

# Evaluating Interface Designs

- Regardless of any insights, theories, or forethought that goes into an interface design, it must ultimately be evaluated and tested
- The major talent with testing is the broad spectrum of choices, each with its own +/- balance
- Some parameters that determine the most appropriate approach have been offered up: stage of design, novelty of project, number of expected users, criticality of the interface, costs and available funding, time available, and experience of the design and evaluation team
  
- Testing and evaluation ranges from informal, inexpensive tests to elaborate, expensive plans, especially for life- or mission-critical applications — increasingly in industry it has shifted from “want” to “need,” becoming more and more intrinsic to standard engineering practice
- In the end, tests still cannot provide guarantees:
  - ◆ Problems *will* be found even after release, so mechanisms must be in place to handle these
  - ◆ Prioritization balances the need to meet deadlines and the need to reach a certain degree of quality
  - ◆ Some systems have so many variables and possible usage scenarios that absolute comprehensiveness before real-world use is difficult

# Expert Reviews

- *Expert reviews* are a relatively low-overhead, time-friendly approach to evaluation; put simply, show the design to acknowledged experts, and have them provide feedback, which can range from informal discussions to a full formal report with recommendations
  - ◇ *Application domain* experts know the real-world tasks and objects in the application
  - ◇ *User interface domain* experts know the general principles of interaction design; they may need a significant amount of training on the application domain
- Expert reviews should aim for comprehensiveness, not pinpoint critiques; they should also supply priorities and estimated effort
- Many expert-review methods are possible:
  - ◇ *Heuristic evaluation* looks at a design based on some established heuristics (e.g., Nielsen's 10, Shneiderman's 8 golden rules, Tognazzini's first principles)
  - ◇ *Guidelines review* checks a design against established guidelines documents
  - ◇ *Consistency inspection* focuses on consistency at various levels: terms, fonts, colors, etc.; this is amenable to automation
  - ◇ *Cognitive walkthrough* is user- and task-oriented: "a day in the life" scenarios have experts go through typical user tasks
  - ◇ *Formal usability inspection* involves a piecemeal presentation of a design in front of experts, with designers justifying design decisions and experts indicating problems
- In the end, expert reviews are produced by individuals, and as such are subject to subjective pitfalls:
  - ◇ Insufficient knowledge of the application domain and user base
  - ◇ Conflicting opinions among experts ("For every PhD, there is an equal and opposite PhD.")
  - ◇ Experienced experts may lose sight of how first-time users might behave

# Usability Testing and Laboratories

- Once considered a nice luxury in the presence of extra time and resources, in-house testing procedures are increasingly integral to the development process
- Distinction between traditional controlled-experiment testing (i.e., the scientific method) and advertising- or marketing-influenced approaches — in one case, the goal is to validate/invalidate a hypothesis; in the other, the goal is to find areas for “improvement”
- In the end, lab testing is still *lab* testing — it doesn’t replace real-world environments and sustained use
  
- Large development shops may maintain a general-purpose *usability laboratory* that can test the full spectrum of possible products, equipped with:
  - ◆ One-way mirror for live observation
  - ◆ Video-recording equipment for later study, particularly to capture users “thinking aloud”
  - ◆ Software instrumentation, also for later study
- Different types of usability test are also possible:
  - ◆ *Paper mockups* can be performed early, at little cost
  - ◆ *Discount usability testing*, proposed by Nielsen, is a scaled down approach based on an 80-20 rule of sorts — the most glaring flaws can be found even by the smallest tests
  - ◆ *Competitive usability testing* compares previous versions or competing products
  - ◆ *Universal usability testing* emphasizes diversity of users, platforms, devices, etc.
  - ◆ *Field tests and portable labs* try to perform tests in a more realistic environment
  - ◆ *Remote usability testing* tries to use the Internet to facilitate online tests
  - ◆ *Can-you-break-this tests*, pioneered by game designers, focus on finding fatal flaws

# Survey Instruments

- Surveys are relatively inexpensive, even for large numbers of respondents; surveys may also feel familiar and “comfortable” to managers and other stakeholders
- Statistical rigor and avoidance of bias are important, as with any survey activity
- Shneiderman et. al. developed the *Questionnaire for User Interaction Satisfaction (QUIS)*, which has since been widely adopted as a survey instrument
  - ◆ Short and long forms, depending on users — short form focuses on general questions, while the long form adds questions on specific design details

# Acceptance Tests

- Adapted from standard engineering procedures, *acceptance tests* consist of *verifiable, quantifiable* assertions on the performance of a design, such as:
  - ◆ Time to learn specific functions
  - ◆ Time to perform a task
  - ◆ Maximum acceptable error rate
  - ◆ Retention over time
  - ◆ Subjective satisfaction (collected, perhaps, by QUIS)
- Avoids generic, subjective language like “user-friendly;” sets easy benchmarks for success/failure

# Active-Use Evaluation

- The previous approaches generally perform evaluation *before* final product release — an ounce of prevention
- But continued evaluation *after* release is also valuable:
  - ◆ *Interviews and focus groups* can reveal new information
  - ◆ *Logging* facilitates coverage of *all* active users, and provides quantitative data; just watch out for privacy concerns
  - ◆ *Technical support* (phone, online, e-mail) not only provides direct assistance but also new insights for improving a product's design
  - ◆ *Suggestion boxes or feedback forms* help elicit direct comments from users
  - ◆ *Discussion groups* allow interactions among users, in addition to participation by support staff — again, another possible source of new insights from real-world use

# Controlled Experiments

- Finally, we come full-circle to traditional scientific method and controlled tests, with primary influences from psychology, since in the end, we are still studying people in a particular environment
- Controlled experimental results tend to be narrow in scope, but more reliable (and thus replicable) — instead of a single test making a sweeping generalization, multiple tests form a big-picture “mosaic”
- Successful controlled experimentation requires a great degree of training and experience