Evaluation

Joseph Barbosa Beck
Computer Science
Learning Sciences & Technology

Summary

- Evaluate early when failure is cheap
  - Paper and mockups are fine
- Easy: is this project doomed?
- Hard: does it "really work?"
- A lot of subtle issues in running (human!) studies
  - Psychologists have a lot of experience

When to evaluate?

- As early as possible
- You're going to make mistakes
  - Find out quickly so they're as cheap as possible
- Seems silly to evaluate whether software is effective when it's still being developed
  - So don't do that

Early evaluation

- Can be with a paper prototype (common in HCI)
- Can answer some useful questions
  - Can users figure out what to do? Is the interface evocative?
  - Features users wish were available?
- Sometimes called *formative evaluation* (guide development)

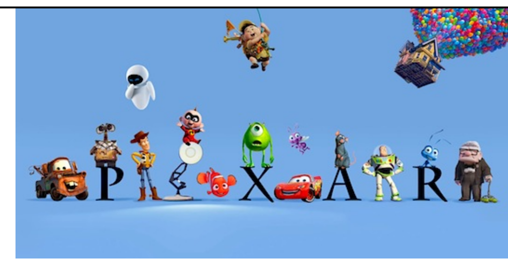
Movie making

- Why are bad movies made?
  - People (usually) don't set out to make bad movies
  - Usually obvious that a movie will be bad
  - So why does it happen?
- Director builds it and hopes the audience will like it

Movie making

- Why are bad movies made?
  - People (usually) don't set out to make bad movies
  - Usually obvious that a movie will be bad
  - So why does it happen?
- By time the movie is complete, money and time are largely gone
  - Not much you can do with editing

Does any studio tend to make movies that are not bad?



- Has generally had good success with movies
- Why?
  - One senior member of company credits prototyping approach

How Pixar makes movies

- Start with a storyboard: low fidelity (cheap!) prototype
- Successively expand the story board to flesh out scenes and give sense of artistic style
- Very quickly can tell if movie is going wrong
  - Fix it or bail
  - ***Not a lot of resources are spent on failure***

JOE: GOOGLE "BLUE SKY STUDIOS"

Get early information!

- Need to get *initial, cheap* evaluations done
  - before trying to answer questions about effectiveness (let alone ROI)
- Skipping this step is more expensive in the long run

Ok it works, but is it *effective?*

- Generally easy to determine if users understand your software and can operate it
- A bit harder: is productivity software *efficient*? (memos / minute, reports / hour)
- Much harder: does the game change user knowledge or *behavior*? (goal of serious games)

Challenges

- What would have happened without the serious game?
  - Hard to know
- (I have a great serious game for physical development. Kids start using it in 8th grade and use it through high school. On average…)

A scenario

- You have a serious game to get people to exercise more
- Give them the game to use for 6 months
- Measure how many push ups they can do
- Problems?

Problems (generated by class)

- If the game is general exercise, why evaluate on just 1 exercise?
- Did anyone actually improve? No beginning evaluation (pretest)
- Outside effects
  - Was he already exercising?
  - Could there be another reason?

Problems (generated by last year's class)

- Was the game about more than pushups?
  - Much more to fitness than that
- Need a baseline
- What about usage? (dosage effect)
- Was it really the game? Some outside factor?

Big problem

- No idea if people got more fit, less fit, or stayed the same
- General idea: giving a **pre-test** gives a much better idea of what is happening

A scenario

- Pretest: Measure how many push ups they do
- Give them the game to use for 6 months
- Posttest: Measure how many push ups they do
- Problems?

Remaining problems

Remaining problems (identified by last year's class)

- There is more to fitness than pushups
- Still don't know if it's the game

Are pushups a good measure of fitness?

- Can think of many components: upper body strength, lower body strength, aerobic endurance, flexibility…
- General idea: want your **measurement** to be a good representation of the task
  - (sounds obvious, but…)

How would we test physical fitness?

- Have control group
- Come up with some biometrics (e.g., heart rate), run through a series of tasks
- Elementary schools have physical fitness test (encompasses a few things)
- "[intelligent!] Laziness is next to Godliness"
  - Grab it. Cite it (!!!)

Spanish vocabulary test

- Si
- No
- Taco
- Espana
- Congratulations! You're an expert in Spanish
  - Gave this test to experts on the Spanish language and you got the same score they did

Concepts in testing

- Ceiling effect: if the test is too easy, everyone does well so cannot detect changes
  - Floor effect: is everyone doing poorly?
- Reliability: do you get the same score if you administer it again?
- Validity: does it measure what it claims to measure?

Where does our "pushups test" fall short?

- Ceiling effect: if the test is too easy, everyone does well so cannot detect changes
  - Floor effect: is everyone doing poorly?
- Reliability: do you get the same score if you administer it again?
- Validity: does it measure what it claims to measure?

Where does our "pushups test" fall short?

- Ceiling effect: if the test is too easy, everyone does well so cannot detect changes
  - Floor effect: is everyone doing poorly?
- Reliability: do you get the same score if you administer it again?
- **Validity**: does it measure what it claims to measure?

Validity is **hard**

- There is no statistical test to validate an measure
  - Although it should be statistically related
- Also need an expert who can confirm it really is measuring the same thing

Improve things

- Pretest: Measure wide range of physical fitness
- Let people use serious game
- Posttest: Measure wide range of physical fitness
- Problems?

Problems (identified by class)

Problems (found last year)

- Did you lock them in a room with no outside interaction? No, IRB frowns on such
- Don't know if it's the game

Is the game responsible?

- Perhaps it was national fitness month?
- Or the weather had warmed up and people became more active?
- If they were already committed to exercise, perhaps game had no influence

Fourth attempt

- Pretest: Measure wide range of physical fitness
- Let people use serious game
  - **Record how long people use the game**
- Posttest: Measure wide range of physical fitness
- **See if people who used the game more had greater increases in fitness**
- Problems?

What type of people would use the game more?

- Those who had resolved to become more fit
- If the game didn't exist, probably would have done *something*
  - Giving credit to the game is inappropriate
- General issue: highly motivated subjects are different in many ways

Random assignment

- Would like to randomly have some people use your game and some people not

Recipe for running an experimental study

- Pre-test
- Randomly assign subjects, to either
  - use your serious game,
  - or be part of the *control condition*
- Post-test

What does the control group do?

- Let them do what they want
- Get a personal trainer
- Let them invent their own exercise regimen
- Use competing software

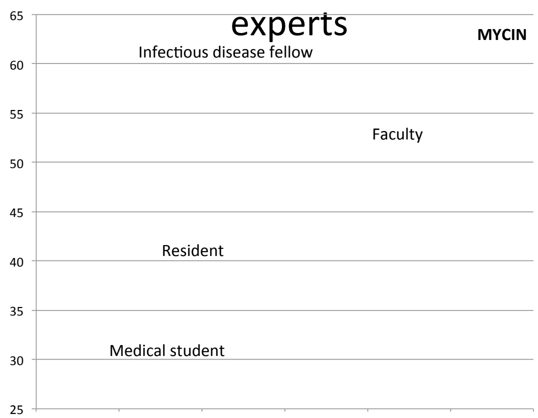
What does the control group do? – it varies!

- If there is existing training materials, using those would make sense
- For something to get you more fit, could use
  - Nothing. Is the software better than no support?
  - A membership at a local gym. Is the software better than gift certificates?
  - Two hours a week with a personal trainer. Is the software better than a very strong control?

Advanced: multiple control groups

- Imagine comparing your software to improve fitness to: nothing, a gift certificate for a gym, and a personal trainer
- Could see how well it does
  - Perhaps better than a gift certificate, less well than a personal trainer
  - More informative than "it's effective"

What multiple control groups gets you: can compare MYCIN to various experts



Big divergence from the book (last time. Same book?)

- Book does not like pre- and post-test designs
- Very common methodology
- They were attacking a straw man
  - Pretest does not have to be a simple test identical to the game

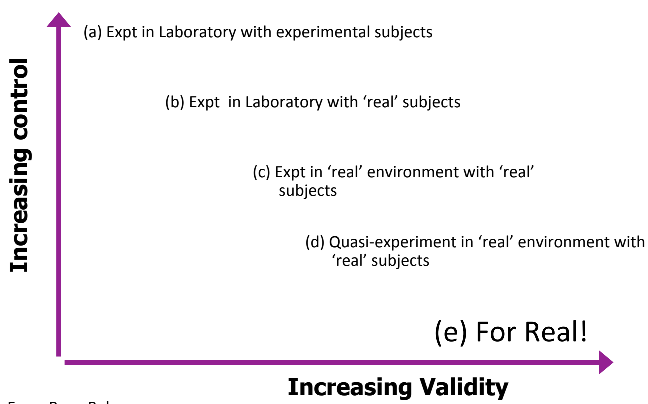
What makes a measurement good?

- Should not be similar to the environment in the game (unless the game environment is the actual context)
- Pre- and post-tests should be equally hard
- Delaying the post test to see if information is *retained* makes for a stronger result

How similar should measurement be to actual environment?

- Imagine a fitness game for professional athletes
- Should you assess them during an actual game? In a stadium? At a gym? In a lab setting?
- All examples of context

Context



From Ryan Baker

Choosing a context

- There is no "perfect" context! Real is not necessarily better.
- Pick depending on access and nature of question
  - Classrooms can be hard to gain access to
  - Precise synchronization of measures is difficult in classrooms
  - Motivation is not natural in artificial settings
    - When you tell the student "get back on-task", they will!
    - Boring systems often beat fun systems in lab studies but lose to them in real-life studies

From Ryan Baker

How many subjects?

- More is better!
- My *lower bound* is 20 per condition
  - Professor Ryan Baker prefers 30 to 60
- Too few subjects mean you can't tell what is going on (saying "I'm not sure." is always hard)

Key bits to remember

- Pre- and post-tests are good
  - Your test should correspond to what you are trying to affect
  - A delayed post-test is even better
- A *random* control group is very powerful

If you want to learn more…

- SS 2400. Methods, Modeling, and Analysis in Social Science
  - Prof. Jeanine Skorinko, SSPS
- CS 567. Empirical Methods for Human Centered Computing