The Evaluation of Adaptable Multimodal System Outputs

Thoughts for the future

Erin Panttaja

David Reitter

Fred Cummins

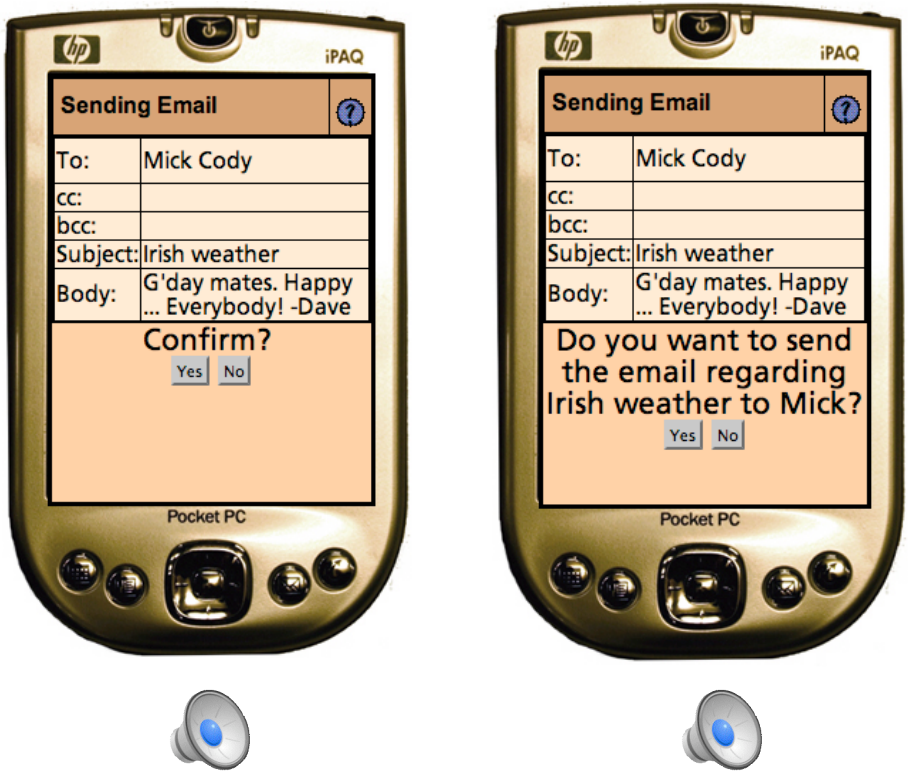
Media Lab Europe
European Research Partner of MIT Media Lab

Coordinated Multimodality

- What is it?
  - Input, output
- Why bother?
- Issues:
  - Dynamic production
  - Natural language

MUG system

- FASiL Consortium
- Multimodal Unification Grammar
- Adaptable email client
- Not a dialogue manager



Testing multimodal systems

- It's hard!
  - Especially when you get into adaptive systems
- What does it mean to be good?
  - Meet the specification
  - Accessible
  - Enjoyable
  - Helpful

Beyond quality

- Acceptance
- Experience
- Users don't like changing paradigms
- But sometimes they surprise you

Recent work in evaluation

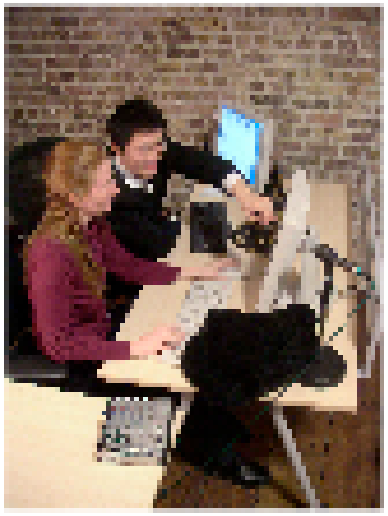
- Experiments for design (Feiner + McKeown)
- Full user-based testing can only be done with a full system.
- Cognitive walkthroughs (Lewis et. al.)

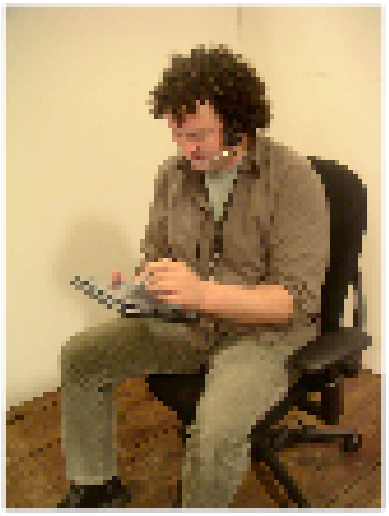
Testing

- Qualitative versus quantitative (Maybury and Wahlster)
  - User perceptions
  - Time to perform, accuracy, percent agreement of systems
- Direct versus indirect metrics
  - Success, time to complete
  - Walking speed, ability to do outside tasks (Pirhonen)
- Heuristic evaluation/rules of thumb (Cockton et. al.)

Modeling the user

- Simulated user
- Wizard of Oz Testing
  - Wizard of Oz Operating System (WOzOS)

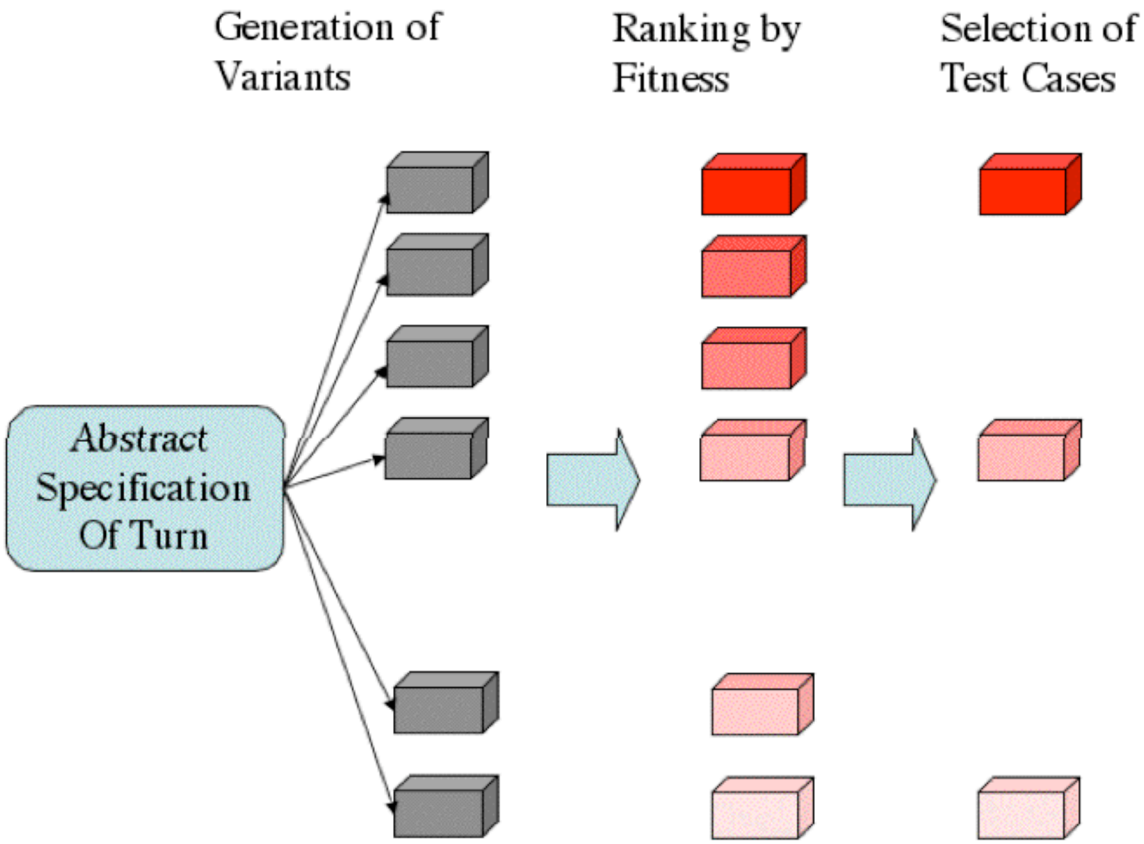




Scales

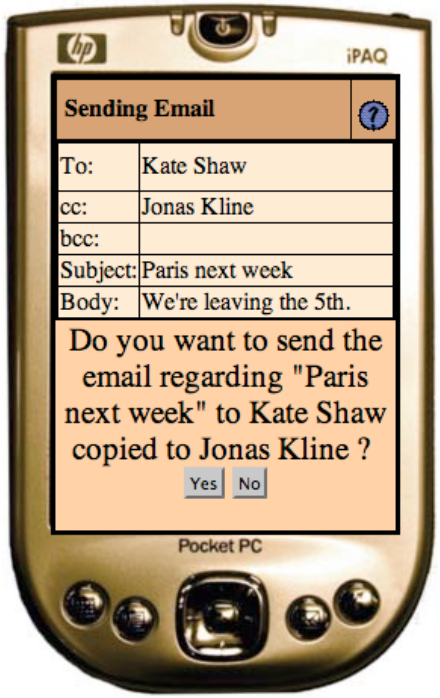
- COMFORT (Knight et. al.)
  - Emotion, attachment, harm, perceived change, movement, anxiety
  - Mobile systems
- NASA/TLX (Hart and Staveland)
  - Mental demands, physical demands, temporal demands, own performance, effort, frustration

Fitness Functions



MUG

- Good/bad
- Our fitness function
  - Cognitive load
    - Reading time
    - TTS time
- Components
  - Compositional
  - Required data

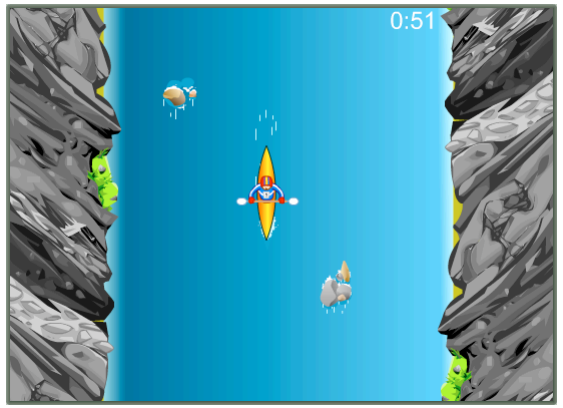


good

bad

Our Experiment

20 users
8 interactions
Half distracted (per user)
Half of the interactions good (overall)
Half of the interactions had errors (overall)



To: Kate Shaw
From: Susan Smith
Subject: Paris next week

We're leaving the 5th.

Does this represent the email that is going to be sent now?
Yes No

Did the computer in this dialogue seem...

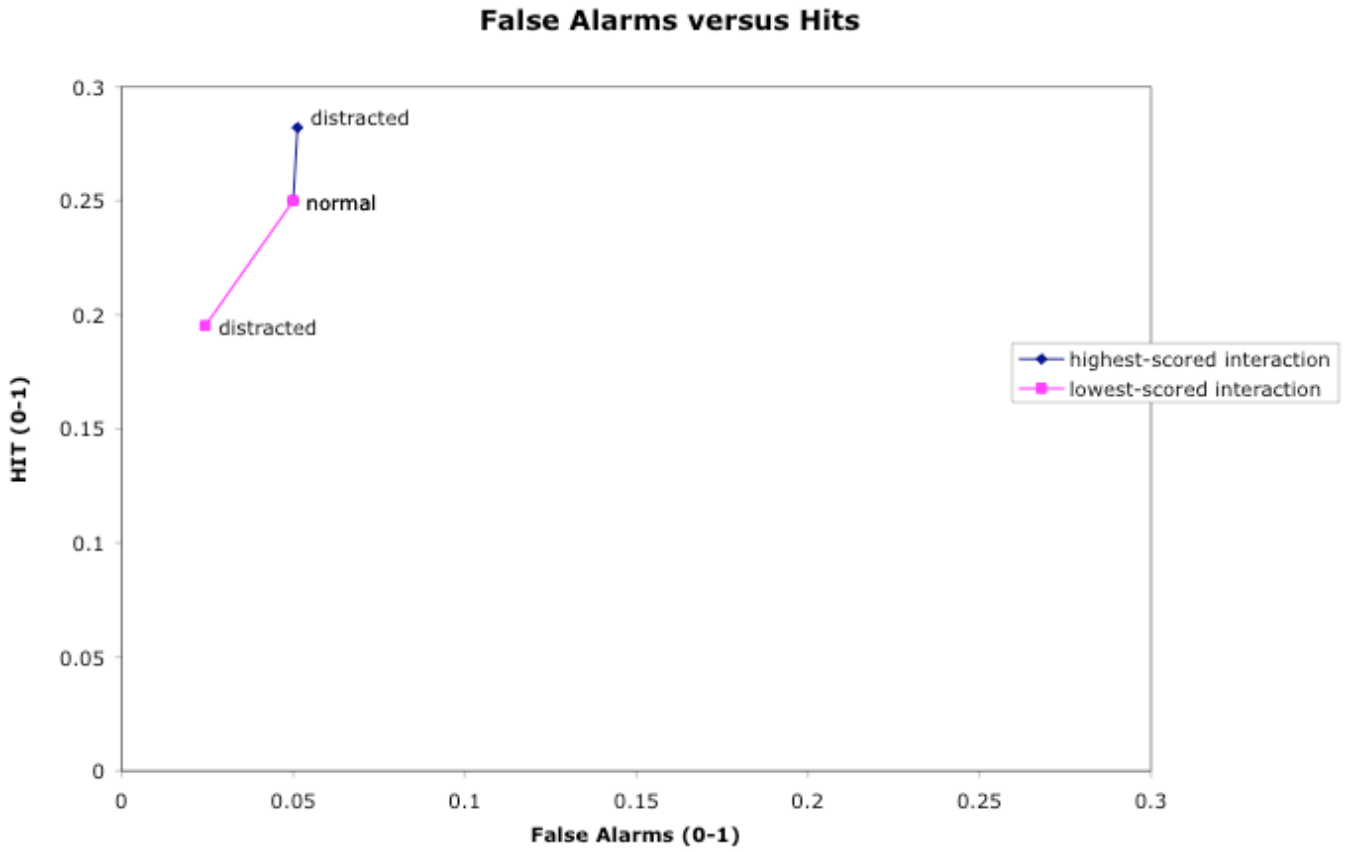
(inefficient) (efficient)

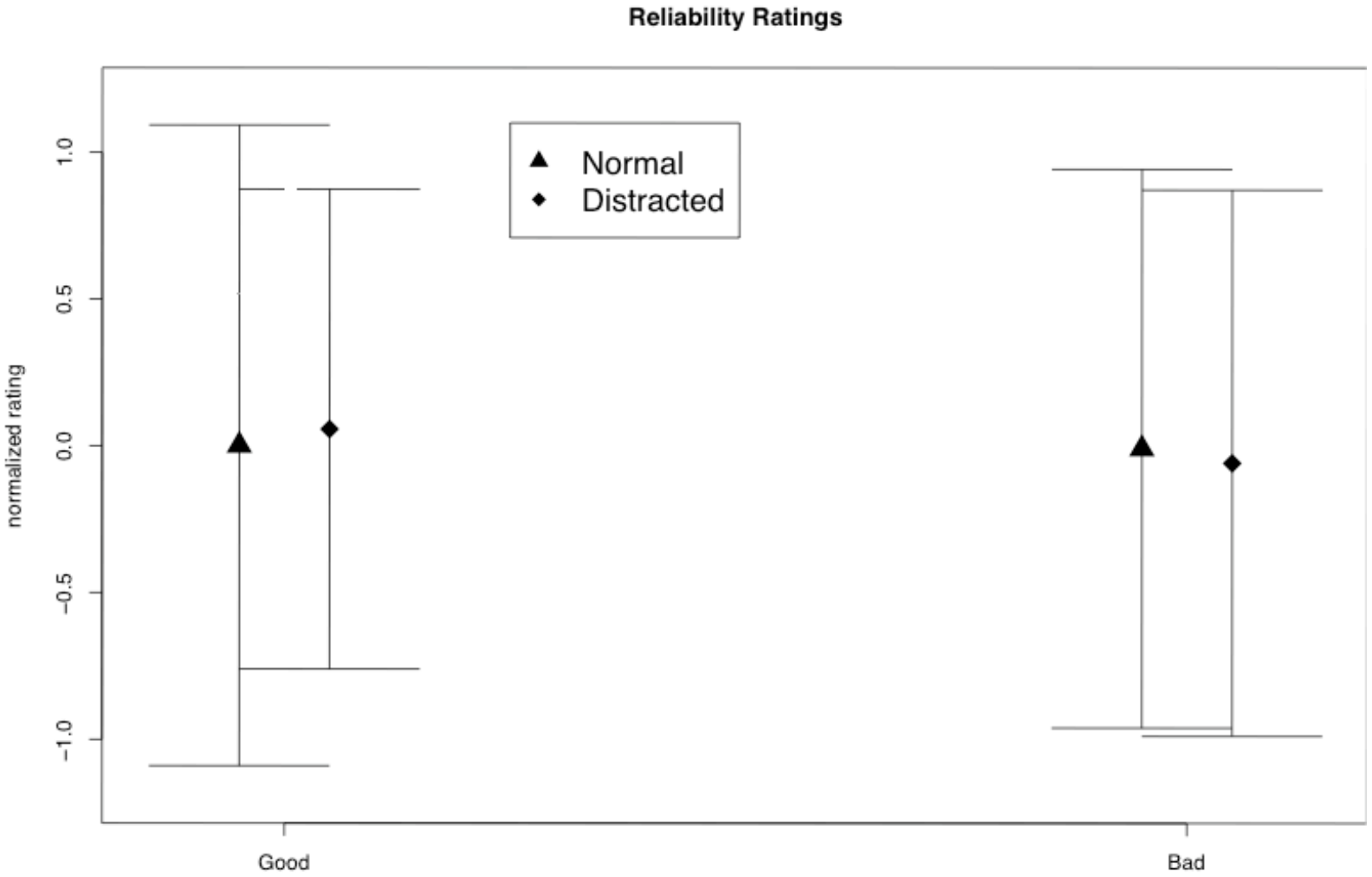
(reliable) (unreliable)

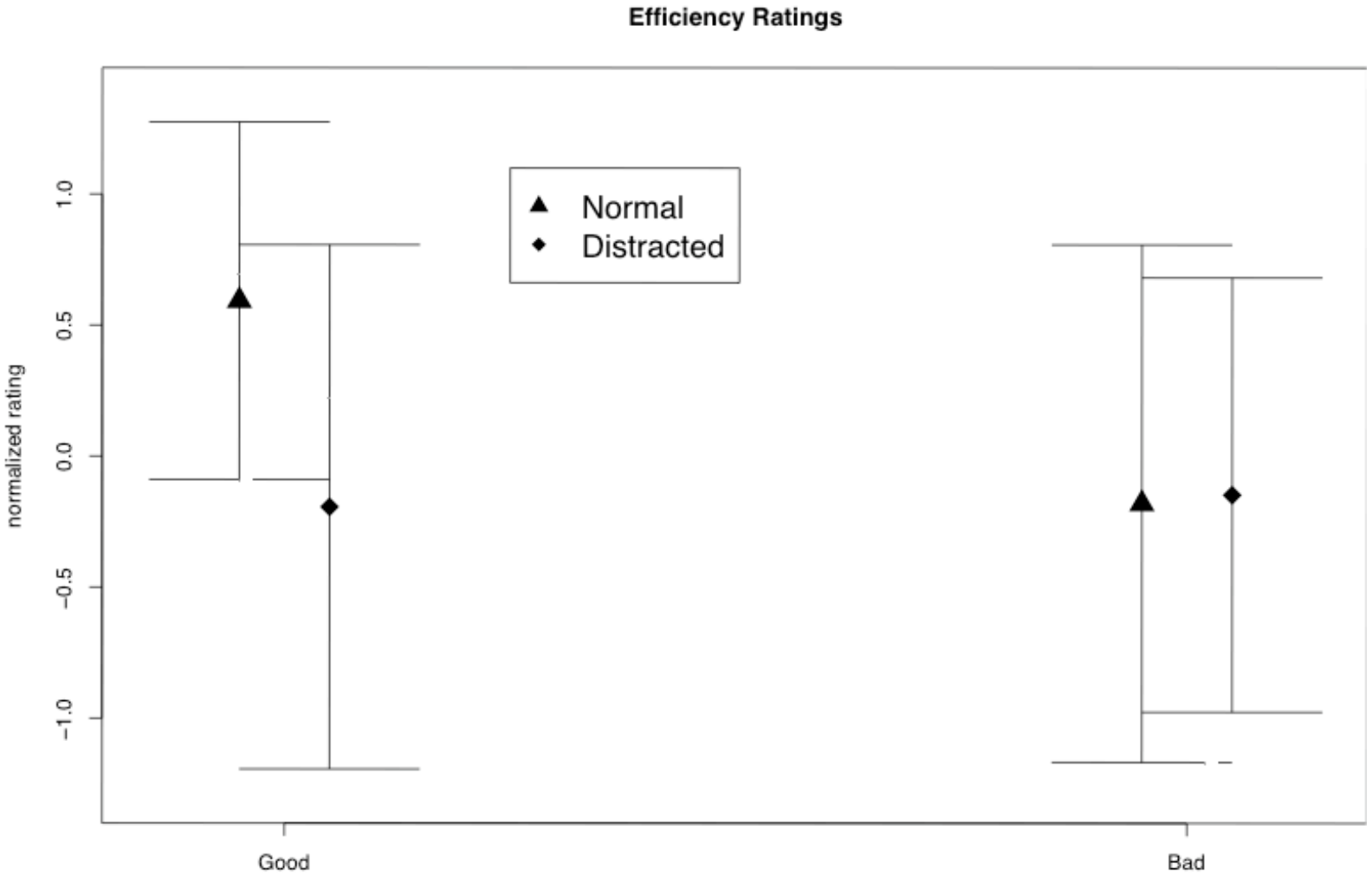
Finished

Evaluating Adaptable Multimodal System Outputs
August 29, 2004

Media Lab Europe
European Research Partner of MIT Media Lab


False Alarms versus Hits
HIT (0-1)
False Alarms (0-1)
0.3
0.25
0.2
0.15
0.1
0.05
0
0
0.05
0.1
0.15
0.2
0.25
0.3
distracted
normal
distracted
highest-scored interaction
lowest-scored interaction


Reliability Ratings
Normal
Distracted
normalized rating
1.0
0.5
0.0
−0.5
−1.0
Good
Bad


Efficiency Ratings
normalized rating
1.0
0.5
0.0
−0.5
−1.0
Normal
Distracted
Good
Bad

This is *hard*.

- Web-based
  - More convenient, but harder to control
- Audio
- Timing (network delay)
- Coordination of a distraction task
- Need a way to do it without a dialogue module

Open Questions

- What makes a good evaluation?
- How can we evaluate parts of the system in isolation?
- Is any testing other than a full user trial really meaningful/predictive?

Supplemental

- GOMS (Kieras)
  - Goals Operators Methods and Selection Rules
- SUPPLE (Gajos and Weld)
  - Adaptable
  - Experts evaluate autogenerated and hand-generated systems