in the 80's the industry overhauls the production processes
  applying the notion of *total quality*
  from Armand Feigenbaum's 1951 book
- by the late 80's Japan builds the most reliable cars in the world
- in the 90's the world industry
  catches up to total quality
- software industry: big push in defense contracts SEI's CMM
  Software Engineering Institute, Capability Maturity Model

costs of quality
invest where it matters most

- many total quality attempts subside
  in the software industry
  because of costs of trying to get everything right
- fact:
  a small portion of the functionality
  gets used most of the time
  - in engineering this is called the *80-20* or *Pareto rule*
- given a limited budget for quality
  where do you place your chips?

under limited budgets
know practices with the most impact

most used practices	found to have most impact
1. visit customer site	1. iterative design
2. iterative design	2. user & task modeling
3. participatory design mockups	3. empirical studies
4. prototyping	4. participatory design
5. analysis of competition	5. visit customer site
	6. post-release follow-up

covered in SWE 632

practitioners survey

this gave rise to
the usability lifecycle

- pre-design
  - model the user, context & tasks
- design
  - participatory design: paratypes, prototypes, Wizard of Oz
  - analysis of current practice and competition
  - coordinated design & guidelines
- post-implementation
  - functional testing
  - empirical studies: lab, in situ, in the wild
- revise design for future releases

evaluation

participatory design
involve the end-user

- multidisciplinary teamwork
  - UI experts propose designs
  - users and stakeholders give feedback
- formative evaluation
  - paratypes
    - mockup device placed in real/realistic situations
      e.g., wooden PDA, voice recording phone
  - prototypes
    - minimally functional product:
      mostly UI, functional components stubbed
  - Wizard of OZ
    - fully functional product,
      but complex functions done by human "behind the curtain"
      e.g., automatic translation

empirical studies
depend on available time and budget

- in the lab
  - typical duration: one day
  - a few representative users, typically ~5-15
    - ideally a random sample: not your friends
- in situ
  - typical duration: a few days, maybe scattered
  - random sample of representative situations
- in the wild
  - typical duration: weeks or months
  - possibly entire user base
    - gather statistics of use
      mostly aggregated data but may drill down on cases of interest

which is the most conclusive evaluation?

empirical studies
different roles for the researcher

- in the lab
  - researcher provides training and guidance
- in situ
  - researcher is present but stays out of the way,
    may tape & make notes
  - ethnographic studies are in situ observations of natural behavior
- in the wild
  - researcher releases product
    - instrumented with mechanisms to collect usage data
  - users entirely left alone to explore at will
    - decide when and how and whether to use product

in the lab studies
technical steps

- explain goals & train participants on the app syntax
- provide concrete scenarios
  and ask users to perform concrete tasks
- verify the success criteria for each task
  - instrument the app, as needed
- record users' action and difficulties for later analysis
  - think aloud protocol
  - screen/video capture tools

empirical studies
gather data

- subjective satisfaction: questionnaires
  - Likert scale
    q: how easy did you find X?
    a: very easy / easy / ok / hard / very hard
  - open questions
    q: what did you find the hardest?
    q: what would you change?

empirical studies
gather data

- quantitative data
  - average and variance
    single variables, e.g., user speed
  - correlations and significance tests
    un/related variables, e.g., # items on menus vs. user speed
  - scatter plots/histograms
    bimodal distributions, e.g., user speed for experienced vs. novices; may also help with Likert scales…

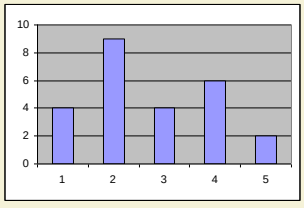
discussion
gathering data

- suppose your team is debating two design alternatives
  - you evaluate one with user A and the other with user B
  - A performed much better than B, what do you conclude?
    - difference may be due to user variability as much as 10x
    - have users (prefb. more users) test both designs and compare performance *diff for each* user

- suppose you evaluate some x of interest and the average x for a group of users is much worse than you expected, what do you conclude?

example
survey on context-aware reminders

- question: would you like to have the app remind you to take your laptop if you'll need it during the day, before leaving home?
  - answers: 3 4 2 4 2 5 1 2 2 4 4 3 2 3 2 3 1 1 1 4 2 4 5 2 2
    (1 - no, 2 - not really, 3 - maybe, 4 - yes, 5 - absolutely)
  - 25 respondents, average 2.72, mode 2
  - how do you interpret the results?
  - do an histogram:
    - subgroups of users with diff reactions
      personae



- also: why did you get those reactions?
  use disambiguation questions
  - do you normally take your laptop to work/school?
  - are you ok with always taking the laptop, even if you don't need it?
  - would you like to get a reminder...?

examples of evaluation

- [Consolvo 05] Lin
- [Marcu 11, Bardram 10] Rasheed