



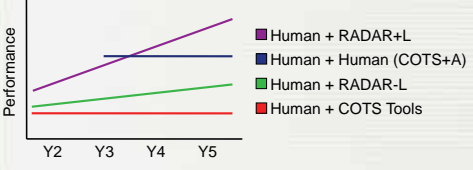
Evaluation

Faculty: Dan Siewiorek & Aaron Steinfeld
Staff: Rachael Bennett, Matt Lahut, Pablo-Alejandro Quinones, Jialiang Wang

Roy's 4-Point Scale

- Validity
 - Internal: Controlled experiment
 - Between-subjects with common participant pool
- Repeatability
 - Common control condition (COTS)
- Reproducibility
 - Shared materials & tools
- Reporting
 - Thorough documentation

RADAR Goals (call your shot)







1. In Year 2 RADAR will beat humans using conventional tools
2. RADAR+L will always outperform RADAR-L, showing that learned knowledge plays a significant positive role in performance
3. In Year 3 a human with a human assistant might beat RADAR+L, but by Year 4 RADAR+L will outperform a human assistant

Evaluation Scenario

The organizer for an academic conference mysteriously vanishes...

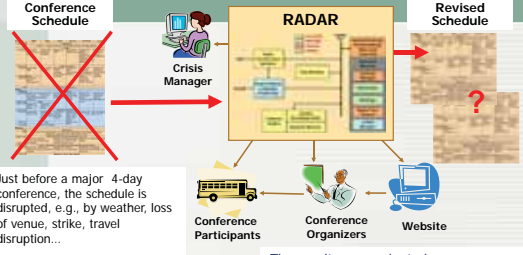
But his email 'inbox' has been recovered, intact. Test subjects take over his job using either RADAR or equivalent COTS tools. Then things start going wrong...

Subjects are scored on a range of tangible work products:

			
Score Weight: 66%	16%	16%	

Criteria: completeness, accuracy, conformance, optimality,...

RADAR Conference Planning Task



Just before a major 4-day conference, the schedule is disrupted, e.g., by weather, loss of venue, strike, travel disruption...

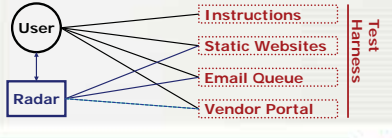
The user (and RADAR) must assess the situation, develop a revised plan, get the word out about the new plan, and deal with queries and requests during the crisis.

The results are evaluated on:

- Quality and completeness of the new plan
- Successful completion of related tasks
- Costs of the solutions

Test Harness

- External to Radar and usable manually in COTS condition
- Non-trivial, realistic, interlocking simulated world designed to show off impact of learning
 - Static websites: building details, instructions, etc
 - Vendor e-commerce site
 - Email stack (stimulus and existing vendor orders)



Working & Realistic Environment

Session A3: Cooperative Cruise Control
 String Stability for 20-Car Cooperative Cruise Control Platoons
 ACC vs. CCO: Monte Carlo Simulations of String Stability
 Modulated LED Tail Light Inter-Vehicle Communications
 Wireless Requirements for Cooperative Cruise Control

Corpus: Real vs. Simulated

- Real is ideal
- Moving pieces, Institutional Review Board (IRB), and core questions may not permit
 - Privacy & Identity
 - Wrong problems for experiment
 - Tying to other tasks (e.g., vendor site)
- Simulated
 - Hired English majors
 - Provided script outlines and character bios
 - Thorough review for adherence and quality

----- Original Message (SIGNAL) -----
 From: jlee@ardra.org [mailto:jlee@ardra.org]
 Sent: Monday, January 30, 2006 8:27 PM
 To: bor@cs.cmu.edu
 Subject: correct title

Please change my first name from Jun to "Dr. Jun" on the conference website. I am a new PhD and I want to be sure that it is clear I have finished my degree. Thank you very much.

Dr. Jun Lee

----- Participant Response -----
 It still doesn't make you a good person.

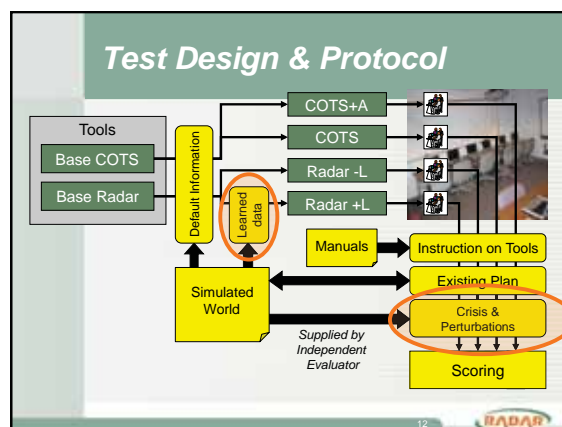
----- Original Message (NOISE) -----
 Ha--you wanna go? It's in Philadelphia this year--it could be a flashback to college when we roadtripped to D.C. for that Star Wars exhibit (except we don't have to sleep in the car this time). We could give Jim a call and see if we could crash with him and Martha for a night. Let me know if you're interested!
 Mark (a.k.a. darth shut-the-hell-up-and-drive)

----- Participant Response -----
 I guess I'm down to go, as long as it gets me out of planning this damn conference. Although, given the choice between going to the Sci Fi Convention and repeatedly beating myself in the face with the business-end of a claw hammer, I'd probably take the hammer.

Peace.
 Blake


Stimuli Must be High Quality

- Calendar Study
 - Stimuli did not match constraints
 - Slot from 5-6 was open
- Pilot test to find holes
 - GIGO (garbage in, garbage out)
 - Allot recovery time in your schedule
 - Look for software problems (e.g., 15 minute launch times, sub-modules dying silently)
 - Subtle data & code bugs



Learning Data

- Large quantity of email loaded into the system
 - No crises, only everyday tasks
- “Experts” use RADAR to process these activities
 - Trained project members
 - Not allowed to train their own component
- Some training in parallel, some serial

13 


A New Marble Every Year

- Third party evaluator compiled the email stack, associated crisis, and constraints
 - Based on negotiated boundaries (e.g., can only wipe out X% of rooms)
 - Pilot tests to try out impact of specific ideas
 - Pilots restricted to small group (lead evaluator and integrator, key software programmers)
 - Initial delivery a week in advance to key personnel & sanity checked for errors and boundary violations
 - Loaded into the test a day or two in advance (final software check)
- Major crisis with widespread ramifications
 - E.g., Primary building unusable
- Perturbations
 - Many short, acute injected problems/constraints
 - E.g., exhibitor requests briefing, keynote speaker requests roses, etc
- Other subtle constraints
 - Room availability, instructional material

14 

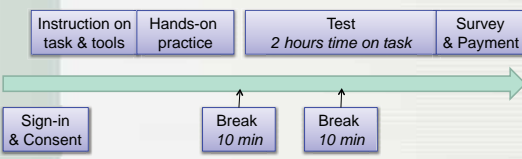
Participant Recruitment

- University-wide site
- Post days in advance
 - Time of day
- Enrollment variable
- Dropouts
 - Over-recruit
 - Small payments for extra people (1st half hour)




15 

Session Timeline




- Instruction and practice with similar, but not identical stimuli
- Breaks as a group

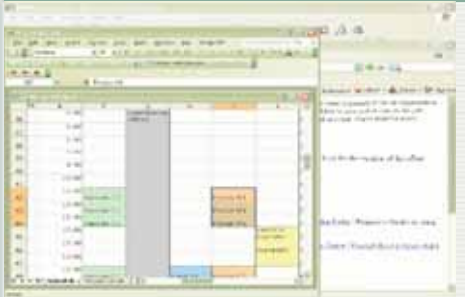
16 


Training & Manuals

- Trained from the manual
- No note taking allowed during training
 - No handwritten notes allowed during test so notes would be captured in software for post processing
- Example task movies also in manual
 - Important for complex and unusual tasks
 - Keep movie lengths consistent across conditions (no time bias)

17 

Example Movie, COTS



18 

Example Movie, RADAR+L



19

Participant Motivation

- Milestone payments not usually necessary
- Important for:
 - *Hard* tasks
 - Risk/reward studies
 - Priming specific behavior
 - Multi-day studies
- In RADAR
 - Three milestone payments totaling up to \$20
 - Schedule, briefing, website updating
 - Thresholds specific to each tool in order to be fair

20

-----Subject Briefing-----

- * Made all necessary contact information and presentation changes to website
- * Ensured that participants would be able to contact each other
- * Ensured that at least one task was appropriately completed
- * Ruined the rest of the conference
- * Delayed preparations
- * Failed in securing room
- * Quit Job
- * Too Embarrassed By Failure
- * Unable to appropriately multitask
- * Ruined Position and credibility of Blake

21

Excluding Participants

- Watch during sessions
 - Sleeping, very low computer skills, violation of entry criteria (e.g., fluency), “seat-filling”
 - Straying out of bounds (e.g., dating sites)
- Take notes
 - *Alert/Exclude*: Context for flag
- When to exclude
 - Pull immediately if biasing others with behavior
 - Keep till end and set data aside
 - Triage each *Alert* case as a team

22

Looking for Outliers

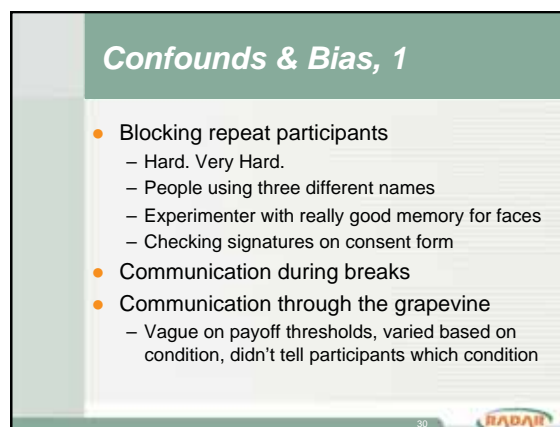
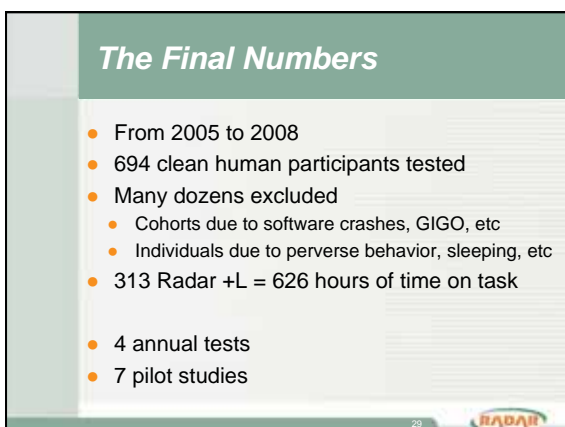
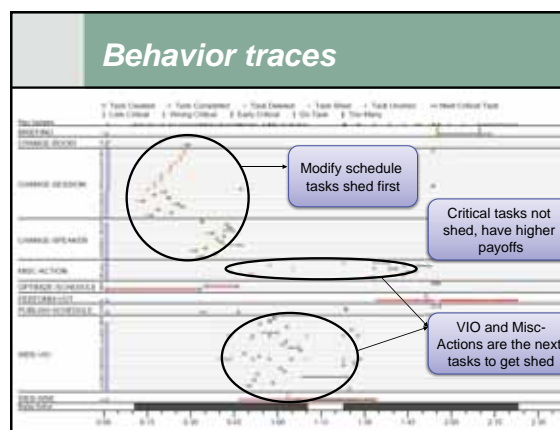
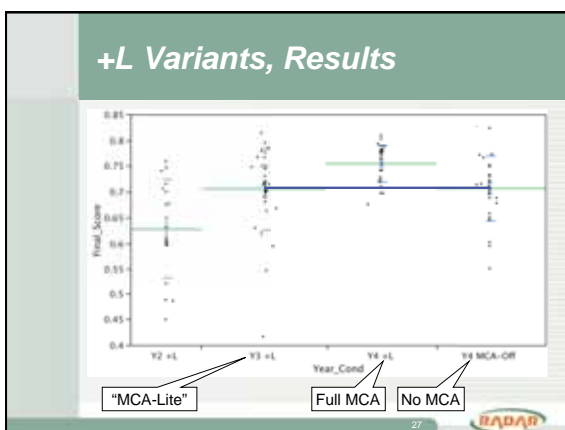
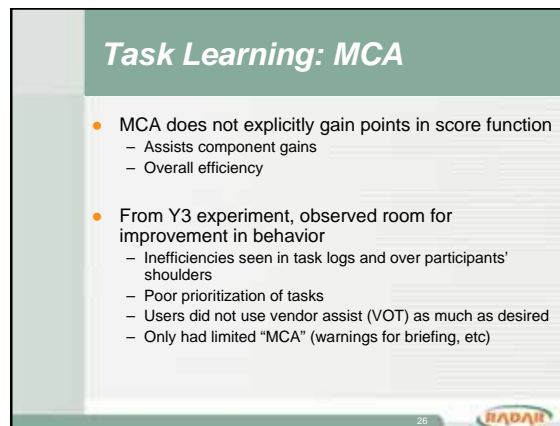
- Some sneak past the flags and should be excluded
 - Perverse behavior (e.g., bizarre banquet orders)
 - Random clicking
 - Objective scores many sigma out
 - Both positive and negative

23

Role of External Evaluators

- Prepared stimuli
 - Verify not training/coding to the test
- Final say on participant exclusions
 - Verify not cherry picking
- Independent statistical analyses
 - Verify no funny business
- *Note*: external evaluators not necessary for most studies

24



Confounds & Bias, 2

- Participant pool
 - Vacations, classmates, coming as a group
 - Multiple cohorts
- Confirming repeatability
 - Common control condition (COTS)
- Time & Day
 - Tried to balance sessions for day/night
 - Days & class schedules

31



General Lessons, 1

- Humans are like water: if there is a hole in the system, they'll find it
- The Robots Institute mantra holds true
 - Test, test, and test before the real test
- Never underestimate the value of good user interaction design
 - Corollary is true too: bad UI can kill a good system (and put a program on life support!)



General Lessons, 2

- Even after research no longer cutting edge, a paper detailing good, reusable methods will be regularly cited as a methods paper
- Sister tech report with a lot more details than present in peer-reviewed paper
- Sharing tools and stimuli
 - <http://www.cs.cmu.edu/~airspace/>
 - Stimuli useful for other types of studies

33



-----Subject Briefing-----

Subject: RE: Brief me, please

I wear boxers, not briefs but here is the brief anyways:

*I have rescheduled all the people who are coming in late or would like to have an earlier time
 *conflicting sessions of interest have been put at alternate times to assuage the fears of those who wish to see both
 *attendance for all events has been modified to fit the newest data
 *I hereby quit being your conference planner and hope that Ardra goes down with you, the captain, faster than the Lusitania



Test Plan

- Before subjects: RADAR wargamed for +L condition
 - E.g., Over 750 email messages prepared for Classifier training
- 2 cohorts of 15 human subjects per day (3hr each)
- Instruction on tools (no hands-on experience)
- Post-instruction quiz
- Inbox has unread crisis email stack (107 messages)
- Backstory email in separate IMAP folders
 - ~30 high value emails from corpus in a folder
 - ~80 emails for 50 original vendor orders in a folder
- Subject works the problem for 2 hours
- User experience exit survey

35




36




Test from Subjects' POV

- Exposure to manuals, tools, and vendors during Instruction
- Backstory, injections, crisis, and injections in mailbox at start
- Websites, backstory, and manuals used for situation awareness
- Conference schedule adjusted (Excel; STP GUI)
- Room availability investigated and reservations obtained (manual website/automatic)
- Existing vendor orders modified and new ones placed
- Website updated
 - Publish schedule (manually; WbE)
 - Other web changes (manually; VIO and manually)
- Random requests handled (e.g., maps, food restrictions, etc)
- Briefing compiled (manually for all conditions)

37 

Metrics


- Final_Score
 - Schedule, with cost and bump penalties
 - Website updating
 - Briefing
- Post-test survey
- A lot of low level logging data
 - Raw performance data: constraint satisfaction, etc
 - Behavioral data: tool use, task selection, etc
 - Should be easier to collate and analyze in Y3

38 

+L Variants, Rates (%)

- Saturated optimize & publish
- Higher briefing rate, lower quality

Condition	Optimize	Publish	Brief
Y3 +L	97	92	94
Y4 +L MCA-off	100	96	89
Y4 +L	100	100	100

39 

+L Variants, Vendor Repair

- MCA seems to improve VOT use
- Money spent on unused vendors
 - Y3 +L: \$27,400*
 - Y4 +L MCA-off: \$36,400*
 - Y4 +L: \$16,400*
 - (0 is ideal)

40 