Contentious Claim?

AIED systems die, the only thing you can hand on to the next generation is information about the success (or lack of) of a current system

Without evaluation, there is no point in doing anything.....
Questions to answer

- What do I want to do with the information
  - Informing design
  - Choosing between alternatives
  - Credit assignment problem
  - Informing about context of use
- What are appropriate forms of measurement?
- What is an appropriate design?
- What is an appropriate form of comparison?
- What is an appropriate context

Two main types

- To inform design
  - Formative evaluation
  - E.g. Heuristic Evaluation, Cognitive Walkthrough
  - http://www.psychology.nottingham.ac.uk/staff/see/clicker/handout.pdf
  - Should the same usability heuristics be used in educational systems as are used in other computer-based systems
- To assess end product
  - To assess end product or discover how it should be used
  - Summative evaluation
  - E.g. Experimental, Quasi-experimental

Questions to answer

- What do I want to do with the information
  - Informing design
  - Choosing between alternatives
  - Credit assignment problem
  - Informing about context of use
- What are appropriate forms of measurement?
- What is an appropriate design?
- What is an appropriate form of comparison?
- What is an appropriate context

Common Measures (Dependent Variables)

- Learning gains
  - Post-test – Pre-test
    - (Post-test – Pre-test)/Pre-test: to account for high performers
- Learning efficiency
  - IE does it reduce time spent learning
- How the system is used in practice (and by whom)
  - ILEs can’t help if learners don’t use them!
  - What features are used
- User’s attitudes
  - Beware happy sheets
- Cost savings
- Teachbacks
  - How well can learners now teach what they have learnt

Learning Gains: Effect Size

(Gain in Experimental – Gain in Control)/ St Dev in Control

<table>
<thead>
<tr>
<th>Comparison</th>
<th>Ratio</th>
<th>Effect</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classroom teaching v Expert Tutoring</td>
<td>1:30 v 1:1</td>
<td>2 sd</td>
</tr>
<tr>
<td>Classroom teaching v Non Expert Tutoring</td>
<td>1:30 v 1:1</td>
<td>0.4 sd</td>
</tr>
<tr>
<td>Classroom teaching v Computer Tutoring</td>
<td>1:30 v C:1</td>
<td>?</td>
</tr>
</tbody>
</table>

A 2 sigma effects means that 98% of students receiving expert tutoring are likely to do better than students receiving classroom instruction

Interaction Data

- Time on task
- Progression through curriculum
- Use of system features (e.g. glossary, notepad, model answers)
- Question Performance (right, wrong, number of attempts..)
- Amount of help sought or provided
DEMIST (Van Labeke & Ainsworth, 2002)

Users' Traces

Unit 1 – 09:34

- Controller
- Value: N
- Chart: N
- Table: N
- New Terms
- Dynamic Equations
- Graph: N v T (log scale)
- Graph: N v (dN/dT)

- Controller
- Action & Hypothesis
- Map Relation
- Experimental Set

Process Data

- Protocols
- Dialogue turns
- Gesture and Non-verbal behaviour
- Eye movement data
- Poor men's eye tracker (e.g. Conatt & Van-Lehn, Romero, Cox & Du Boula)

Galapagos (Luckin et al, 2001)

Introduction

- About Islands
- South
- North
- Map

- Model
- Model

DV Summary

- Rarely the case that a single DV will be sufficient
- Could look for more innovative outcome measures (e.g. learn with complex simulation but then multi-choice post-test)
- Beware the Law of Gross Measures
  - Subtle questions require subtle DVs which may be impossible in many situations
  - Interaction data often got for free and it's a crime not to look at it! Process data hard work but often worth it.
- Capturing interaction data rarely changes learners' experiences, but capturing process data often does.

Questions to answer

- What do I want to do with the information
  - Informing design
  - Choosing between alternatives
  - Informing about context of use
- What are appropriate forms of measurement?
- What is an appropriate design?
- What is an appropriate form of comparison?
- What is an appropriate context

Two Types of Experimental Design

- Experimental
  - State a causal hypothesis
  - Manipulate independent variable
  - Assign subjects randomly to groups
  - Use systematic procedures to test hypothesised causal relationships
  - Use specific controls to ensure validity

- Quasi – experimental
  - State a causal hypothesis
  - Include at least 2 levels of the independent variable
  - We may not be able to manipulate it
  - Cannot assign subjects randomly to groups
  - Use specific procedures for testing hypotheses
  - Use some controls to ensure validity
Potential Biases in Design

- **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
- **Subject biases**
  - Hawthorne effect
  - A distortion of research results caused by the response of subjects to the special attention they receive from researchers

Choosing Between Designs

- **Validity**
  - Construct validity
    - Is it measuring what it’s supposed to?
  - External validity
    - Is it valid for this population?
  - Ecological validity
    - Is it representative of the context?

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

Prototypical designs

- (intervention) post-test
- Pre – (intervention) - post-test
- Pre – (intervention) – post-test – delayed post-test
- Interrupted time-series
- Cross-over

Post-test

- Advantages
  - Quick
- Disadvantages
  - A lot!
  - Need random allocation to conditions
  - Can’t account for influence of prior knowledge on performance or system use

Pre-test to Post-test
**Pre-test to Post-test**

- **Advantages**
  - Better than just measuring post-test as can help explain why some sorts of learners improve more than others.
  - Can show whether prior knowledge is related to how system is used.
  - If marked prior to study can be used to allocate subjects to groups such that each group has a similar distribution of scores.

- **Disadvantages**
  - No long term results.
  - Can not tell when improvement occurred if long term intervention.

**Pre-test to Post-test to Delayed Post-test**

- **Advantages**
  - Does improvement maintain?
  - Some results may only manifest sometime after intervention (e.g. Metacognitive training).
  - Different interventions may have different results at post-test and delayed post-test (e.g. individual and collaborative learning).

- **Disadvantages**
  - Practical.
  - Often find an across the board gentle drop off.

**Interrupted Time-Series Design**

- **Advantages**
  - Time scale of learning.
  - Ceiling effects.

- **Disadvantages**
  - Time-consuming.
  - Effects of repeated testing.

**Full Cross-over**

Gain A  Gain B
**Full Cross-over**

- **Advantages**
  - Controls for the (often huge) differences between subjects
  - Each subject is their own control
  - May reveal order effects
- **Disadvantages**
  - Four groups of subjects rather than two!
  - Statistically complex – predicting at least a 3 way interaction
  - Never come across one yet in AIED!

**Partial Cross-over**

- **Pre-test A**
- **Post-test A**
- **Pre-test B**
- **Post-test B**

**Partial Cross-over**

- Same as full cross over but
  - **Advantages**
    - less complex and subject hungry
  - **Disadvantages**
    - less revealing of order effects

**Some Common Problems**

**Questions to answer**

- What do I want to do with the information
  - Informing design
  - Choosing between alternatives
  - Informing context of use
- What are appropriate forms of measurement?
- What is an appropriate design?
- What is an appropriate form of comparison?
- What is an appropriate context

**Nature of Comparison**

- ILE alone
- ILE v non-interventional control
- ILE v Classroom
- ILE\(_{a}(b)\) v ILE\(_{b}(a)\) (within system)
- ILE v Ablated ILE
- Mixed models
ILE alone

- Examples
  - Smithtown — Shute & Glaser (1990)
  - Cox & Brna (1995) SWITCHER
  - Van Labeke & Ainsworth (2002) DEMIST

- Uses
  - Does something about the learner or the system predict learning outcomes?
    - E.g. Do learners with high or low prior knowledge benefit more?
    - E.g. Does making help messages lead to better performance?

- Disadvantages
  - No comparative data – is this is good way of teaching??
  - Identifying key variables to measure

Smitownt - Shute & Glaser (1990)

- Guided discovery environment to scientific enquiry skills and principles of basic economics
  - Notebook, grapher, hypothesis maker
  - Explorations & experiments
  - Issue-based tutoring to detect and remediate scientific method
  - Students who did well with Smithtown (n = 530) engaged in goal or hypothesis driven activity.

SwitchER – Cox & Brna (1995)

- Solving constraint satisfaction problems by constructing representations.
  - N = 16
  - Learners tended to switch between representations, particularly at impasses
  - Idiosyncratic representations associated with poorer performance
  - (Performance on system in this case is the learning measure)


- Learners (N = 20) using a multi-representational simulation to learning population biology
- Free Discovery with minimal exercises

  - No significant relationship between use of representations and
    - Pre-test scores, Post-test scores, Prior experience with maths/biology
    - Stated preference as to visualisation/verbal
  - Conclusion: Inappropriate method as can’t answer "WHY"
  - What does spending a lot of time with a representation mean?
  - Need for protocols

ILE v non-interventional control

- Examples

- Uses
  - Is this a better way of teaching something than not teaching it at all?
  - Rules out improvement due to repeated testing

- Disadvantages
  - Often a no-brainer!
  - Does not answer what features of the system lead to learning
  - Ethical?


- Can children learn to give multiple solutions to the same question (Simplified Design)
- 20 eight to 9 yr olds
COPPERS Results

![Graph showing COPPERS Results](image)

- Children don't get better at this just because they are asked to do it repeatedly.
- A simple intervention can dramatically improve performance.

ILE v Classroom

- **Examples**
  - LISPITS (Anderson & Corbett)
  - Smithtown (Shute & Glaser, 1990)
  - Sherlock (Lesgold et al, 1993)
  - PAT (Koedinger et al, 1997)
  - ISIS (Meyer et al, 1999)

- **Uses**
  - Proof of concept
  - Real-world validity

- **Disadvantages**
  - Classrooms and ILEs differ in some ways, what can we truly conclude?

LISPITS Anderson

- **Classic Model and Knowledge tracing tutor: the ITS!**
- **Novices with LISPITS or conventional teaching or just textbook (N = 30)**
  - Learning Outcomes: All groups did equivalently well on post test, but some subjects on own not complete test.
  - Learning Efficiency: LISPITS (11.4 hrs): Teacher (15 hours), Textbook (26.5 hours)
- More experienced beginners on Lisp course: exercises vs. LISPITS (N = 20)
  - Learning Outcomes LISPITS group did 43% better on post-test
  - Learning Efficiency: LISPITS group finished 30% faster

SHERLOCK — Lesgold et al (1992)

- **Intelligent training system**
  - Airforce technicians
  - Complex piece of electronics test gear
- **Model of student under instruction — adjust level of and specificity of feedback**
- **Comparisons with conventional training**
  - Air force evaluation — 20-25 hours on SHERLOCK similar 4 years job experience
  - Pre/post comparison over 12 days (N = 64)
    - Learning outcomes: experimental group solved significantly more problems in post test
    - quality of problem-solving judged more expert

Smithtown V Class Teaching

- **Comparison with class teaching (n = 30)**
  - Learning Outcomes: Did as well as conventionally taught student
  - Learning Efficiency: Finished in about half the time (Shrs vs. 11hrs)

Evaluation of SHERLOCK

- **Comparisons with conventional training**
  - Airforce evaluation — 20-25 hours on SHERLOCK similar 4 years job experience
  - Pre/post comparison over 12 days (N = 64)
    - experimental group solved significantly more problems in post test
    - quality of problem-solving more expert

---

8

- Cognitive Tutor with Model & Knowledge tracing
  - Practical Algebra System
  - Pittsburgh Urban Mathematics Project
- Detailed model of student under instruction
  - Extensive prior analysis of learning algebra

<table>
<thead>
<tr>
<th>Task</th>
<th>Control Group Mean (SE)</th>
<th>PAT Group Mean (SE)</th>
<th>P value significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>SSS (SSS)</td>
<td>31.14 (15)</td>
<td>34.16 (15)</td>
<td>F(1, 287) = 17.6</td>
</tr>
<tr>
<td>Mult-SAT Subset</td>
<td>64</td>
<td>127</td>
<td>0.0001</td>
</tr>
<tr>
<td>Problem Situation Test</td>
<td>3.21 (2.01)</td>
<td>3.27 (2.01)</td>
<td>0.01</td>
</tr>
<tr>
<td>Representations Test</td>
<td>1.53 (1.61)</td>
<td>1.57 (1.62)</td>
<td>F(1, 287) = 13.4</td>
</tr>
</tbody>
</table>

ISIS Meyer et al (1999)

- Simulation-based tutor for scientific enquiry skills
- Generating hypotheses, designing and conducting experiments, drawing conclusions, accepting/rejecting hypotheses
- Quasi-expt. 3 studies: N = 1553, N = 1594, N = 488
- Learning Outcomes: ISIS generally better than classroom
  - The further through the ISIS curriculum the greater the learning gains
  - Effective time on task? ability?
- Mistakes
  - Too many subjects!
  - Not sophisticated enough analyses – huge wasted opportunity

ILE(a) v ILE(b) (within system)

- Examples
  - Galapagos — Lucken et al (2001)
- Uses
  - Much tauter design, e.g. nullifies Hawthorne effect
  - Identifies what key components add to learning
  - Aptitude by treatment interactions
- Disadvantages
  - Identifying key features to vary – could be very time consuming!


- Guided practice environment to teach 10-12 yr old children the role of number sense in estimation
- Issue explored – what format of representation best supports learning

Which do you think will be best?

- Pictures
- Maths
- Mixed


- Another CMU cognitive tutor - Geometry
- Two versions – a Self-Explanation v Answer only
- Expt 1 (N = 23) – Significantly greater gains for SE group
- Expt 2 (N = 43) – Overall suspect non significant interaction! But SE students doing better on harder problems.
MEN0 – Luckin et al (2001)

- To investigate the role of narrative in the comprehension of educational interactive media programmes (e.g. Galapagos)
- Principles of Darwin’s theory of natural selection.
- Task: use the notepad to construct an explanation of the variations in the wildlife on the islands.
- Three versions: same content different structure

‘Galapagos’: three version

<table>
<thead>
<tr>
<th>NARRATIVE GUIDANCE</th>
<th>SUPPORT FOR NARRATIVE CONSTRUCTION</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear</td>
<td>• recognisable linear structure</td>
</tr>
<tr>
<td></td>
<td>• easy navigation</td>
</tr>
<tr>
<td></td>
<td>• limited interaction</td>
</tr>
<tr>
<td></td>
<td>• implicit guidance in interface design</td>
</tr>
<tr>
<td></td>
<td>(e.g. order of items)</td>
</tr>
<tr>
<td>Notepad</td>
<td>• notepad</td>
</tr>
<tr>
<td>Model Answer</td>
<td>• model answer</td>
</tr>
<tr>
<td>RBL</td>
<td>• no explicit narrative guidance</td>
</tr>
<tr>
<td></td>
<td>• implicit guidance in interface design</td>
</tr>
<tr>
<td>GDLD</td>
<td>• easily accessible statement of task</td>
</tr>
</tbody>
</table>

Findings

- Twice as much CONTENT as NON-TASK or TASK talk.
- Contentful discussions do not happen while learners are looking solely at the content related sections of the CD-ROM
  - Linear users conducted more CONTENT talk whilst using the notepad whilst viewing the content sections of the CD-ROM, whilst RBL and GDLD learners conducted much more CONTENT talk with the content sections of the CD-ROM themselves.
- The notepad prompts discussion about the practicalities of answer construction

Galapagos Conclusions

- Simple interface design elicited a much higher ratio of on-task to procedural discussion than commercial interfaces;
- Goal, Reminders, Notepad, Model Answer, and Guide Features were all effective, as evidenced by the use all groups made of them, and the high proportion of on-task talk they elicited;
- Model Answer & Notepad prompted learners to discuss answer construction, content features alone did not;
- Learners were much more likely to refer back to other sections as they constructed their answers within the learner-controlled resource-based and guided discovery versions, and therefore tended to use quotes from the material in their notes, which linear users did not.

ILE v Ablated ILE

- Ablation experiments remove particular design features and performance of the systems compared.
- Examples
  - VCR Tutor – Mark & Greer (1995)
  - Luckin & du Boulay (1999)
- Uses
  - What is the added benefit of AI
- Disadvantages
  - System may not be modular
Animal Watch – Arroyo et al

- ITS for teaching arithmetic in the context of biology
- Hint Symbolism (symbolic v concrete) & Hint Interactivity (learning by doing v learning by being told)
- Attitude by treatment exploration Cognitive Development & Gender (n = 60)
- Some results
  - Girls do better with interactive hints
  - High cognitive levels better with symbolic & interactive hints


VCR Tutor — Mark & Greer

- Intelligent tutoring system to teach operation of (simulated) Video Tape Recorder
- Four versions: ‘Dumb’ to ‘Clever’
  - Conceptual as well as procedural feedback
  - Model-tracing to allow flexibility of problem solution
  - Recognize and tutor certain misconceptions
- Compare pre/post test (N = 76)
- Increasing intelligence produced in post-test
  - Solutions with fewer steps
  - Solutions with fewer errors
  - Faster performance


StatLady — Shute (1995)

- Tutoring system for elementary statistics
- Unintelligent version
  - Same curriculum for all learners
  - Fixed thresholds for progress
  - Fixed regime of feedback messages on errors
- Intelligent version
  - More detailed knowledge representation
  - Individualized sequence of problems
  - More focused feedback and remediation
- Unintelligent version produced learning outcomes as good as experienced lecturer (N = 103)
- Learning outcomes greater with intelligent version produced but lesser learning efficiency (N = 100)


Evaluation of StatLady

- Unintelligent version produced pre/post tests differences as good as experienced lecturer (N = 103)
- Intelligent version produced better pre/post test differences than unintelligent version, but with longer time on task (N = 100)


Dial-A-Plant – Lester et al.

- Botanical anatomy
- Pedagogical agent - Herman the Bug
- Advice response types
  - Muted
  - Task-Specific Verbal (concrete)
  - Principle-Based verbal (abstract)
  - Principle-based Animated / Verbal
  - Fully Expressive


Evaluation of Dial-A-Plant

- Reduced errors on complex problems
  - Fully expressive agent did best
  - Task specific verbal did next best
  - Benefit of agent increases with problem complexity


- Vygotskian inspired: Fundamental Feature = collaboration or assistance from another more able partner.
- 3 forms of assistance
  - Vygotskian
  - Wood
  - None

Empirical Evaluation: Structure

n=8

- Pre-test T=9
- WIS
- WIS
- WIS
- post-test T=9
- WIS
- WIS
- WIS

1 x 20 min orientation with demo
2 x 30 min ‘Ecolab’ sessions

Week 1
Week 2 - 3
Week 4
Week 14

Learning with the Ecolab

Mean Test Score %

- VIS
- WIS
- NIS

- Pre-test
- Time
- Post-test

Mixed Comparisons

- REDEEM – Ainsworth & Grimshaw (2002)
- Within system (5 versions) + ablated version

Tests MC
10 RED
10 ST
10 Non
Up to 5 sessions over three weeks
N = 84

Differentiated REDEEM ITSs

- Group A
- Group B
- Group C
- Group D
- Group E

<table>
<thead>
<tr>
<th>Context Difficulty Amount</th>
<th>Group A</th>
<th>Group B</th>
<th>Group C</th>
<th>Group D</th>
<th>Group E</th>
</tr>
</thead>
<tbody>
<tr>
<td>Difficult 44 &amp; 56 pages</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Extra-difficult 44 &amp; 56 pages</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Easier 32 &amp; 44 pages</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Easier 32 &amp; 44 pages</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Easier 32 &amp; 44 pages</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Easier 32 &amp; 44 pages</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Questions

- Types
- Difficulty
- Amount

- All types med. & hard
- med. & hard
- med. & hard
- med. & hard
- med. & hard

Strategy

- Authority
- Help
- Assistance:
- Multiple attempts at? 3
- Multiple attempts at? 3
- Multiple attempts at? 3
- Multiple attempts at? 3
- Multiple attempts at? 3

Results

- RED and ST better in REDEEM
- Post-test totals were greater than pre-test totals but no sig. interaction
Results: Category by Learning Outcomes

Significant effects of time and category
No significant interactions

Process Measures

- Analysis showed that students improved but the amount was neither substantial nor influenced by the type of system.
- A great deal of variability in improvement
- Hence, we explored a number of measures of system use to determine how learners were using the system which influenced what they learned.

Question performance on the system

<table>
<thead>
<tr>
<th></th>
<th>Pre-test</th>
<th>Post-test</th>
<th>Improvement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Right 1st Time</td>
<td>+0.327</td>
<td>+0.636</td>
<td>+0.433</td>
</tr>
<tr>
<td>(p&lt;0.004)</td>
<td>(p&lt;0.005)</td>
<td>(p&lt;0.005)</td>
<td></td>
</tr>
</tbody>
</table>

Results: Process Measures

- Time (adjusted by number of pages) correlated with improvement for REDEEM not CBT
  - REDEEM gen1, r = 0.314, p = 0.021
  - REDEEM gen2, r = 0.262, p = 0.067
  - CBT gen1, r = 0.099, p = 0.288
  - CBT gen2, r = 0.043, p = 0.397

- Significant correlation between word count of notes in on-line tool and post-test performance (r = 0.314, p = 0.006).

Summary of Four Studies

<table>
<thead>
<tr>
<th>Study</th>
<th>Subjects</th>
<th>ITSs</th>
<th>Gain</th>
<th>Effect size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genetics at Uni</td>
<td>86, 14-16 yrs</td>
<td>3 ITSs, different context &amp; strategies</td>
<td>RED = 10%</td>
<td>0.21</td>
</tr>
<tr>
<td>Genetics in School</td>
<td>15, 14-16 yrs</td>
<td>3 ITSs, different context</td>
<td>CBT = 8%</td>
<td>0.82 *</td>
</tr>
<tr>
<td>Undergrad</td>
<td>25, 20-28 yrs</td>
<td>1 ITS</td>
<td>RED = 7%</td>
<td>1.11 *</td>
</tr>
<tr>
<td>RAF</td>
<td>16, 20-45 yrs</td>
<td>1 ITS</td>
<td>CBT = 32%</td>
<td>0.88 *</td>
</tr>
</tbody>
</table>

Questions to answer

- What do I want to do with the information
  - Informing design
  - Choosing between alternatives
  - Informing about context of use
- What are appropriate forms of measurement?
- What is an appropriate design?
- What is an appropriate form of comparison?
- What is an appropriate context

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!
Choosing a context

• There is no “perfect” context! Real is not necessarily better.
• I try to avoid (a) but can’t always…(e.g. this conference!)
• Pick depending on access and nature of question
  • E.g. beware expts which need effort in artificial situations
  • Why should subjects who have no need to learn something apart from payment or course credit, work hard at learning?
  • Remember the Law of Gross Measures, time data often impossible in classrooms contexts

Miscellaneous Issues

• Other sorts of design/comparisons
• Evaluating other sorts of AIED systems
  • Authoring Tools
  • Part of Systems

Other designs

• Bystander Turing Test
  • Useful when outcome data not possible
  • Can you tell the difference between a human and a computer?
  • May be particularly useful for examining specific components
  • But susceptible to poor judgement
  • E.g. Auto-tutor (Person & Graesser, 2002)
• Simulated Students
  • E.g. Evaluating the effectiveness of different strategies/curriculum by running on simulated students
  • Unlimited number of patient, uncomplaining subjects!
  • But, how valid are the assumptions in your Sim Students
  • Still rare
  • E.g. see Van Lehn et al (1994), McClaren & Koedinger (2002)

Other comparisons

• Predicted outcomes and norms
  • Fitz-Gibbons ALIS, YELLIS
  • valued added analyses of individual performance (educational history, attitude, gender, ses) with predictive power
  • (see http://cem.dur.ac.uk/software/files/durham_report.pdf)
• MUC Style evaluations
  • The Learning Open
    (http://go260.sp.cs.cmu.edu/LearningOpen2003/default.htm)

Authoring Tools: Evaluation criteria

• the diversity of the subject matter and teaching styles that an authoring environment supports;
• the cost effectiveness of those tools
• the depth and sophistication of the ITs that the result from the authoring process
• the ease with which the tools can be used.
• the learning outcomes and experiences of students with the ITS
• the way the tools support articulation and representation of teaching knowledge
• the way that results from evaluations can inform the science base.
Problems in Evaluating ITSATs

- Evaluating an ITS Authoring Tool is particularly difficult.
- Need to evaluate the author’s experiences as well as the students
- If your tool is to be accepted, it must be usable, functional and effective.
- But the effectiveness of an ITS created with an ITSAT depends on the author, authoring tools and ITS shell.
  - E.g. if your ITS is not effective, is this because of the constraints provided by the ITSAT, decisions that an author made within those constraints, or the Shell’s interpretation of those results
- Massive credit assignment problem

Parts of System

- E.g. Dialogue component, Student Model
- Particularly difficult as many system features are co-dependent
  - E.g. Effectiveness of new Student modelling technique may depend upon remediation
- Wizard of Oz
- Sensitivity Analysis

Summary

- What not to do
  - Issues to beware
- What to do
  - Good habits
- Lessons Learned

Beware of...

- Evaluating on an inappropriate population
  - E.g. Bernhard & Sandberg (1996) evaluated a system to encourage learners to understand the tidal system by self-explanation.
  - Their subjects wouldn’t self-explain! Problem with the system or with evaluating on 14-16 yr material on undergrads who need not learn this
  - Two many or two few subjects
    - Normally see too few (try to keep a minimum of 12 per cell) but this will change depending on variability
    - Too many also a problem – want to find differences that are educationally as well as statistically significant
  - Inappropriate control
    - Most of the time comparison with traditional teaching/non-intervention control not helpful – huge credit assignment problem

Beware of... Inappropriate Generalisations

Learner Characteristics
- Ability levels
- Prior knowledge
- Developmental levels
- Gender
- Attitudes
- Motivation

Task Characteristics
- Procedural v conceptual learning
- Collaborative v Individual
- Time on task
- Timescale of intervention
- Frequency of use
  - e.g. 10 minutes a day v 1 hour a week

Beware of...

- Evaluating something else
  - Murray et al (2001) Make sure system features are visible if you want to see what their effects are.
  - Inappropriate DVs/lack of data
  - E.g. why were some DEMEST learners successful and some not?
  - Context effects
    - ILES are only one part of a complex system
    - It’s the whole shbang!
  - Relying only on attitude data
    - E.g. teachers and pupils very positive in ILS studies but in some cases actually harming exam performance
  - Inappropriate outcomes measures
    - If your system gives truly individualised experiences, how do you design a post-test?
Good habits

- More use of formative evaluation in development
- Multiple independent variables with matched learning outcomes measures to system goals
- Use of process and interaction measures
- Pre-testing
  - Both for allocation of subjects to condition and for ATI
- Effect size analysis
  - To compare your results to others

Good habits

- Build lots of time in
  - A variant of Hofstadter’s law “Evaluation takes four times as long as you think it is going to, even when you’ve taken Hofstadter’s law into account”.
- Conduct multiple evaluation studies
- Consider designs other than just pre to post
- Recognise the value of evaluation studies
- Multi-disciplinary teams
- Publishing negative as well as positive data
- Running longer evaluation studies with increased periods of intervention and delayed post-tests

AIED Evaluations: Lessons Learned

- Some evidence for value of “I” in “AIED”
- Reduces time on task, e.g. Anderson
- Produces better learning outcomes
  - than conventional teaching e.g. Lesgold, Anderson, Shute, Meyer, Koedinger
  - than less clever systems e.g. Airsworth, Shute, Luckin, Lester, Mark & Greer
  - For certain types of learner, e.g. Shute, Luckin, Arroyo
  - In certain contexts, e.g. Koedinger, Wood

- Why
  - Micro-adaptation
  - Macro-adaptation
  - Interactivity