Understanding and serving users

Cognitive walkthroughs

User testing issues

Surveys, interviews and protocols

Cognitive walkthrough

An alternative evaluation method
for you to apply

Cognitive Walkthrough (CW)

- Alternative expert-based evaluation method
- Assumes learning by exploration on part of the user
- Not based on interface heuristics
- Applied to detailed design specifications or prototypes

Simple Model of CW

- What is the user trying to do?
- What actions does the interface support?
- Do user's intentions suggest which actions to take with the interface?
- Does the interface provide feedback suggesting the user's goal is closer?

Cognitive Walkthrough in 5 stages

- 1 Define the inputs
  - Who are the users
  - Sample tasks
  - Action sequences for completing the tasks
  - What does the interface look like at each stage?

- 2 Start the walkthrough (the easiest stage!)

Where to get these inputs:

- Design team must agree and articulate a clear definition of the user
- The tasks for evaluation should be selected as relevant, appropriate and typical for these users

Stage 3 - The analysis activities

- Walkthrough the action sequences required
- Tell a credible story considering...
  - Will the user try to achieve the desired result?
  - Will the user notice what is the right action?
  - Will the user relate the right action to their goal?
  - Will the user see the progress being made?

How well can you do this?

- You need:
  - Insights into user thinking and reasoning,
  - Ability to empathize with user expectations
  - Knowledge of interaction variables that matter to users
  - A willingness to explore the ‘what-ifs’ of user experiences

Stage 4 - Record critical information

- What does the user have to know?
- What assumptions are we making about the user population?
- What should user learn while performing task?
- What side issues emerged?
- Develop credible success or failure story

Stage 5 - Revision

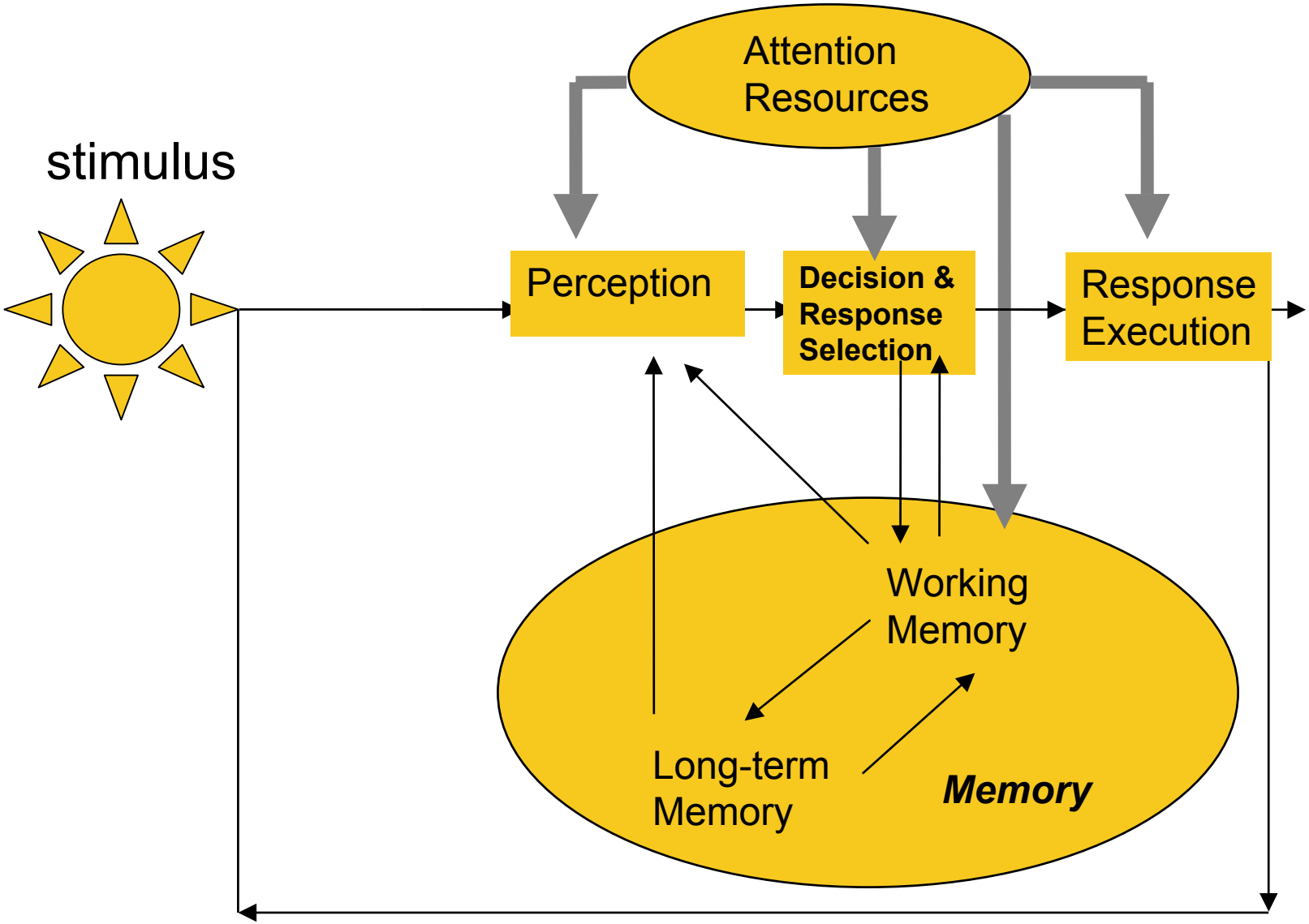
- Re-design and start again!
- Change assumptions about users and their tasks?

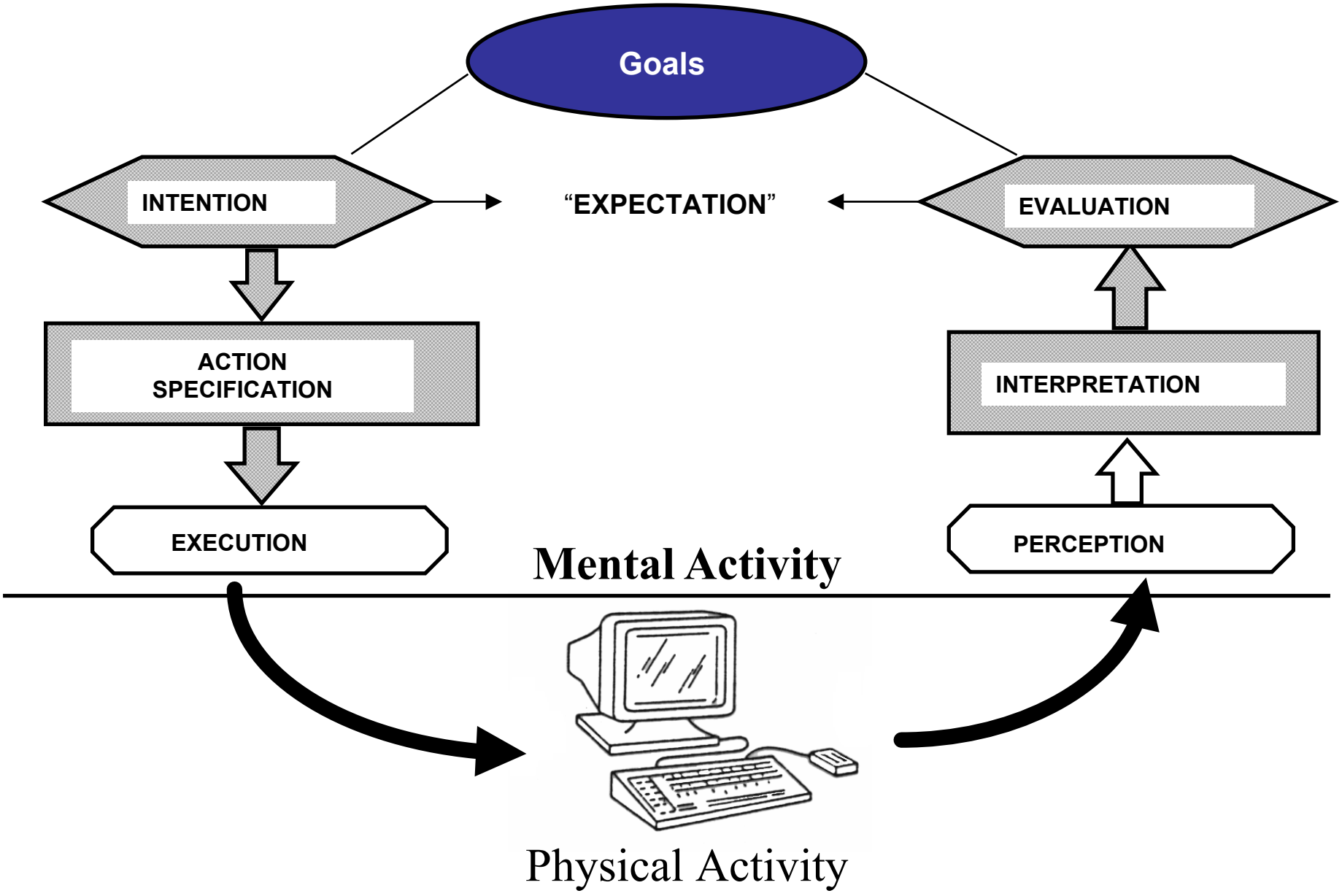
OUTPUT of CW

- Step by step description of the task to be performed
- Success or failure ‘story’ for each step
- Failures listed for possible re-design

Assumptions for use

- Usability analyst must understand cognition
  - human information processing
  - skill acquisition
- Distinguish GOALS from ACTIONS
- Stay on track: Always assume that the step you are currently evaluating is the right one to be at and don't follow errors.


Attention
Resources
stimulus
Perception
Decision &
Response
Selection
Response
Execution
Working
Memory
Long-term
Memory
Memory



Seven stages of user activities involved in the performance of **a task**

Don Norman (1987) ***The Design of Everyday Things.***

Task analysis questions

- What does the user see?
- What decisions does a user make?
- What must a user know?
- How does a user get help?
- How does a user recover from errors?
- What physical acts must be accomplished?

Team walkthroughs

- Given evaluator performance variation, use more than 1 evaluator
- Pool results for comparison
- Check assumptions and failure stories
- Quick version: use several evaluators at once, in real time.

Team ‘jogthroughs’

- Have several evaluators perform the CW collectively
- Discuss impressions in real time
- Fast, but prone to group dynamic effects
- Useful for bringing non-evaluators into the process

Streamlining CW (Spencer 2000)

- Social constraints exist in real evaluation contexts:
  - Time Pressure
  - Lengthy design discussions (instead of evaluating)
  - Design defensiveness

Changing the questions

- Streamlined CW asks only 2 questions
  - Will the user know what to do at this step?
  - If user does right thing, will they know that it was the right thing and are making progress?
- These carry most of the points raised in CW standard form

Ground rules for streamlined CW process meeting

- No designing
- No defending a design
- No debating cognitive theory
- Session is led by usability specialist

Tips for good Walkthroughs

- Step 1 is vital - without it, the CW is useless or misleading
  - Relate this to user-centered design theory
- Offer detailed justifications for success or failure stories
- Focus on factors that influence task flow
- Make suggestions for re-design
- Use more than one evaluator

Possible problems with CW

- Evaluator unable to place self in position of user
  - Evaluator labeling every transition a ‘success’
- Evaluator failing to justify success or failure story
  - “I just don’t like it” or “it works for me” is not useful design input
- Finding relevant tasks that real users do

Output

- Story for each stage of task
- Highlight problems
- Rate severity of problems
- Suggest re-design
- Identify side-issues noted

Case study - John & Mashyna (1997)

- Single evaluator
- Working alone
  - Introductory lecture on methods and HCI
  - Researched technique
  - Taught self how to apply it
  - Evaluated specification of multimedia system
  - Output from CW compared with user tests

Time commitment

- 3 hrs preparatory reading
- 4.5 hours on preparatory task analysis
- 3.5 hours to do “short task” walkthrough
- 2.5 hours reading/analysis for “long task”
- 16 hours doing long walkthrough including write up

Quote

"the overall sense of the timeline is that 'A1' steadily progressed from being very dependent on the documentation for CW to doing the analysis without reference to it" (p. 136)

Output

- 46 unique problems, 6 duplicate problems
- A1 reported:
  - 61% of problems found by CW method
  - 12% as a side-effect
  - 15% from reading the spec
  - 10% - other (?)
- Problem severity: ‘trivial’ to ‘must change’
  - 5% must be changed, 49% very serious, 29% moderately serious

Reported difficulties using CW

- Initial worry about expertise required
  - eroded by later in the evaluation
  - little theory needed by evaluator
- Concern with methods relevance to tool
- Concern with generating relevant tasks
- Tedious activities required for completion
- Might miss global issues by locking on task
- Estimating frequency and severity of problem

CW and User test compared

- John performed user test (on 4 users) to compare results
- A1's CW predicted 5% (2/37) of observed user problems
- A1's CW predicted 15 problems not found in user test

So?

- Selection of tasks is important
  - remember the contextual determination of usability
- CW may focus too much on local issues
- Expert found two problems in A1's CW
  - A1 was assumed too much knowledge on the part of the user
  - A1 too strict in identifying failures in one feature

Intro to user testing

- User data collection methods
  - Behavioral data
  - Think-alouds
  - Interviews & Surveys
  - Planning and running a user test

Before running a user test, establish:

- Scenario of use:
  - users, tasks, environments
- Procedural issues
  - How, where and by whom will the evaluation be run
- Question, Data, Method:
  - What are the “Dependent Variables”?
  - How will data be observed/captured?
  - Will these data answer your question?

Then:

- Get access to as many relevant types of user as possible
- Collect background data on users
  - Including previous test experience
- Allow users to interact naturally
- Let the design team see/know what is happening
- Use surveys, interviews and questionnaires to capture data after interaction
- Try to combine questions in one study

LAW: Develop a test protocol

- Ensure all users are treated identically
- Ensure others can follow test method and replicate it if necessary
- Provide consistent response for possible deviations from the test plan
- Clearly state the dependent variables being measured

LAW: Pilot your test

- All test designs yield unexpected results
- Be sure you are getting the data that answers your question
- Change the test design as indicated by the pilot

General classes of data (1)

- Performance measures -
  - counts of actions and behaviors that are observed
- Subjective measures
  - users' perceptions, opinions and judgments
    - (See Dumas and Redish. 1994)
- Represents distinction between "what users do" and "what users think"

General classes of data (2)

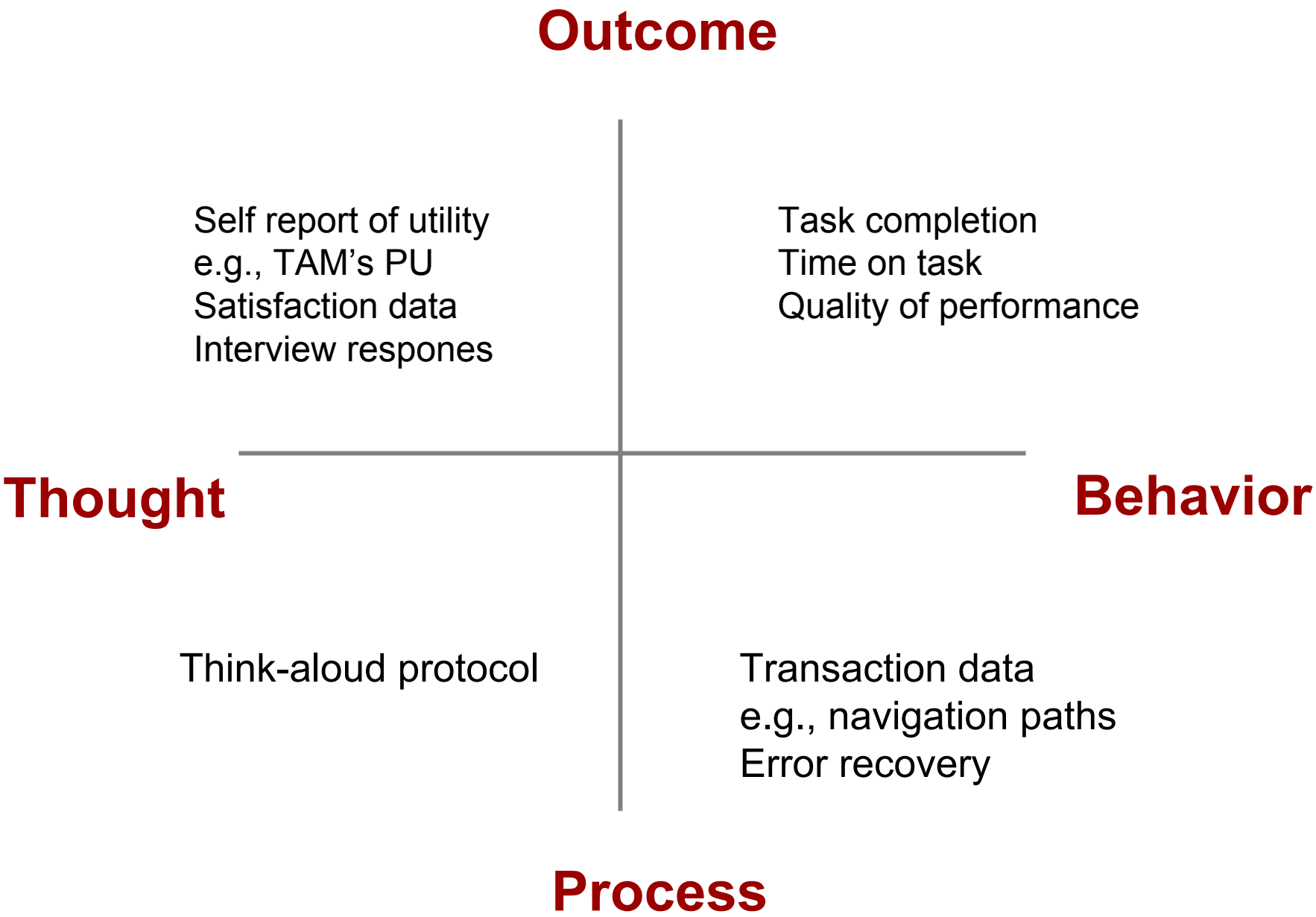
- Outcome measures
  - Measures of user performance or user rating taken upon completion of task or sub task
  - Speed, satisfaction scores etc.
- Process measures
  - Measures of the interaction which reflect the users actions and choices while performing the task
  - Navigation paths, menu selections
- Distinguishes “what they finally do/think” from “how they do/think it”

Usability Outcome Measures

- Common usability outcomes
  - Task completion
  - Quality of output
  - Speed of output
  - User rating of tool
- Each can be related to effectiveness, efficiency or satisfaction in usability terms
  - *See Bevan & Macleod papers in readings*

Usability Process measures

- Common usability process measures:
  - Transaction logs/ path analysis
  - Verbal protocols
  - Time spent in help, in menus etc.
  - Number of screens visited/revisited etc.
  - Observations of confusion/frustration in users
- Process measures relate most to efficiency and to problem identification


Outcome
Self report of utility
e.g., TAM's PU
Satisfaction data
Interview respones
Task completion
Time on task
Quality of performance
Thought
Behavior
Think-aloud protocol
Transaction data
e.g., navigation paths
Error recovery
Process

Methods are your toolkit….

- Select methods that will collect the data to answer your question
  - Time and error data are standard
  - But alone tell you nothing about satisfaction
  - Global outcome measures tell you nothing about process
  - General user comments tell you little about performance

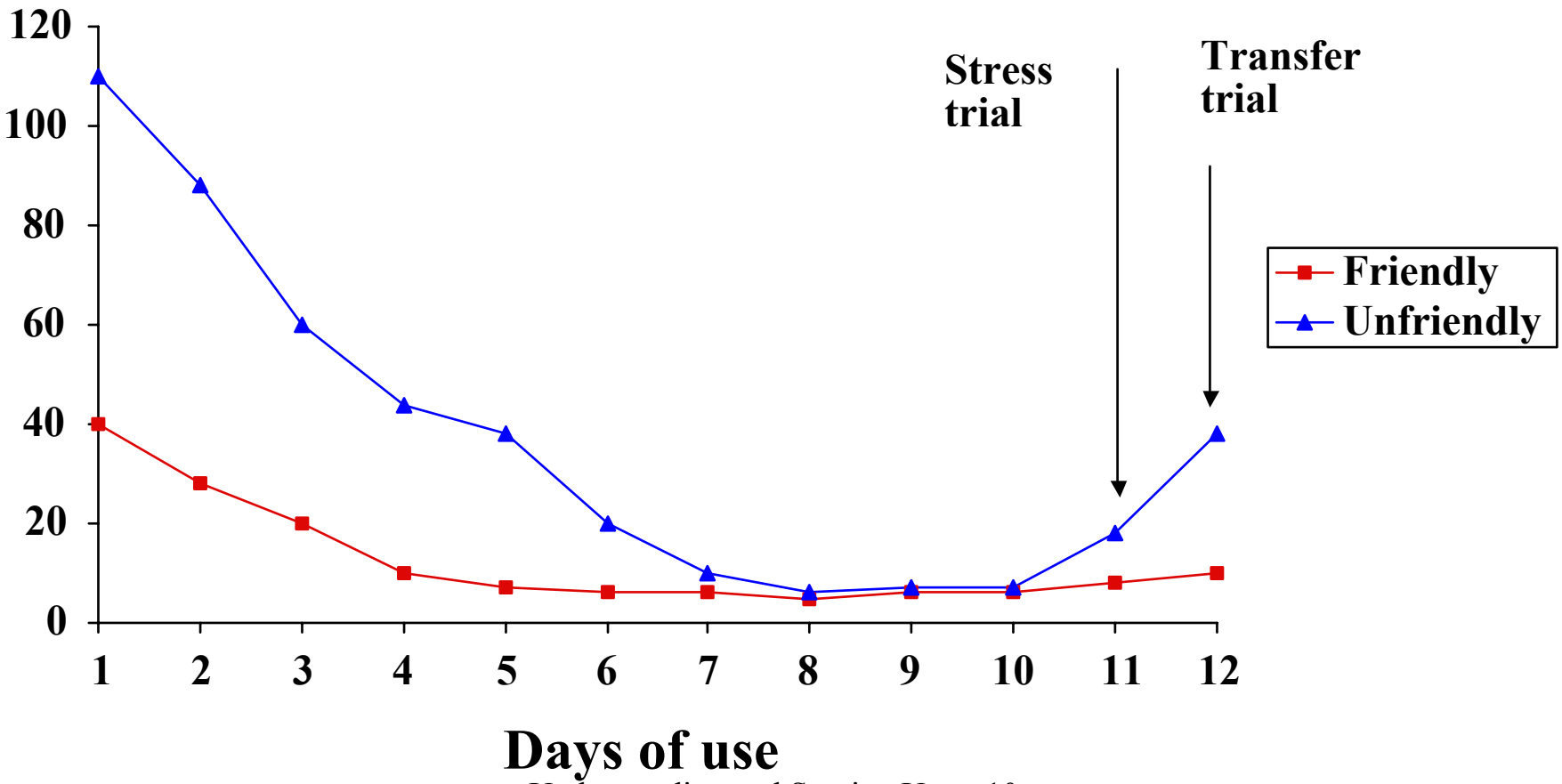
Starting out

- Identify tasks that you will use as basis for evaluation
- Determine the variables that will provide measures of:
  - effectiveness
  - efficiency
  - satisfaction
    - THESE AT A MINIMUM..you may have other issues

Learnability or Usability

- Will users be provided with training?
- Learnability =
  - rate of change in user data over time
  - number of task cycles required to meet target scores
  - hours of training required to meet target
  - Number of errors and repeat errors
- Many evaluations confuse these variables

Error scores over time…..



Capturing user's thoughts

- Process data:
  - Verbal protocol analysis
- Outcome data
  - Ratings and surveys
  - Interviews

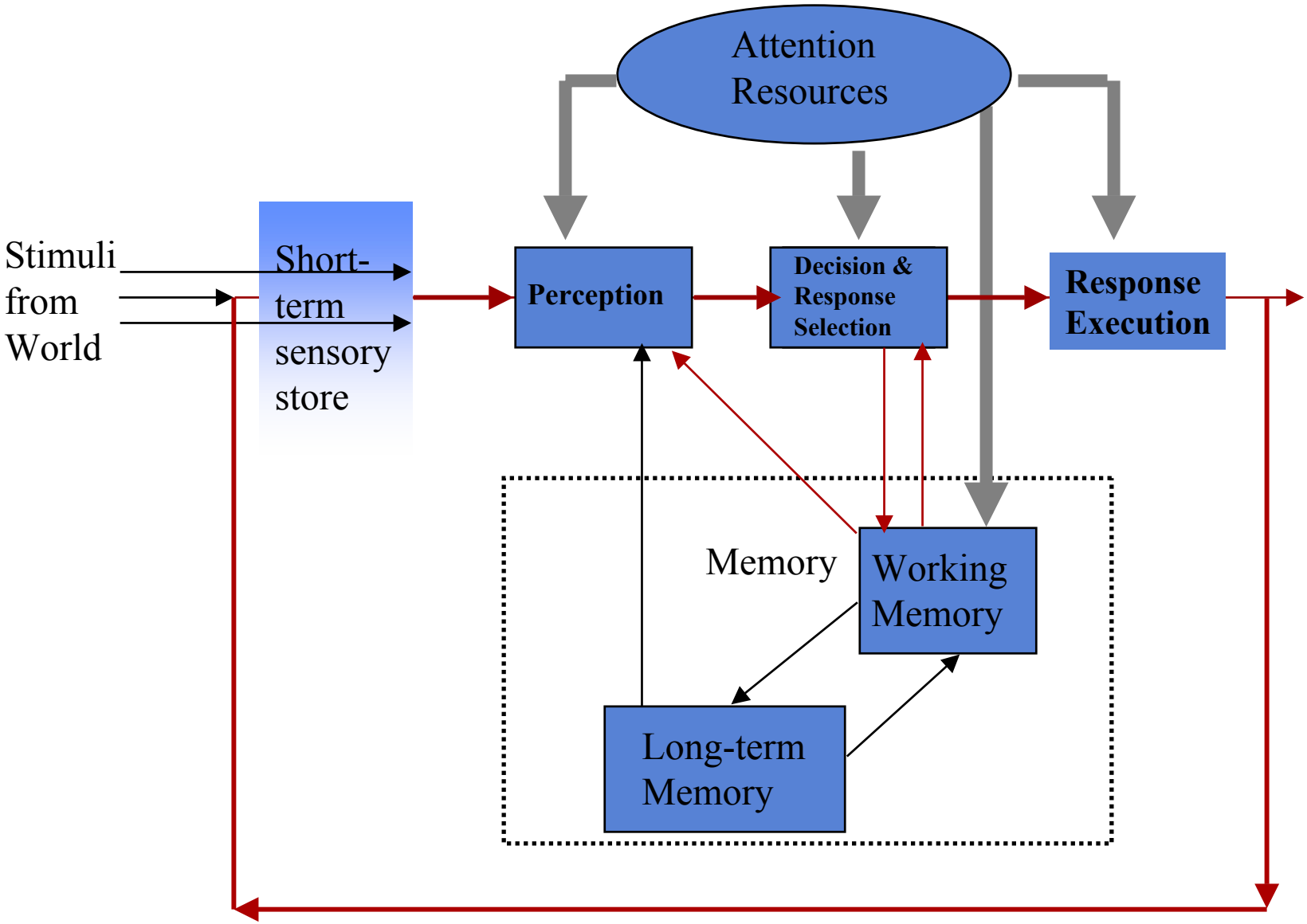
Verbal reports as data

- Verbal protocols are the user's account of what she is thinking while performing a task
  - Concurrently elicited while working
  - Not answers to questions
  - Should reflect cognitive processes of user
  - Key insight into reasoning of user

Concurrent v. retrospective VPs

- Concurrent protocols reflect immediate contents of short-term memory

- Retrospective protocols reflect memory of contents and are subject to distortion


Attention Resources
Stimuli from World
Short-term sensory store
Perception
Decision & Response Selection
Response Execution
Memory
Working Memory
Long-term Memory

TIME	COMMENT	ACTION
00:10	'Right, I'm going to go to the Contents and try and get some idea of where this might be.....'	Drags SCROLL bar up Reading Contents
00:24	'I think it's probably in 'The Making of Wine''	Still reading CONTENTS
00:38	'Oh is it Sweetness and Body?'	Refers to CONTENT items
00:42	'I'll just go there and check if they're the two.....'	Scrolls down the text first then clicks in SCROLL bar until she reaches relevant section. Reads them.
00:58	'So the two things are.....' 'Sweetness and Body.....'	Finds Answer

Practical concerns - TRY IT!

- People stop talking unless prompted
- Practice providing a protocol helps
- Pronouns and incomplete info are typical
- Scoring the output
  - General 10:1 ratio of analysis to data

Analysis of VPA

- Coding structure provided by task analysis
- Segment protocol into phrases/units
- Order segments by time
- Relate protocols to state of interface
- Apply perspective driving your research
  - systematic modeling of user cognition?
  - general identification of user difficulties?
  - Location of user interest throughout interaction?

Reliability

- Inter-rater reliability is primary
  - Can two raters see the same things in the protocol?
  - If yes, to what degree
    - Reliability index of scored unit agreements
    - {% agree} -{% disagree}
  - If no (or low), then data are not acceptable as evidence in scientific studies

In a typical usability test

- Protocol divided into task units
  - e.g. screen by screen actions in Web
- Verbal comments related to screen
  - e.g., comments about layout, wording,
- Comments related to self
  - references to own inability to act at any point
- General comments about system
  - positive, negative etc.

Example web site protocol

Time	Comment	Action
0.32:	"What do I choose here....looks like there is no direct link....and I don't like the colors here, too bright...er..	(SELECTS TEACHING)
0:41:	'teaching and courses' sounds right	(SCREEN CHANGES)..
0.46:	oh this is all about missions and stuff… hang on....	(HITS BACK BUTTON)
0.53:	well.....that looks the best of these, you know."	

Example web site protocol

Time	Comment	Action
0.32:	"What do I choose here....looks like there is no direct link....and I don't like the colors here, too bright...er..	(SELECTS TEACHING)
0:41:	'teaching and courses' sounds right	(SCREEN CHANGES)..
0.46:	oh this is all about missions and stuff… hang on....	(HITS BACK BUTTON)
0.53:	well.....that looks the best of these, you know."	

VPA output

- Exhaustive synthesis of user comments by topic or theme
- General clustering of comments by theme or application unit
- Note of agreement between raters
- Data with which to interpret user behavior

Variations on VPA

- Walkbacks
  - User talks aloud while reviewing use with evaluator
- Partial protocol analysis
  - Evaluator has pre-defined categories and only captures relevant comments as they occur
- Group protocol
  - Several users discuss their likely use with evaluator in real-time