ADAPTIVE LEARNING ENVIRONMENTS: Summative Evaluation

Contents

Some material based on Ainsworth's AIED 2003 tutorial on Evaluation Methods for Learning Environments, see AILE course web page and link:

http://www.psychology.nottingham.ac.uk/staff/sea/Evaluationtutorial.ppt

1. Design of Experiments

Role of Experiment in Design

Often experiments are used to guide new designs or the help understand existing design

Programs are not themselves experiments but are part of the basis for conducting experiments (on an algorithm or a system or a group of people)

Three types of activity:

Exploratory: where we are wondering what to design

Formative Evaluation: we experiment with a preliminary design with the aim of building a better one

Summative Evaluation: where a final design is analysed definitively

Formative v. Summative Evaluation

Formative Evaluation:

- iterative, throughout design and implementation
- test preliminary designs for usability etc
- assessing impact of changes
- make decisions about later project stages
- frequently qualitative

Summative Evaluation:

- on completion of each stage
- assessing effectiveness
- frequently quantitative

Qualitative v. Quantitative Data

Qualitative

Descriptive data Subjective

Based on system behaviour or user experience

Obtained from observation, questionnaires, interviews, protocol analysis, heuristic evaluation, cognitive and post task walkthrough

Quantitative

Numerical data Objective

Based on measures of variables relevant to performance or user experience

Obtained from empirical studies, e.g. experiments, also questionnaires, interviews

Amenable to statistical analysis

Systems and Experiments

When we talk about **experiments**, generally talking about...

- stating specific hypotheses
- identifying and manipulating variables
- systematic procedures to TEST our hypotheses
- some degree of control (often limited in "real world" settings)

Not all studies we do in ALE are "experiments" in a strict sense

- May do a survey about how system was used in class
- May observe participants using a system
- May mine data afterwards doing post hoc analysis, looking for general patterns

Also, difference between a **true experiment** with randomised group assignment and so forth, versus a **quasi-experiment,** where less control, may not be able to randomly assign groups, etc.

Typical Questions

Having gone through a number of iterations of formative evaluation, you think that the system is finally ready.

You need to see now how well it works….

Does it do what it was claimed it would do?

Is it effective?

Such questions need to be made more precise.

A number of methods can be used, e.g.

- an experimental set-up with alternative versions of the tool - perhaps without a crucial feature
- a control group for comparison.

Methodology has to be tight for strong claims to be made.

Common Measures (Dependent Variables) *(from Ainsworth, 2003)*

Learning gains

Post-test – Pre-test

Learning efficiency

i.e. does it reduce time spent learning

How the system is used in practice (and by whom)

ILEs cannot help if learners do not use them!

What features are used

User attitudes

Cost savings

Teachbacks

How well can learners now teach what they have learnt

Prototypical designs *(Ainsworth, 2003)*

1. (intervention) post-test
2. Pre – (intervention) - post-test
3. Pre – (intervention) - post-test – delayed post-test
4. Interrupted time-series
5. Cross-over

Look at Ainsworth (2003) tutorial for examples of these (see web page)

Nature of Comparison *(Ainsworth, 2003)*

1. ILE alone
2. ILE v non-interventional control
3. ILE v Classroom
4. $ILE_{(a)}$ v $ILE_{(b)}$ (within system)
5. ILE v Ablated ILE
6. Mixed models

Again, see Ainsworth (2003) tutorial for examples of these (see web page)

ILE alone *(Ainsworth, 2003)*

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?
  - Does reading help messages lead to better performance?

Disadvantages

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

ILE v non-interventional control
(Ainsworth, 2003)

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 ?

ILE v Classroom *(Ainsworth, 2003)*

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?

ILE$_{(a)}$ v ILE$_{(b)}$ (within system) *(Ainsworth, 2003)*

Examples

- PACT – Aleven 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!

ILE v Ablated ILE *(Ainsworth, 2003)*

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

Context *(Ainsworth, 2003)*

(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!

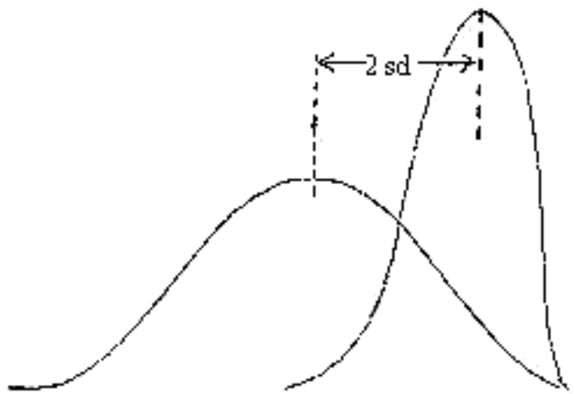
Increasing control

Increasing Validity

Learning Gains: Effect Size *(Ainsworth, 2003)*

(Gain in Exp Condtn– 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 98% of students receiving expert tutoring are likely do to better than students receiving classroom instruction

Choosing Between Designs
(Ainsworth, 2003)

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?

Some issues and problems

Natural environment v ability to control variables

e.g. test in classroom v. bring into laboratory

Interference with participants - ethical issues

* Should you use a method of teaching that you do not think is going to work on your participants?
* Should everyone get the opportunity to use the best approach?
* Will getting poor scores on a test that is not relevant to the curriculum affect student's morale and consequently their other work?
* Should you use teaching time to do experiments?

Problems of measurement:

* What is improvement?
* How long does it last?
* Does it generalise?

2. Evaluating the Design and Effectiveness of a Maths Tutoring System

Maths Tutoring System Example

Goal: ***intelligent computer tutor for university maths students to practice calculus***

- How do human tutors teach calculus?
- What can we infer from human tutors behaviour to inform tutor design?
- What is feasible to incorporate in system and what not?

Questions we might consider to inform design:

1. What errors do students typically make?
2. What should the system do when students make errors?

Methods for collecting maths errors

Task analysis
Cognitive Walkthrough
Protocol analysis
Video Recording
Questionnaire
Sensitivity Analysis
Post-hoc analysis
Dialogue mark-up and analysis
Manipulation experiment
Self Report

Observation
Mock-ups
Wizard of Oz
Interview
Focus groups
Expert evaluation
Logging use
Sentient analysis

What errors do students typically make?

1. **Interview** teachers about errors that target users frequently make (*error types and examples*)
2. Devise a **set of test calculus examples**
3. Give target user group test set and **observe**, **collect log of** their **interaction** (*example errors*)
4. **Analyse** results to see most frequent errors
5. Give **questionnaire** to teachers with example errors and ask what feedback they would give (*feedback types in relation to each error*)
6. **Observe** tutor teaching student through chat interface + **record interaction** (*example errors*)
7. **Analyse interaction** in relation to student errors and actions taken by teacher (*feedback types*)
8. **Cognitive walkthrough** by tutor (*when feedback type given and general feedback strategies*)

What should the system do when students make errors?

Using these methods you find that human tutors usually use one of the following feedback options:

1. *give feedback immediately*
2. *just flag to the student that they have made an error*
3. *let the student realise they have made a mistake and ask for help*

You want to see which works best…

Do some experiments with the tutoring system, with some students.....

[Based loosely on a experimental study described in Corbett, A.T. and Anderson, J.R., 1990]

Other Evaluation Questions...

Does interface A to the Maths tutor work better than interface B?

Does student enjoyment correlate with learning?

Does student enjoyment correlate with learning?

Assessing student enjoyment - affective measures:

- Observe facial expressions
- Self-report of enjoyment: sliders
- Questionnaire
- Verbal Protocol
- Expert observation

Assessing Learning - performance measures:

- Number of errors
- Time to learn to mastery
- Amount of materials covered in set time

Does interface A to the Maths tutor work better than interface B?

Could use **various methods**:

- Questionnaire
- Observation
- Interviews
- Logging use
- …

but considering **experimental methods** here…..

General Experimental Design: Overview

1. Testing Hypotheses
2. Experimental Design
3. Method
   - Participants
   - Materials
   - Procedure
4. Results
5. Discussion and Conclusions

Testing Hypotheses

"Immediate Feedback is best!"

Hard to test - we need to be more specific

"Differences in performance on a specific test will be shown between students given no feedback and students given immediate feedback."

= the experimental hypothesis

"There will be no difference in performance shown by students given immediate feedback or no feedback."

= the null hypothesis

Possible Variables

* **Whether or not feedback is given**
* **When it is given** -- immediately? after 3 errors of same type? after certain types of errors? at the end of session?
* **What is given as feedback** -- correct or incorrect; detailed explanation; further examples
* **How much control** does student have over feedback?
* **How long does the student take** to complete a task?
* What is the student's **level of performance**?
* **How does the student feel** about different types of feedback -- which do they prefer? Which do they feel they learn most from? Which helps them learn most quickly?
* **How good are students at estimating** their performance on a task?

Experimental Design

Experimental conditions:

1. immediate error feedback and correction
2. immediate error flagging but no correction
3. feedback on demand

Control condition: to eliminate alternative explanations of the data obtained

- no feedback

Experimental Variables

Independent Variable - manipulated by experimenter

Dependent Variable - not manipulated, but look to see if manipulating the independent variable has an effect on it (but not necessarily a causal relationship)

Independent Variable: *type of feedback*

Dependent variable: *time to complete the exercises; post-test performance*

Keep what is taught constant, so all learners cover the same material

Other factors are **Extraneous Variables** - things that vary without our wanting them to…

Results: Test Scores and Completion Time
(from Corbett and Anderson, 1990)

Mean post-test scores (% correct) and Mean Exercise Completion Times (minutes) for 4 versions of the tutor.

	Immediate feedback	Error flagging	Demand feedback	No tutor
Post-test Scores	55%	75%	75%	70%
Exercise Times	4.6	3.9	4.5	4.5

- We could then compare the sets of scores across

Results: Table 3 from Corbett and Anderson, 1990

Questionnaire 1 Mean Ratings

	Imm fdbk	Error flag	Demand fdbk	No tutor
1. How difficult were the exercises? (1 = easy, 7 = challenging)	4.1	3.9	3.4	2.8
2. How well did you learn the material? (1 = not well, 7 = very well)	5.4	4.6	5.4	5.8
3. How much did you like the tutors? (1 = disliked, 7 = liked)	5.2	4.5	4.8	4.9
4. Did the tutor help you finish more quickly? (1 = slower, 7 = faster)	5.1	4.6	4.7	4.5
5. Did the tutor help you understand better? (1 = interferred, 7 = helped)	5.3	4.9	4.7	4.7
6. Did you like the tutor's assistance? (1 = disliked, 7 = liked)	5.3	5.0	4.7	4.7
7. Would you like more or less assistance? (1 = less, 7 = more)	4.3	4.9	4.5	4.6

Discussion and Conclusions

The effect of tutor type, as measured by post-test scores and mean exercise completion times, is not statistically significant.

- So there would be no evidence in this case that feedback manipulation affected learning.

There were no significant differences among the four groups in rating:

* how much they liked working with the tutor
* how much help the tutor was in completing the exercises
* how well they liked the tutor's assistance
* whether they would prefer more or less assistance

Correlational design

If this study had showed that immediate feedback was best, we might want to follow it up by looking at the relationship between:

* performance on later maths tests
* the amount of time spent using the tutor over the year

Does spending more time on the tutor correlate well with best performance on later tests?

Warning: correlation is not causation

e.g. if it doesn't rain, reservoirs dry out

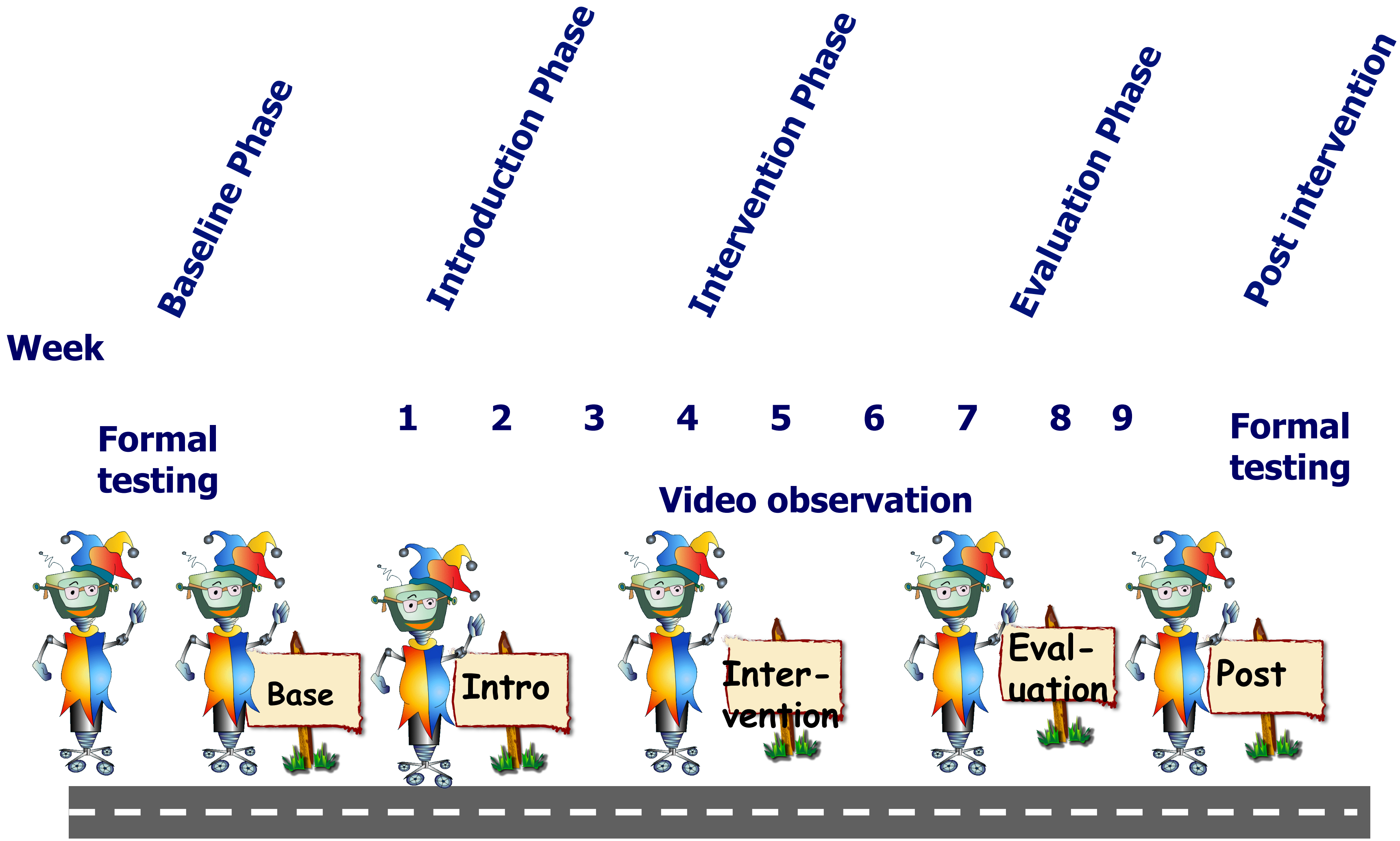
if it doesn't rain, people stop using umbrellas

….. So using umbrellas stops reservoirs drying out? (NO)

A correlation between use of umbrellas and dry reservoirs is likely, but one does not ***cause*** *the other*.

3. Summative evaluation of Standup

Evaluation with children with CCN



The evaluation study

1. 9 participants from independent special school
2. 14 sessions c. 30 minutes over 9 weeks (April/May/June),
3. Consent obtained from parents and children
4. Pre-testing with standardised tests
5. Children shown how to use the software weeks 1 and 2
6. Intervention period exploring software weeks 3 to 6
7. Level of support and guidance reduced, and task complexity increased, as sessions went on
8. Use of system video-recorded for study
9. Favourite jokes stored in paper folder and on AAC devices
10. Evaluation period weeks 7 and 8
11. Further standardised testing
12. Structured interviews and questionnaires for feedback from staff and parents
13. Talking mats to collect feedback from children

Use with typically-developing children March/April 2007

Participants profiles

For all participants: Aetiology: Cerebral Palsy
Mobility: Wheelchair
Literacy: Emerging and assisted

Level	Participant	Communication	Access
Early primary	S1, female; age: 8y4m	Dynavox DV4 user: PCS	**Head switch**
Middle primary	S2, female; age: 10y10m	Intelligible speech: poor articulation	Direct
Middle primary	S3, female; age: 10y9m	Communication book: gross fist & eye gaze	**Head switch**
Middle primary	S4, male; age: 10y3m	Communication Board: PCS, TechSpeak	Direct
Middle primary	S5, male; age: 10y3m	Clear speech	Direct
Senior primary	S6, male; age: 11y3m	Dynavox DV4 user: PCS	**Head switch**
Senior primary	S7, male; age: 12y9m	Speech: poor intelligibility uses PCS	**Head switch**
Senior primary	S8, male; age: 11y10m	Dynavox DV4 user: PCS	Direct
Senior primary	S9, female; age: 11y3m	Intelligible speech	Direct

Evaluation Instruments

CELF Clinical Evaluation of Language Fundamentals (Semel, Wiig, Secord, 1995)

- **CELF Linguistic concepts** (participants are asked to point to…: "the blue line", "the line that is not yellow"; participants must point to a stop sign if they think they cannot do what they are asked to do.)
- **CELF Sentence structure** (e.g. show me…: "The girl is not climbing", "The dog that is wearing a collar is eating a bone")
- **CELF Oral directions** (e.g. point to…: "The black circle", "The last white triangle and the first black square")
- **CELF Word classes** (participants choose two related items from a set of four, e.g. "**girl boy** car table", "slow **nurse doctor** rain")

PIPA Preschool and primary inventory of phonological awareness (Frederickson, Frith and Reason, 1997)

EM tells AL one of ‘her’ jokes Week 3 (intervention)

NI exploring to get 'any joke' Week 8 (evaluation)

Results

Videos transcribed, annotated and analysed:

- Determine task achievement, degree of participant's initiation, response and anticipation
- Good inter-rater reliability
- Transcripts and interview also coded by SLTs

All children benefited

- nearly all able to locate name; exit program; generate and tell, and store and retrieve jokes by end of study
- some participants in exploring system discovered different ways to accomplish tasks and worked out shortcuts
- all gave feedback using talking mats
- reported increase in self-confidence and maturity in all
- carry-over to day-to-day use of AAC
- participants distinguished between generating and telling joke
- joke folders used to tell jokes to others
- jokes liked even when poor

Task Difficulty: progress

	Description	Train	Inter	Eval
A1 A 2	Find name (log onto the system) End program (log off from the system)			
B1 B2 B3 B 4	Generate any joke from new jokes Speak a joke using speech synthesis Save a joke to favourites Choose a joke from favourite s	 P1,3,7,8,9 P5 P2,4,6	 P 7 , 8	 P5 P8
C1 C2 C3 C4 C 5	Generate a joke on specified topic (e.g. about an animal) Generate a joke on a specified sub topic (e.g. about a wild animal) Choose a joke from old joke collection not saved to favourites. Generate a joke of a particular Joke Class Generate a joke by keyword, from topics		P3 P 1 ,2, 4,5,9 P6	P9 P 2 , 7
D1 D 2	Generate a joke by keyword, using alphabet Generate a joke by keyword, typing in wo r d			 P4
E1	Generate a joke appropriate to a current conversation to p i c .			P 1 , 3 , 6,

Pre-post test results

CELF WC: choose 2 related items from 4,
e.g. “girl boy car table”
PIPA Rhyme: Phonological awareness

		CELF word classes (max. 27)		PIPA Rhyme (max. 12)	
Level	**Participant**	**Pre-test**	**Post-test**	**Pre-test**	**Post-test**
Early primary	S1, female; age: 8y4m	**19**	**25**	**10**	**11**
Middle primary	S2, female; age: 10y10m	**11**	**18**	**3**	**3**
Middle primary	S3, female; age: 10y9m	**23**	**26**	**11**	**11**
Middle primary	S4, male; age: 10y3m	**0**	**2**	**10**	**9**
Middle primary	S5, male; age: 10y3m	**17**	**26**	**11**	**11**
Senior primary	S6, male; age: 11y3m	**1**	**4**	**1**	**8**
Senior primary	S7, male; age: 12y9m	**17**	**24**	**12**	**11**
Senior primary	S8, male; age: 11y10m	**9**	**8**	**5**	**3**
Senior primary	S9, female; age: 11y3m	**12**	**13**	**10**	**11**

CELF scores significantly higher on post-test (t-test, 8df, $p < 0.01$)

Results: Feedback

Unexpected Outcomes impact on school curriculum

Questionnaires with parent, teachers and Classroom assistants (not significant issues raised but all positive)

Semi-structured interviews with SLTs

Participant Feedback using Talking Mats S1

Good:

Jester character

Way screen changes

Way of telling jokes

OK

Jokes

Scanning

Bad

Voice

Bad OK Good

Participant Feedback using Talking Mats S8

Good:

Jester character

OK

Touchscreen

OK/Bad

Way screen changes

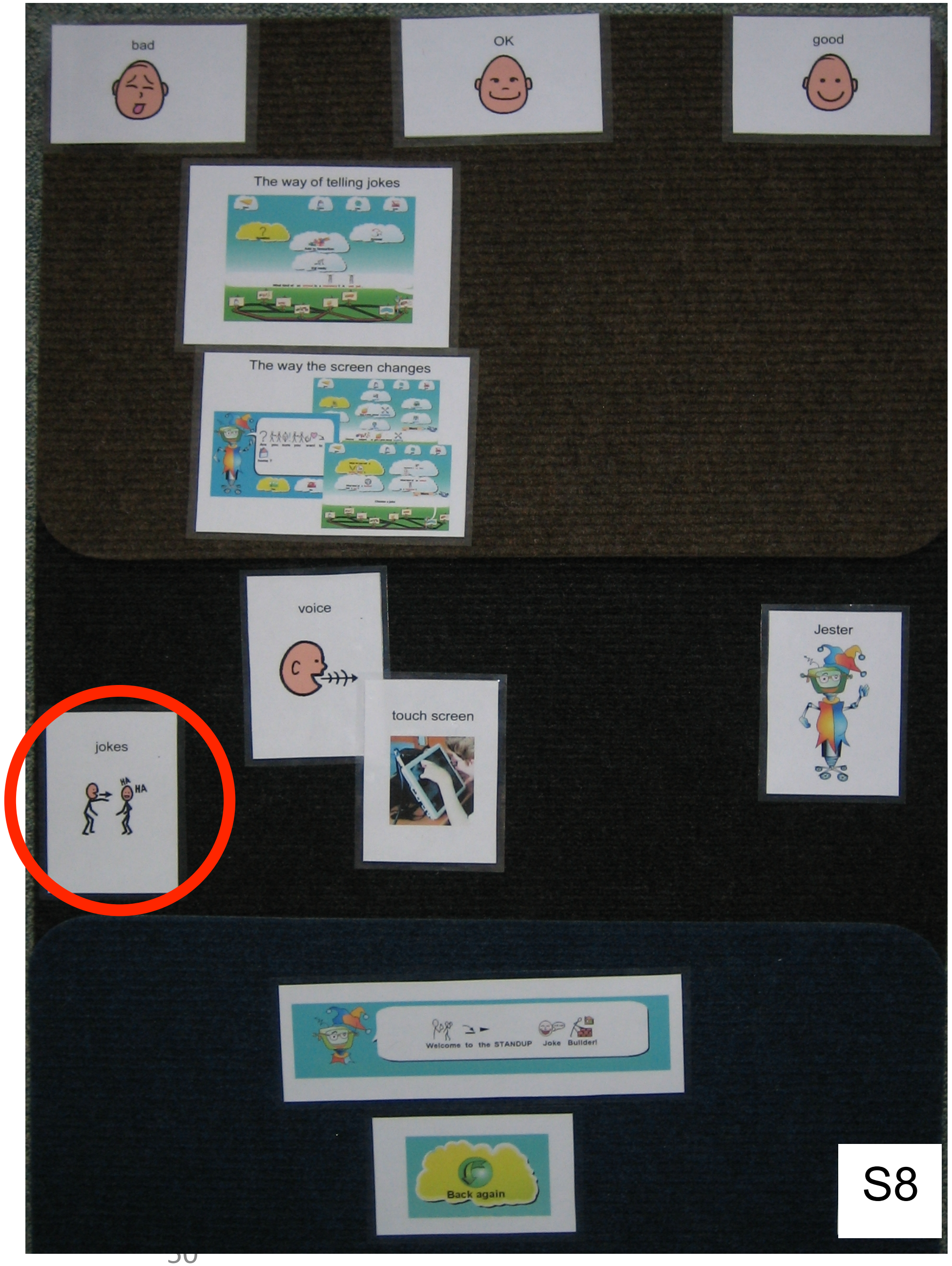
Way of telling jokes

Voice

Bad

Jokes

Bad OK Good



STANDUP: some initial conclusions

Interfaces CAN be designed which provide children with CCN with successful access to complex underlying technology

Using STANDUP:

- the generative capabilities allows opportunity for natural language development, cf DA choosing punchline first
- the generative capabilities allows novel explorative learning, cf NI searching subjects

All children benefited

- enhanced desire to communicate
- knock on effect on other AAC usage
- illustrated children's abilities and potential of AAC

Illustrated use of technology within a wider environment

STANDUP: some initial conclusions

Issues with interface design

- scanning
- voice output
- improved appropriateness of vocabulary

The telling of the joke is important - what is the impact of STANDUP:

- on interactive conversation?
- on joke comprehension and vocabulary acquisition?

Do we want better jokes? (yes)

Use with speaking children with language impairment and other user groups

4. Writing up Experiments and Empirical Studies

Writing-up empirical studies 1

Abstract:

Short summary of the problem, the results and the conclusion.

Introduction:

What is the problem? What related work have other people done?

[Should go from general statement of the problem to a succinct and testable statement of the hypothesis].

Method:

Participants: state number, background and any other relevant details of participants

Materials: exactly what test materials, teaching materials, etc. were used, giving examples

Procedure: clear and detailed description of what happened at each stage in the experiment

[Someone reading should be able to duplicate it from this information alone. Should also clearly indicate what data was collected and how.]

Writing-up empirical studies 2

Results:

Give actual data, or a summary of it.

Provide an analysis of data, using statistical tests if appropriate.

Use tables and graphs to display data clearly.

[Interpretation of results goes in discussion section, NOT here].

Discussion:

Interpretation of results; restating of hypothesis and the implications of results; discussion of methodological problems such as weaknesses in design, unanticipated difficulties, confounding variables, etc.

Wider implications of the work should also be considered here, and perhaps further studies suggested.

Conclusion:

Statement of overall conclusion of the study.

5. References

References - Methodology

Cohen, P. (1995) *Empirical Methods for Artificial Intelligence*, MIT Press, 1995.

Corbett, A.T. and Anderson, J.R., (1990) The Effect of Feedback Control on Learning to Program with the Lisp Tutor, *Proceedings of the 12th Annual Conference of the Cognitive Science Society*, LEA, New Jersey, 1990

Dix, A., Finlay, J., Abowd, R. and Beale, R. (2004) *Human-Computer Interaction.* Prentice Hall

Preece, J., Rogers, Y., Sharp, H., Benyon, D. Holland, S. and Carey, T. (1994). *Human-Computer Interaction*. Addison-Wesley

References – various studies

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.

References

Cohen, P. (1995) *Empirical Methods for Artificial Intelligence*, MIT Press, 1995.

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. 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.

Conlon, T. and Pain, H. (1996). Persistent collaboration: a methodology for applied AIED, *Journal of Artificial Intelligence in Education,* 7, 219-252.

Conlon, T. (1999). Alternatives to Rules for Knowledge-based Modelling. *Instructional Science* Vol 27 No 6, pp 403-430.

Corbett, A.T. and Anderson, J.R., (1990) The Effect of Feedback Control on Learning to Program with the Lisp Tutor, *Proceedings of the 12th Annual Conference of the Cognitive Science Society*, LEA, New Jersey, 1990 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.

References

Dix, A., Finlay, J., Abowd, R. and Beale, R. (2004) *Human-Computer Interaction.* Prentice Hall ***(Evaluation chapter in particular)***

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.

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.

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.

References

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.

Luger, G. F. and Stubblefield, W. A., (1989) *Artificial Intelligence and the Design of Expert Systems,* Benjamin Cummings, 1989.

MacLaren, & Koedinger, K (2002): When and Why Does Mastery Learning Work: Instructional Experiments with ACT-R "SimStudents". ITS 2002 355-366

Mark, M.A. and Greer, J.E. (1993). Evaluation methodologies for intelligent tutoring systems, *Journal of Artificial Intelligence in Education,* 4, 129-153.

Mark, M., & Greer, J. E. (1995). The VCR tutor: Effective instruction for device operation. The Journal of the Learning Sciences, 4(2), 209-246.

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).

Murray, T. (1993). Formative Qualitative Evaluation for "Exploratory" ITS research. Journal of Artificial Intelligence in Education, 4(2/3), 179-207.

References

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.

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.

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