Introduction to Human-Computer Interaction

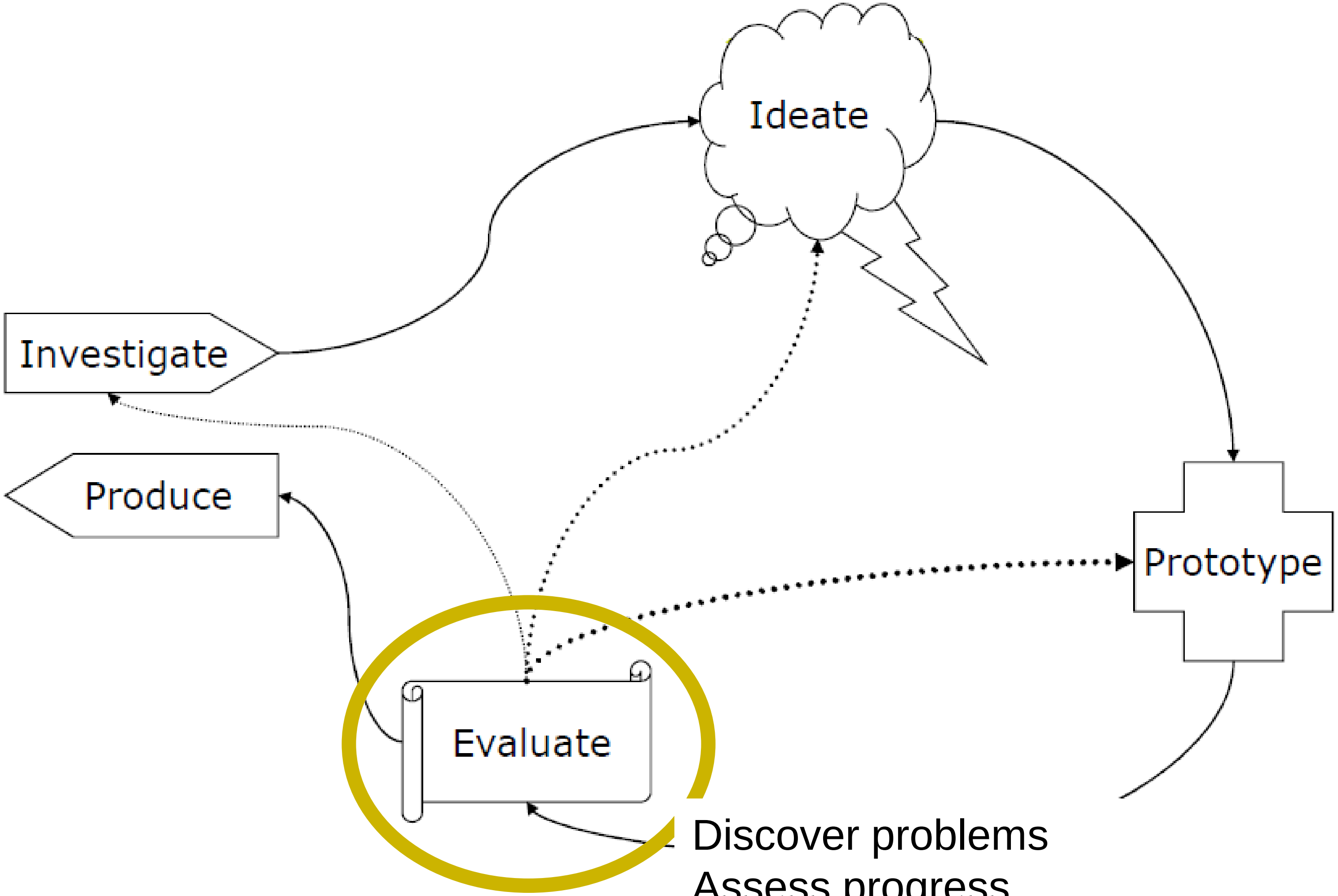
User Interface Design

Lecture 6 –Evaluation

Nadia Boukhelifa
nadia.boukhelifa@inria.fr

with acknowledgements to:
Petra Isenberg, Scott Klemmer, Anastasia Bezerianos, Anthony Tang, Nic Marquardt, Tobias Isenberg, Raimund Dachselt

User Centered Design Cycle



Discover problems
Assess progress
Determine next steps

usability

- ***usability***: ease with which ***people*** (users) can use a particular tool or object to achieve a specific goal

- aspects of usability:
  - *learnability:* how easy to accomplish tasks the first time?
  - *efficiency:* once learned, how quickly to complete tasks?
  - *memorability:* how easy to reestablish proficiency after not having used a design for a period of time?
  - *errors:* how many, how severe, how easy to recover?
  - *satisfaction:* how pleasant to use the design?

usability test

- a usability test is a "formal" method for evaluating whether a design is learnable, efficient, memorable, can reduce errors, and meets users' expectations.
  - users are not being evaluated
  - the design is being evaluated

why, what, where, when to evaluate?

- **why**: to check that users can use a product and that they like it
- **what**: conceptual modes, prototypes (early & late)
- **where**: in natural or laboratory settings
- **when**: throughout design or after to inform new products

when should I use a usability test?

- any time.
- early:
  - exploring potential possible designs
- late:
  - close to end stage to determine possible showstoppers
- after:
  - investigate reported problems

usability evaluation

- has become an established and accepted part of the design process
- might be anywhere from
  - an ambitious two-year test with multiple phases for a new national air-traffic–control system
  - to a three-day test with six users for a small internal web site

Bruce Tognazzini…

“Iterative design, with its repeating cycle of design and testing, is the only validated methodology in existence that will consistently produce successful results.

If you don’t have user-testing as an integral part of your design process you are going to throw buckets of money down the drain.”

How to

DESIGN AN EVALUATION

1| choosing an evaluation goal

what usability goals could we be interested in studying for a new smart watch?

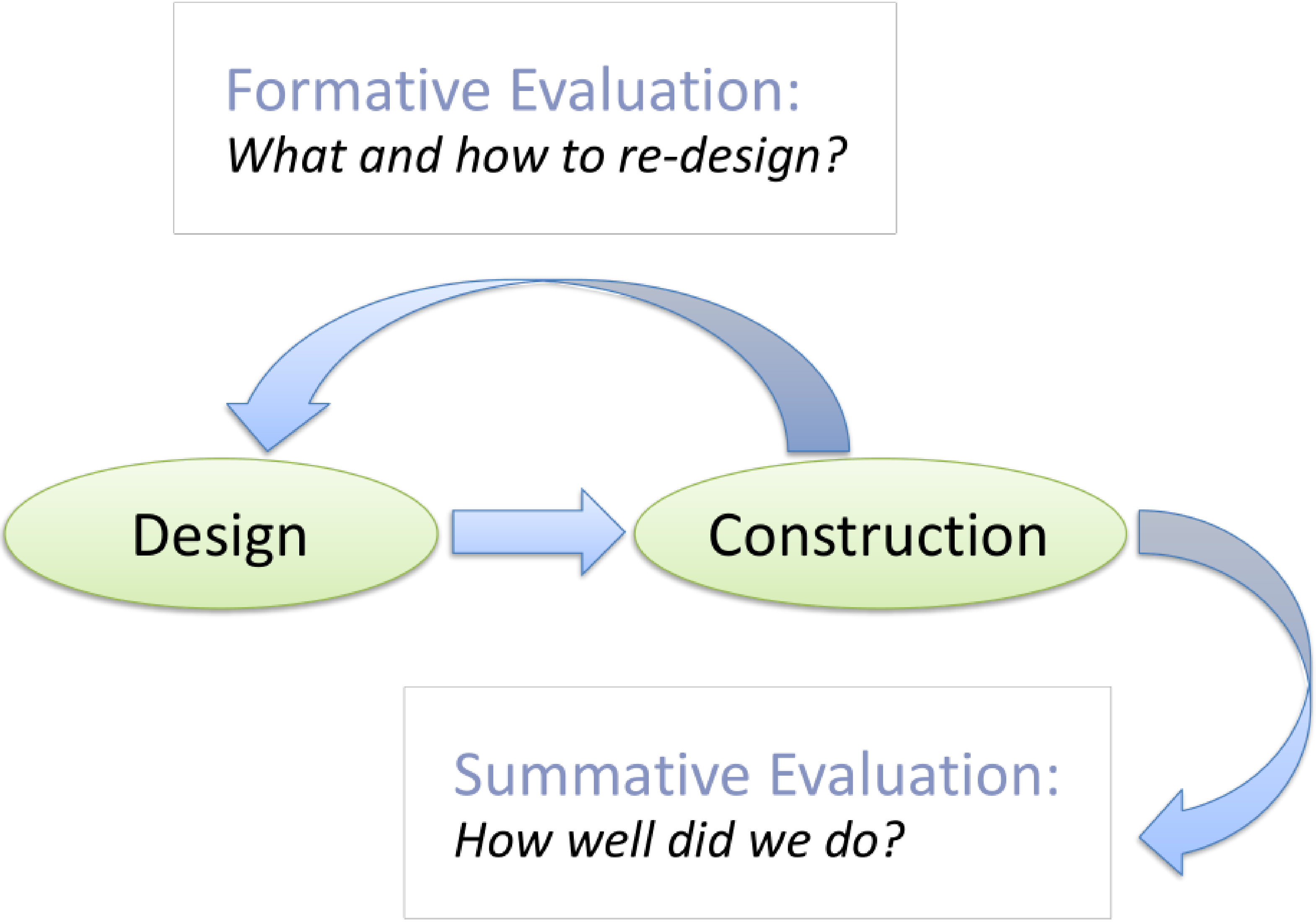
variations on usability tests

- how well do people learn the interface?
- does the interface work with people's actual real-life interactions?
- how well does this interface work when people are busy with other things?
- how well does this interface work with only a few seconds of interaction at a time?

things you care about in a usability test

- learnability / discoverability:
  - how easy is it for users to accomplish basic tasks the first time they encounter the design?
- efficiency
  - once users have learned the design, how quickly can they perform the tasks?
- memorability
  - when users return to a design after a period of not using it, how easily can they reestablish proficiency?
- errors
  - how many errors do users make, where are these errors occurring, and how easy is it to recover from these errors?
- Satisfaction
  - how pleasant is it to use?

evaluation goals



Formative Evaluation:

What and how to re-design?

Design

Construction

Summative Evaluation:

How well did we do?

[Rosson02]

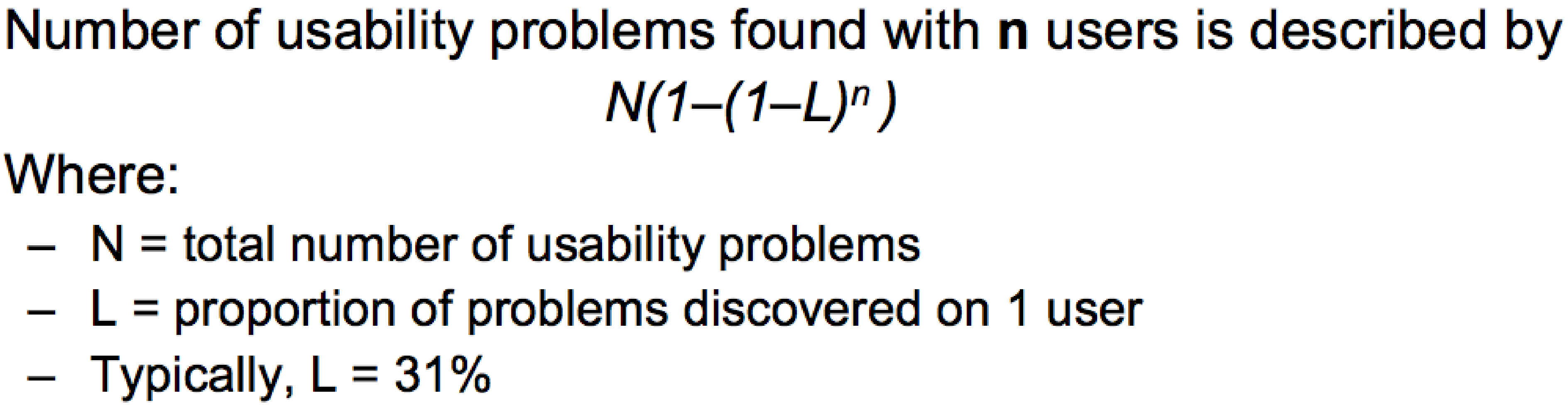
formative vs. summative evaluation

- formative: during development, guides process
  - find problems for next iteration of design
  - evaluates prototype or implementation, in lab, with chosen tasks
  - often qualitative observations and inspections of usability problems)
- summative: after development, or at a checkpoint
  - aimed at measures of quality
  - quantitative measures, e.g. performance times and error rates

2| *choosing users*

- who?
  - depends on your needs
  - goal: get the people that will be using it, or people that represent those that will be using it
- how many?
  - considerable debate in the community. Rule of thumb: ~5

usability tests: how many users?



Number of usability problems found with **n** users is described by

$$N(1-(1-L)^n)$$

Where:

- N = total number of usability problems
- L = proportion of problems discovered on 1 user
- Typically, L = 31%

Main argument: If you have 15 people, it’s better to test three designs with 5 users each, rather than one design with 15 people. ➔ Pragmatics, bang for buck

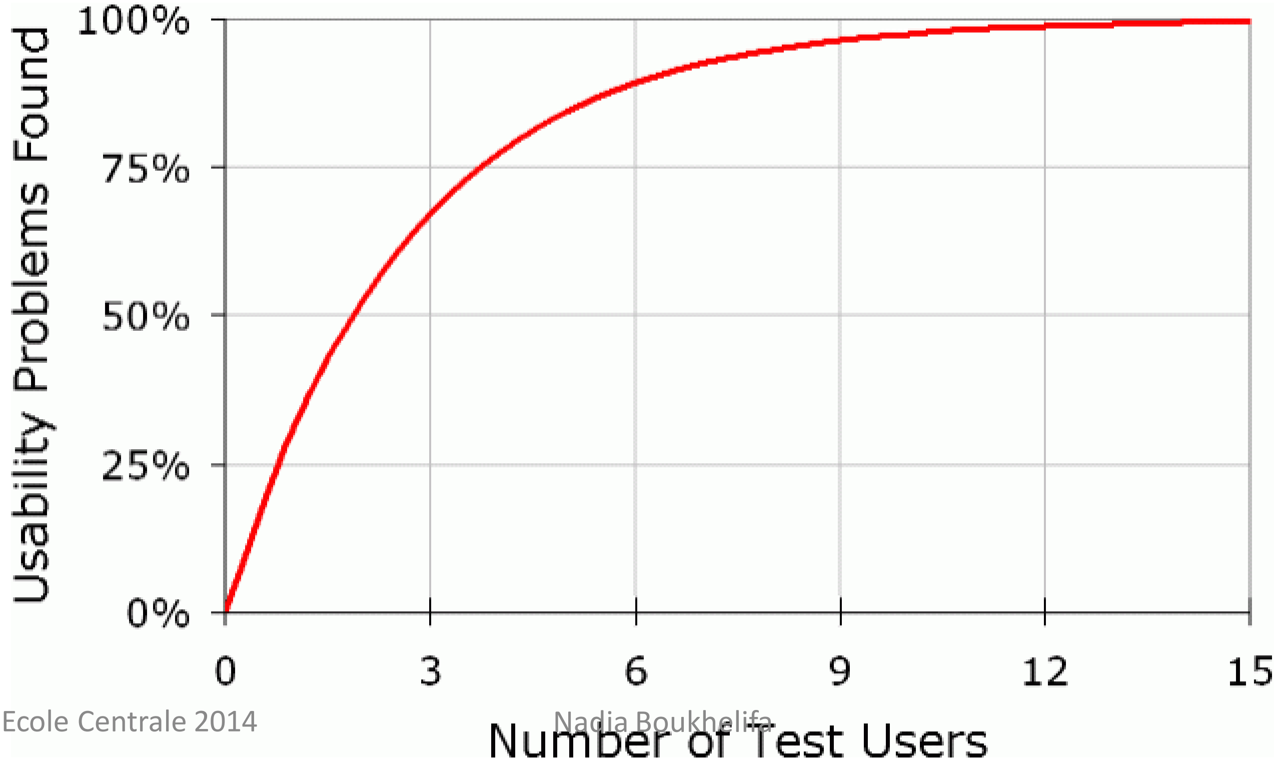
how many users?

Number of usability problems found with **n** users is described by

$$N(1-(1-L)^n)$$

Where:

- N = total number of usability problems
- L = proportion of problems discovered on 1 user
- Typically, L = 31%



3| choosing the right testing method

analytical vs. empirical methods

- analytical: theory, modeling, guidelines (from experts)
  - investigations that involve modeling and analysis of a system's features and their implications for use
  - produces many interpretations, but no solid facts
- empirical: observations, surveys (from users)
  - investigations that involve observations or other data collection from users
  - from informal to very systematic

analytics vs. empirical

- *“If you want to evaluate a tool, say an axe, you might study the design of the bit, the weight distribution, the steel alloy used, the grade of hickory in the handle, etc., or you might just study the kind and speed of the cuts it makes in the hands of a good axeman.”*

http://pixabay.com/en/war-fight-axe-battle-arms-weapon-33991/

evaluation methods

	Formative	**Summative**
Analytical	• claims analysis • **task analysis** • **usability inspection**	• theory-based design rationale • cognitive model • **expert review**
Empirical	• **think aloud observation** • critical incidents • **cognitive walk-through** • **user interviews / surveys** • **field studies**	• **competitive analysis** • usability specifications • **controlled experiment** • model testing

other terms that characterize methods

- qualitative vs. quantitative
- objective vs. subjective
- hypothesis testing vs. exploratory

INSPECTION TECHNIQUES

recap: Jakob Nielsen's Heuristics

1. Visibility of system status
2. Match between system and real world
3. User control and freedom
4. Consistency and standards
5. Error prevention
6. Recognition over recall
7. Flexibility and efficiency of use
8. Aesthetic and minimalist design
9. Help users recognize, diagnose, and recover from errors
10. Help and documentation

heuristic evaluation

- heuristics by Jakob Nielsen (1994) and others
- use of design principles/heuristics to inspect an interface for usability problems
- general approach:
  take the interface and check for the interface guidelines/heuristics
- number of evaluations
  - single inspector
  - multiple inspectors

heuristic evaluation – procedure

- inspect UI thoroughly
- compare UI against heuristics
- list usability problems

heuristic evaluation works on:

- sketches, paper prototypes, unstable prototypes, prototypes, systems

single-inspector heuristic evaluation

- example:
  average over six case studies of heuristic evaluation
  - 35% of all usability problems found
  - 42% of the major problems found
  - 32% of the minor problems found

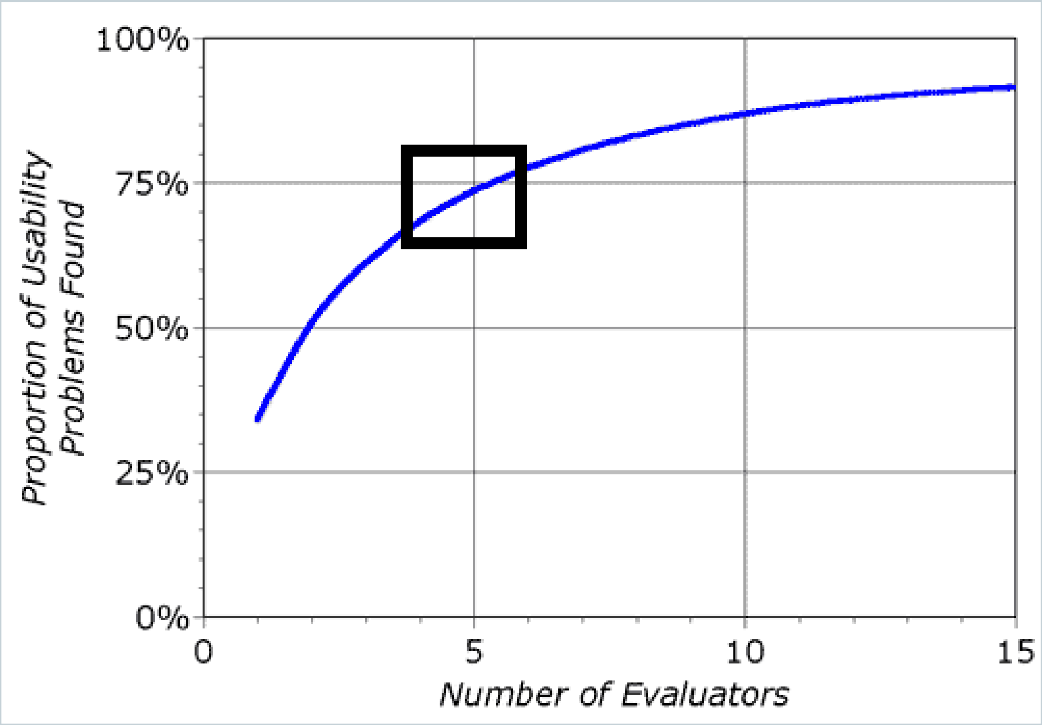
- score not great, but finding some problems with one inspector is better than finding no problems with no evaluators ...

single-inspector heuristic evaluation

- results vary according to:
  - difficulty of the interface being evaluated
  - expertise of the inspectors
- average percentage of problems found:
  - 22% – novice evaluators (no usability experience)
  - 41% – regular specialists (expertise in usability)
  - 60% – double specialists (expertise in both usability and the particular type of interface being evaluated; also find domain-related problems)
- tradeoff:
  novices yield poorer results, but are cheaper

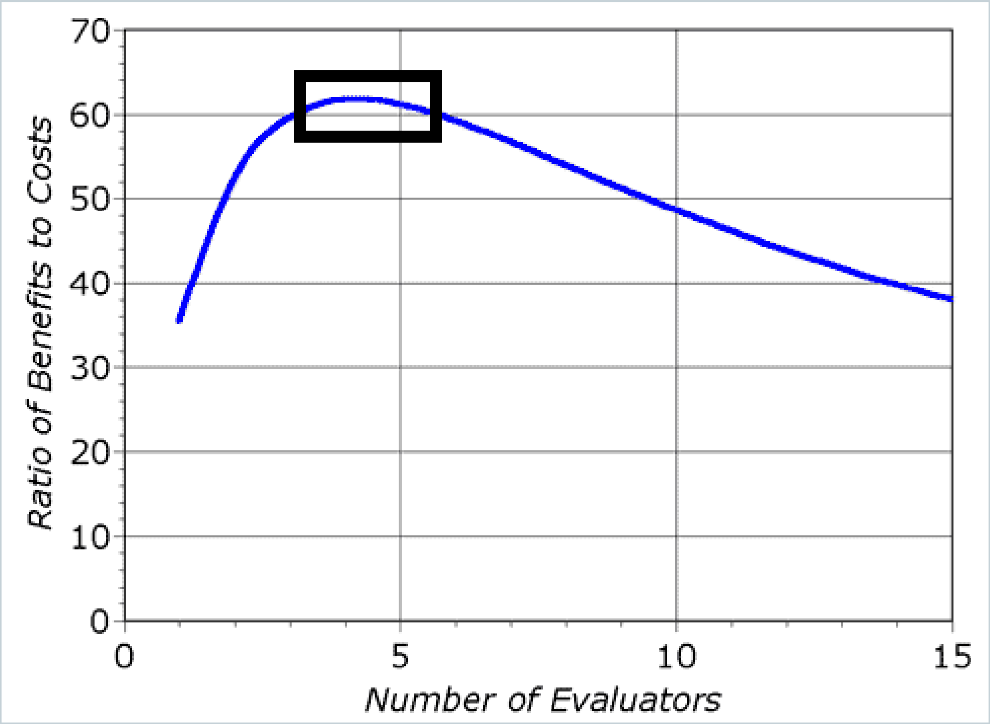
multiple-inspector heuristic evaluation

- 3–5 evaluators find 66%–75% of usability problems
  - different people find different usability problems
  - only modest overlap between the sets of problems found



multiple-inspector heuristic evaluation

- where is the best cost/benefit?
- depends on the costs, but:



heuristic evaluation: individuals vs. teams

- individual inspectors who look at an interface alone are recommended (according to Nielsen)

- reasons:
  - evaluation not influenced by others
  - independent and unbiased
  - greater variability in the kinds of errors found
  - no overhead required to organize group meetings

- problem: some interfaces require groups, then use several independent groups

heuristic evaluation

- benefits
  - can be difficult & expensive to find experts
  - many trivial problems are often identified
  - best experts have knowledge of application domain AND users
- biggest problems
  - few ethical & practical issues to consider because users not involved
  - important problems may get missed
  - inspections do not reveal validity of the findings
  - experts have biases, especially if they are the developers

self-guided vs. scenario exploration

- self-guided exploration
  - open-ended exploration
  - not necessarily task-directed
  - good for exploring diverse aspects of the interface, and to follow potential pitfalls

- scenario exploration
  - step through interface using a number of representative end user tasks (remember task-centered design)
  - ensures problems found in relevant parts of the interface
  - ensures that specific features of interest are evaluated
  - limits scope of evaluation – problems may be missed

MODEL-BASED ANALYSIS

predictive models

- out of established theories in science and engineering
- provide a way of evaluating products or designs without directly involving users
- less expensive than user testing
- usefulness limited to systems with predictable tasks - e.g., telephone answering systems, mobiles, cell phones, etc.
- based on expert error-free behavior
- classical example: GOMS analysis

GOMS analysis *(Card, Moran, Newell 1983)*

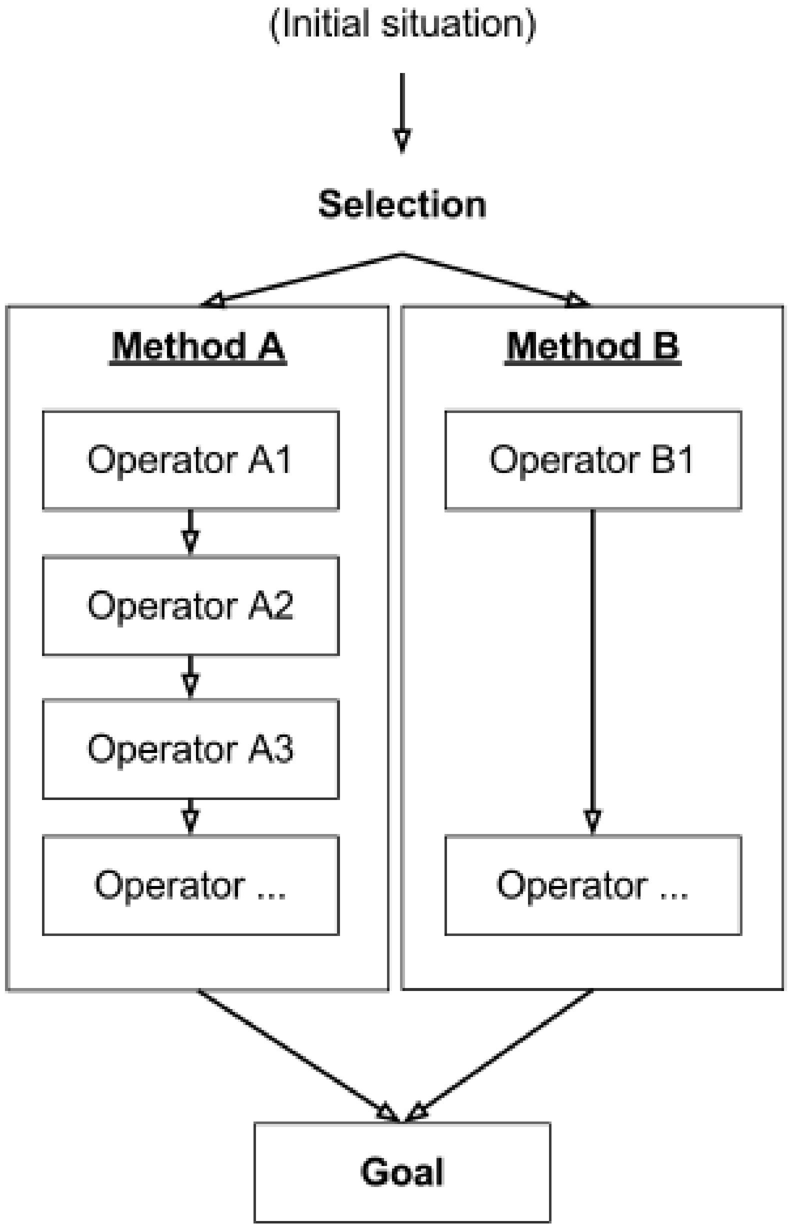
- GOMS: Goals, Operators, Methods, Selection Rules
  - Goals - what the user wants to achieve
    >> find a website
  - Operators - the cognitive processes & physical actions needed to attain the goals
    >> decide which search engine to use
  - Methods - the procedures for accomplishing the goals
    >> drag mouse over field, type in keywords, press the go button
  - Selection rules – describe when a user would choose a certain method over others
    (*selections rules are often not described in typical GOMS analysis)*

Selection

Method

Operator

Goal

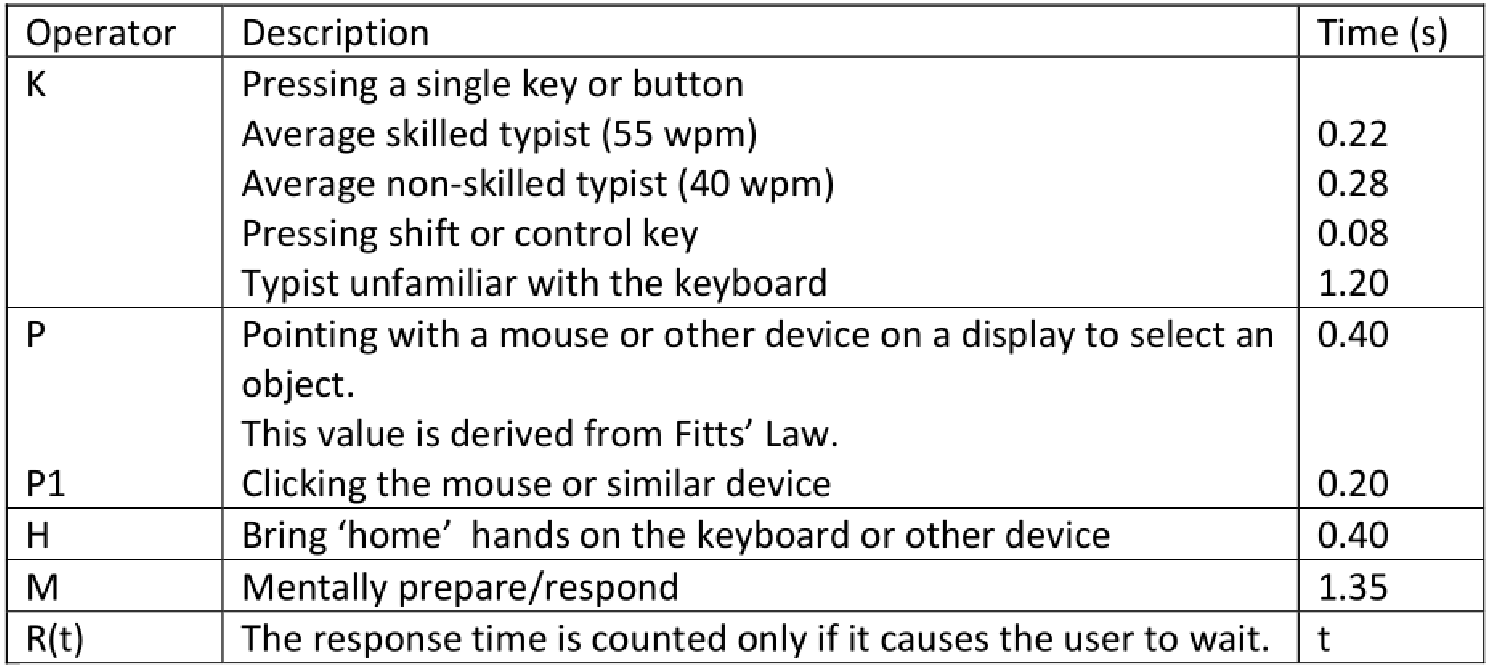


GOMS analysis

- build predictive model using scientific knowledge about human memory and behavior
  - like HTA, can analyze for complexity, consistency or build computational version, to estimate task times for design alternatives
- extends general techniques of HTA
  - goals, subgoals, plans, actions
  - BUT adds model elements for mental activities such as goal creation and selection, memory retrieval, etc.

keystroke level model

- quantitative GOMS model
- allows predictions to be made about how long it takes an expert user to perform a task
- response times for keystroke level operators (Card et al., 1983):



Operator	Description	Time (s)
K	Pressing a single key or button	
	Average skilled typist (55 wpm)	0.22
	Average non-skilled typist (40 wpm)	0.28
	Pressing shift or control key	0.08
	Typist unfamiliar with the keyboard	1.20
P	Pointing with a mouse or other device on a display to select an object. This value is derived from Fitts' Law.	0.40
P1	Clicking the mouse or similar device	0.20
H	Bring 'home' hands on the keyboard or other device	0.40
M	Mentally prepare/respond	1.35
R(t)	The response time is counted only if it causes the user to wait.	t

model-based approaches

- model-based approaches have good scientific foundation, are credible, can be very powerful
  - but current theories have limited scope, and developing the models takes time/expertise
- GOMS, Keystroke Level Model, & Fitts' Law only predict expert, error-free performance

EMPIRICAL METHODS

empirical evaluation

- what happens when people use the system in real situations?
- usability testing
  - involves ‘measuring’ of typical users doing typical tasks
  - data is used to calculate performance and to identify & explain errors

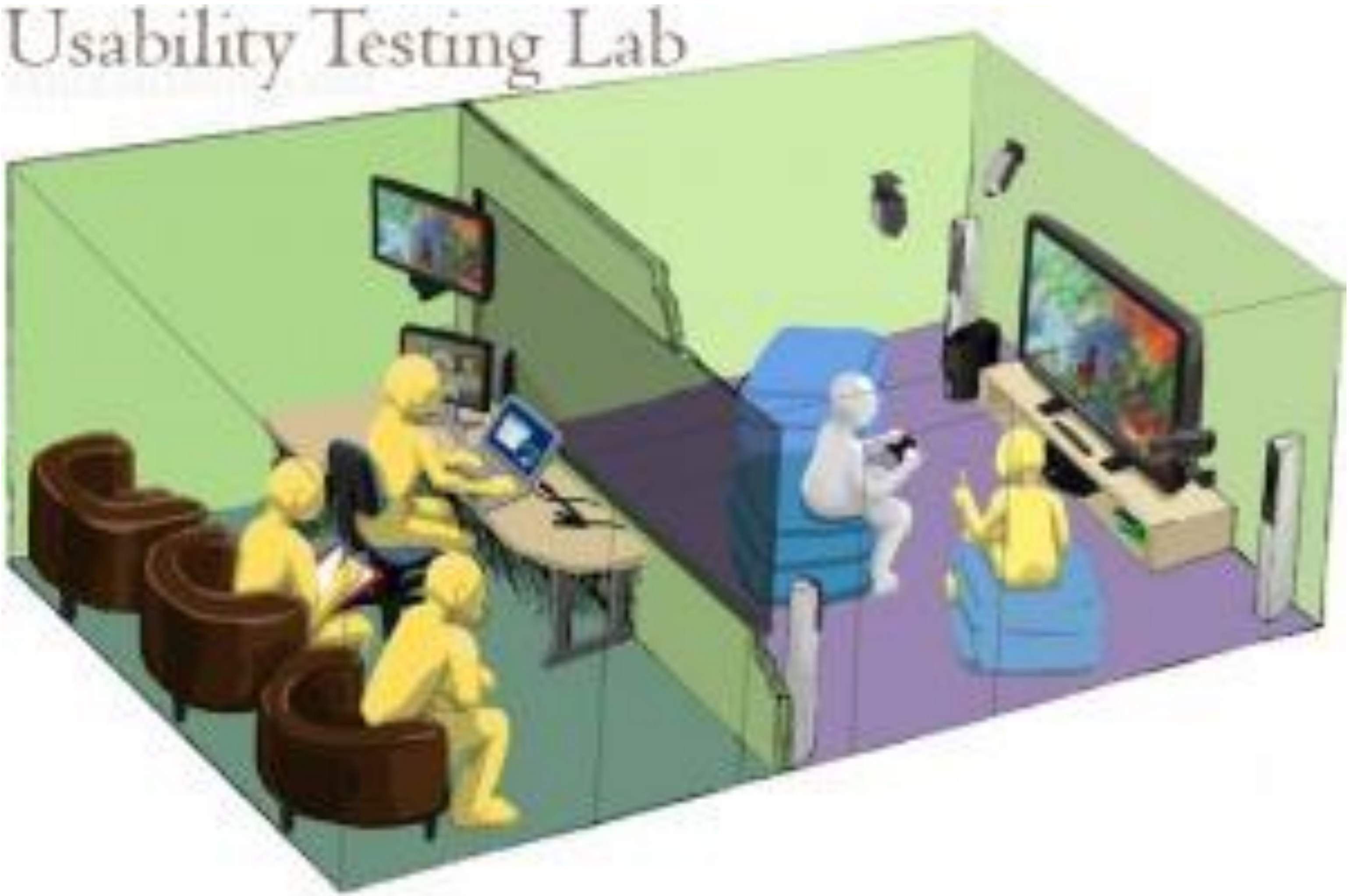
LAB-STUDIES

lab-based usability test: essentially…

- bring in real users
- have them complete tasks with your design, while you watch with your entire team
- use a think-aloud protocol, so you can “hear what they are thinking”
- measure
  - task completion, task time
  - satisfaction, problem points, etc.
- identify problems (major ones | minor ones)
- provide design suggestions to design/engineering team
- iterate on the design, repeat

testing environments…

testing environments…



testing environments…

testing environments…

empirical usability tests: HOWTO

- determine goals of usability test
- determine target audience & recruitment plan
- develop testing plan
  - what are the most important things you want to know? (top 10)
  - conceptual model extraction
  - provide non-leading questions or tasks
  - simple/realistic scenarios
  - prepare any written materials (audience-specific, if necessary)
- determine testing timeframe
- run a pilot study
- run your test with real participants

quantitative evaluation techniques

- quantitative evaluation
  - precise measurements
  - results in form of numeric values
  - bounds on how correct these statements are
- methods
  - user performance data collection
  - controlled experiments

collecting performance data

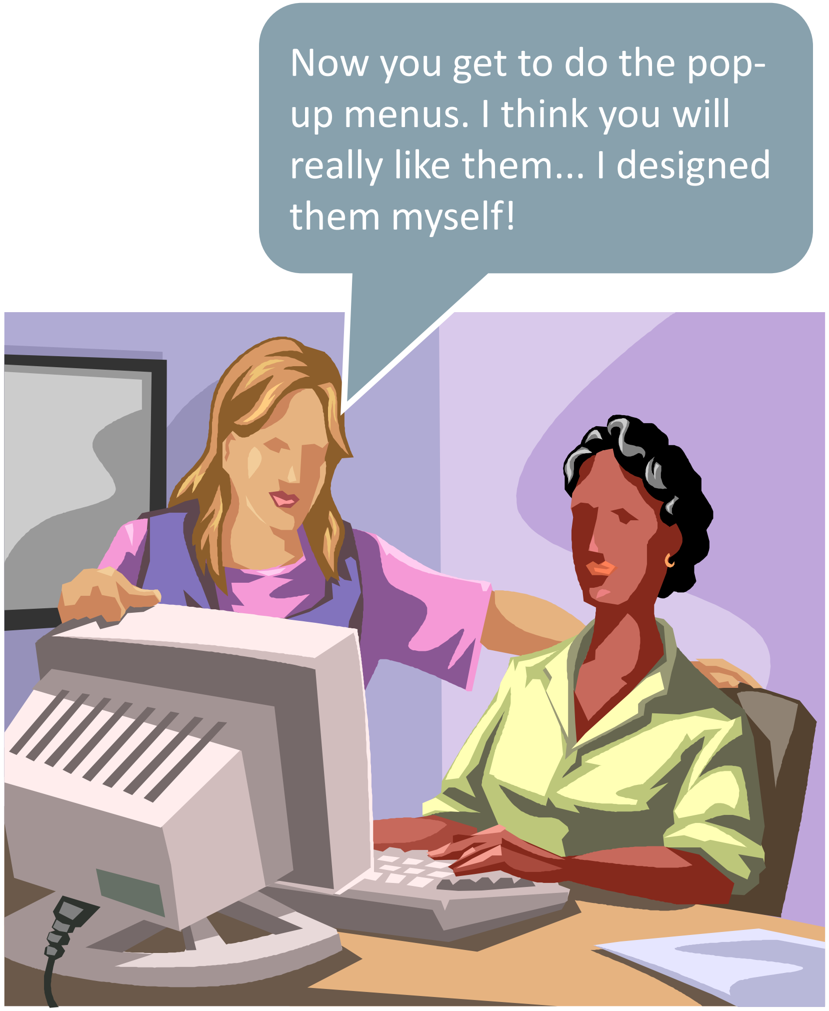
- people using a system (often lots of data)
- exploratory data collection
  - hope something interesting shows up
  - difficult to analyze
- targeted data collection
  - look for specific information, but may miss something
  - e.g., frequency & type of request for online assistance
  - e.g., frequency of use of different parts of the system
  - e.g., number of errors and where they occurred
  - e.g., time it takes to complete some operation
  - all these tell you something about the usability

controlled experiments

- traditional scientific method
  - obtaining a clear & convincing result on specific issues
  - in human-computer interaction (HCI)
    - insights into cognitive processes, human performance limitations …
    - results allow system comparison, fine-tuning of details …
- striving for
  - removal of experimenter bias
  - clear and testable hypothesis
  - control of variables and conditions
  - quantitative measurement
  - replicability of experiment
  - measurement of confidence in obtained results (statistics)

removal of experimenter bias

- unbiased instructions
- unbiased experimental protocols, for instance, by preparing scripts ahead of time
- unbiased subject selection


Now you get to do the pop-up menus. I think you will really like them... I designed them myself!

clear and testable hypothesis

- hypothesis: statement about the world
- examples – valid hypotheses?
  - The French are great football players.
  - The French are better football players than the Germans.
  - The French have won more soccer games than the Germans in the last four years.
  - The French have won more matches at the World and European Championships for men and women than the Germans in the last four years.
- hypotheses need to be clear, specific, and testable statements about the world/about our experiment

null hypothesis

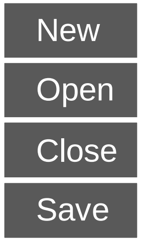
- "… is a *pinpoint statement* as to the unknown quantitative value of the *parameter* in the *population[s]* of interest" [Huck, S.W. *Reading Statistics and Research*]
- i.e., assigns a specific value to the parameter (= real value), use equality statements
- population & parameter vs. sample/participants & statistic

- goal: ***disprove*** the null hypothesis

null hypothesis

- example 1:
  There is no difference in the number of cavities in children and teenagers using Colgate and Elmex toothpaste when brushing daily over a one year period.

- example 2:
  There is no difference in user performance (time & error rate) when selecting a single item from a pop-up or pull down menu of 4 items, regardless of the participant's previous expertise with mice or using different menu types.



independent variables (factors)

- variables that are to be altered
  - ***independent*** of the participants' behavior
  - modification to the conditions the participants undergo
  - could also be the classification of participants into groups

- example 1: toothpaste

– toothpaste	Colgate or Elmex
– age	≤11 years or >11 years

- example 2: menus

– menu type:	pop-up or pull-down
– menu length:	3, 4, 5, 6, 7
– subject type:	expert or novice

New

Open

Close

Save

dependent variables (measures)

- variables that will be measured
  - ***depend*** on the participants' reactions to the independent variables in the experiment, included in hypothesis
  - specific things that will be measured quantitatively
- example 1: toothpaste
  - number of cavities
  - frequency of brushing
  - preference of toothpaste
- example 2: menus
  - time to select an item
  - selection errors made
  - time to learn to use it to proficiency

New

Open

Close

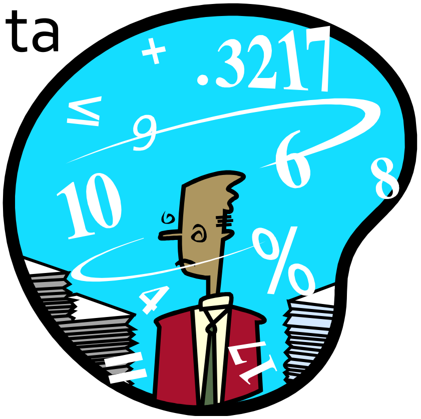
Save

usability testing metrics

- performance
  - task success, time on task, errors, efficiency
- issue metrics
  - identify issue, issue severity
- behavioural
  - observe verbal behaviour, issue severity
- self-reported
  - ease, satisfaction, clarity, comprehension, etc.

statistical analysis

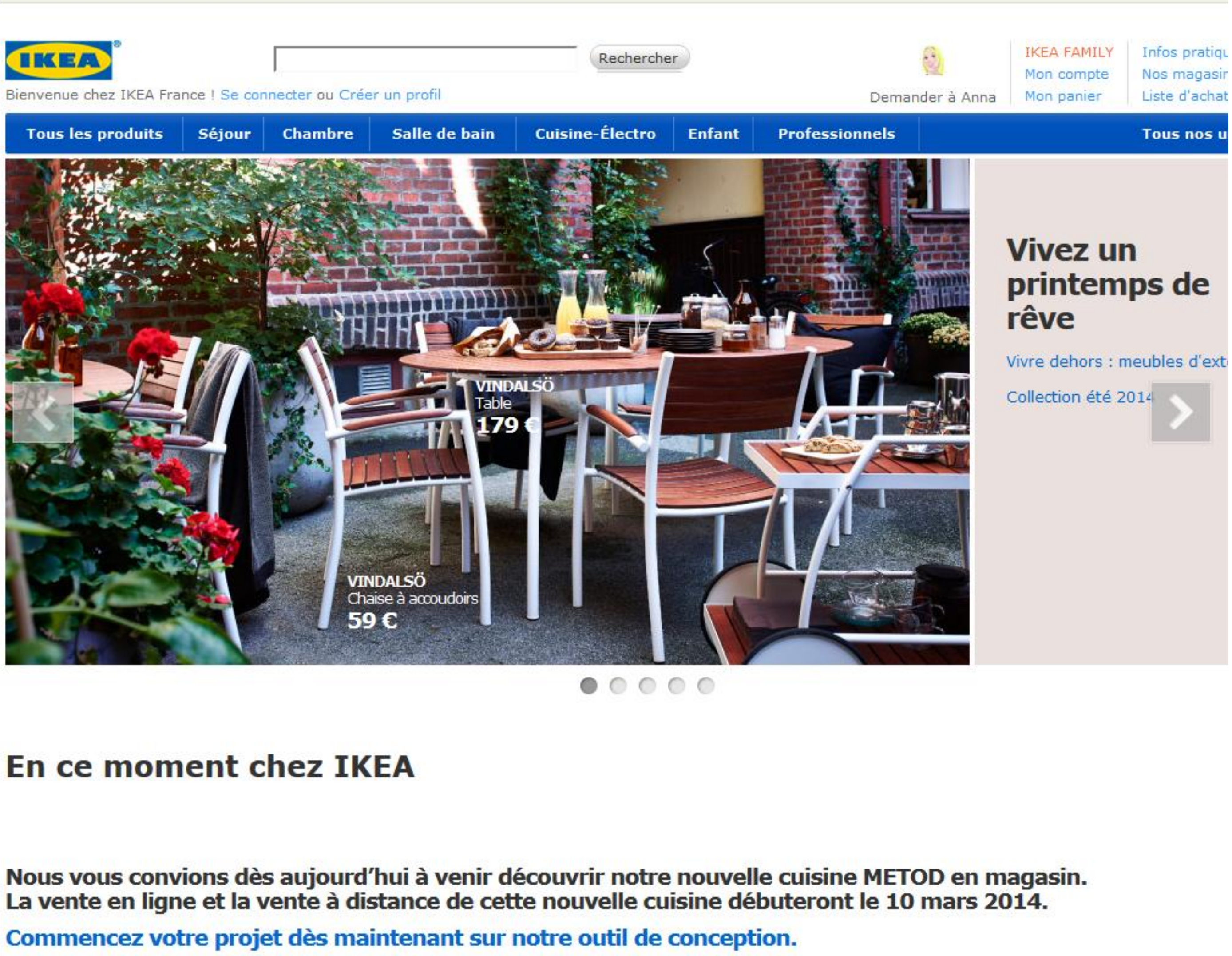
- statistical methods for analyzing collected data
  - mathematical attributes about collected data: mean (average), amount of variance, …
  - how data sets relate to each other
  - *p-value:* the probability that finding a difference when there is actually none (thinking that null hypothesis is false when it is true)
  - *statistical significance:* is achieved when the probability *p* is low, i.e. we can safely reject the null hypothesis
  - *confidence limits/intervals:* confidence that our conclusion is correct and that the findings are statistically significant, i.e., that we accept or reject the hypothesis (very likely) without making a mistake



statistical vs. practical significance

- statistical significance: probability that we rejected the null hypothesis wrongfully is low
  - popular levels: ≤5%, ≤1%, or ≤0.1% (for *p*-value)
- ***caution****:* when *n* is large, even a trivial difference may show up as a statistically significant result:
  - mean selection time of menu a: 3.00 seconds
  - mean selection time of menu b: 3.05 seconds
- statistical significance ***does not mean or imply*** that difference is important
  - matter of interpretation
  - statistical significance often abused and used to misinform

example: ikea website



Bienvenue chez IKEA France ! Se connecter ou Créer un profil

Vivez un printemps de rêve

En ce moment chez IKEA

Nous vous convions dès aujourd'hui à venir découvrir notre nouvelle cuisine METOD en magasin. La vente en ligne et la vente à distance de cette nouvelle cuisine débuteront le 10 mars 2014.

Commencez votre projet dès maintenant sur notre outil de conception.

usability tasks: IKEA example

- "find a bookcase"
  - search for "bookcase."
- "you have 200+ books in your fiction collection, currently strewn around your living room. find a way to organize them."
  - clicking on catalogues looking for a storage solution
  - searches (a few) were for "shelves" and "storage systems"

usability tasks

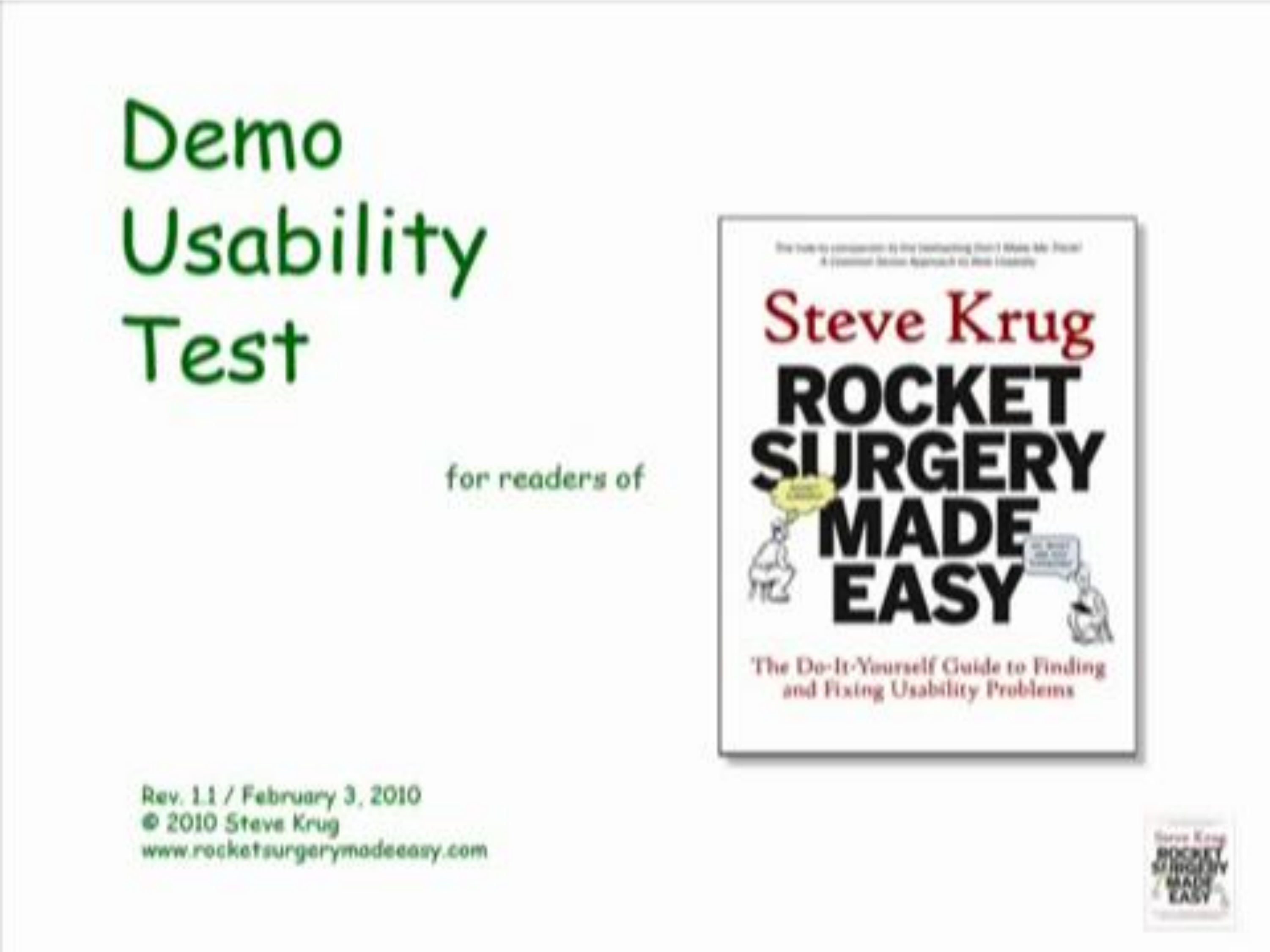
- again, depends a lot on what you're looking for
  - **specific**: does a task flow work?
  - **broad**: does your language match the user's mental model/language?
- consider "the context of use"
  - what would someone need to do with your tool?
  - under what circumstances would they be in? (relaxed vs. under pressure; non-interrupted vs. interrupted constantly)
  - etc.

usability tasks: Netflix.com

- “rate a few movies”
- “it’s a Friday night, and you’re looking for a movie to watch. What do you do?”
- “you’re about to watch `Batman 3’, but want to watch the first two, first. How do you do this?”
- “you want to watch Batman 1 through Netflix in your living room with your xbox. How do you set that up?”
- “what do you think about the site?”

think-aloud protocol

- as participants complete a task, you ask them to report
  - » what they are thinking
  - » what they are feeling
  - » rationale for their actions and decisions
- **idea**: rather than interpret their actions/lack of action, you can actually understand why they are doing what they are doing



Demo Usability Test

for readers of

Steve Krug

ROCKET SURGERY MADE EASY

The Do-It-Yourself Guide to Finding and Fixing Usability Problems

Rev. 1.1 / February 3, 2010
© 2010 Steve Krug
www.rocketsurgerymadeeasy.com

https://www.youtube.com/watch?v=QckIzHC99Xc

think-aloud protocol

- what’s weird:
  - people are not normally used to saying things out loud as they work.
  - they may also be embarrassed to say things out loud.

co-discovery learning protocol

- main idea: remove the awkwardness of think-aloud

- two people sit down to complete tasks
- only one person is allowed to touch the interface
- monitor their conversation

- variation: use a semi-knowledgeable “coach” and a novice (only the novice gets to touch the design)

making sense of your data

- statistics for quantitative measures
- affinity diagrams for qualitative results

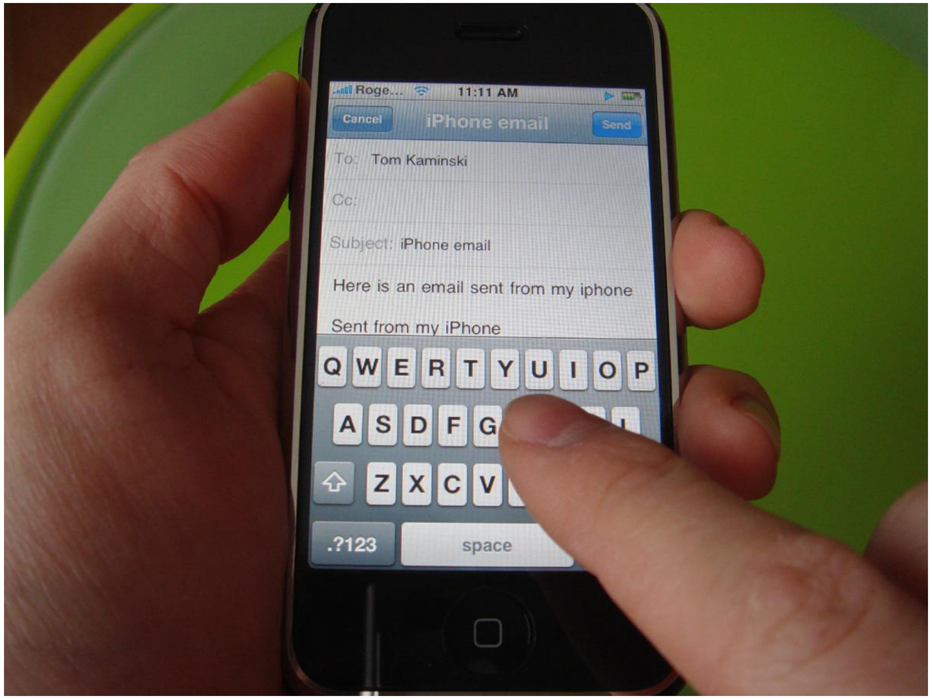
VALIDITY OF RESULTS

experimental validity

- external validity » realism
  - across situations
  - across people
- internal validity » integrity
  - confound
  - selection bias
  - learning effects
  - priming
  - experimenter bias

imagine this test…

- design a typing interface for use while running.
- bring people into the lab, put them at a desk.
- ask them to write an email, and time how long it takes.



To: Tom Kaminski

Cc:

Subject: iPhone email

Here is an email sent from my iphone

Sent from my iPhone

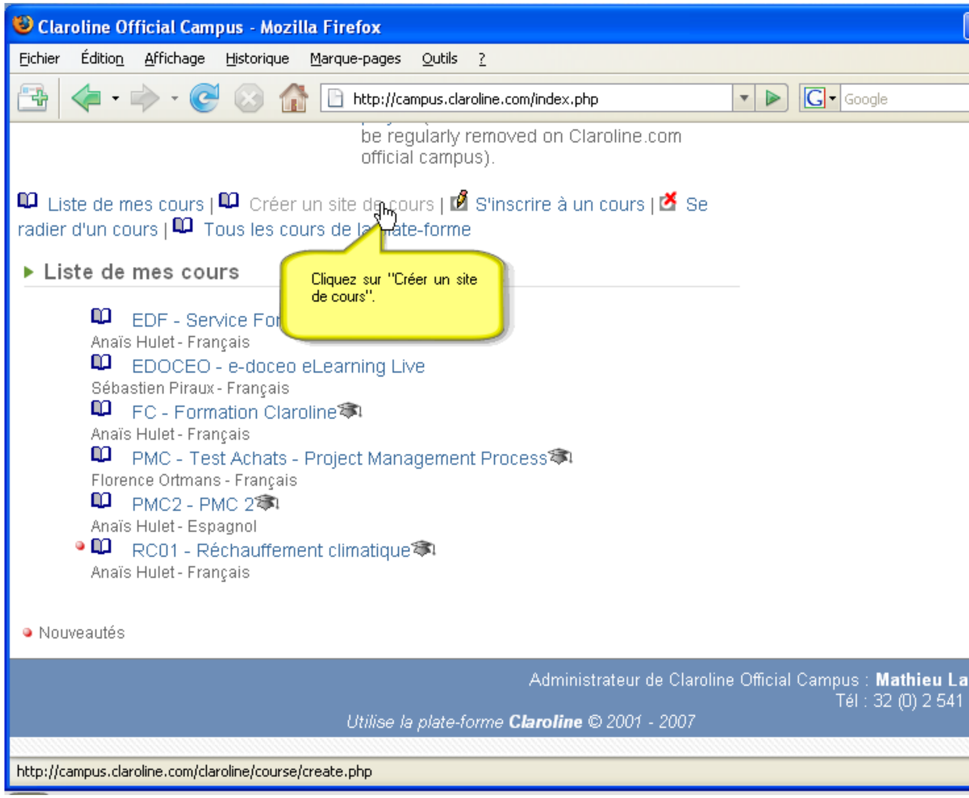
external validity » across situations

- does the test situation match the situation that the design will be used in?
- does it match at least in critical ways?
- what are aspects that are different?

- **artificiality**

imagine this test…

- recruiting developers of Claroline, ask them to register for courses.
- because they can register for their courses within 5 minutes, the interface is deemed usable.


Claroline Official Campus - Mozilla Firefox
Fichier Édition Affichage Historique Marque-pages Outils ?
http://campus.claroline.com/index.php
Google

be regularly removed on Claroline.com official campus).

Liste de mes cours | Créer un site de cours | S'inscrire à un cours | Se radier d'un cours | Tous les cours de la plate-forme

Cliquez sur "Créer un site de cours".

Liste de mes cours

EDF - Service For
Anaïs Hulet - Français

EDOCEO - e-doceo eLearning Live
Sébastien Piraux - Français

FC - Formation Claroline
Anaïs Hulet - Français

PMC - Test Achats - Project Management Process
Florence Ortmans - Français

PMC2 - PMC 2
Anaïs Hulet - Espagnol

RC01 - Réchauffement climatique
Anaïs Hulet - Français

Nouveautés

Administrateur de Claroline Official Campus : **Mathieu La**
Tél : 32 (0) 2 541

*Utilise la plate-forme **Claroline** © 2001 - 2007*

http://campus.claroline.com/claroline/course/create.php

external validity » across people

- are test subjects representative of the target user population?
- is it a randomly selected group, or are there constraints on how the group is selected that may affect test results?

- **generalizability across a population**

imagine this test…

- you design two computer games for children, and bring it to a school to test.
- the first 10 students that complete their homework are sent to your testing office for the first game.
- the next 10 students that complete their homework are sent to play the second game.

imagine this test…

- you design two computer games for children, and bring it to a school to test.
- the first 10 students that complete their homework are sent to your testing office for the first game.
- they find the game easy to play.

internal validity » confound

- when you are testing something, and changing one aspect of the test (i.e. a variable), if something else changes along with that variable, then you have a confound.
- this means that you cannot tell what is causing the difference.

- e.g. When you do not eat cereal in the morning, you are fine, but if you do, then you get sick. You conclude that you are allergic to cereal.

internal validity » selection bias

- systematic, non-random sampling of the population distorts your ability to generalize from the results.

imagine this test…

- you have designed two new interfaces for PeopleSoft. You recruit students to test your interface.
- for each participant, you give them your least favourite interface first to complete the task, and then you give them your favourite interface second.

https://prdrps3.ehs.ucalgary.ca/psp/saspr

Anthony Hoi Tin Tang

Faculty Center

My Schedule

Fall 2012 | University of Calgary

Select display option: ● Show All Classes ○ Show En

imagine this test…

- you are designing a colour scheme for your interface, and recruit participants for the entire day. For morning participants, you use interface A; for afternoon participants, you use interface B.
- morning participants seem to have no problems with the interface, but participants take a lot more time to complete the task

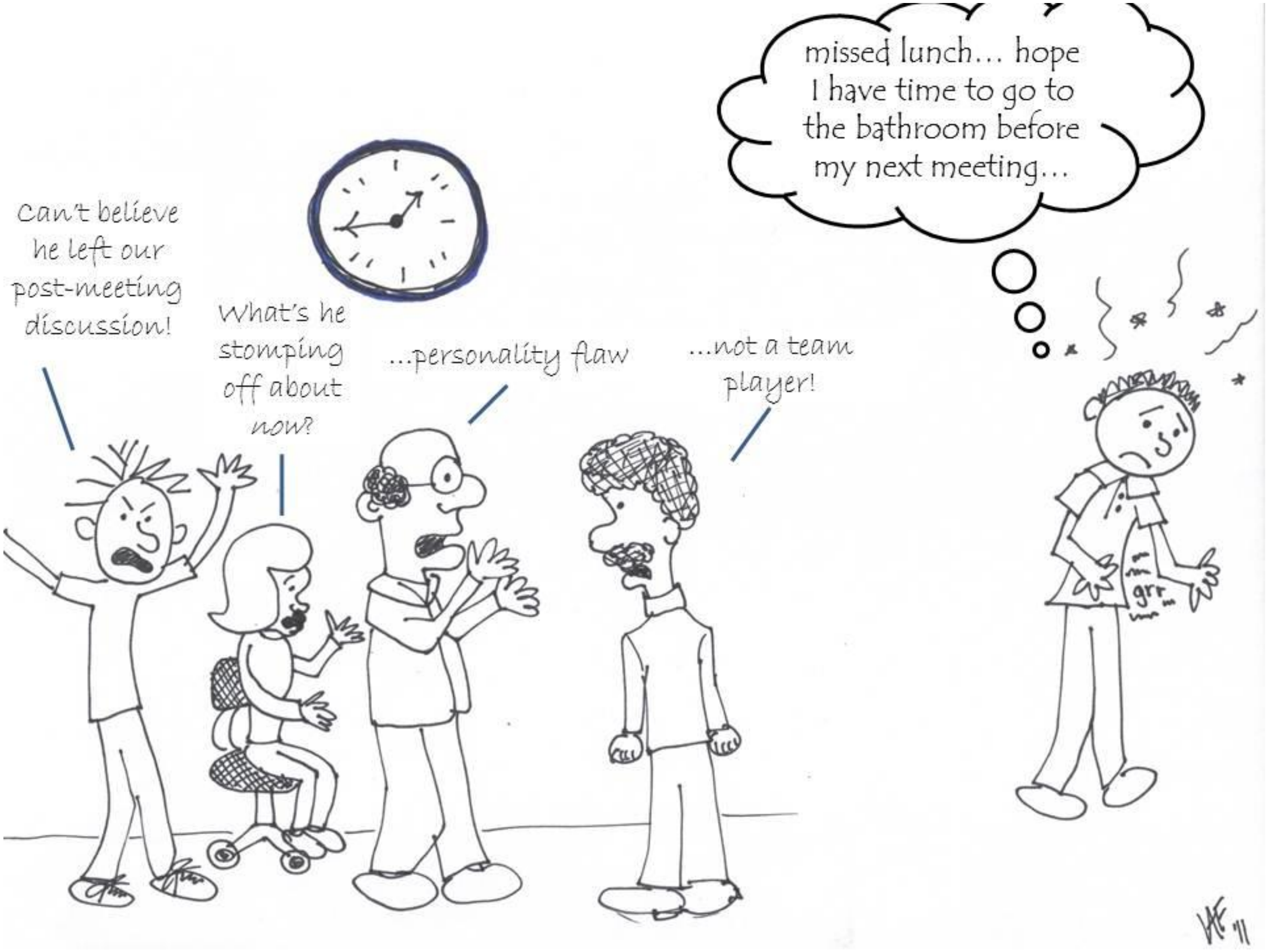
internal validity » learning | fatigue effects

- experience gained from using the first interface (to conceptual model) affects how they think about and use the second interface.

- too much testing means participants get tired of testing.

- mix it up: for some participants, A then B; for others, B then A.

internal validity » experimenter bias


Can't believe he left our post-meeting discussion!
What's he stomping off about now?
...personality flaw
...not a team player!
missed lunch... hope I have time to go to the bathroom before my next meeting...
grr

internal validity » demand characteristics

- if participants know what your hypothesis is, they will actively try to be “good participants” and help you.

ways to overcome some of these problems…

- “double-blind” experiment
  - » neither participant nor experimenter know the hypothesis

- active deception
  - » tell participants you’re expecting the opposite of what you expect

ways to overcome some of these problems…

- randomized assignment to conditions
  - » reduces systematic assignment biases

- randomized ordering of conditions
  - » normalizes the effect of order/learning/fatigue

- large sample size
  - » reduces effect of “randomness”

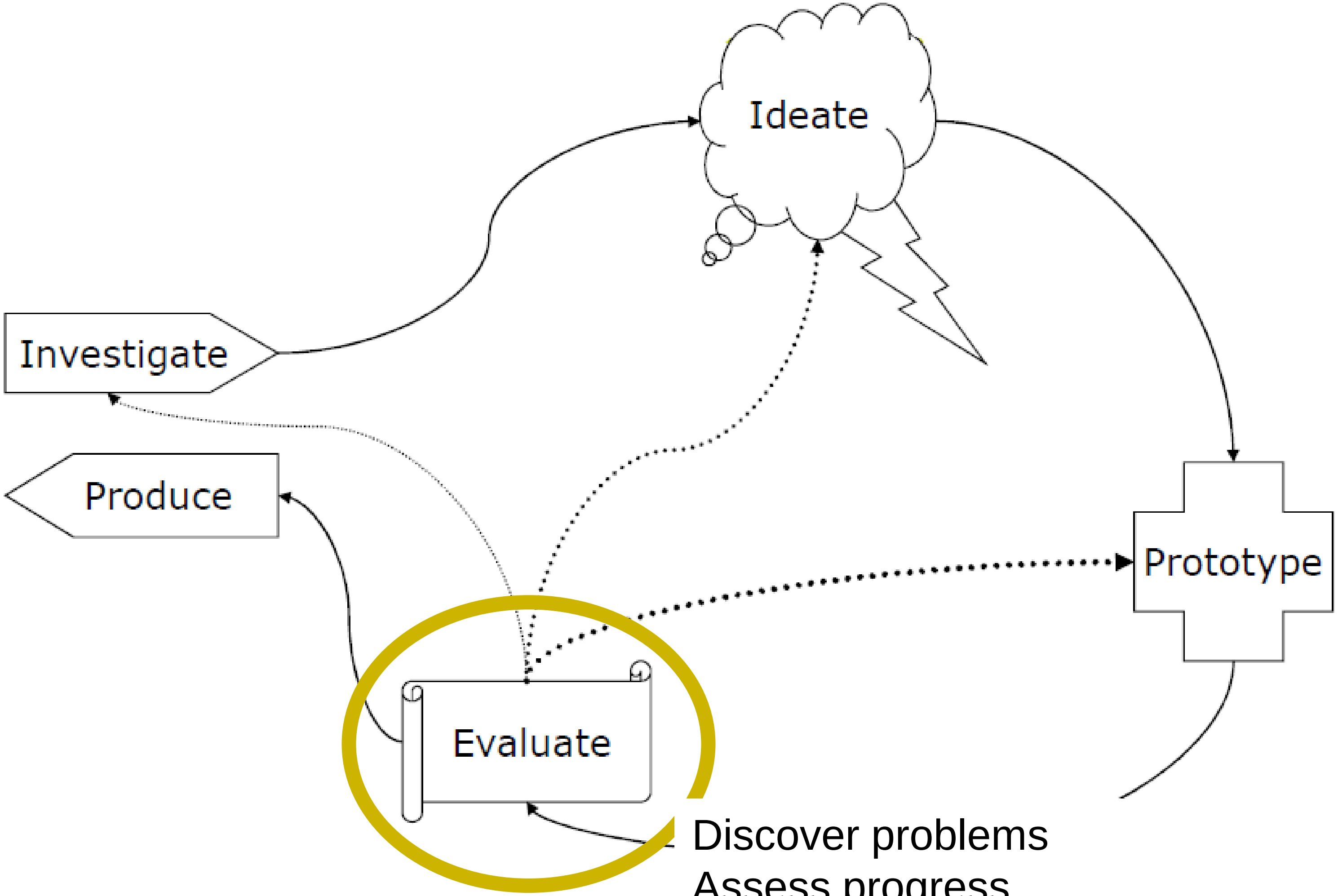
experimental validity

- external validity » realism
  - confidence that results applies to real situations
- internal validity » integrity
  - confidence in our explanation of experimental results

Summary

- usability tests can take different forms
- usability test design takes care and expertise

User Centered Design Cycle



Discover problems
Assess progress
Determine next steps