Empirical Methods for Evaluating a Serious Game

Ryan Shaun Joazeiro de Baker

Some slides adapted from Shaaron Ainsworth, University of Nottingham

Why evaluate?
Why evaluate?

Because, quite honestly, many serious games suck
Why evaluate?

- For every serious game that helps improve learning, attitudes, and affect

- There are dozens that have no benefits whatsoever (see O’Neil, Wainess, & Baker, 2005 for a review of hundreds of serious games)

Evaluation...

- Can be conducted **formatively**, during development, to find out what’s not working right, and help you make a serious game that is genuinely beneficial

- Can be conducted **summatively**, after development, to prove (hopefully) that your serious game is genuinely beneficial
Empirical methods

- All center around one thing

- What happens when learners use your system?

- There’s a lot of details beyond that, of course...

Empirical methods

- What kind of study should I run?
- What data should I collect?
- What should I do with that data?
What kind of study should I run?

- Many potential types of studies

Types of studies

- Experiments
- Quasi-experiments
- Single-condition study
- Design experiments/design research
- Case studies
- Ethnography
- Instructor-led role playing
"Fixed Method" studies

- Experiments
- Quasi-experiments
- Single-condition study
- Design experiments/design research
- Case studies
- Ethnography
- Instructor-led role playing

More conclusive, more objective
- There’s a place for all of these methods, but the evidence from fixed method studies is considered more conclusive

A few specific words on

- Instructor-led role playing

- The authors claim that this is a really good method

- I am *not* a fan

- This method is very easy to screw up, and very expensive to conduct
  - Testing method and assessment are both very difficult to standardize
  - Can you react the same way every time?
  - Can you assess the same way in real-time, without taking personality, likability, etc. into account

- Only military labs/contractors with huge budgets use this method
Single-Condition Study

- Have students use your serious game
- Measure learning, affect, liking, etc. as I will discuss later

- Good for formative studies; can show you what is not working out and needs improvement

- Not good for summative studies; even if your evaluation shows “good learning”, it might be worse than every other possible approach

Experiments

- State a causal hypothesis
- Assign subjects randomly to groups
  - Typically: Experimental group is your game, Control group is some valid comparison curriculum teaching the same thing
- Manipulate independent variable(s)
- Use systematic procedures to test hypothesized causal relationships
- Use specific controls to ensure validity
Choosing a comparison curriculum

- Should be something that teaches the same material as your serious game
- Should be something that is not obviously terrible
- For instance, Cordova & Lepper (1996) compared their game to a version of the same game called “Math Game”, which removed all fantasy and choice from the game
- Possible comparison curricula: Intelligent Tutor, Computer-Aided Instruction, Classroom Activity run by teacher

Quasi–Experimental Methods

- State a causal hypothesis
- Include at least 2 levels of the independent variable
- Subjects are not assigned randomly to groups
- Use specific procedures for testing hypotheses
- Use some controls to ensure validity
- Less conclusive results than a true experiment
- But more feasible in “natural” settings such as schools, where subjects are already in natural, non-randomly-selected groups
Things to beware

- **Selection bias** - individuals assigned to conditions happen to be different at the outset of the experiment in ways that might erroneously be attributed to the treatment.
  - This is more of a problem for quasi-experiments
  - But you still need to check for it in experiments

- **Attrition** – People drop out, or drop out differently in the two conditions

---

Things to beware

- **Resentful Demoralization** – Participants in the control condition see that they get a less fun system, become angry, and work less hard (a good reason not to give a game at the same time and in the same room as some more boring curriculum)

- **Compensatory Rivalry** (“John Henry effect”) – Participants in the control condition decide that they want to beat the participants in the experimental condition
Things to beware

- Hawthorne Effect – the sheer novelty of the serious game makes it seem better to the subjects (over a brief period of time) than it really is

Expectancy Effects

- Experimenter effects
  - Expectancy effects during intervention
    - E.g. Inadvertently supporting students in your “preferred” condition
  - Expectancy effects on analysis
    - E.g. throwing away outliers inappropriately

![Expectancy Effects Graph](image)
Issues

Validity
- Construct validity
  - Are you measuring what you’re supposed to?
- External validity
  - Is your sample valid for the broader population?
- Ecological validity
  - Is your study representative of the actual context of use?

Reliability
- Would the same test produce the same results if
  - Tested by someone else?
  - Tested in a different context?
  - Tested at a different time?

Data collection in experiments, quasi-experiments, single-cond studies
Pre-test and post-test of knowledge/skill

- The VERY VERY VERY most important thing to collect
  - Note that I disagree with the authors of your text on this

- What did the learner know before using the system?
- What did the learner know afterwards?

Good tests...

- Measure the knowledge that the student should have gained by using the system, in a concrete fashion

- Measuring factual knowledge is good
Better tests...

- Measure the knowledge that the student should have gained by using the system, in a concrete fashion
- Measuring factual knowledge is good
- Measuring ability to apply it is better

Even Better tests...

- Measure the knowledge that the student should have gained by using the system, in a concrete fashion
- Measuring factual knowledge is good
- Measuring ability to apply it is better
- Measuring ability to apply it 2 weeks later is even better
  - Is learning maintained?
Even Still Better tests...

- Measure the knowledge that the student should have gained by using the system, in a concrete fashion
- Measuring factual knowledge is good
- Measuring ability to apply it is better
- Measuring ability to apply it 2 weeks later is even better
- Measuring transfer of knowledge to new contexts is even still better
  - If the learner can only apply the newly learned mathematics in the exact type of problems presented in the system, will a difference be made?

Of course...

- Better tests are harder to conduct and more work
Of course...

- Better tests are harder to conduct and more work
- The authors praise “Advanced Simulation Assessments”
- Note their example, though
- “The Army, for example, runs trainees through a series of computer simulation war games and then gets the trainees all suited up and has them carry out an assessment in real space with real gear. Several large military bases have built cutaway towns, and troops simulate close combat training in these mockup urban environments.”

Things not to mess up
Things not to mess up

- DON’T use the same items for both pre-test and post-test for any given student

“Gee, this looks familiar...”

Things not to mess up

- DON’T use the same items for both pre-test and post-test for any given student
- But DO counterbalance items between the pre-test and post-test to avoid having unequal difficulty

Pre-Test: What is the capital of Madagascar?
Post-Test: What is the capital of France?
Things not to mess up

★ DON’T use the same items for both pre-test and post-test for any given student
★ But DO counterbalance items between the pre-test and post-test to avoid having unequal difficulty
★ DON’T use items from your learning environment in your post-test

“Gee, this looks familiar…”

Things not to mess up

★ DON’T use the same items for both pre-test and post-test for any given student
★ But DO counterbalance items between the pre-test and post-test to avoid having unequal difficulty
★ DON’T use items from your learning environment in your post-test
★ DON’T let students “help” each other during the tests

It will happen
Things not to mess up

DON’T use the same items for both pre-test and post-test for any given student
But DO counterbalance items between the pre-test and post-test to avoid having unequal difficulty
DON’T use items from your learning environment in your post-test
DON’T let students “help” each other during the tests
DON’T let the teacher give any student the answer during the test

I’ve seen it happen

Things not to mess up

DON’T use the same items for both pre-test and post-test for any given student
But DO counterbalance items between the pre-test and post-test to avoid having unequal difficulty
DON’T use items from your learning environment in your post-test
DON’T let students “help” each other during the tests
DON’T let the teacher give any student the answer during the test
DON’T let the students steal your pencils

Or just buy extra pencils
Learning Gain Measures

- t-test on Post-test – Pre-test for each group
- t-test on \( \frac{(Post-test - Pre-test)}{Pre-test} \) for each group
  - accounts for high performers...
  - but has weird effects if anyone does worse on post-test than pre-test
- Regression set up as
  \[
  Post-test = \alpha_0 \text{ Pre-test} + \alpha_1 \text{ Condition} + \alpha_2
  \]
  - allows you to find mean difference in conditions while controlling for each student’s pre-test score
  - but you need to check that condition differences are not actually pre-test differences between conditions using
    \[
    Pre-test = \alpha_0 \text{ Condition} + \alpha_1
    \]
- Effect Size: \( \frac{(Mean \text{ Gain in Experimental} - \text{ Mean Gain in Control})}{\text{St Dev in Control}} \)
  - statistical significance is important but it isn’t everything

Details

- In some contexts, a pre-test is not appropriate
  - Domains where the learners are all expected to be at floor (i.e. zero knowledge) on the pre-test
  - YodelLand, SwordfightingSim, iCricket (in USA)
  - But are you sure none of your participants know anything?
Other Important Measures (Dependent Variables)

**Learning efficiency**

- Even if both conditions produce equivalent learning outcomes, does your system reduce time spent learning?
- Usually not the case in serious games – serious games typically are less time-efficient than other forms of instruction
  - They have other benefits – improved engagement, better match to target situation leading to better transfer, etc.

Other Important Measures (Dependent Variables)

**User’s attitudes**

- Easy to do wrong
- “Do you like my wonderful system?”
- Of course they do... (they probably don’t want to disappoint you)
Other Important Measures (Dependent Variables)

- A better way to do it (one of many)

- Pre-test
  “How much do you like learning mathematics?”

- Post-test
  “How much do you like learning mathematics with Zombie Division?”

Other Important Measures (Dependent Variables)

- Recommended by text

  “Would you recommend this game to a friend?”
Other Important Measures (Dependent Variables)

- Be careful
  - Just because students liked a system does not imply they learned better from it
  - But still a useful measure for a serious game
  - If they don’t like it better than the comparison curriculum, you’ve really messed up

Other Important Measures

- Data on usage, affect, student choices
  - Mining system log files
  - Time on task, off-task behavior
  - Progression through curriculum
  - Use of system features
  - Question Performance (right, wrong, number of attempts..)
  - Amount of help sought or provided

- Corresponds to “Advanced Simulation Assessments within the Game”
  - Major limitation – may not assess transfer of skill outside of the assessment (a problem if students learn to “game the system”)

Other Important Measures

Data on usage, affect, student choices
- Quantitative field observations – rigorized observations of the emotion/affect a student displays while playing the game (cf. Baker et al., 2004)
  - Were the students bored and frustrated less often when playing the game?

Data on usage, affect, student choices
- Video logs – similar to quantitative field observations
- More time-consuming to analyze
- Easier to re-check, therefore more trustworthy
Other Important Measures

- Data on usage, affect, student choices
  - FMRI?
Context

(a) Expt in Laboratory with experimental subjects
(b) Expt in Laboratory with ‘real’ subjects
(c) Expt in ‘real’ environment with ‘real’ subjects
(d) Quasi-experiment in ‘real’ environment with ‘real’ subjects
(e) For Real!

Increasing control

Increasing Validity

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
Beware of...

Evaluating on an inappropriate population
- Are kids in the one Worcester school you found typical of American students in general?
- Are WPI engineering PhD students typical of American university students?

Beware of...

Inappropriate Generalisations
- Your results may depend on prior knowledge, gender, attitudes, teacher, small-scale details of your study (40 minutes or 45?), experimenter presence (students behave differently the first day they see an experimenter than later), etc etc etc
Beware of…

Too few subjects
- Typically, I try to have 30-60 students per condition, a sample size that gives good statistical power (ability to demonstrate that results are present) while being feasible to conduct

Conclusions
Evaluation of your serious game

- There are a lot of ways to do it
- Benefits and drawbacks to each
- But absolutely essential
  - For improving your serious game
  - For validating its effectiveness

To learn more...

- Take
  - SS 2400. METHODS, MODELING, AND ANALYSIS IN SOCIAL SCIENCE
    - Prof. Jeanine Skorinko, SSPS
  - PSY 503. RESEARCH METHODS FOR THE LEARNING SCIENCES
    - Prof. Ryan Baker, SSPS
- Both offered in Fall 2011
The End

❖ Now get out there and evaluate!