Human-Computer Interaction

Objective, Behavioral, & Physiological Measures

Professor Bilge Mutlu

Today's Agenda

» Topic overview: *Objective, Behavioral, & Physiological Measures*

» Hands-on activity

Recap: *What are different kinds of measurements we can take?*

1. **Objective**: Measurement from participants against an objective standard, e.g., performance in a test 👈 **our focus today**
2. **Behavioral**: Measurement of the actions and behaviors of participants, E.g., how much eye-contact participants maintain with a robot 👈 **our focus today**
3. **Subjective**: Measurement of self-report data on subjective evaluations, e.g., preferences, personality
4. **Physiological**: Measurements taken directly from participants' bodies, e.g., body temperature, GSR, EEG, EMG, fMRI 👈 **our focus today**

Objective Measurements

What are objective measurements?

Definition: Measurements of variables that can be determined objectively through direct observation, a.k.a., *performance* measurements.

Another way to think about it: Measurements of user behaviors/actions contextualized in a domain task.

*What are types of **objective measurements**?*[1]

1. Task success
2. Time-on-task
3. Errors
4. Efficiency
5. Learnability

[1] Albert & Tullis, 2013, Measuring the User Experience

*What is **task success**?*

Definition: Task success measures capture how effectively users are able to complete a given set of tasks.

Task success can be measured as a **binary** variable or as a **level** of success.

Alternatively, task success can be measured in terms of task **failure**.

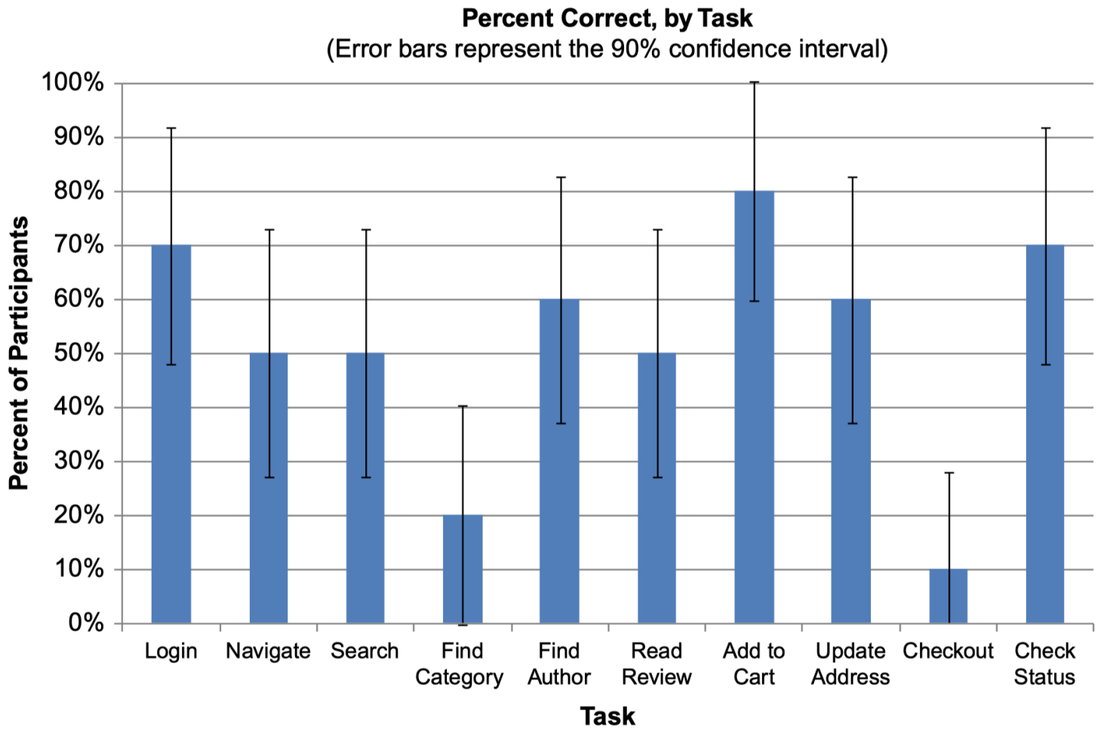
Binary success

Definition: A measure of whether participants successfully completed the task or failed to complete the task. Progress in the task is not captured in binary success measures.

When multiple tasks are used, we can calculate average task success per task or per participant.

1

	Task 1	Task 2	Task 3	Task 4	Task 5	Task 6	Task 7	Task 8	Task 9	Task 10	Average
Participant 1	1	1	1	0	1	1	1	1	0	1	80%
Participant 2	1	0	1	0	1	0	1	0	0	1	50%
Participant 3	1	1	0	0	0	0	1	0	0	0	30%
Participant 4	1	0	0	0	1	0	1	1	0	0	40%
Participant 5	0	0	1	0	0	1	0	0	0	0	20%
Participant 6	1	1	1	1	1	0	1	1	1	1	90%
Participant 7	0	1	1	0	0	1	1	1	0	1	60%
Participant 8	0	0	0	0	1	0	0	0	0	1	20%
Participant 9	1	0	0	0	0	1	1	1	0	1	50%
Participant 10	1	1	0	1	1	1	1	1	0	1	80%
Average	70%	50%	50%	20%	60%	50%	80%	60%	10%	70%	52.0%



[1] Albert & Tullis, 2013, Measuring the User Experience

Level of success

Definition: A measure of to the extent to which participants performed the task successfully and independently.

Different ways to formulate:

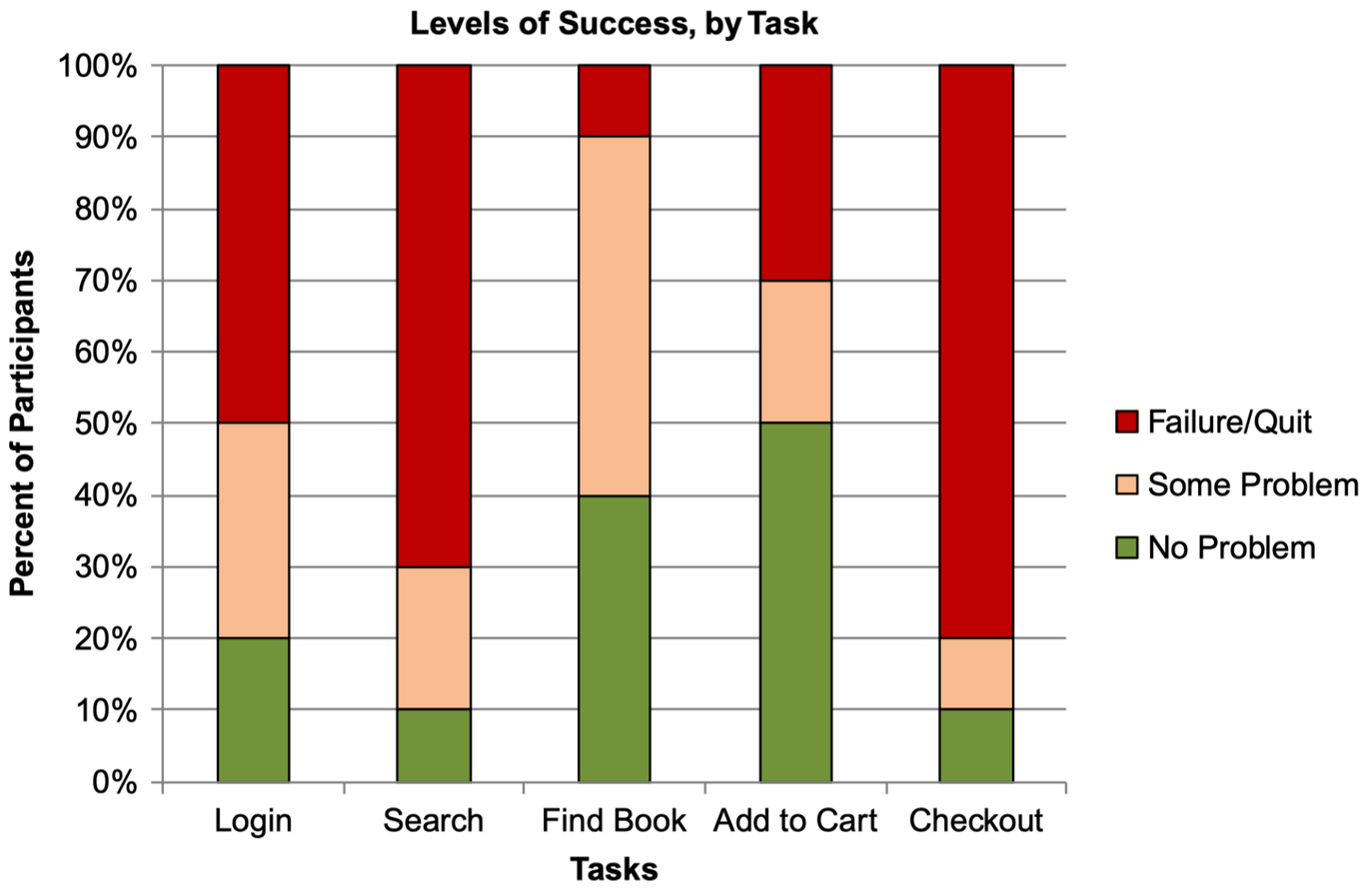
» *Discrete levels of success:* complete success, partial success, failure

» *Discrete levels of failure:* no problem, minor problem, major problem, failure/gave up

» *Subtask success:* The number of steps/subtasks completed successfully

» *Success with/without assistance:* complete success (with/without assistance), partial success (with/without assistance), failure (user thought it was complete, but it wasn't/user gave up)

1



[1] Albert & Tullis, 2013, Measuring the User Experience

Task failure

Definition: Measures of the rate at which users fail to successfully complete the task.

Different forms of failure:

- User gives up
- Experimenter ends the session, e.g., because the user is stuck/clearly not making progress
- User takes longer than the allotted time for the task
- Incorrect outcome where the user thought that the task was successful, but it wasn't

What is time on task?

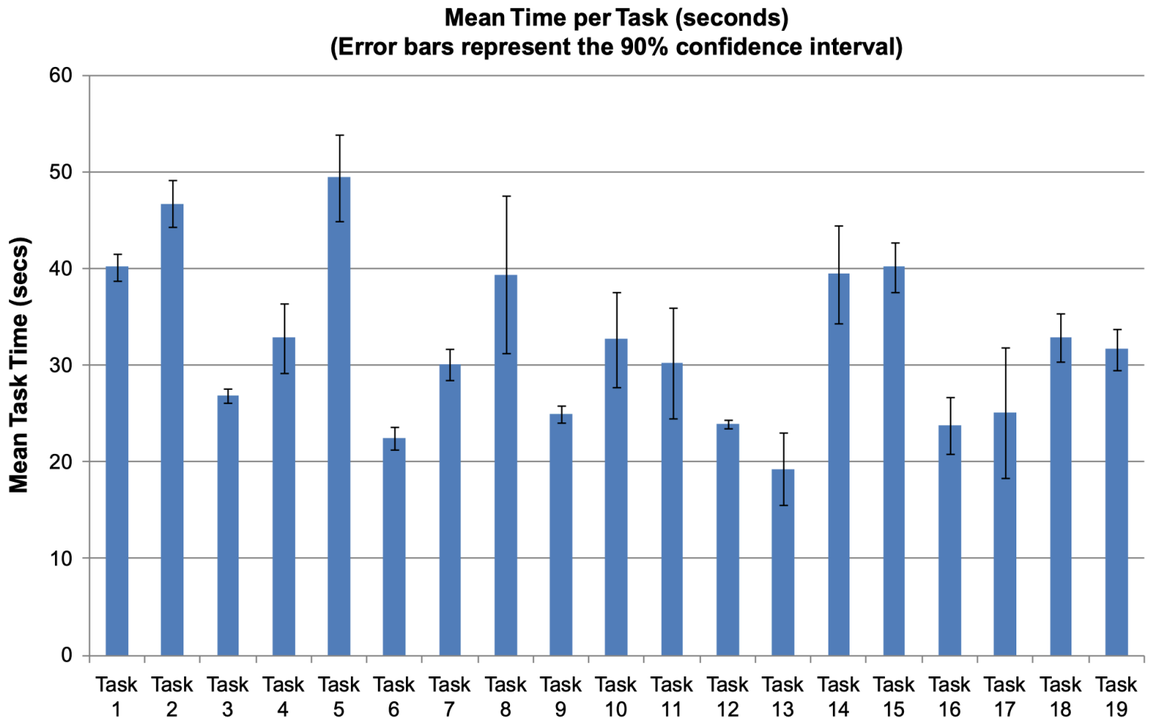
Definition: Time on task measures how much time is needed by users to perform the experimental task.

Different summary statistics, e.g., geometric mean, might be more appropriate for time on task measures.[2]

[2] McNichol, 2018, On Average, You're Using the Wrong Average

1

	Task 1	Task 2	Task 3	Task 4	Task 5
Participant 1	259	112	135	58	8
Participant 2	253	64	278	160	22
Participant 3	42	51	60	57	26
Participant 4	38	108	115	146	26
Participant 5	33	142	66	47	38
Participant 6	33	54	261	26	42
Participant 7	36	152	53	22	44
Participant 8	112	65	171	133	46
Participant 9	29	92	147	56	56
Participant 10	158	113	136	83	64
Participant 11	24	69	119	25	68
Participant 12	108	50	145	15	75
Participant 13	110	128	97	97	78
Participant 14	37	66	105	83	80
Participant 15	116	78	40	163	100
Participant 16	129	152	67	168	109
Participant 17	31	51	51	119	116
Participant 18	33	97	44	81	127
Participant 19	75	124	286	103	236
Participant 20	76	62	108	185	245
Mean	86.6	91.5	124.2	91.4	80.3
Median	58.5	85.0	111.5	83.0	66.0
Geometric mean	65.2	85.2	105.0	73.2	60.3
90% confidence interval	31.1	15.4	33.1	23.6	28.0
Lower bound	55.5	76.1	91.1	67.7	52.3
Upper bound	117.7	106.9	157.3	115.0	108.3



[1] Albert & Tullis, 2013, Measuring the User Experience

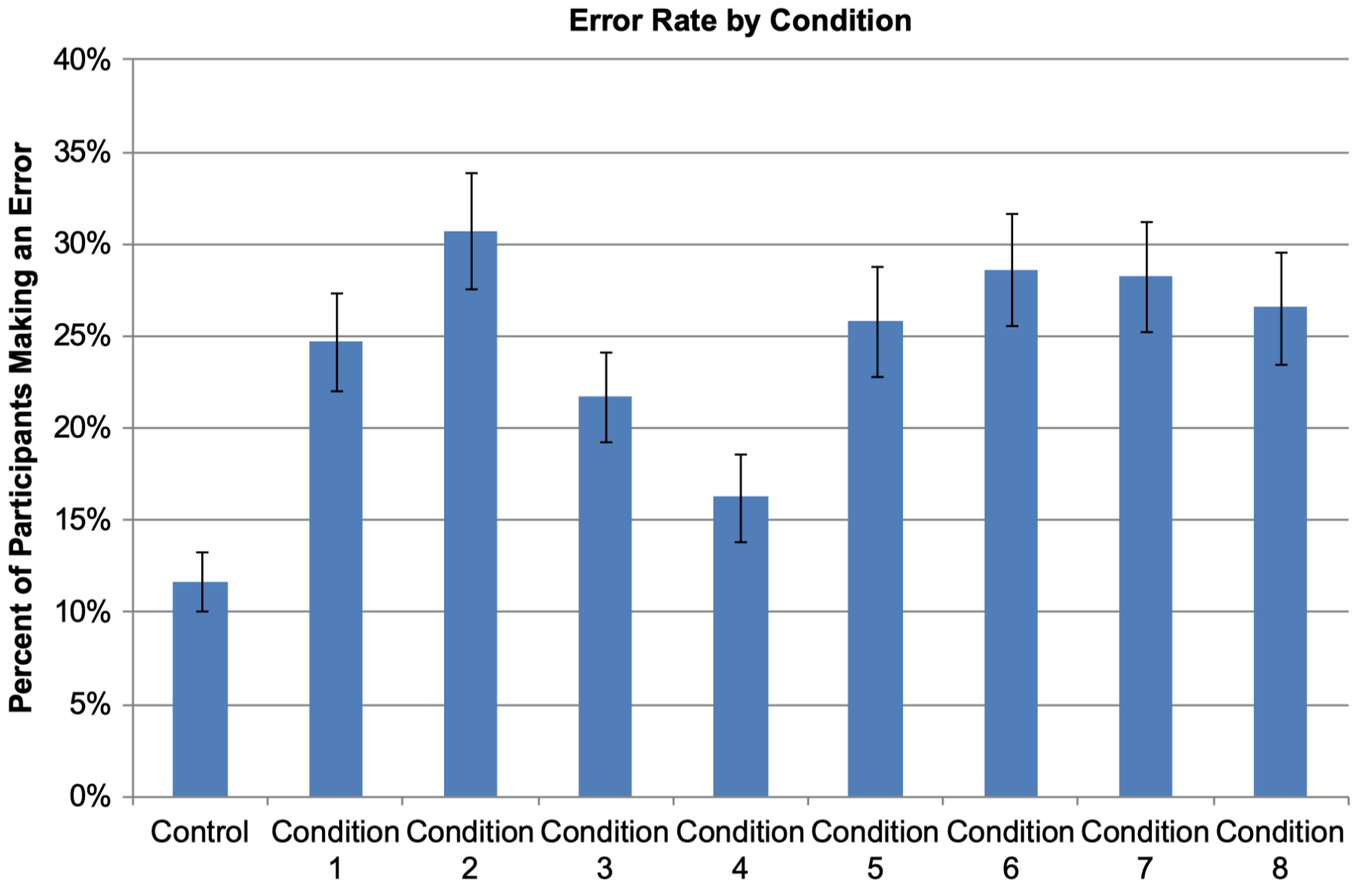
What is task error?

Definition: Task error is a measurement of various errors that users make in performing the task. Task errors are usually incorrectly performed steps within a task, e.g., pressing the wrong button.

Most error measurements involve simple *counts* of the number of errors observed.

Errors can also be rated for *severity* to create a composite error measure.

1



[1] Albert & Tullis, 2013, Measuring the User Experience

What is task efficiency?

Definition: Task efficiency measures the *effort* that users exert to successfully complete the task.

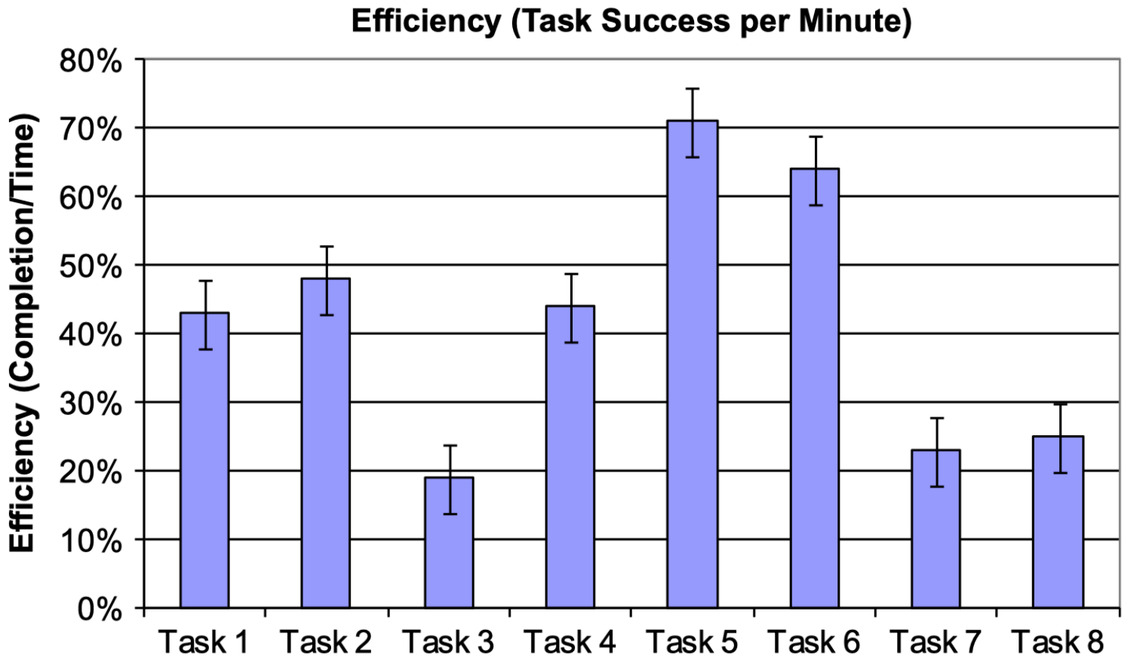
Can be formulated to capture *cognitive* (e.g., understanding instructions) or *physical* (e.g., performing copy and paste) effort.

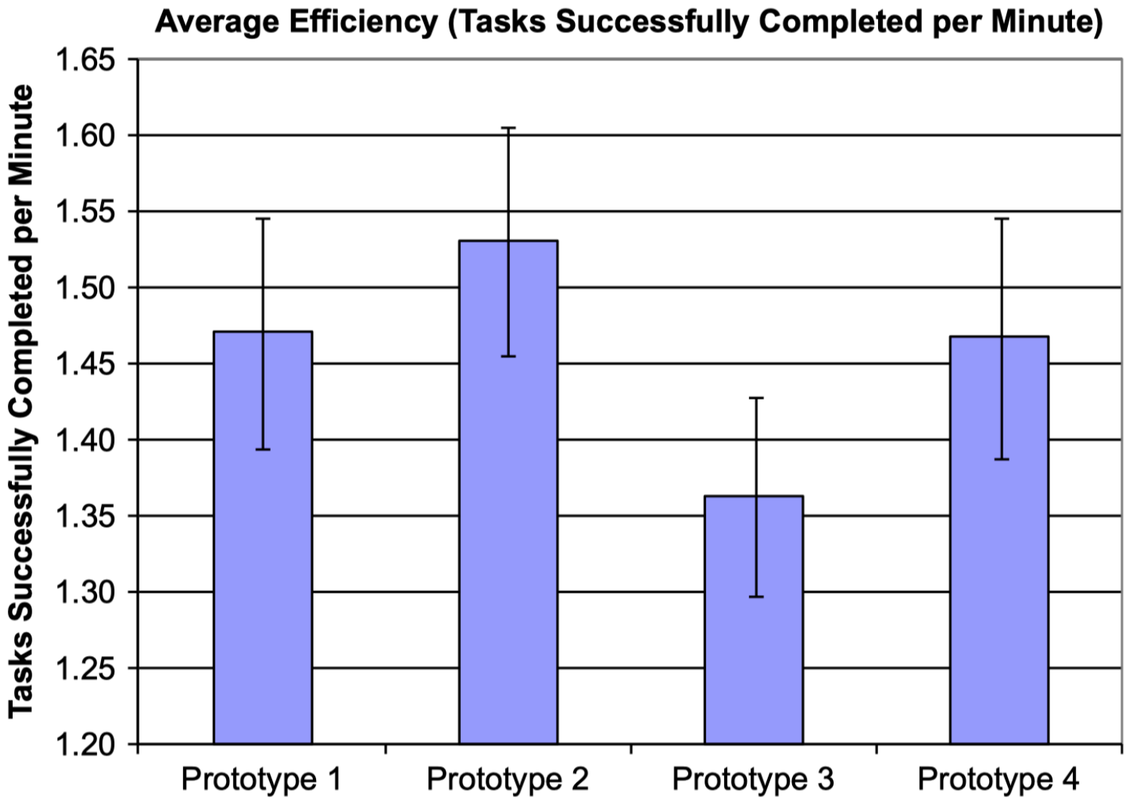
Efficiency is commonly a *composite* measure of *task success* and *task time*.[1]

	Task Completion Rate	Task Time (min)	Efficiency (%)
Task 1	65%	1.5	43
Task 2	67%	1.4	48
Task 3	40%	2.1	19
Task 4	74%	1.7	44
Task 5	85%	1.2	71
Task 6	90%	1.4	64
Task 7	49%	2.1	23
Task 8	33%	1.3	25

[1] Albert & Tullis, 2013, Measuring the User Experience

1

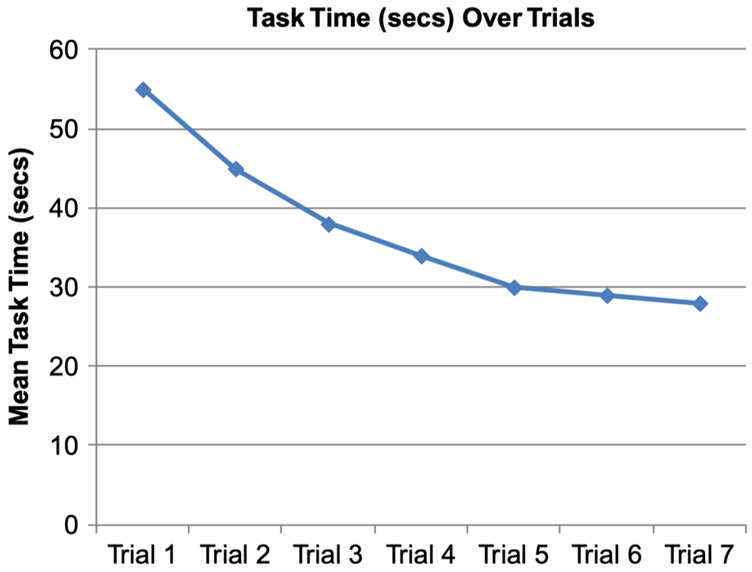


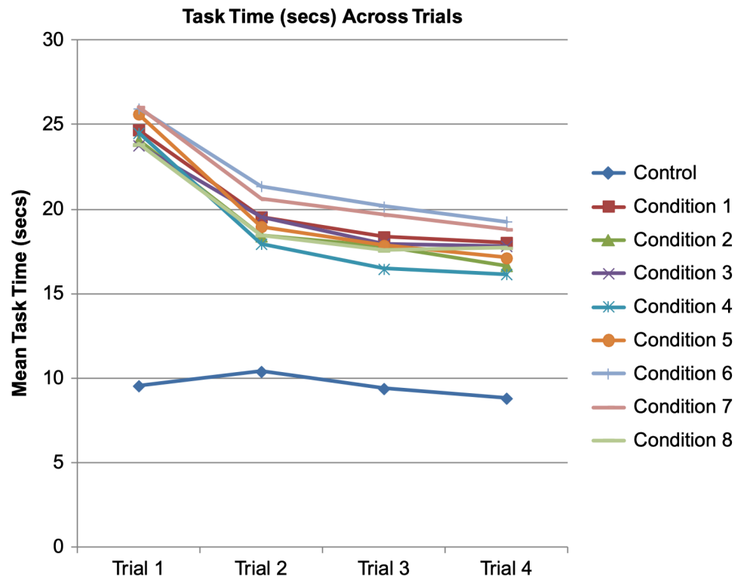


[1] Albert & Tullis, 2013, Measuring the User Experience

What is task learnability?[1]

Definition: Task learnability captures how user performance improves (e.g., in task learning) or degrades (e.g., fatigue) over time. Usually measured by repeating measurements across **trials**.





[1] Albert & Tullis, 2013, Measuring the User Experience

Single, multiple, & composite measures

Single: When a single observable measurement is taken to assess the dependent variable.

Multiple: When multiple measures assess the same high-level concept to understand capture multidimensional phenomenon (speed, error frequency, error amount can be considered different dimensions of **performance**) or tradeoffs (e.g., speed-accuracy tradeoff).

Composite: When multiple measurements are combined into a single measurement (e.g., efficiency is a combination of time and success).

Behavioral Measurements

What are behavioral measurements?

Definition: Measurements of participant task actions or behavior that are observed directly or through instrumentation.

What's the difference between behavioral and objective measures?

Not all behavioral measures are inherently objective. If objectivity can be established (e.g. through inter-coder reliability), then behavioral variables can be used as objective measures.

What are different kinds of behavioral measures?

High-level behaviors, e.g., task/goal-related behaviors

Low-level behaviors, e.g., verbal (e.g., frequency of word use), nonverbal (e.g., frequency of mutual gaze), psychophysical (e.g., heart rate)

How do we measure behaviors?

An example: gaze behavior

» Number of fixations — *overall*

» Gaze % — *proportion of time* — on each area of interest

» Fixation duration mean — *overall*

» Number of fixations on each area of interest

» Gaze duration mean, on each area of interest

» Fixation rate overall — *fixations/s*

» Many others…

Challenges in behavioral measurements

Effort: It is a significant amount of effort

Measurement quality: Difficult to ensure objectivity and reliability

Some good news...[3]

Behavior	1-minute slice	Two 1-minute slices	Three 1-minute slices
Slice(s) contained in 15-minute interaction			
Gaze	.73**	.81**	.91**
Gesture	.95**	.91**	.92**
Nod	.79**	.75**	.69**
Self-touch	.41	.76**	.83**
Smile	.62*	.68*	.76**
Mean *r*	.79	.80	.84
Slice(s) removed from 15-minute interaction			
Gaze	.64*	.72**	.85**
Gesture	.93**	.80**	.77**
Nod	.70**	.60*	.37
Self-touch	.32	.68**	.73**
Smile	.52*	.52*	.56*
Mean *r*	.68	.67	.68

[3] Murphy, 2005, Using thin slices for behavioral coding

Physiological Measurements

What are physiological measurements?

Definition: Measurements taken directly from the participant's body as an indicator of a physiological state. Common measures include:

» Eye tracking, pupil dilation

» Galvanic skin response (GSR)

» Muscle movements — Galvanic Skin Response (GSR)

» Muscle activity — Electromyogram (EMG)

» Brain activity — Electroencephalogram (EEG), Functional Magnetic Resonance Imaging (fMRI), Functional Near-infrared Imaging (fNIR)

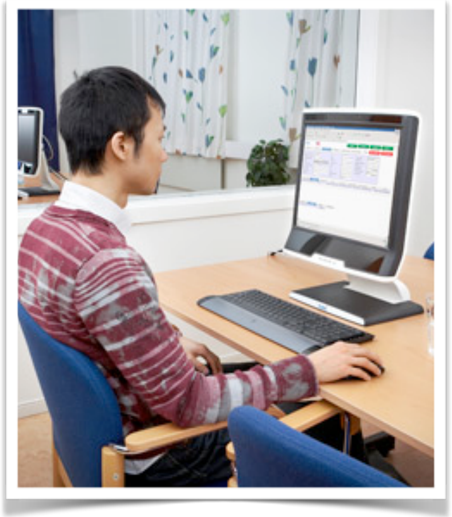
4

Data Source	Technique	Signal Type	Possible Locations	Sensors
Electrodermal activity	Galvanic skin response (GSR) (Scheirer et al., 2002; Mandryk and Inkpen, 2004)	Electrical	Fingers, toes	Surface electrodes
Cardiovascular data	Blood-volume pressure (Scheirer et al., 2002)	Light absorption	Finger	Surface electrodes
	Electrocardiography (Mandryk and Inkpen, 2004)	Electrical	Chest, abdomen	Surface electrodes
Respiration	Chest contraction and expansion (Mandryk and Inkpen, 2004)	Physical	Thorax	Stress sensor
Muscular and skeletal positioning	Pressure or position sensing (Brady et al., 2005; Dunne et al., 2006a,b; Dunne and Smyth, 2007)	Physical or electrical	Varied	Pressure sensor, fiber optics, others
Muscle tension	Electromyography (Mandryk and Inkpen, 2004)	Electrical	Jaw, face	Surface electrodes
Brain activity	Electroencephalography (Lee and Tan, 2006)	Electrical	Head	Electrodes in helmet
	Evoked responses (Stern et al., 2001)	Electrical	Head	Surface electrodes

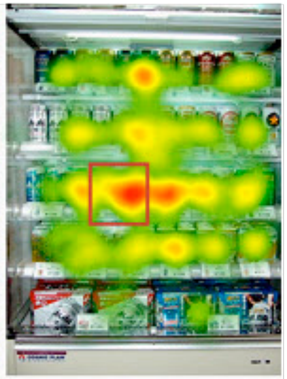
[4] Lazar et al., 2017, Chapter 13

eye tracking

- Head or desk mounted equipment tracks the position of the eye
- Eye movement might indicate the amount of cognitive processing a display requires
- Measurements;
  - Fixations: Eye maintains stable position. Number and duration indicate level of difficulty with display
  - Saccades: Rapid eye movement from one point of interest to another or around a point of interest
  - Scan paths: Moving straight to a target with a short fixation at the target is optimal



Heat map showing male respondents' view of Plan G incl. Ginger Draft Green.

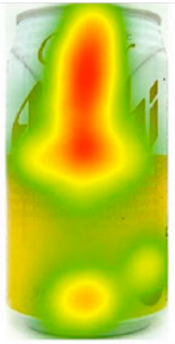


Heat map showing male respondents' view of Plan Z incl. Kirin Zero.

Heat map showing male respondants' view of Plan G.

Heat map showing female respondants' view of Plan G.

Kin-Mugi
Gaze concentration in a straight line in the middle; brand logo and message "Rich Malt" viewed.

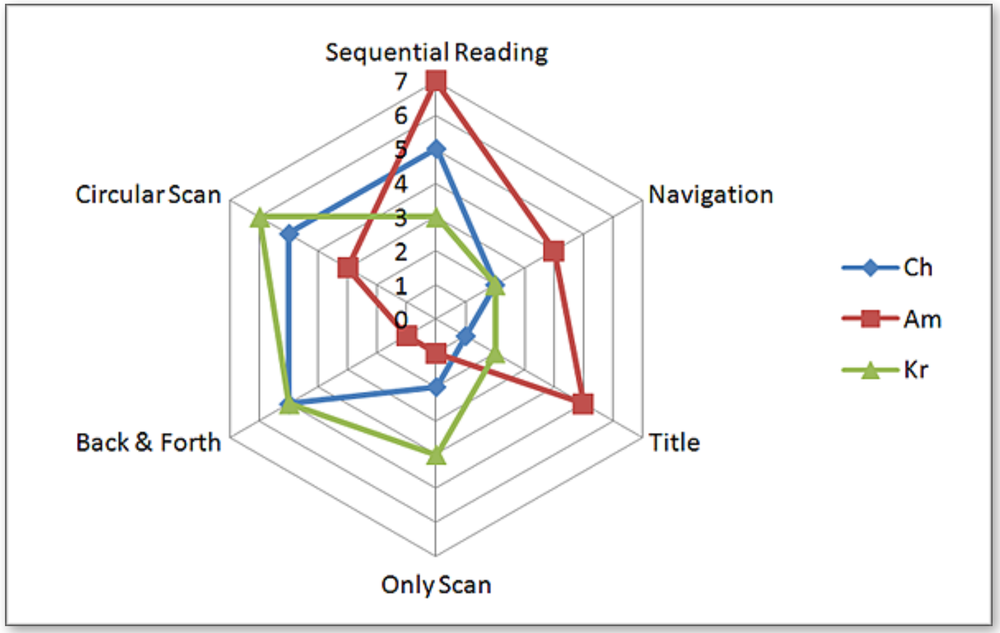


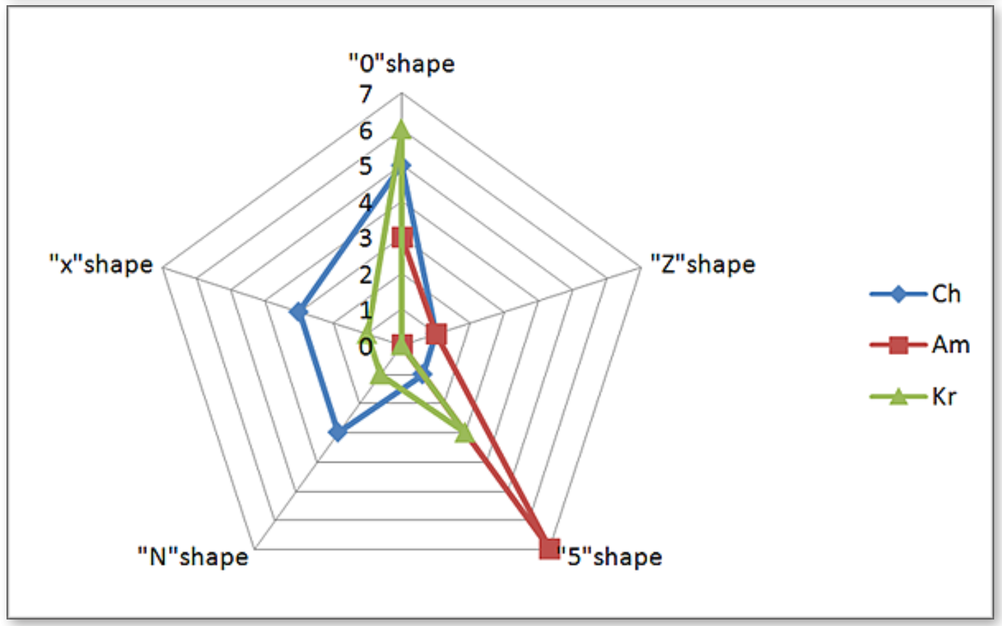
Clear Asahi
Gaze concentration from mid-to-upper; indicates viewing of the brand logo and beer bubbles motif in the center, which includes a brand message.

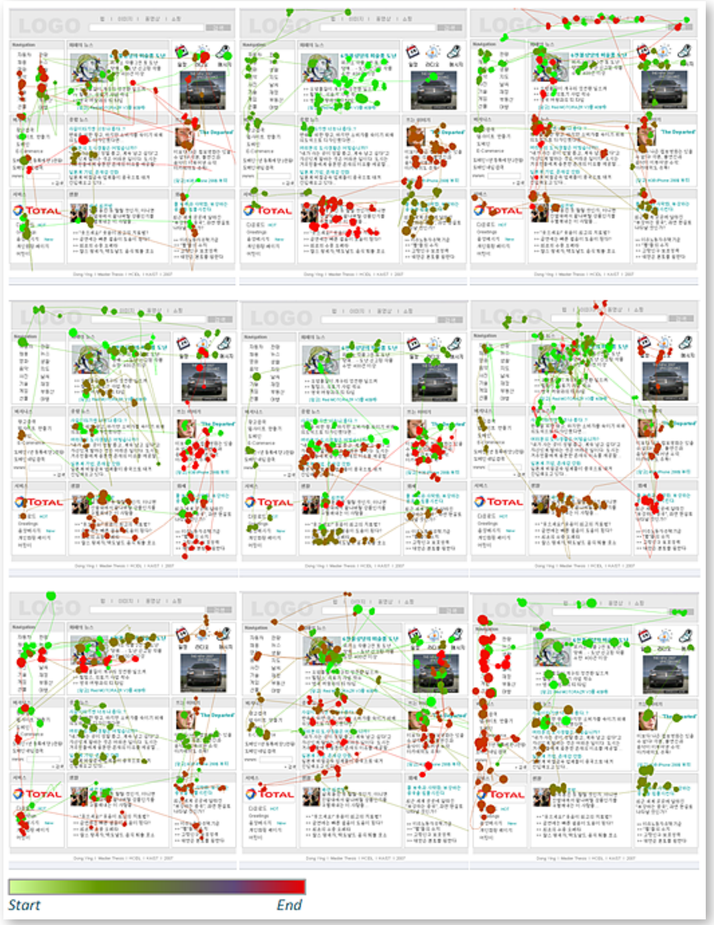
Mugi-to-Hop
Gaze concentration at center; brand logo and the entirety of the message text seen.

BEFORE

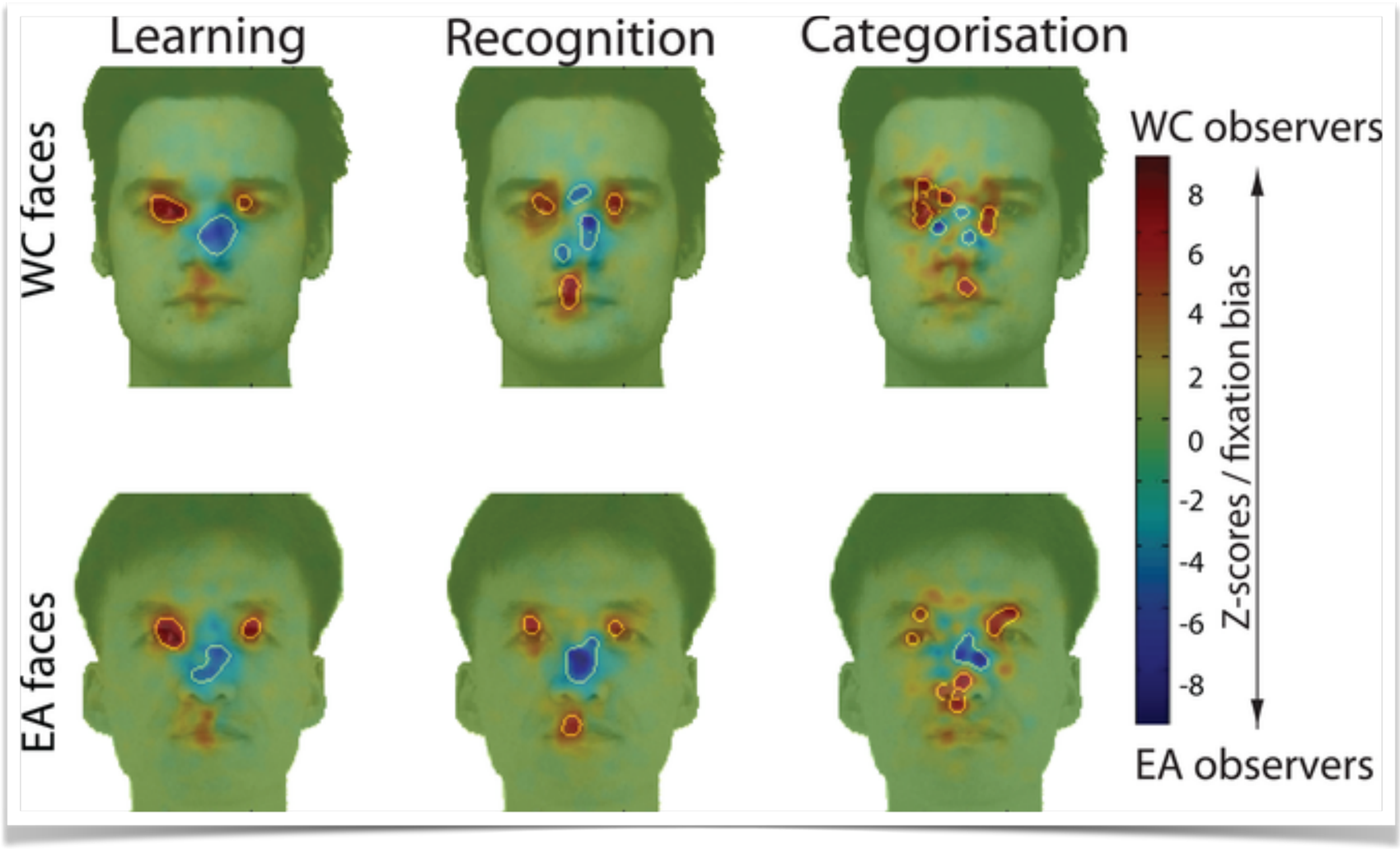
AFTER







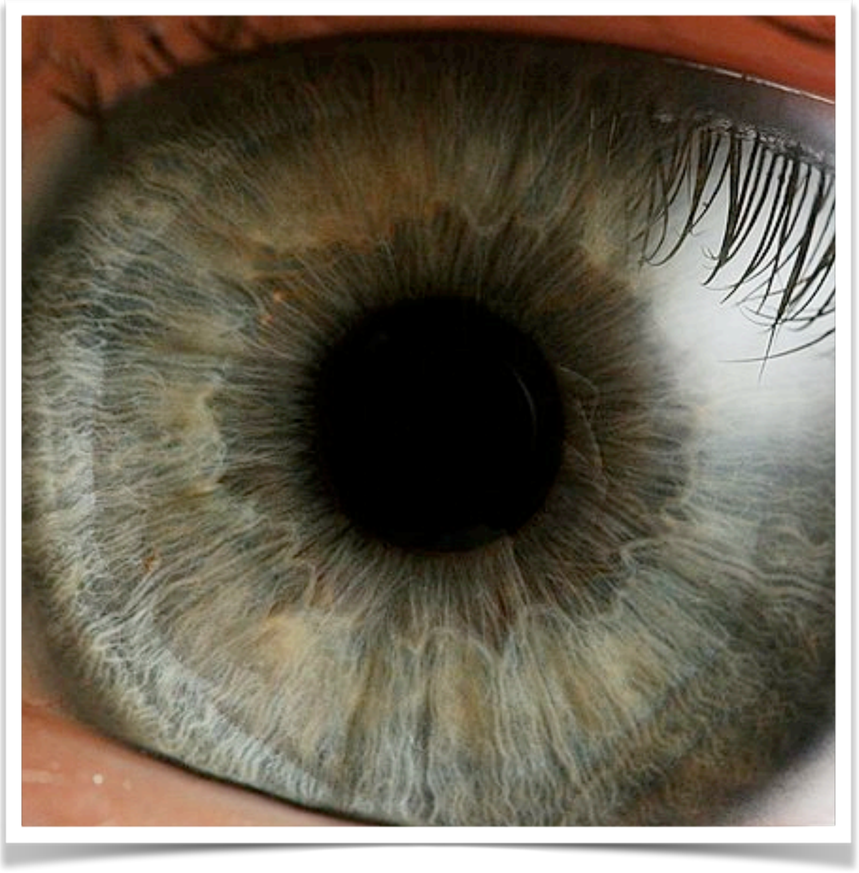
Dong, Y., Lee, K. P. (2008). A cross-cultural comparative study of users' perceptions of a webpage: With a focus on the cognitive styles of Chinese, Koreans and Americans. *International Journal of Design*, 2(2), 19-30.

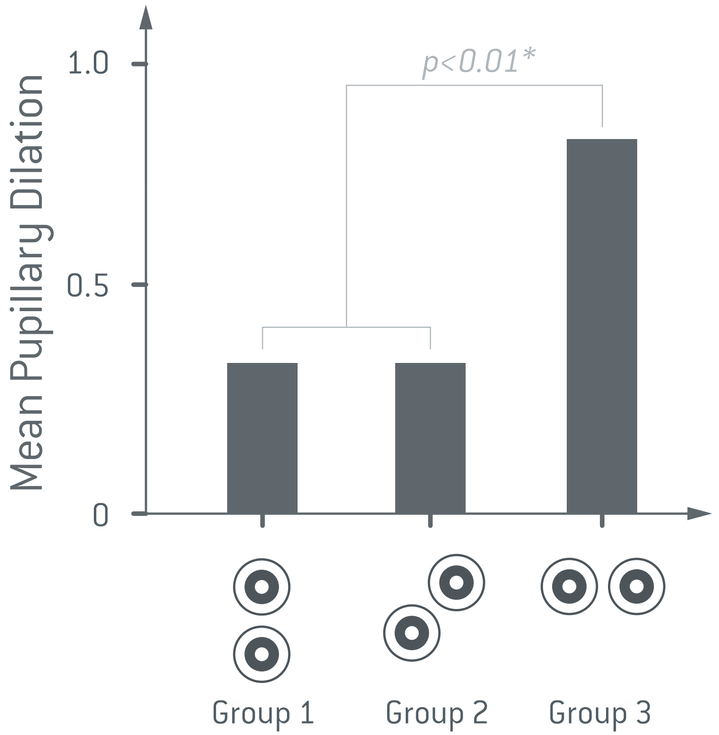


Blais C, Jack RE, Scheepers C, Fiset D, and Caldara R, (2008). Culture Shapes How We Look at Faces. *PLoS ONE,* 3(8): e3022.

pupil dilation

- Pupils dilate in response to;
  - Extreme emotional situations (fear, pain, contact with nerves)
  - Loads on working memory, increased attention, sensory discriminations, cognitive load
- Intensifies perceptions of emotional states (sadness)
- Mirror responses to other people's pupil sizes
  - Might indicate empathy





Pupillary dilation in response to eye-shaped patterns

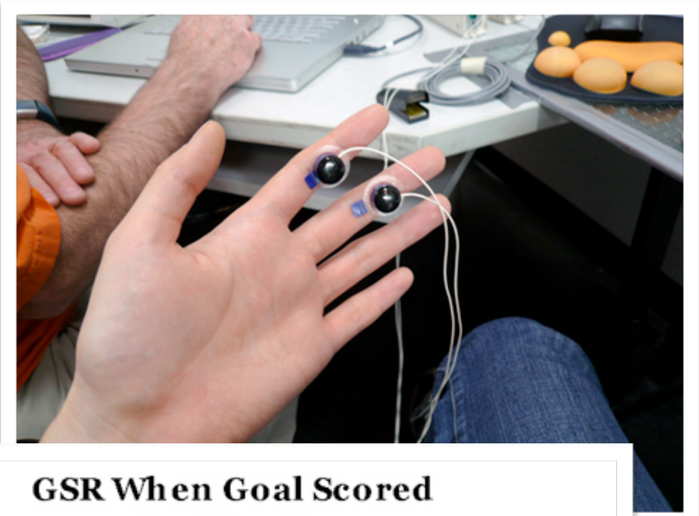
(Coss, 1972)

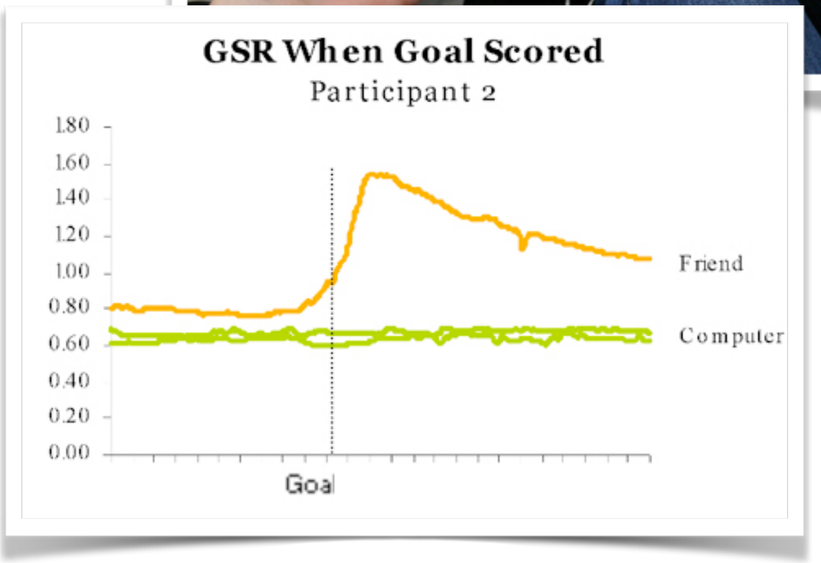
other physiological signals

- Emotional response linked to physical changes
- These may help determine a user's reaction to an interface
- Measurements:
  - Heart activity, including blood pressure, volume and pulse
  - Activity of sweat glands: Galvanic Skin Response (GSR)
  - Electrical activity in muscle: electromyogram (EMG)
  - Electrical activity in brain: electroencephalogram (EEG)
- Difficult to interpret some physiological responses

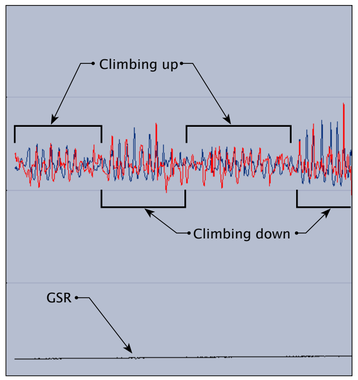
GSR

- Galvanic skin response (GSR)
  - Also called Electodermal response (EDR)
    - A.k.a. “sweat”
  - Sensitive to affective stimulation and attention
  - May measure long-term recall and effectiveness

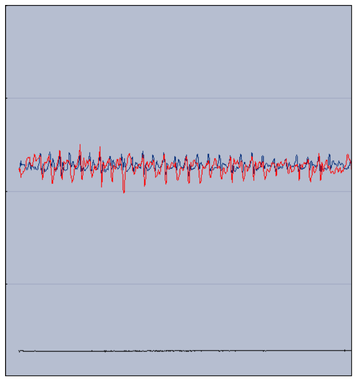




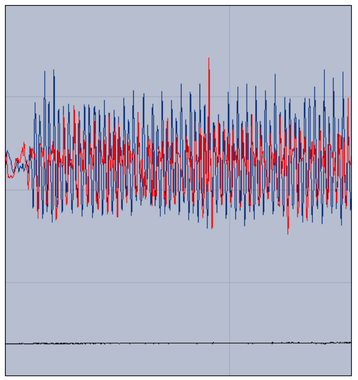
Mandryk, et al. 2006



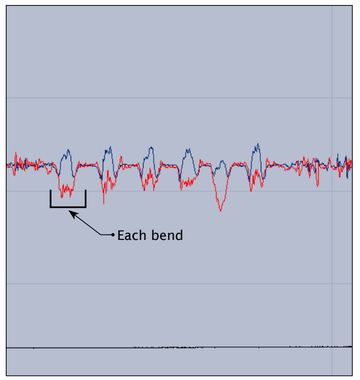
Climbing up / down the stairs



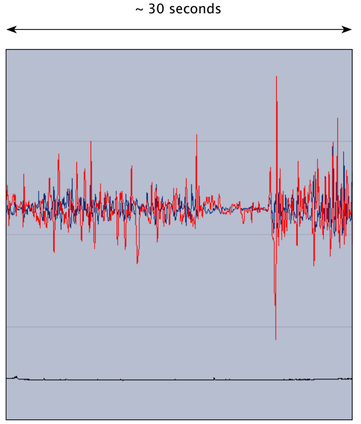
Medium pace walking



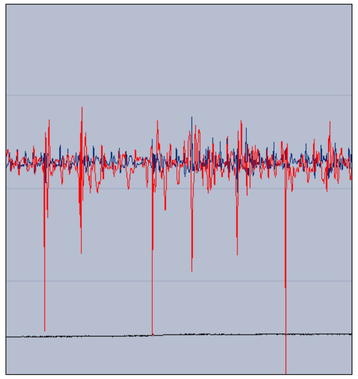
Running



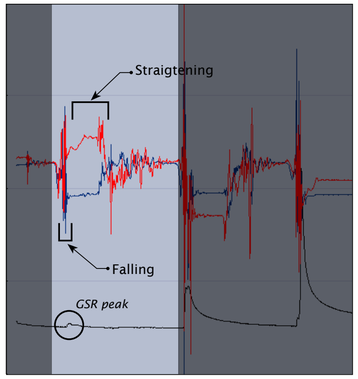
Bending down



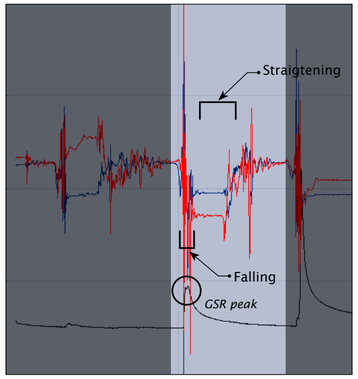
Loosing orientation
(oscillatory movement)



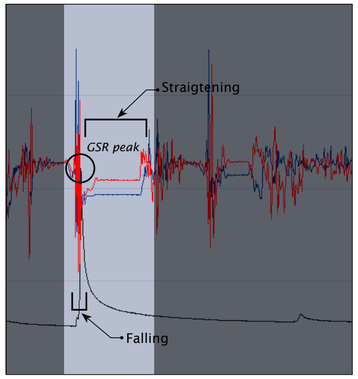
Hitting an obstacle
(wall, door handle, etc.)



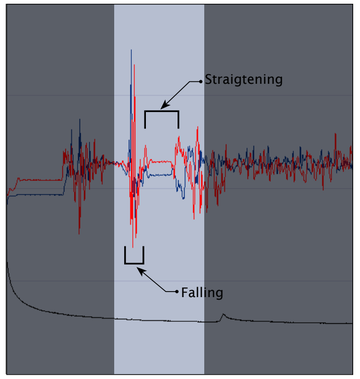
Falling sideways
(to the left - armband side)



Falling forward



Falling backward

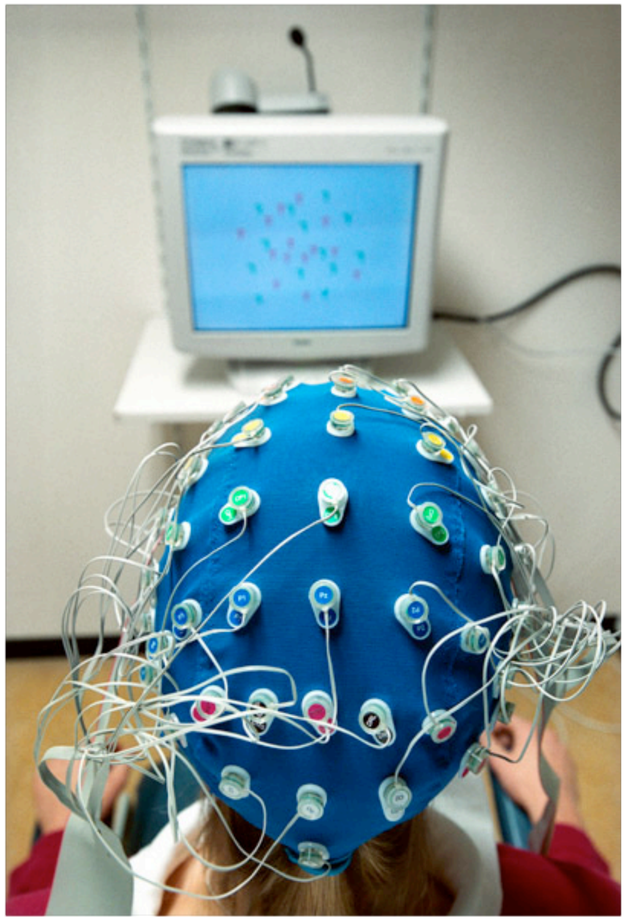


Falling sideways
(to the right)

Longitudinal acceleration (average)
Transverse acceleration (average)
GSR (average)

brain activity

- The electroencephalograph (EEG)
  - Alpha activity
    - Degree of activation
    - Associated with *inactivity, resting, & sleeping*
  - Hemispheric lateralization
    - Distinguishes between activity in the:
      - Left hemisphere of the brain — processes visual stimuli
      - Right hemisphere of the brain — processes verbal stimuli

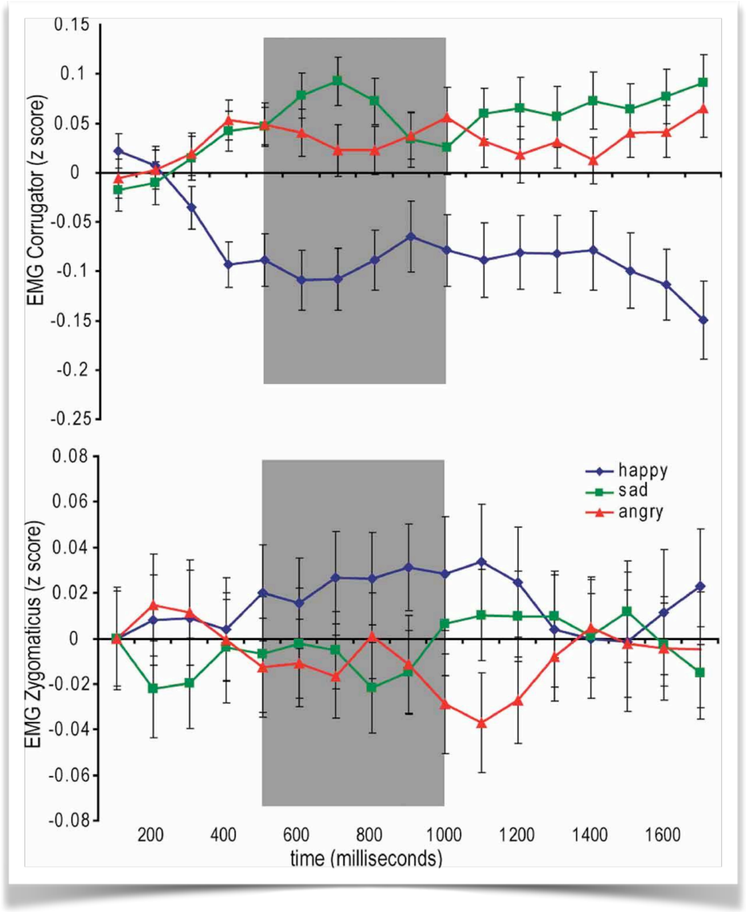


$$E = \frac{\beta}{(a + \theta)}$$

Szafir, D. & Mutlu, B. (2012). Pay Attention! Designing Adaptive Agents that Monitor and Improve User Engagement. In *Proceedings of the 30th ACM/SigCHI Conference on Human Factors in Computing (CHI 2012)*, Austin, TX.

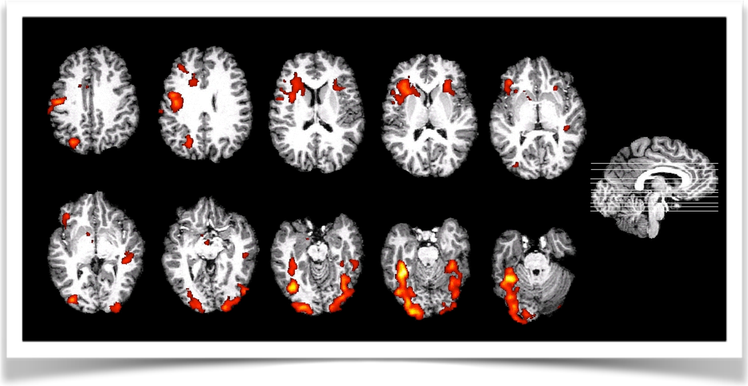
EMG

- Electromyography (EMG)
  - Measures electrical activity in the muscles
  - Facial expressions



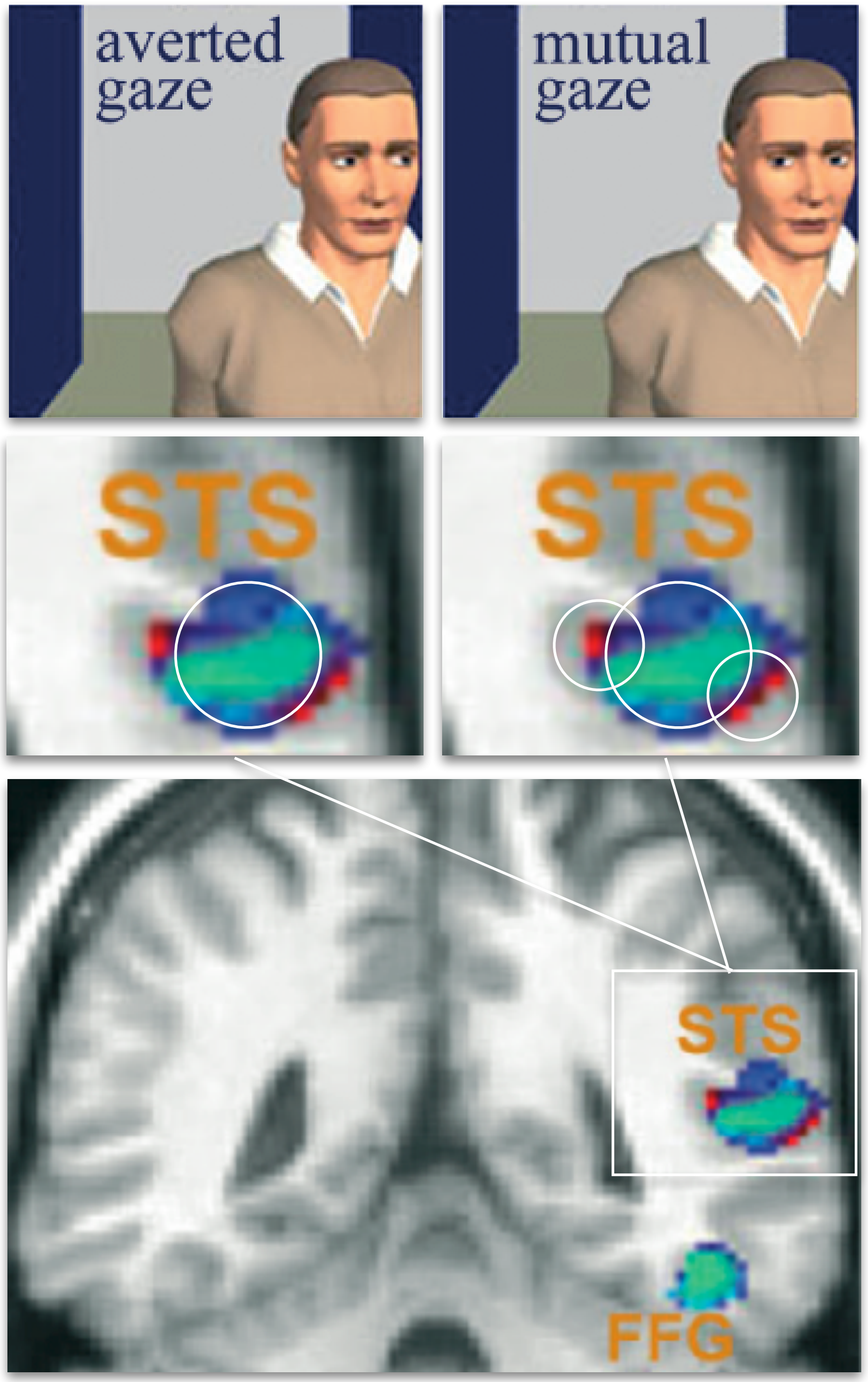
fMRI

- Functional Magnetic Resonance Imaging (fMRI)
  - Measures in changes in blood flow related to neural activity
  - Blood-oxygen-level dependence (BOLD)
  - Measures activation in all regions with high spatial resolution
  - Particularly useful in emotion research



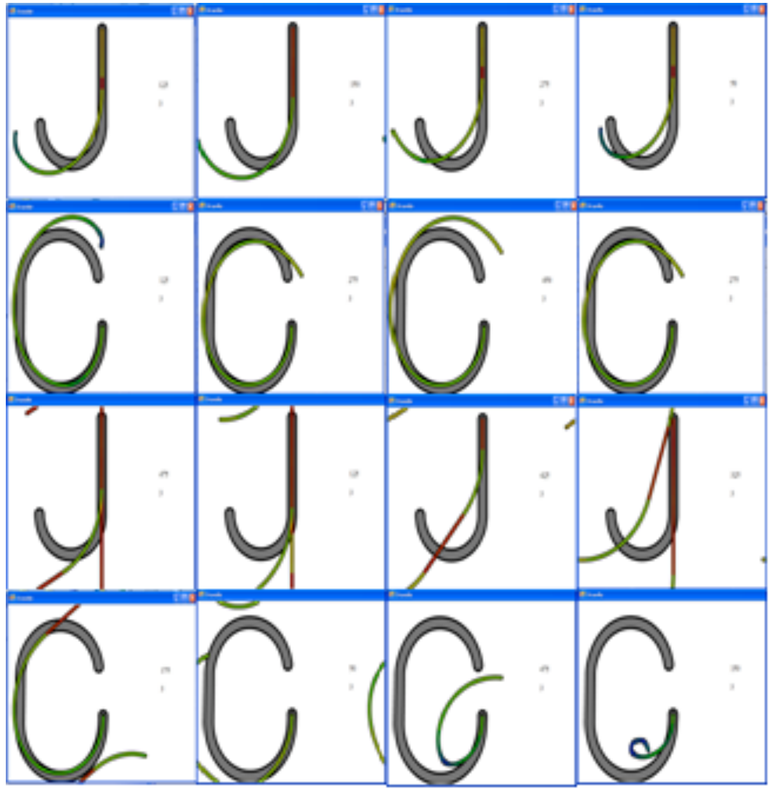
Mutual and averted gaze stimuli & response in the brain

(Pelphrey et al., 2004)



fNIR

- Functional near-infrared imaging (fNIR)
  - Measures oxygen levels
  - Indicates cognitive activity



Mappus, R. L., Venkatesh, G. R., Shastry, C., Israeli, A., and Jackson, M. M. 2009. An fNIR based BMI for letter construction using continuous control. In *Proceedings of Human Factors in Computing Systems*, 3571-3576.