Evaluating Assessments of Novice Programming Environments
Paul Gross and Kris Powers
Tufts University

How did we get into this?
- Wanted to know: “Do novice programming environments really help students learn? And, if so, how?”
- Literature/web search uncovered plethora of environments…
  - ~40 new tools in the last 5 years
  - …but assessments of their impact sparse, disjoint
- No consistent questions or methods
- Little recognition of other assessments

What can we learn from them?
- Focus on empirical assessments
  - Hard, objective data
  - Repeatable and generalizable
  - Multiplicity of approaches complicates comparison
    - Different observation and analysis methods, data observed, questions, etc.
  - Need evaluation tool!

Evaluating Assessments
- Objective variable coding
  - E.g., population, study duration, questions asked, conclusions, etc.
  - Allows comparison of study methods and pragmatics
- Subjective evaluation
  - Critique of study design and reporting
  - Fairness managed by rubric of 8 questions
  - Adapted from Long & Godfrey (2004)

Evaluation Conclusions
- Questions asked are often too vague
- Studies often only conducted by developer or those closely associated
- Approaches tend towards outcome-based rather than process-based
- Data collected is naturally occurring, rarely explicitly intended for assessment study
- Observation instruments used are not validated
- Reporting of practices incomplete

Our Assessment Evaluations
- Evaluation of 5 environment assessments
  - Alice, BlueJ, Jeliot 2000, Lego Mindstorms with Ada, RAPTOR
    - Represent a cross-section of environment types
    - Variety of approaches to assessment
- Evaluated using
  - Objective variable coding
  - Rubric of 8 questions
1. How appropriate is the question asked and is the question of reasonable scope?
   - Example (Alice; Moskal et al., 2004)
     - Does exposure to the Alice course improve student performance in CS1?
   - Evaluation
     - Appropriate as Alice course expected to prepare students for CS1
     - Reasonable as question addresses very specific, measurable effect

2. What theoretical framework guides or informs the study and how is it reflected in the methodology?
   - Example (Jeliot 2000; Levy et al., 2003)
     - Authors cite previous results showing that animation’s impact is more noticeable in labs rather than exams
   - Evaluation
     - Study incorporates previous results by deliberately integrating Jeliot 2000 into lab assignments

3. Is the reporting of the observation and analysis methods adequate?
   - Example (BlueJ; Ragonis & Ben-Ari, 2005)
     - Investigated teaching objects-first approach to young novices, BlueJ chosen tool
     - Analyzed audio/video recordings and student artifacts to identify “difficulties” with program flow
   - Evaluation
     - Inadequate reporting of the analysis methods
     - “Difficulties” are said to occur “frequently” with no discussion about how difficulties were recognized or what might constitute frequent occurrence

4. Are the observation and analysis methods valid and appropriate to the question?
   - Example (Lego Mindstorms with Ada; Fagin & Merkle, 2002)
     - Asked what is the impact of using robotics on student exam performance?
     - Observed midterm and final exam scores
     - Analyzed scores for statistically significant difference between robotics and control sections
   - Evaluation
     - Both valid and adequate as question is simple and straightforward to satisfy

5. Do the authors outline potential sources of bias?
   - Example (RAPTOR; Carlisle et al., 2005)
     - Treatment group performed worse than control group on exam question for one semester
   - Evaluation
     - No, sources of bias not adequately addressed
     - Performance result attributed to difficult lab
     - No discussion about other possible factors including lack of grading standardization, instructor bias, or other variables between courses and semesters

6. To what degree is the study generalizable and repeatable?
   - Example (Alice; Moskal et al., 2004)
     - Study determines “at risk” students, intervenes with Alice course, measures CS1 grades, retention, and attitudes
   - Evaluation
     - Easily generalizable as observation and analysis methods are not explicitly dependent on Alice
     - Mostly repeatable as most of the methods are discussed (not “at risk” measure and focus group methods) and materials are available
Evaluative Rubric (7/8)

7. Is there a coherent chain of reasoning from the analysis results to the assessment conclusions?
   - Example (Jeliot 2000; Levy et al., 2003)
     - Concludes animation students used a different and better vocabulary describing solutions in interview questions than control students
   - Evaluation
     - Not particularly strong
     - Need to clarify interview methodology and criteria for how the solution descriptions were classified and evaluated

Evaluative Rubric (8/8)

8. Do the conclusions answer the original questions?
   - Example (Lego Mindstorms with Ada; Fagin & Merkle, 2002)
     - Ask what is the effect of using robotics in CS1 course on exam performance?
     - Concludes that robotics had negative effect on exam performance
   - Evaluation
     - Yes and to a great degree as they account for other factors that could bias exam performance data

Future Work

- Refine questions asked in assessments
  - Consider individual features
  - Ask how and why impact occurs
- Develop validated instruments
- Multi-institutional studies of a single environment

Thank you!

- Questions?