



Evaluation Methods for Learning Environments

Shaaron Ainsworth
School of Psychology & Learning Sciences Research Institute
University of Nottingham

Acknowledgements
Ben du Boulay, Claire O'Malley
Participants at AIED03 tutorial

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.....

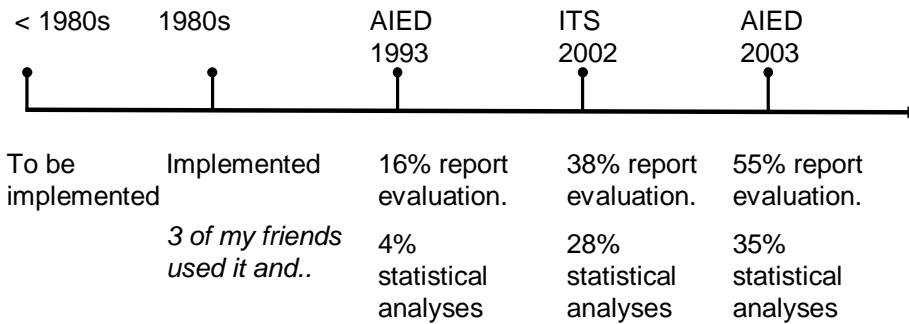
Today

- ◆ Why evaluate
- ◆ What questions should you ask to design an evaluation
 - What do I want to achieve
 - What can I measure
 - What is an appropriate methodology
 - What is an appropriate experimental design?
 - What should I compare my system to?
 - What is an appropriate context in which to evaluate
- ◆ Misc issues
- ◆ Summary and Conclusions

Why Designer's Don't Evaluate

- ◆ Assumption that designer's personal behaviour is representative
- ◆ Implicit unsupported assumptions about human performance
- ◆ Acceptance of traditional/standard interface design
- ◆ Lack of time/costs built into the project
- ◆ Lack of expertise in analysing experiments/Lack of multi-disciplinary teams.

Times they are a changing



Evaluation Methods for Learning Environments: A tutorial for AIED 2005.

Shaaron Ainsworth

Questions to answer

- ◆ **What do I want to do with the information**
- ◆ What are appropriate forms of measurement?
- ◆ What is an appropriate methodology
- ◆ What is an appropriate experimental design?
- ◆ What is an appropriate form of comparison?
- ◆ What is an appropriate context

Evaluation Methods for Learning Environments: A tutorial for AIED 2005.

Shaaron Ainsworth

Two main types of answer

- ◆ To inform design
 - Formative evaluation
 - E.g. Heuristic Evaluation, Cognitive Walkthrough
 - <http://www.psychology.nottingham.ac.uk/staff/sea/c8cxce/handout4.pdf>
 - Should the same usability heuristics be used in educational systems as are used in other computer-based systems
 - E.g. Squires & Preece (1999), Gilmore (1996)

- ◆ To assess end product
 - To assess end product or discover how it should be used
 - Summative evaluation
 - E.g. Experimental, Quasi-experimental, Case Studies, Ethnography....

Evaluation Methods for Learning Environments: A tutorial for AIED 2005.

Shaaron Ainsworth

Why perform summative assessment?

- ◆ To increase user acceptance
- ◆ To compare alternative systems
- ◆ To compare alternative design features
- ◆ To identify appropriate end users
- ◆ To sell your ILE
- ◆ To enhance learning outcome/ efficiency
- ◆ To inform theory
- ◆ To understand how to support your system and context influences use
- ◆ To help develop the field

Evaluation Methods for Learning Environments: A tutorial for AIED 2005.

Shaaron Ainsworth

Questions to answer

- ◆ What do I want to do with the information
- ◆ **What are appropriate forms of measurement?**
- ◆ What is an appropriate methodology
- ◆ What is an appropriate experimental design?
- ◆ What is an appropriate form of comparison?
- ◆ What is an appropriate context

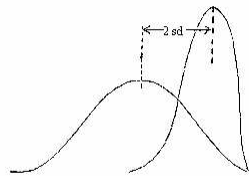
Common Measures (Dependent Variables)

- ◆ Learning gains
 - Post-test – Pre-test
 - (Post-test – Pre-test)/Pre-test: to account for high performers
- ◆ Learning efficiency
 - I.E 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

Comparison	Ratio	Effect
Classroom teaching v Expert Tutoring	1:30 v 1:1	2 sd
Classroom teaching v Non Expert Tutoring	1:30 v 1:1	0.4 sd
Classroom teaching v Computer Tutoring	1:30 v C:1	?



A 2 sigma effects means that the average tutored student performed as well as the top 2% of those receiving classroom instruction

Evaluation Methods for Learning Environments: A tutorial for AIED 2005.

Shaaron Ainsworth

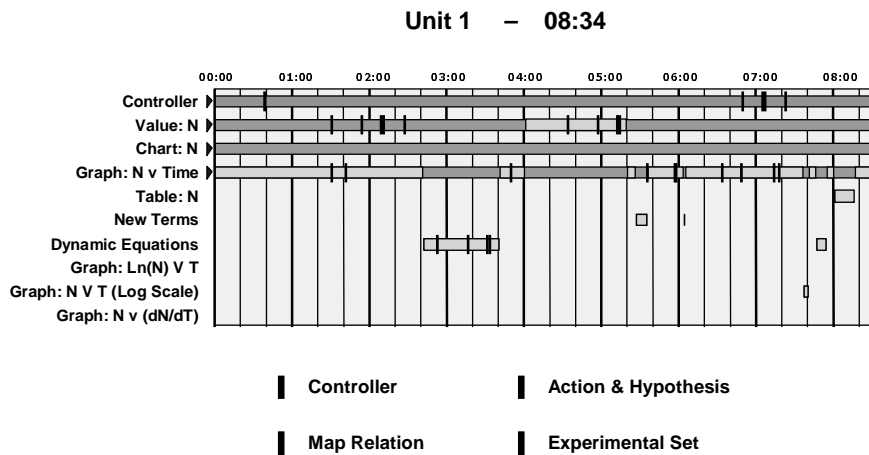
Interaction Data

- ◆ 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

Evaluation Methods for Learning Environments: A tutorial for AIED 2005.

Shaaron Ainsworth

DEMIST (Van Labeke & Ainsworth, 2002) Users' Traces



Evaluation Methods for Learning Environments: A tutorial for AIED 2005.

Shaaron Ainsworth

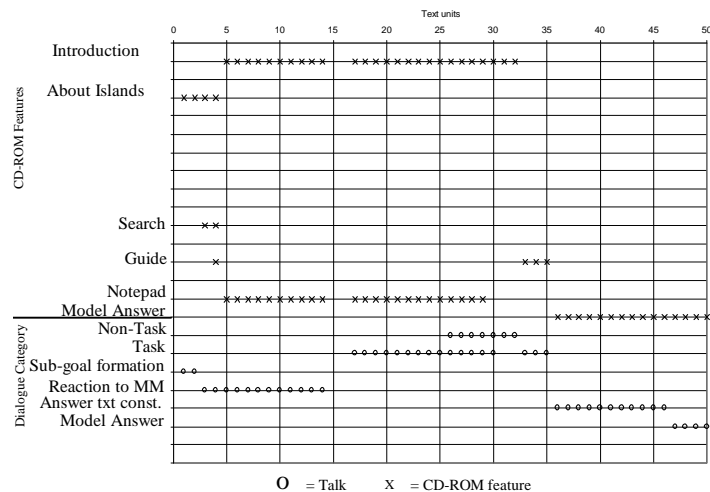
Process Data

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

Evaluation Methods for Learning Environments: A tutorial for AIED 2005.

Shaaron Ainsworth

Galapagos (Luckin et al, 2001)



Evaluation Methods for Learning Environments: A tutorial for AIED 2005.

Shaaron Ainsworth

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! But it does not always mean what you think it does....
- ◆ Process data hard work but often worth it.
- ◆ Capturing interaction data rarely changes learners' experiences, but capturing process data often does.
- ◆ Future: more integration of quantitative data with process data (such as video) as e-science and e-social science provides us with the tools...

Evaluation Methods for Learning Environments: A tutorial for AIED 2005.

Shaaron Ainsworth

Questions to answer

- ◆ What do I want to do with the information
- ◆ What are appropriate forms of measurement?
- ◆ **What is an appropriate methodology**
- ◆ What is an appropriate experimental design?
- ◆ What is an appropriate form of comparison?
- ◆ What is an appropriate context

Fixed v Flexible Evaluation Methods

- ◆ Fixed: tight pre-specification, normally quantitative, focussed on outcomes
 - Experimental or Quasi-experimental
- ◆ Flexible: evolving, often qualitative, focussed on processes
 - E.g. case studies, ethnography, grounded theory
- ◆ Can have multiple methods over the life time of a project. Typically:
 - Flexible for exploration
 - Fixed for explanation
- ◆ No bun fights! No method is better than others, only more appropriate to particular questions

Experimental Methods

- ◆ 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 Methods

- ◆ 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
 - Surprisingly few examples in AIED
 - See my talk this conference for an example

Trust in Fixed 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?

Case Studies

- ◆ Develop detailed, intensive knowledge about a small numbers of cases
- ◆ Study the case(s) in context
- ◆ Data collection is varied including observation, interviews, analysis of physical/virtual artefacts
- ◆ Generally found across social sciences
- ◆ Common in ILE research though slightly less in AIED

Ethnography

- ◆ Aims to capture, interpret and explain experiences
- ◆ Researcher becomes immersed in the setting
- ◆ Participant observation
- ◆ Originates from cultural anthropology/sociology
- ◆ Rarely seen in AIED (maybe Schofield an exception)
as systems become more common this may change

Grounded Theory

- ◆ Aims to develop theory from data collected during study.
- ◆ Commonly interview-based
- ◆ Originates from sociology
 - Find conceptual categories (open coding)
 - Find relationships (axial coding)
 - Conceptualise and account for these relationships through finding core categories (selective coding)

Trust in Flexible Designs

Validity

- ◆ Description
 - Inaccurate or incomplete data
- ◆ Interpretation
 - Imposition of inappropriate meaning
- ◆ Theory
 - Not considering alternatives

Reliability

- ◆ Issues often associated with standardised tests
- ◆ Observation

Evaluation Methods for Learning Environments: A tutorial for AIED 2005.

Shaaron Ainsworth

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
 - John Henry effect: Compensatory rivalry

Evaluation Methods for Learning Environments: A tutorial for AIED 2005.

Shaaron Ainsworth

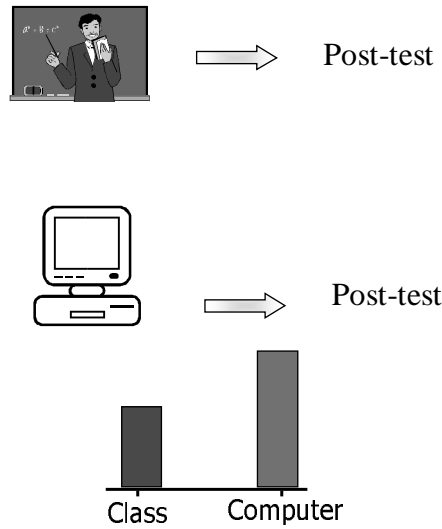
Questions to answer

- ◆ What do I want to do with the information
- ◆ What are appropriate forms of measurement?
- ◆ What is an appropriate methodology
- ◆ **What is an appropriate experimental design?**
- ◆ What is an appropriate form of comparison?
- ◆ What is an appropriate context

Prototypical designs

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

Post-test



Evaluation Methods for Learning Environments: A tutorial for AIED 2005.

Shaaron Ainsworth

Post-test

◆ Advantages

- Quick

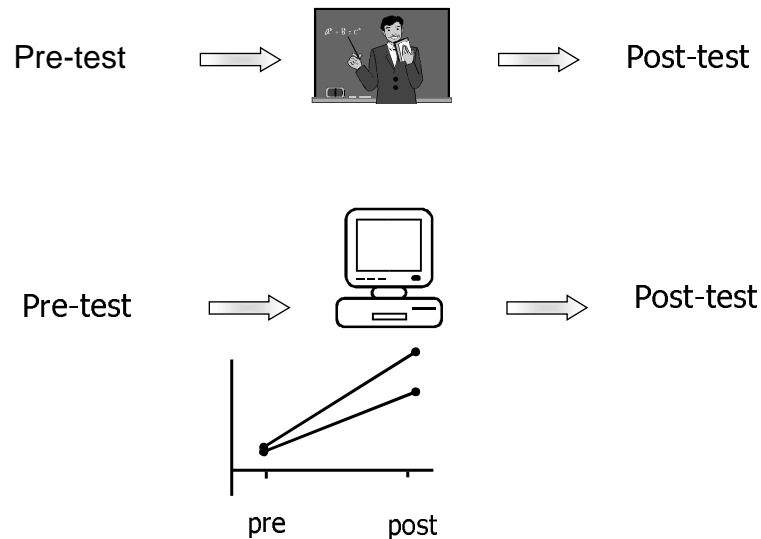
◆ Disadvantages

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

Evaluation Methods for Learning Environments: A tutorial for AIED 2005.

Shaaron Ainsworth

Pre-test to Post-test



Evaluation Methods for Learning Environments: A tutorial for AIED 2005.

Shaaron Ainsworth

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

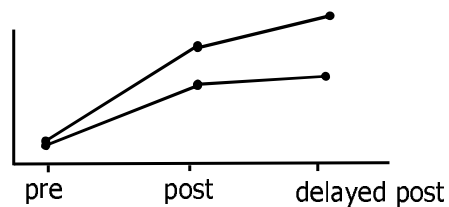
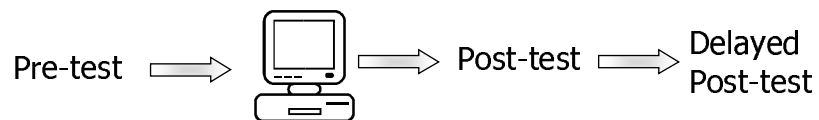
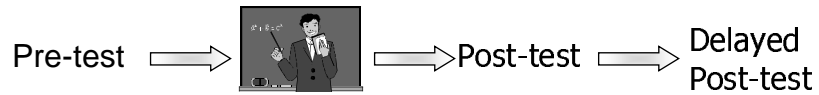
- No long term results
- Can not tell when improvement occurred if long term intervention



Evaluation Methods for Learning Environments: A tutorial for AIED 2005.

Shaaron Ainsworth

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



Evaluation Methods for Learning Environments: A tutorial for AIED 2005.

Shaaron Ainsworth

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)

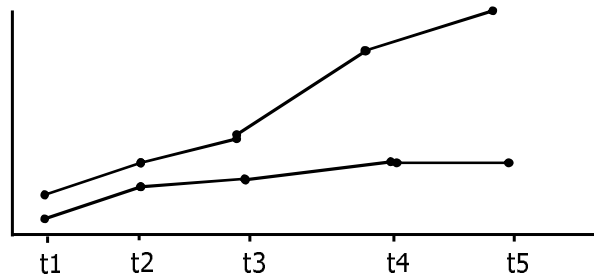
◆ Disadvantages

- Practical
- Often find an across the board gentle drop off

Evaluation Methods for Learning Environments: A tutorial for AIED 2005.

Shaaron Ainsworth

Interrupted Time-Series Design



Evaluation Methods for Learning Environments: A tutorial for AIED 2005.

Shaaron Ainsworth

Interrupted Time-Series Design

◆ Advantages

- Time scale of learning
- Ceiling effects

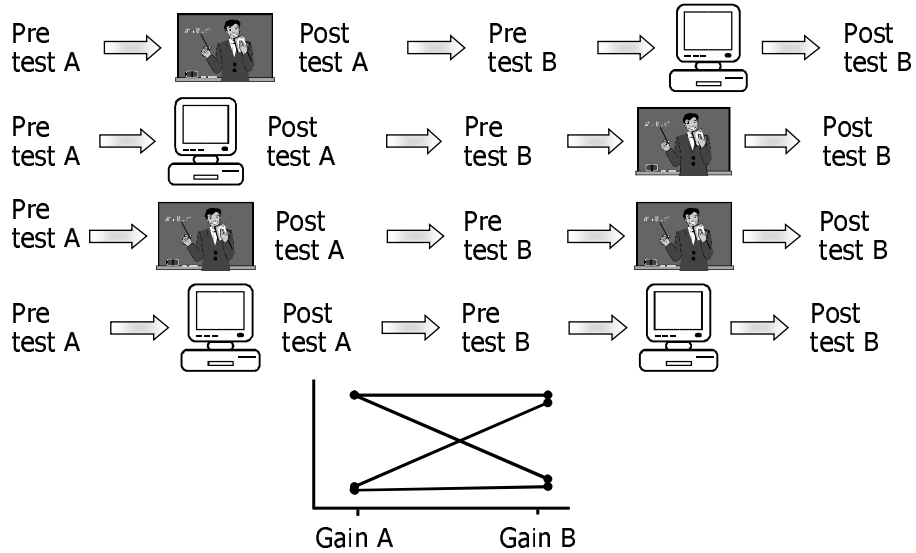
◆ Disadvantages

- Time-consuming
- Effects of repeated testing

Evaluation Methods for Learning Environments: A tutorial for AIED 2005.

Shaaron Ainsworth

Full Cross-over



Evaluation Methods for Learning Environments: A tutorial for AIED 2005.

Shaaron Ainsworth

Full Cross-over

◆ Advantages

- Controls for the (often huge) differences between subjects
 - ◆ Each subject is their own control
- May reveal order effects

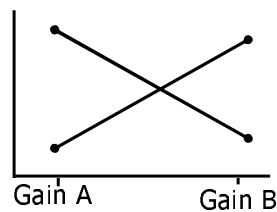
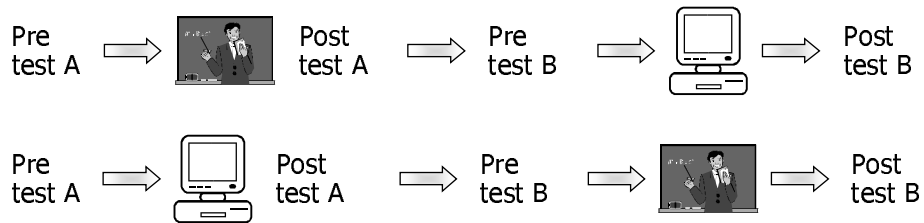
◆ Disadvantages

- Four groups of subjects rather than two!
- Statistically complex – predicting at least a 3 way interaction
- ◆ Never come across one yet in AIED!

Evaluation Methods for Learning Environments: A tutorial for AIED 2005.

Shaaron Ainsworth

Partial Cross-over



Evaluation Methods for Learning Environments: A tutorial for AIED 2005.

Shaaron Ainsworth

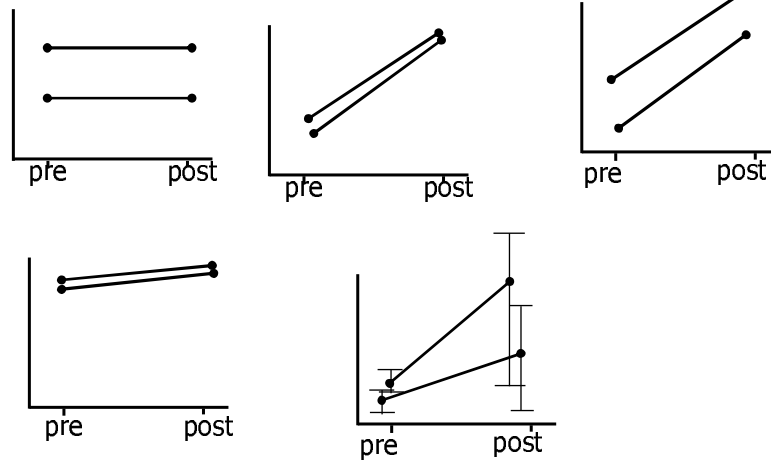
Partial Cross-over

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

Evaluation Methods for Learning Environments: A tutorial for AIED 2005.

Shaaron Ainsworth

Some Common Problems/Results



Evaluation Methods for Learning Environments: A tutorial for AIED 2005.

Shaaron Ainsworth

Questions to answer

- ◆ What do I want to do with the information
- ◆ What are appropriate forms of measurement?
- ◆ What is an appropriate methodology
- ◆ What is an appropriate experimental design?
- ◆ **What is an appropriate form of comparison?**
- ◆ What is an appropriate context

Evaluation Methods for Learning Environments: A tutorial for AIED 2005.

Shaaron Ainsworth

Nature of Comparison

- ◆ ILE alone
- ◆ ILE v non-interventional control
- ◆ ILE v Classroom
- ◆ ILE_(a) v ILE_(b) (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 reading help messages lead to better performance?
- ◆ Disadvantages
 - No comparative data – is this is good way of teaching??
 - Identifying key variables to measure

Smithtown — Shute & Glaser (1990)

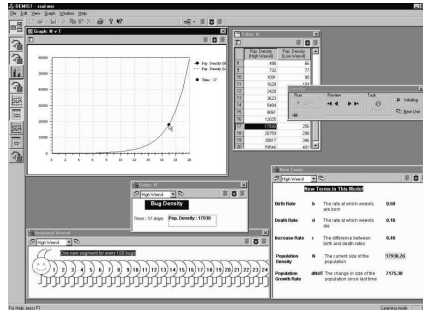
- ◆ Guided discovery environment to scientific enquiry skills and principles of basic economics
 - Notebook, grapher, hypothesis maker
 - Explorations & experiments
- ◆ Issue-based tutoring to detect and remediate scientific method
- ◆ Students who did well with Smithtown (n = 530) engaged in goal or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

DEMIST – Van Labeke & Ainsworth (2002)

- ◆ 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 visualiser/verbaliser
- ◆ Conclusion: Inappropriate method as can't answer "WHY"
 - What does spending a lot of time with a representation mean?
 - Need for protocols

Evaluation Methods for Learning Environments: A tutorial for AIED 2005.

Shaaron Ainsworth

ILE v non-interventional control

- ◆ Examples
 - COPPERS – Ainsworth et al (1998)
- ◆ 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 ?

Evaluation Methods for Learning Environments: A tutorial for AIED 2005.

Shaaron Ainsworth

COPPERS – Ainsworth et al (1998)

The screenshot shows the COPPERS software interface. On the left, under 'QUESTIONS', the problem is: 'How much does this make?' with the equation $2 \times 20p + 2 \times 10p$ and two coins (20p and 10p). On the right, under 'ANSWERS', there are buttons for 'CLEAR', '60 p', and 'TOTAL', and a grid of coin buttons (1p, 2p, 5p, 10p, 20p, 50p, £1). A 'HELP' button is at the bottom left. To the right of the main interface is a 'PREVIOUS' window showing a table of solutions and a list of calculations.

1p	2p	5p	10p	20p	50p	£1	TOTAL
1	2	3	1	1			60p
2	3	1	2	3			60p
5	2	2	2	1			60p

ANSWERS
You're right a correct answer to this problem is

1 x 20 pence = 20p +
5 x 2 pence = 10p +
2 x 10 pence = 20p +
2 x 5 pence = 10p

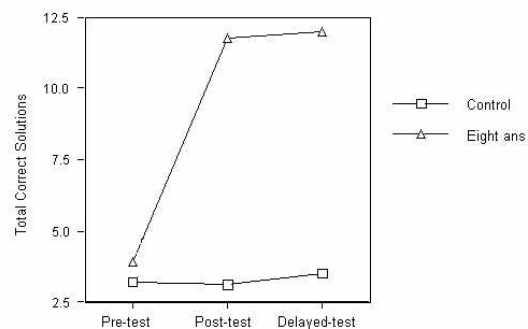
60p

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

Evaluation Methods for Learning Environments: A tutorial for AIED 2005.

Shaaron Ainsworth

COPPERS Results



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

Evaluation Methods for Learning Environments: A tutorial for AIED 2005.

Shaaron Ainsworth

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 many ways, what can we truly conclude?

Evaluation Methods for Learning Environments: A tutorial for AIED 2005.

Shaaron Ainsworth

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

Evaluation Methods for Learning Environments: A tutorial for AIED 2005.

Shaaron Ainsworth

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 (5hrs vs. 11hrs)

SHERLOCK — Lesgold et al (1992)

- ◆ Intelligent training system
 - Airforce technicians
 - Complex piece of electronics test gear
- ◆ Interface & overall training context
- ◆ 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

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

Evaluation Methods for Learning Environments: A tutorial for AIED 2005.

Shaaron Ainsworth

PAT — Koedinger et al (1997)

- ◆ 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

	Control Group	PAT Group	F value significance	sigma
Iowa Algebra Aptitude	.46 (.17) 80	.52 (.19) 287	F(2,398) = 17.0 P < .0001	0.3
Math SAT Subset	.27 (.14) 44	.32 (.16) 127	F(2,205) = 5.1 P < .01	0.3
Problem Situation Test	.22 (.22) 42	.39 (.33) 127	F(2,186) = 5.3 P < .01	0.7
Representations Test	.15 (.18) 44	.37 (.32) 124	F(2,183) = 13.4 P < .0001	1.2

Evaluation Methods for Learning Environments: A tutorial for AIED 2005.

Shaaron Ainsworth

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
 - time on task? ability?
- ◆ Mistakes
 - Too many subjects!
 - Not sophisticated enough analyses – huge wasted opportunity

Evaluation Methods for Learning Environments: A tutorial for AIED 2005.

Shaaron Ainsworth

ILE_(a) v ILE_(b) (within system)

- ◆ Examples
 - PACT – Alevan et al (1999)
 - CENTS – Ainsworth et al (2002)
 - Galapagos – Lucken et al (2001)
 - Animal Watch – Arroyo et al (1999,2000)
- ◆ Uses
 - Much tauter design, e.g. nullifies Hawthorne effect
 - Identifies what key system components add to learning
 - Aptitude by treatment interactions
- ◆ Disadvantages
 - Identifying key features to vary – could be very time consuming!

Evaluation Methods for Learning Environments: A tutorial for AIED 2005.

Shaaron Ainsworth

PACT – Alevan et al (1999, 2002)

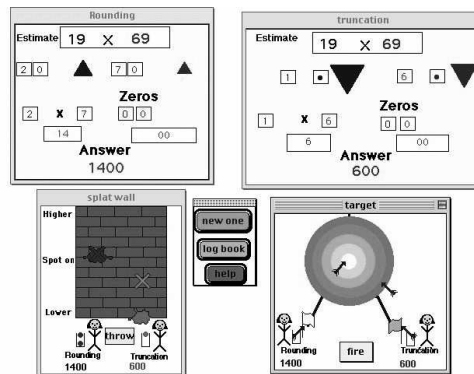
- ◆ Another CMU cognitive tutor - Geometry
- ◆ Two versions – a Self-Explanation v Answer only
- ◆ Expt 1 (N = 23) – Significantly greater gains for SE group
- ◆ Expt 2 (N = 43) – Overall suspect non significant interaction! But SE students doing better on harder problems.

Evaluation Methods for Learning Environments: A tutorial for AIED 2005.

Shaaron Ainsworth

CENTS – Ainsworth et al (2002)

- ◆ 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

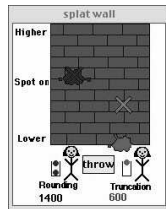


Evaluation Methods for Learning Environments: A tutorial for AIED 2005.

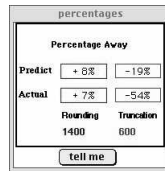
Shaaron Ainsworth

Which do you think will be best?

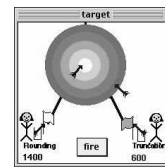
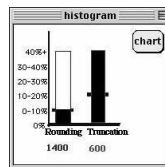
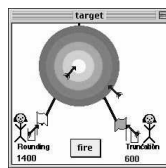
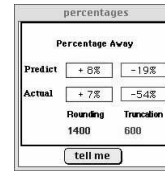
Pictures



Maths



Mixed

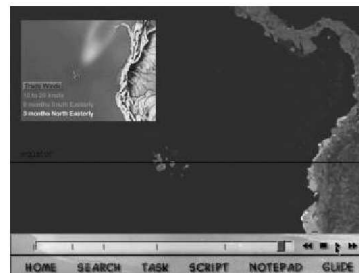


Evaluation Methods for Learning Environments: A tutorial for AIED 2005.

Shaaron Ainsworth

MENO – 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



Evaluation Methods for Learning Environments: A tutorial for AIED 2005.

Shaaron Ainsworth

'Galapagos': three version

	NARRATIVE GUIDANCE	SUPPORT FOR NARRATIVE CONSTRUCTION
LINEAR	<ul style="list-style-type: none"> • recognisable, linear structure • easy navigation • limited interaction • implicit guidance in interface design (eg order of items) 	<ul style="list-style-type: none"> • notepad • model answer
RESOURCE-BASED LEARNING (RBL)	<ul style="list-style-type: none"> • no explicit narrative guidance • implicit guidance in interface design 	<ul style="list-style-type: none"> • easily accessible statement of task
GUIDED DISCOVERY LEARNING (GDL)	<ul style="list-style-type: none"> • three text guides offer routes through material and stimulate enquiry • implicit guidance in interface design 	<ul style="list-style-type: none"> • script

Evaluation Methods for Learning Environments: A tutorial for AIED 2005.

Shaaron Ainsworth

Dialogue Categories

- ◆ Non-Task: Navigational/Operational e.g. "click on one" "play" c
- ◆ Task: Mechanics of getting the task done e.g. "shall I type?"
- ◆ Content
 - Sub-Goal e.g. "why do we want to take notes?"
 - Reaction to Multi Media e.g. "Its really cool"
 - Answer Construction e.g. "Well they are all very similar aren't they, just with slightly different"
 - Model Answer e.g. "so we have missed that massive chunk out"

Evaluation Methods for Learning Environments: A tutorial for AIED 2005.

Shaaron Ainsworth

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 GDL 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.
- ◆ Simple interface design elicited a higher ratio of task to procedural discussion than commercial interfaces

Evaluation Methods for Learning Environments: A tutorial for AIED 2005.

Shaaron Ainsworth

ILE v Ablated ILE

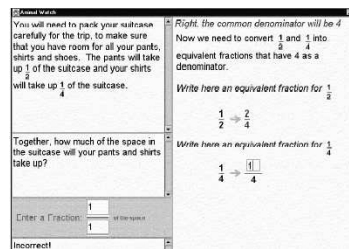
- ◆ Ablation experiments remove particular design features and performance of the systems compared
- ◆ Examples
 - VCR Tutor – Mark & Greer (1995)
 - StatLady – Shute (1995)
 - Dial-A-Plant – Lester et al (1997)
 - Luckin & du Boulay (1999)
- ◆ Uses
 - What is the added benefit of AI
- ◆ Disadvantages
 - System may not be modular

Evaluation Methods for Learning Environments: A tutorial for AIED 2005.

Shaaron Ainsworth

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: cognitive development & gender
- ◆ Some results: Girls do better with interactive hints. High cognitive levels better with symbolic & interactive hints



Evaluation Methods for Learning Environments: A tutorial for AIED 2005.

Shaaron Ainsworth

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
 - recognise 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

Evaluation Methods for Learning Environments: A tutorial for AIED 2005.

Shaaron Ainsworth

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
 - Much 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 Methods for Learning Environments: A tutorial for AIED 2005.

Shaaron Ainsworth

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

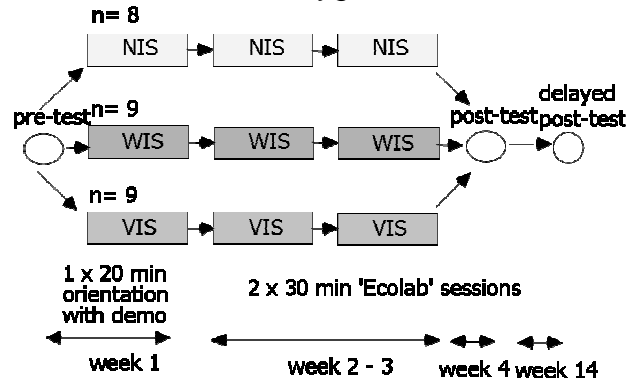


Evaluation Methods for Learning Environments: A tutorial for AIED 2005.

Shaaron Ainsworth

Ecolab – Lucken & Du Boulay (1999)

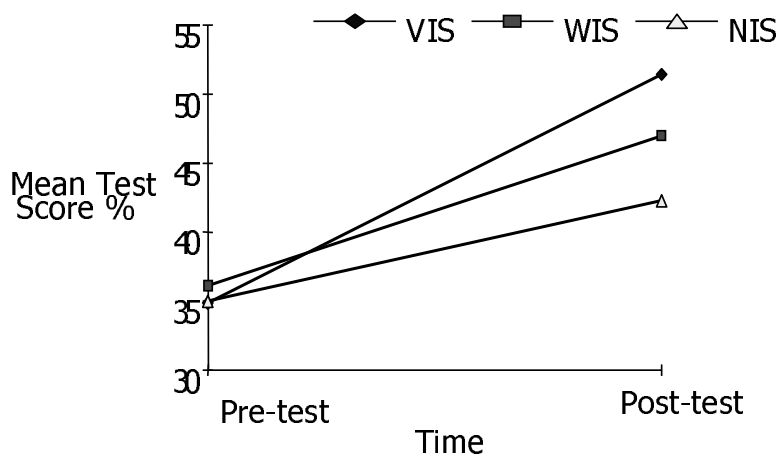
- ◆ Vygotskian inspired: Fundamental Feature = collaboration or assistance from another more able partner.
- ◆ 3 forms of assistance: Vygotskian, Wood or None



Evaluation Methods for Learning Environments: A tutorial for AIED 2005.

Shaaron Ainsworth

Learning with the Ecolab

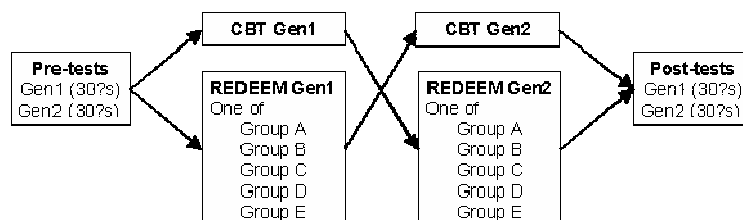


Evaluation Methods for Learning Environments: A tutorial for AIED 2005.

Shaaron Ainsworth

Mixed Comparisons

- ◆ REDEEM – Ainsworth & Grimshaw (2004)
- ◆ Within system (5 versions) + ablated version



Tests MC

10 RED

10 ST

10 Non

Up to 5 sessions over three weeks
N = 84

Evaluation Methods for Learning Environments: A tutorial for AIED 2005.

Shaaron Ainsworth

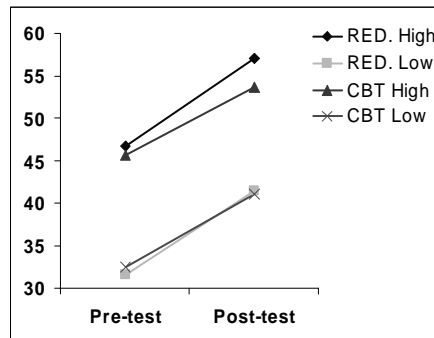
Differentiated REDEEM ITSs

	Group A	Group B	Group C	Group D	Group E
Content					
Difficulty	difficult	quite difficult	easier	easier	easier
Amount	44 & 60 pages	44 & 50 pages	32 & 44 pages	30 & 44 pages	30 & 44 pages
Questions					
Types	all types	all types	all types	no matching	no matching
Difficulty	med. & hard	med. & hard	easy & med.	easy & med.	easy & med.
Amount	36 & 39 ?s all	36 & 39 ?s all	24 & 24 ?s 1 per page	23 & 24 ?s 1 per page	23 & 24 ?s 1 per page
Strategy					
Autonomy	choose sections	choose sections	no choice	no choice	no choice
Help	selects ? type	selects ?s	? after section	? after section	? after page
Answers- deduced	help on error multiple attempts at ?	help on error multiple attempts at ?	help on error multiple attempts at ?	help on error & request 2 attempts at ?	help on error & request 2 attempts at ?

Evaluation Methods for Learning Environments: A tutorial for AIED 2005.

Shaaron Ainsworth

Results: Category by Learning Outcomes



Significant effects of time and category
No significant interactions

Evaluation Methods for Learning Environments: A tutorial for AIED 2005.

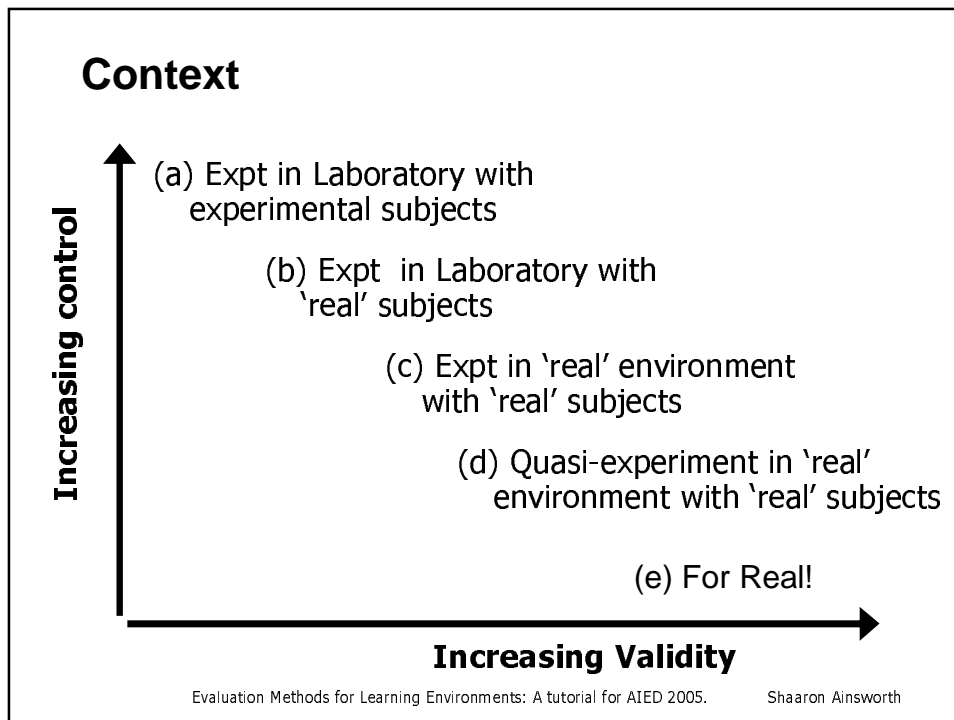
Shaaron Ainsworth

Questions to answer

- ◆ What do I want to do with the information
- ◆ What are appropriate forms of measurement?
- ◆ What is an appropriate methodology
- ◆ What is an appropriate experimental design?
- ◆ What is an appropriate form of comparison?
- ◆ **What is an appropriate context?**

Evaluation Methods for Learning Environments: A tutorial for AIED 2005.

Shaaron Ainsworth



Choosing a context

- ◆ There is no “perfect” context! Real is not necessarily better.
- ◆ I try to avoid (a) but can't always
- ◆ 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

For Real: Integrated Learning Systems Wood, et al (1999)

- ◆ An ILS is made up of two components, CAI modules and a Management System. Individualised learning programme with teacher reports, some remediation and immediate feedback.
- ◆ Evaluation in many schools, very large N
- ◆ Positive learning outcomes in basic numeracy but not for basic literacy, some evidence of gains on more extensive maths tests
- ◆ No transfer to standard educational attainment measures and some evidence of degraded performance
- ◆ Positive attitudes to ILS expressed by teachers & pupils (80%+)
- ◆ Attitudes were not linked to assessed learning outcomes.
- ◆ Patterns of usage had significant effects on outcomes
- ◆ Overall – evaluation probably saved UK from massively investing in inappropriate software

Evaluation Methods for Learning Environments: A tutorial for AIED 2005.

Shaaron Ainsworth

Miscellaneous Issues

- ◆ Other sorts of design/comparisons
- ◆ What not to do
 - Issues to beware
- ◆ What to do
 - Good habits
- ◆ Lessons Learned

Evaluation Methods for Learning Environments: A tutorial for AIED 2005.

Shaaron Ainsworth

(Some) Other Methods

- ◆ 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
 - E.g. see Van Lehn et al (1994), McClaren & Koedinger (2002) (still rare)
- ◆ Microgenetic studies (e.g. Siegler, 1995)
 - involving a high density of observations relative to the rate of change,
 - large number of observation when change is taking place
 - intensive trial-by-trial analysis using both quantitative and qualitative methods with the goal of inferring the processes that give rise to change.
 - See Van-Labeke and Ainsworth (2002) for a study with DEMIST (which is still under analysis!)

Evaluation Methods for Learning Environments: A tutorial for AIED 2005.

Shaaron Ainsworth

Other comparisons

- ◆ Predicted outcomes and norms
 - Fitz-Gibbons ALIS, YELIS
 - 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://gs260.sp.cs.cmu.edu/LearningOpen2003/default.htm>)

Evaluation Methods for Learning Environments: A tutorial for AIED 2005.

Shaaron Ainsworth

Evaluating only 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

Beware of...

- ◆ Evaluating on an inappropriate population
 - E.g. Barnard & 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

Evaluation Methods for Learning Environments: A tutorial for AIED 2005.

Shaaron Ainsworth

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 DEMIST learners successful and some not!
- ◆ Context effects
 - ILES are only one part of a complex system
 - It's the whole shebang!
- ◆ 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?

Evaluation Methods for Learning Environments: A tutorial for AIED 2005.

Shaaron Ainsworth

Good habits

- ◆ No bun fights – accept all methods as valid and interesting if well performed and targeted at appropriate questions
- ◆ Fixed Design: For example
 - Is my system more effective than an alternative..
 - How effective is my system
 - Are my results robust and replicable
- ◆ Flexible: For example
 - Who benefits from learning with ILEs
 - How do people learn with ILEs
 - How does learning with ILEs change over time?
 - How does the wider context influence learning with ILEs?

Evaluation Methods for Learning Environments: A tutorial for AIED 2005.

Shaaron Ainsworth

Good habits

- ◆ Recognising the value of evaluation....
- ◆ More use of formative evaluation in development
- ◆ Multiple dependent 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

Evaluation Methods for Learning Environments: A tutorial for AIED 2005.

Shaaron Ainsworth

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 experimental designs other than just pre to post test
- ◆ Multi-disciplinary teams
- ◆ Publishing negative as well as positive data
- ◆ Running longer evaluation studies with increased periods of intervention and delayed post-tests

Evaluation Methods for Learning Environments: A tutorial for AIED 2005.

Shaaron Ainsworth

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. Ainsworth, 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

Evaluation Methods for Learning Environments: A tutorial for AIED 2005.

Shaaron Ainsworth

Go out and evaluate

Evaluation Methods for Learning Environments: A tutorial for AIED 2005.

Shaaron Ainsworth

Further Reading

- ◆ Robson, C.(2002) Real world research. Blackwell Publishing. Oxford.. "*Great Book*"
- ◆ Journal of AI and Education (1993) 4(2/3)– double issue on evaluation – "*lots of classic papers*"
- ◆ du Boulay, B. (2000). Can we learn from ITSs? In Gauthier, G., Frasson, C., and VanLehn, K., editors, *Intelligent Tutoring Systems: Proceedings of 5th International Conference, ITS 2000, Montreal*, number 1839 in Lectures Notes in Computer Science, pages 9-17. Springer-Verlag. "*I stole half of it for this tutorial*"

Evaluation Methods for Learning Environments: A tutorial for AIED 2005.

Shaaron Ainsworth

References

- ◆ Ainsworth, S. E., Bibby, P., & Wood, D. (2002). Examining the effects of different multiple representational systems in learning primary mathematics. *Journal of the Learning Sciences*, 11(1), 25-61.
- ◆ Ainsworth, S. E., & Grimshaw, S. K. (2002). Are ITSs created with the REDEEM authoring tool more effective than "dumb" courseware? In S. A. Cerri & G. Gouardères & F. Paraguaçu (Eds.), 6th International Conference on Intelligent Tutoring Systems (pp. 883-892). Berlin: Springer-Verlag.
- ◆ Ainsworth, S. E., Wood, D., & O'Malley, C. (1998). There is more than one way to solve a problem: Evaluating a learning environment that supports the development of children's multiplication skills. *Learning and Instruction*, 8(2), 141-157.
- ◆ Arroyo, I., Beck, J. E., Woolf, B. P., Beal, C. R., & Schultz, K. (2000). Macroadapting animalwatch to gender and cognitive differences with respect to hint interactivity and symbolism. In G. Gauthier & C. Frasson & K. VanLehn (Eds.), *Intelligent Tutoring Systems: Proceedings of the 5th International Conference ITS 2000* (Vol. 1839, pp. 574-583). Berlin: Springer-Verlag.
- ◆ Barnard, Y.F. & Sandberg, J.A.C. 1996. Self-explanations: do we get them from our students. In P. Brna, et al. (Eds.), *Proceedings of the AI and Education Conference*, p. 115-121.
- ◆ Conati, C., & VanLehn, K. (2000). Toward Computer-Based Support of Meta-Cognitive Skills: a Computational Framework to Coach Self-Explanation. *International Journal of Artificial Intelligence in Education*, 11, 389-415.

Evaluation Methods for Learning Environments: A tutorial for AIED 2005.

Shaaron Ainsworth

References

- ◆ Corbett, A. & Anderson, J. (1992). LISP intelligent tutoring system: Research in skill acquisition. In J. H. Larkin and R. W. Chabay, editors, *Computer-Assisted Instruction and Intelligent Tutoring Systems: Shared Goals and Complementary Approaches*, pages 73-109. Lawrence Erlbaum
- ◆ Cox, R., & Brna, P. (1995). Supporting the use of external representations in problem solving: The need for flexible learning environments. *Journal of Artificial Intelligence in Education*, 6(2/3), 239-302.
- ◆ Gilmore, D. J. (1996). The relevance of HCI guidelines for educational interfaces. *Machine-Mediated Learning*, 5(2), 119-133.
- ◆ Greer, J.E., McCalla, G.I., Cooke, J.E., Collins, J.A., Kumar, V.S., Bishop, A.S., Vassileva, J.I. "Integrating Cognitive Tools for Peer Help: the Intelligent IntraNet Peer Help-Desk Project" in S. Lajoie (Ed.) *Computers as Cognitive Tools: The Next Generation*, Lawrence Erlbaum, 2000, 69-96.
- ◆ Koedinger, K. R., Anderson, J. R., Hadley, W. H., & Mark, M. A. (1997). Intelligent tutoring goes to school in the big city. *International Journal of Artificial Intelligence in Education*, 8, 30-43.
- ◆ Lesgold, A., Lajoie, S., Bunzo, M., & Eggan, G. (1992). Sherlock: A coached practice environment for an electronics troubleshooting job. In J. Larkin & R. Chabay (Eds.), *Computer Based Learning and Intelligent Tutoring* (pp. 202-274). Hillsdale, NJ: LEA.

Evaluation Methods for Learning Environments: A tutorial for AIED 2005.

Shaaron Ainsworth

References

- ◆ Lester, J. C., Converse, S. A., Stone, B. A., Kahler, S. A., and Barlow, S. T. (1997). Animated pedagogical agents and problem-solving effectiveness: A large-scale empirical evaluation. In du Boulay, B. and Mizoguchi, R., Proceedings of the AI-ED 97 World Conference on Artificial Intelligence in Education,, pages 23–30, Kobe, Japan. IOS Press.
- ◆ Litmann, D., & Soloway, E. (1988). Evaluating ITSs: The cognitive science perspective. In M. Polson & J. J. Richardson (Eds.), Foundations of Intelligent Tutoring Systems. Hillsdale, NJ: LEA.
- ◆ Luckin, R., & du Boulay, B. (1999). Ecolab: The Development and Evaluation of a Vygotskian Design Framework. *International Journal of Artificial Intelligence in Education*, 10, 198-220.
- ◆ Luckin, R., Plowman, L., Laurillard, D., Stratfold, M., Taylor, J., & S, C. (2001). Narrative evolution: learning from students' talk about species variation. *International Journal of AIED*, 12, 100-123.
- ◆ MacLaren, & Koedinger, K (2002): When and Why Does Mastery Learning Work: Instructional Experiments with ACT-R "SimStudents". *ITS 2002* 355-366
- ◆ Mark, M., & Greer, J. E. (1995). The VCR tutor: Effective instruction for device operation. *The Journal of the Learning Sciences*, 4(2), 209-246.
- ◆ Mark, M. A., & Greer, J. E. (1993). Evaluation methodologies for intelligent tutoring systems. *Journal of Artificial Intelligence in Education*, 4(2/3), 129-153.
- ◆ Meyer, T. N., Miller, T. M., Steuck, K., & Kretschmer, M. (1999). A multi-year large-scale field study of a learner controlled intelligent tutoring system. In S. Lajoie & M. Vivet (Eds.), *Artificial Intelligence in Education - (Vol. 50, pp. 191-198)*.

Evaluation Methods for Learning Environments: A tutorial for AIED 2005.

Shaaron Ainsworth

References

- Murray, T. (1993). Formative Qualitative Evaluation for "Exploratory" ITS research. *Journal of Artificial Intelligence in Education*, 4(2/3), 179-207.
- Person, N.K., Graesser, A.C., Kreuz, R.J., Pomeroy, V., & TRG (2001). Simulating human tutor dialog moves in AutoTutor. *International Journal of Artificial Intelligence in Education*. 12, 23-39.
- Rogers, Y., Price, S., Harris, E., Phelps, T., Underwood, M., Wilde, D. & Smith, H. (2002) 'Learning through digitally-augmented physical experiences: Reflections on the Ambient Wood project'. (Equator working paper) (see http://www.cogs.susx.ac.uk/interact/papers/pdfs/Playing%20and%20Learning/Tangibles%20and%20virtual%20environments/Rogers_Ambient_Wood2.pdf)
- Shute, V. J. (1995). SMART evaluation: Cognitive diagnosis, mastery learning and remediation. In J. Greer (Ed.), *Proceedings of AI-ED 95* (pp. 123-130). Charlottesville, VA: AACE.
- Shute, V. J., & Glaser, R. (1990). A large-scale evaluation of an intelligent discovery world: Smithtown. *Interactive Learning Environments*, 1, 51-77.
- Shute, V. J., & Regian, W. (1993). Principles for evaluating intelligent tutoring systems. *Journal of Artificial Intelligence in Education*, 4(2/3), 243-271.
- Squires, D., & Preece, J. (1999). Predicting quality in educational software: Evaluating for learning, usability and the synergy between them. *Interacting with Computers*, 11(5), 467-483.

Evaluation Methods for Learning Environments: A tutorial for AIED 2005.

Shaaron Ainsworth

References

- Van Labeke, N., & Ainsworth, S. E. (2002). Representational decisions when learning population dynamics with an instructional simulation. In S. A. Cerri & G. Gouardères & F. Paraguaçu (Eds.), *Intelligent Tutoring Systems: Proceedings of the 6th International Conference ITS 2002* (pp. 831-840). Berlin: Springer-Verlag.
- Van Labeke, N., & Ainsworth, S. (2003). A microgenetic approach to understanding translation between representations. Paper presented at the 10th EARLI conference, Padova, Italy.
- VanLehn, K., Ohlsson, S., & Nason, R. (1994). Applications of simulated students: An exploration. *Journal of AI in Education*, 5, 135-175.
- Wood, D. J., Underwood, J. D. M., & Avis, P. (1999). Integrated Learning Systems in the Classroom. *Computers and Education*, 33(2/3), 91-108