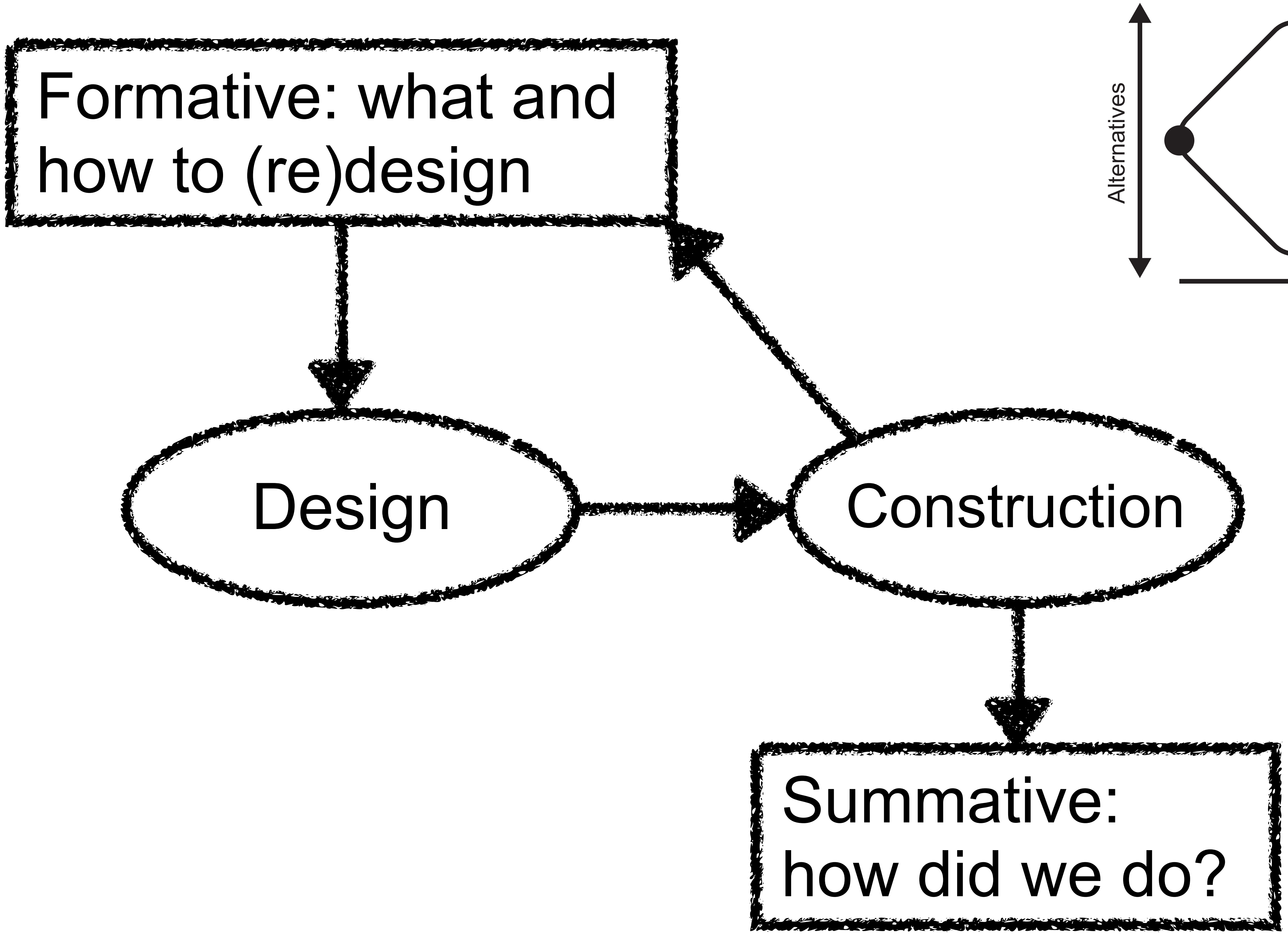
Chapter 5 - Evaluation

- Types of Evaluation
  - Formative vs. Summative
  - Quantitative vs. Qualitative
  - Analytic vs. Empirical
- Analytic Methods
  - Cognitive Walkthrough
  - Heuristic Evaluation
  - GOMS and KLM
  - Motor Functions: Fitt's Law, Steering Law
- Empirical Methods
  - Field Studies und Lab Studies
  - Longitudinal und Diary Studies
  - Usability Scales

Formative vs. Summative Evaluation

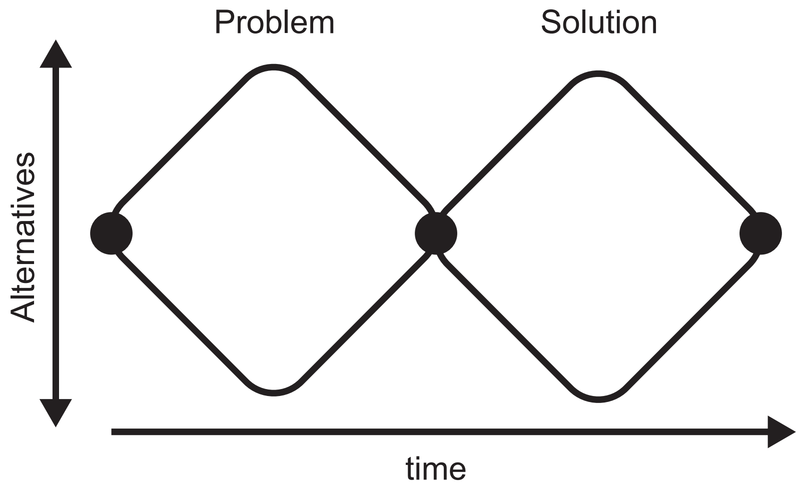


Formative: what and how to (re)design

Design

Construction

Summative: how did we do?

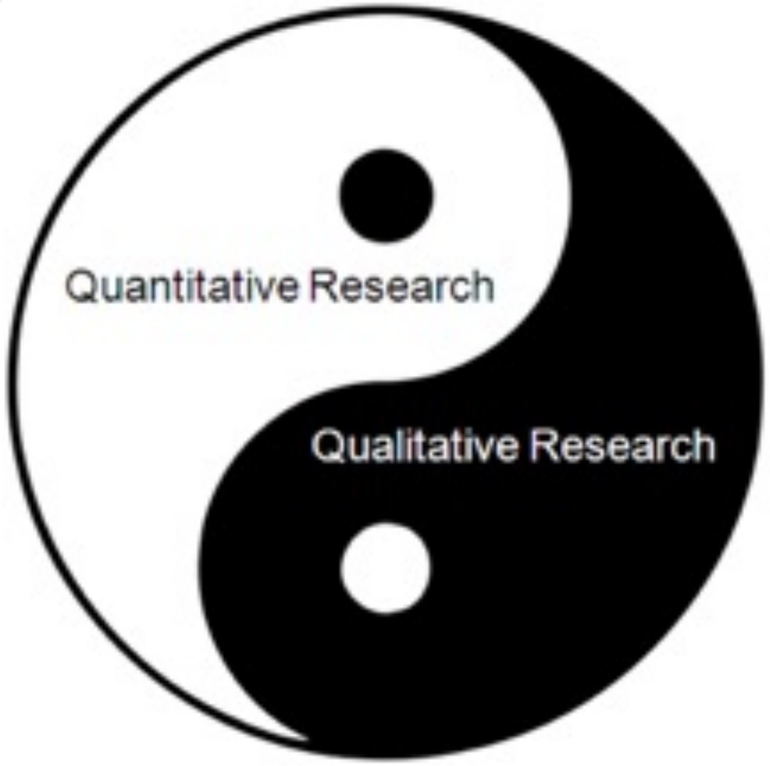


- M. Scriven: The methodology of evaluation, 1967

Qualitative vs. Quantitative Evaluation

http://www.scope-mr.ch/de/dienstleistungen/methoden/

http://www.scope-mr.ch/de/dienstleistungen/methoden/



http://blog.efpsa.org/wp-content/uploads/2012/05/yin_yang.png

Analytic vs. Empirical Evaluation

Scriven, 1967: “If you want to evaluate a tool, say an axe, you might study the design of the bit, the weight distribution, the steel alloy used, the grade of hickory in the handle, etc., or you may just study the kind and speed of the cuts it makes in the hands of a good axeman.”

http://www.wimmer-construction.de/images/product_images/original_images/BE600A.jpg

Empirical and Analytic Methods are Complementary

- Empirical evaluation produces facts which need to be analyzed.
- Analytic evaluation produces facts which need to be tested (empirically).

Chapter 5 - Evaluation

- Types of Evaluation
  - Formative vs. Summative
  - Quantitative vs. Qualitative
  - Analytic vs. Empirical
- Analytic Methods
  - Cognitive Walkthrough
  - Heuristic Evaluation
  - GOMS and KLM
  - Motor Functions: Fitt's Law, Steering Law
- Empirical Methods
  - Field Studies und Lab Studies
  - Longitudinal und Diary Studies
  - Usability Scales

Cognitive Walkthrough

…One or more **evaluators…**

…Step by step…

…along well-defined tasks…

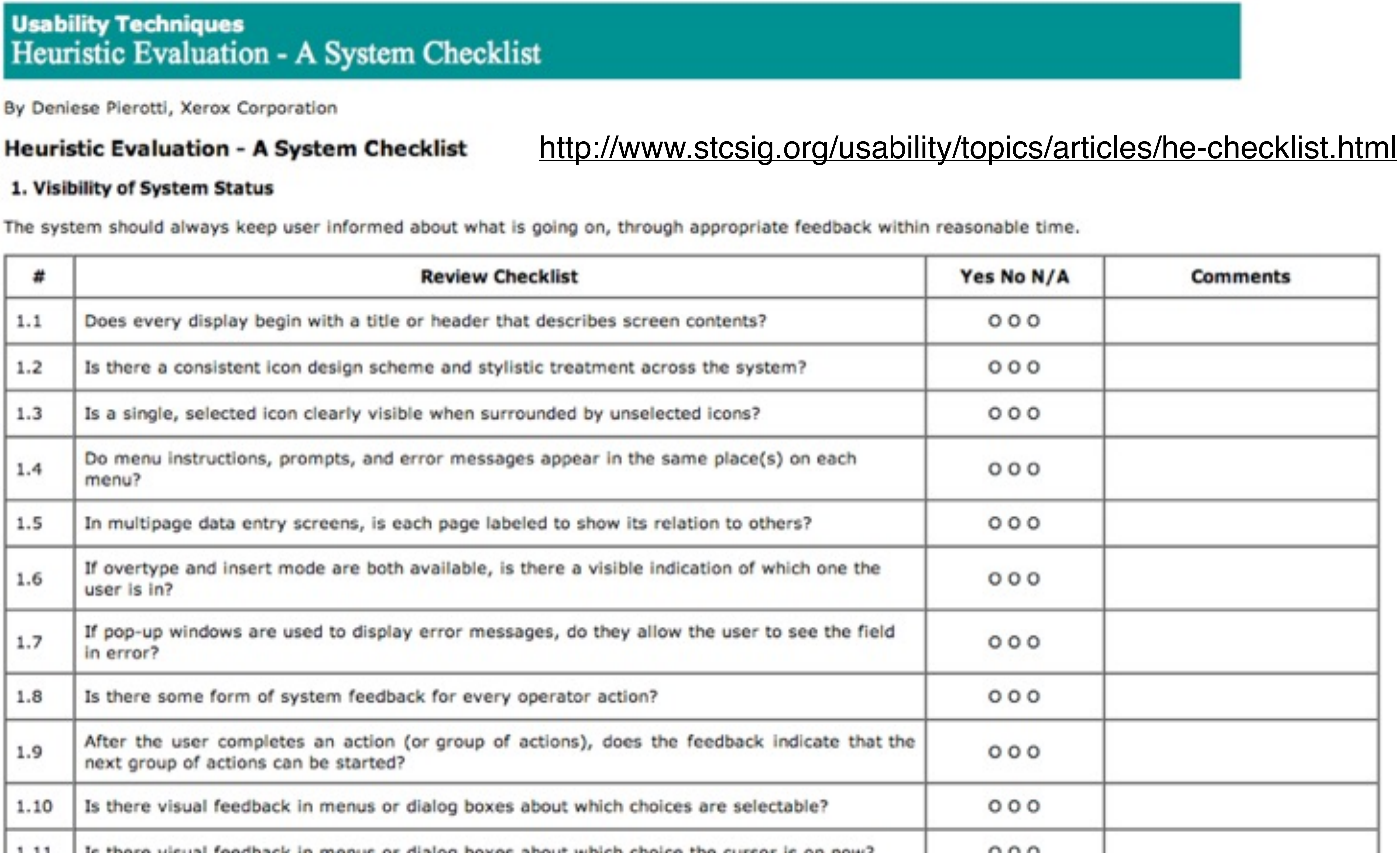
1. Is the **correct action** for executing the next step always clearly defined? Does the user know what to do next?
2. Is the correct action clearly **recognizable**? Does the user actually find it?
3. Does the user receive a sufficient **feedback** after executing the action, such that he can determine whether the action was executed successfully?

10 Usability Heuristics

Jakob Nielsen

- Visibility of system status
- Match between system and the real world
- User control and freedom
- Consistency and standards
- Error prevention
- Recognition rather than recall
- Flexibility and efficiency of use
- Aesthetic and minimalist design
- Help users recognize, diagnose, and recover from errors
- Help and documentation

Detailed Checklist Example



Usability Techniques

Heuristic Evaluation - A System Checklist

By Deniese Pierotti, Xerox Corporation

Heuristic Evaluation - A System Checklist

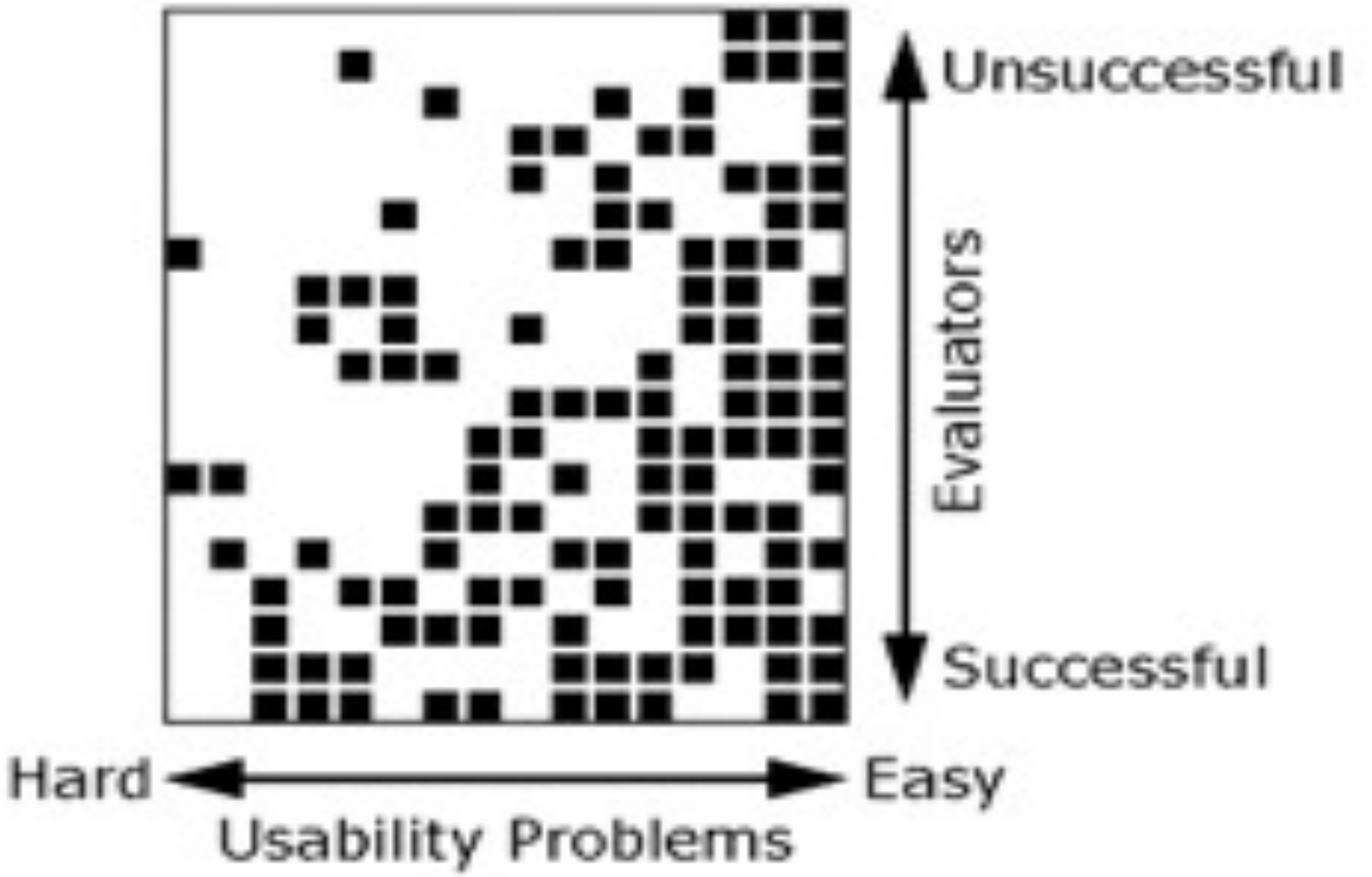
http://www.stcsig.org/usability/topics/articles/he-checklist.html

1. Visibility of System Status

The system should always keep user informed about what is going on, through appropriate feedback within reasonable time.

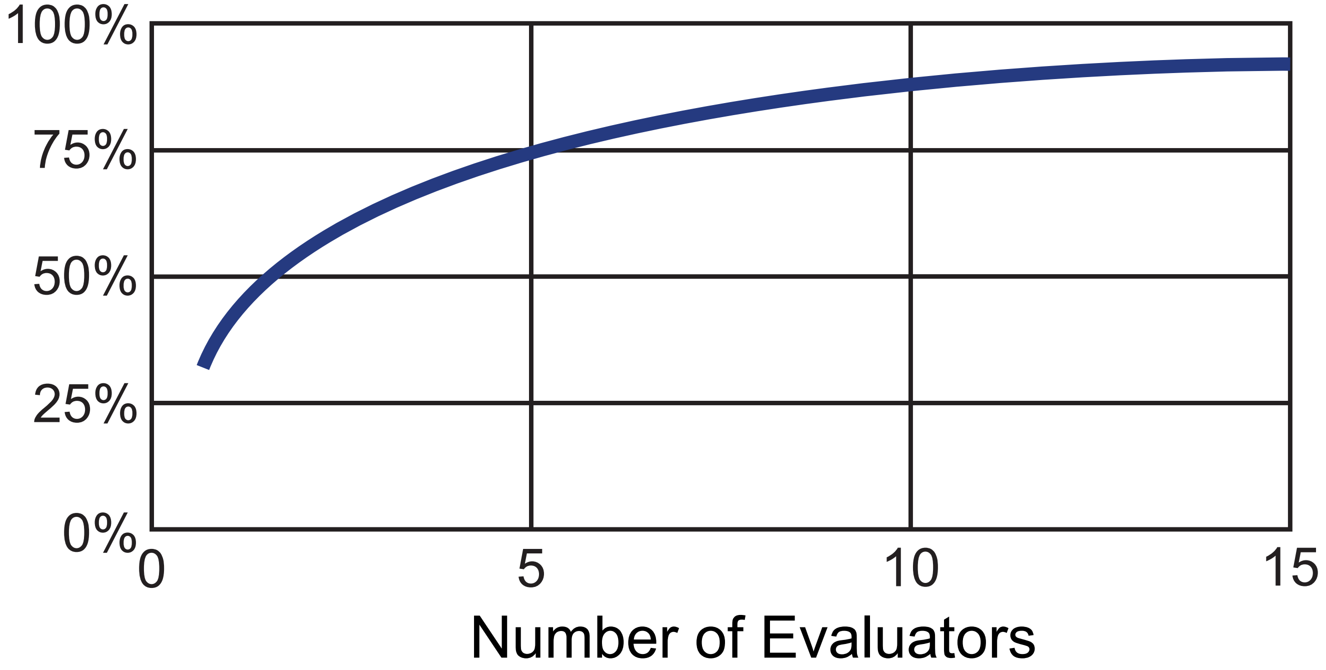
#	Review Checklist	Yes No N/A	Comments
1.1	Does every display begin with a title or header that describes screen contents?	○ ○ ○	
1.2	Is there a consistent icon design scheme and stylistic treatment across the system?	○ ○ ○	
1.3	Is a single, selected icon clearly visible when surrounded by unselected icons?	○ ○ ○	
1.4	Do menu instructions, prompts, and error messages appear in the same place(s) on each menu?	○ ○ ○	
1.5	In multipage data entry screens, is each page labeled to show its relation to others?	○ ○ ○	
1.6	If overtype and insert mode are both available, is there a visible indication of which one the user is in?	○ ○ ○	
1.7	If pop-up windows are used to display error messages, do they allow the user to see the field in error?	○ ○ ○	
1.8	Is there some form of system feedback for every operator action?	○ ○ ○	
1.9	After the user completes an action (or group of actions), does the feedback indicate that the next group of actions can be started?	○ ○ ○	
1.10	Is there visual feedback in menus or dialog boxes about which choices are selectable?	○ ○ ○	
1.11	Is there visual feedback in menus or dialog boxes about which choice the cursor is on now?	○ ○ ○	

http://www.nngroup.com/articles/how-to-conduct-a-heuristic-evaluation/

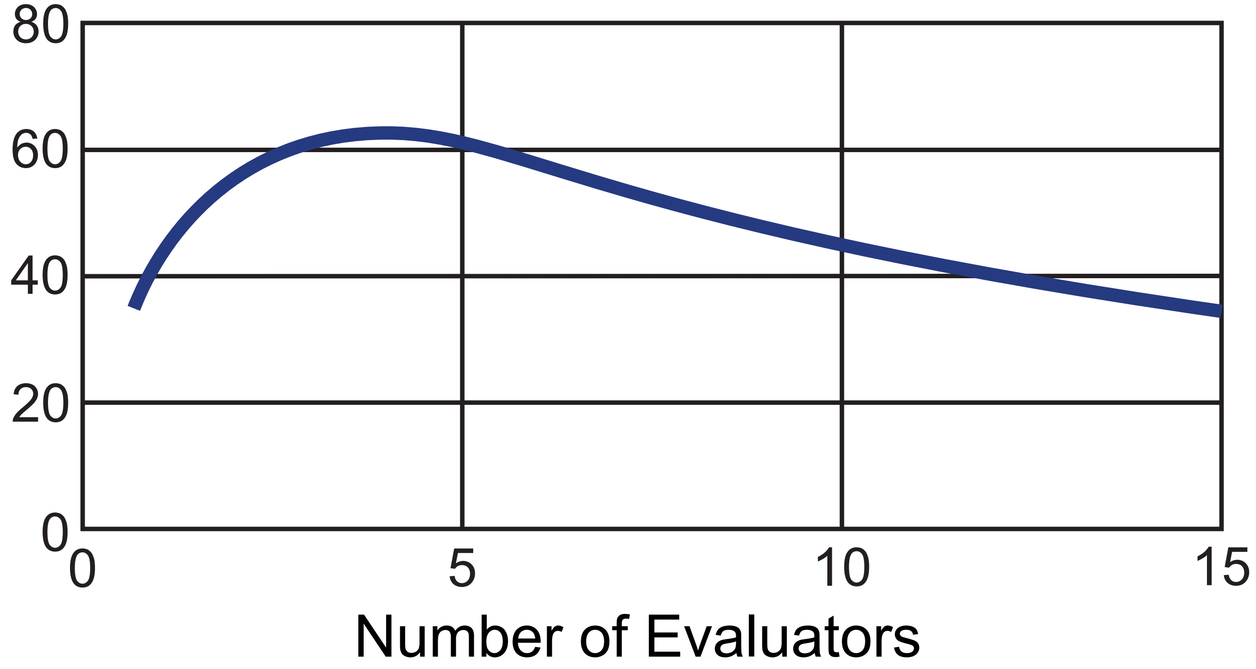


Jakob Nielsen

#Errors



#Errors / Effort



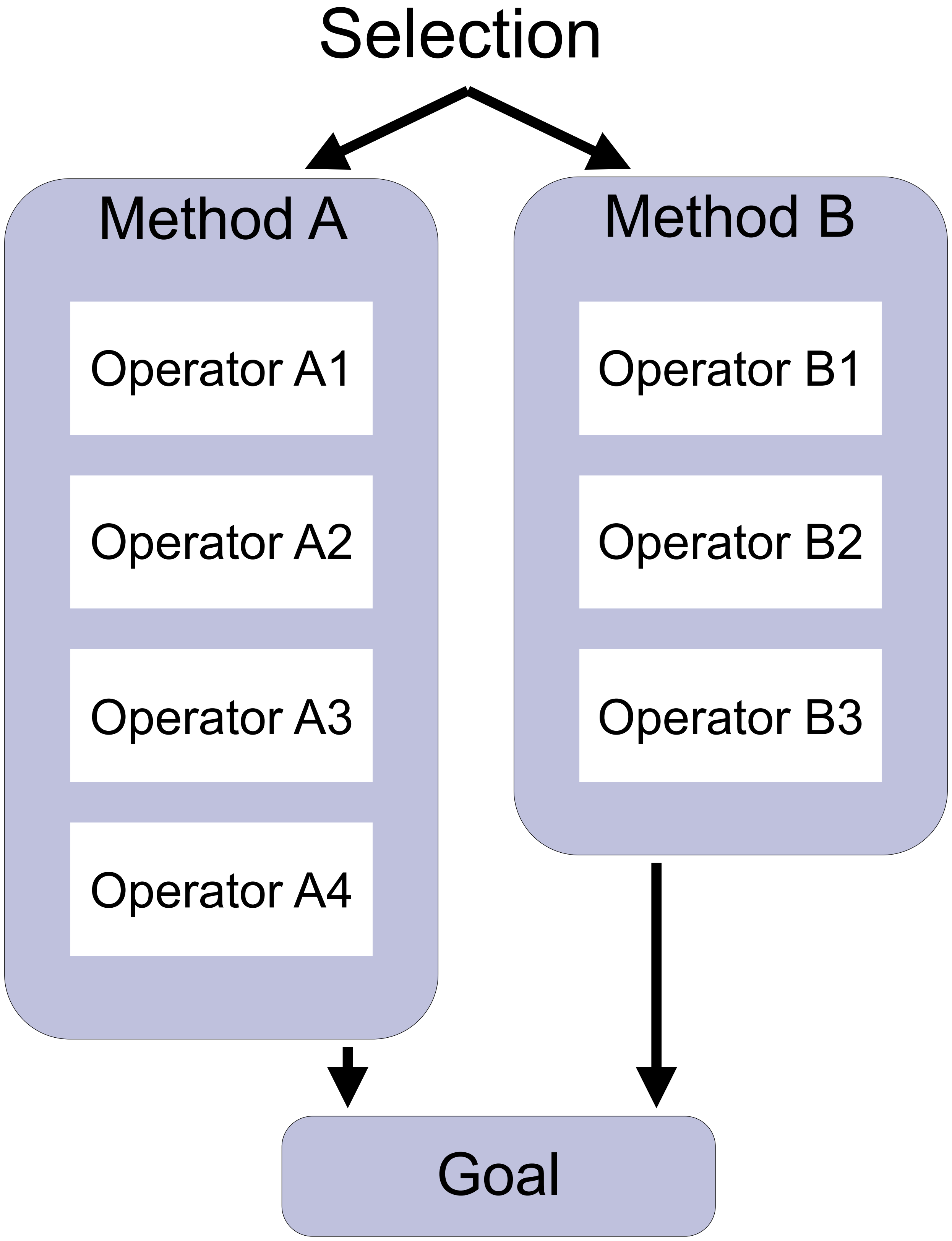
Goals, Operators, Methods & Selection Rules (GOMS)

- **S**election rules

- **M**ethods

- **O**perators

- **G**oals



Card / Moran / Newell: The Psychology of HCI, 1983

Keystroke Level Model (KLM)

Used times in experimental average:

- **K** (Keystroke): Pressing a key: **t_K = 0.28s**
- **P** (Pointing): Pointing to a position on screen: **t_P = 1.1s**
- **B** (Mouse button): Pressing/releasing mouse button: **t_B = 0.1s**
- **H** (Homing): Switch between keyboard and mouse: **t_H = 0.4s**
- **M** (Mental preparation): Mental preparation of successive operation: **t_M = 1.35s**
- **R(t)** (Response time): Response time of the systems (within **t** seconds, system-dependent).

Card / Moran / Newell: The Psychology of HCI, 1983
Data according to D. Kieras (http://courses.wccnet.edu/~jwithrow/docs/klm.pdf)

KLM Example

- Which of the methods M1 or M2 is faster?

- **M1**: Switch to mouse, move mouse pointer to file icon, clicking the icon, dragging to trash icon and release, switch to keyboard
- **M2**: Switch to mouse, selecting the icon, switch to keyboard, press ‘delete’

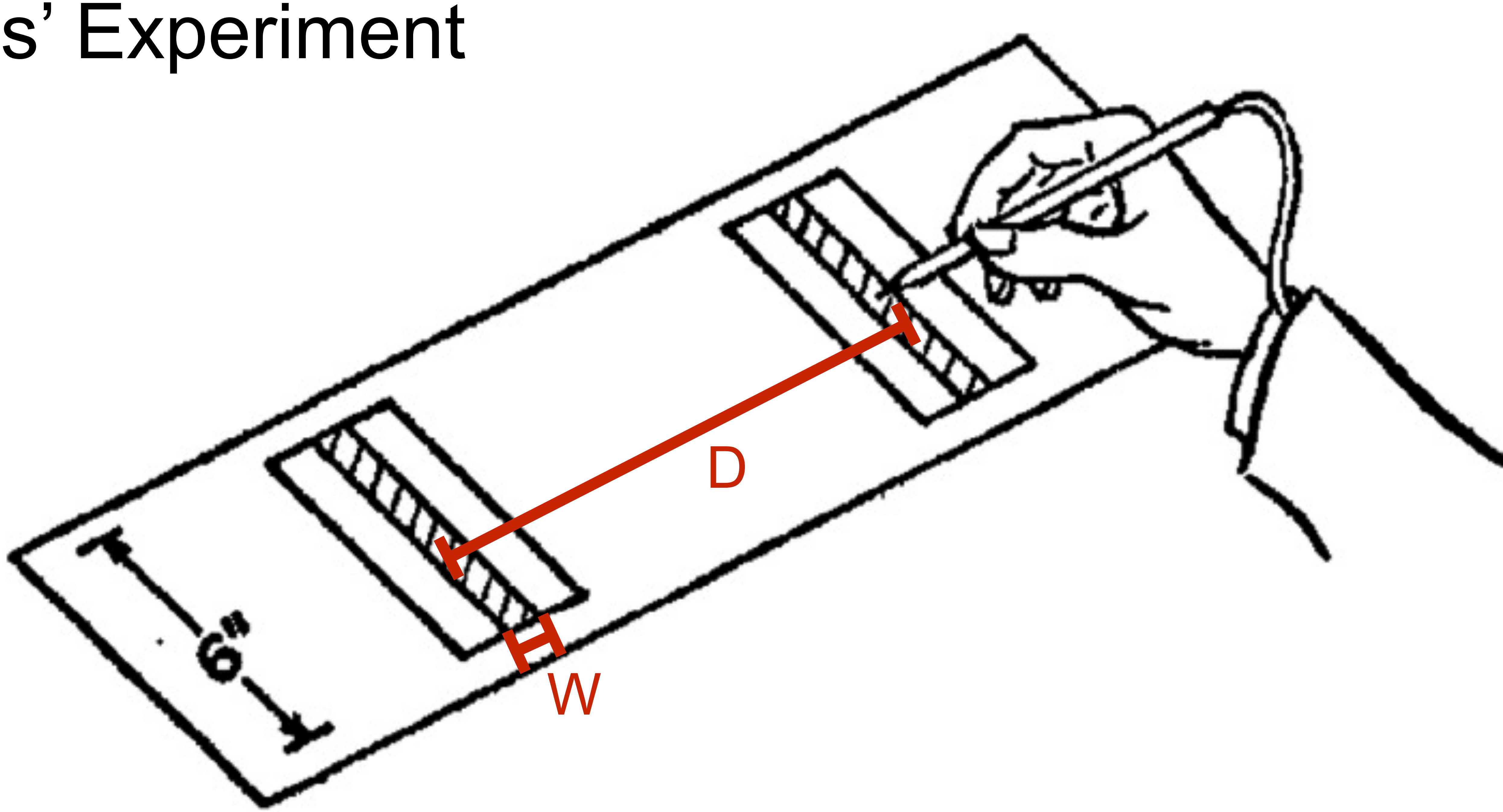
- $\mathbf{t_{M1}} = t_H + t_P + t_B + t_P + t_B + t_H = 0.4 + 1.1 + 0.1 + 1.1 + 0.1$
  $= \mathbf{2.8s}$
- $\mathbf{t_{M2}} = t_H + t_P + t_B + t_H + t_K = 0.4 + 1.1 + 0.1 + 0.4 + 0.28$
  $= \mathbf{2.28s}$

More Sophisticated KLM table

- **K** - Keystroke (.12 - 1.2 sec; .28 recommended for most users).
  - Expert typist (90 wpm): .12 sec
  - Average skilled typist (55 wpm): .20 sec
  - Average nonsecretarial typist (40 wpm): .28 sec
  - Worst typist (unfamiliar with keyboard): 1.2 sec
- **T(n)** - Type a sequence of n characters on a keyboard (n * K sec).
- **P** - Point with mouse to a target on the display (1.1 sec).
  - The actual time required can be determined from *Fitts' law*.
  - For typical situations, it ranges from .8 to 1.5 sec, with an average of 1.1 sec.
- **B** - Press or release mouse button (.1 sec).
- **BB** - Click and release mouse button (.2 sec).
- **H** - Home hands to keyboard or mouse (.4 sec).

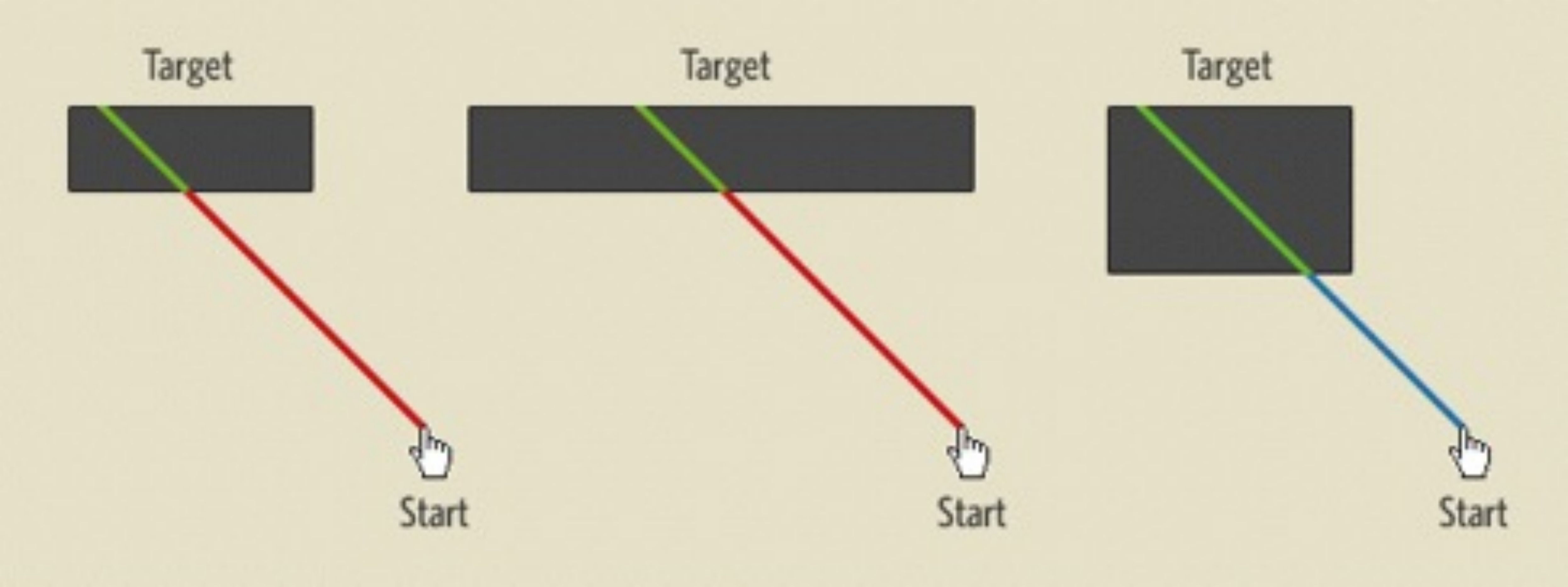
Speed vs. Accuracy

Fitts' Experiment



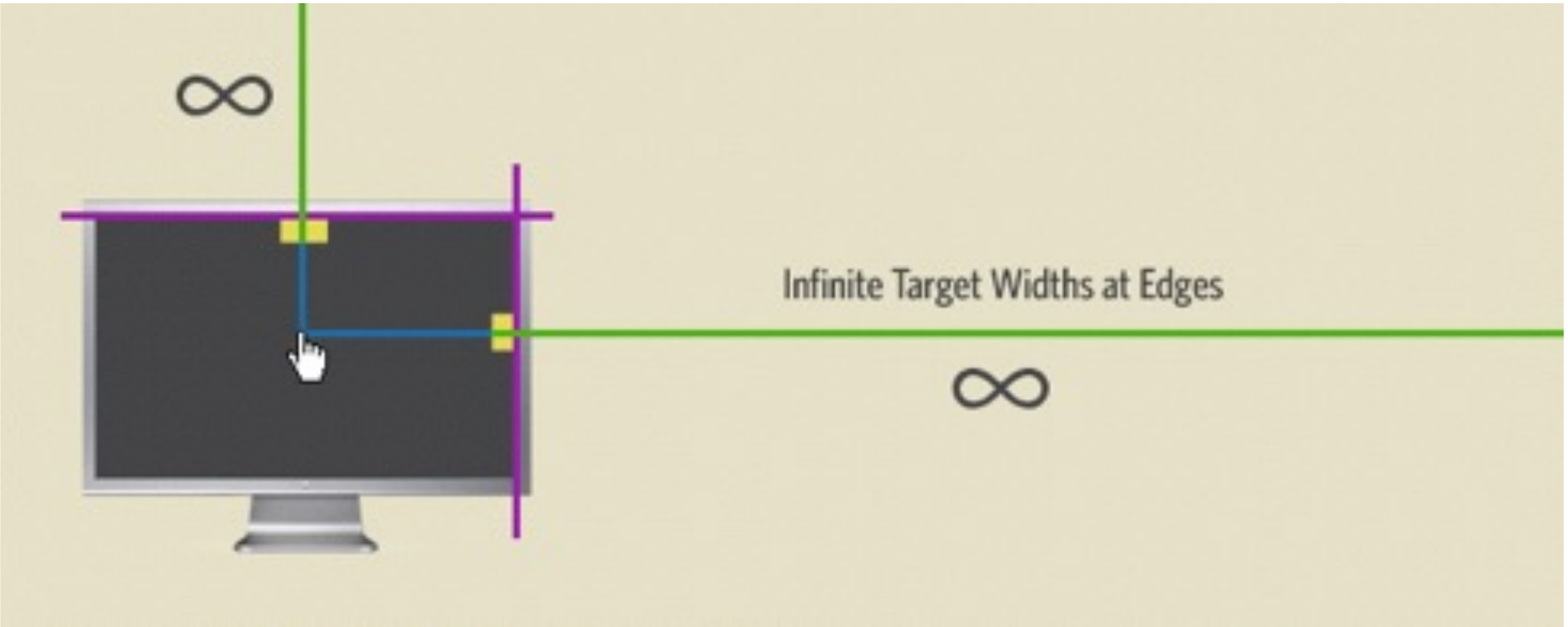
$$MT = a + b * ID = a + b * \log_2(\frac{D}{W} + 1)$$

Enlarge Targets, the Right Way!

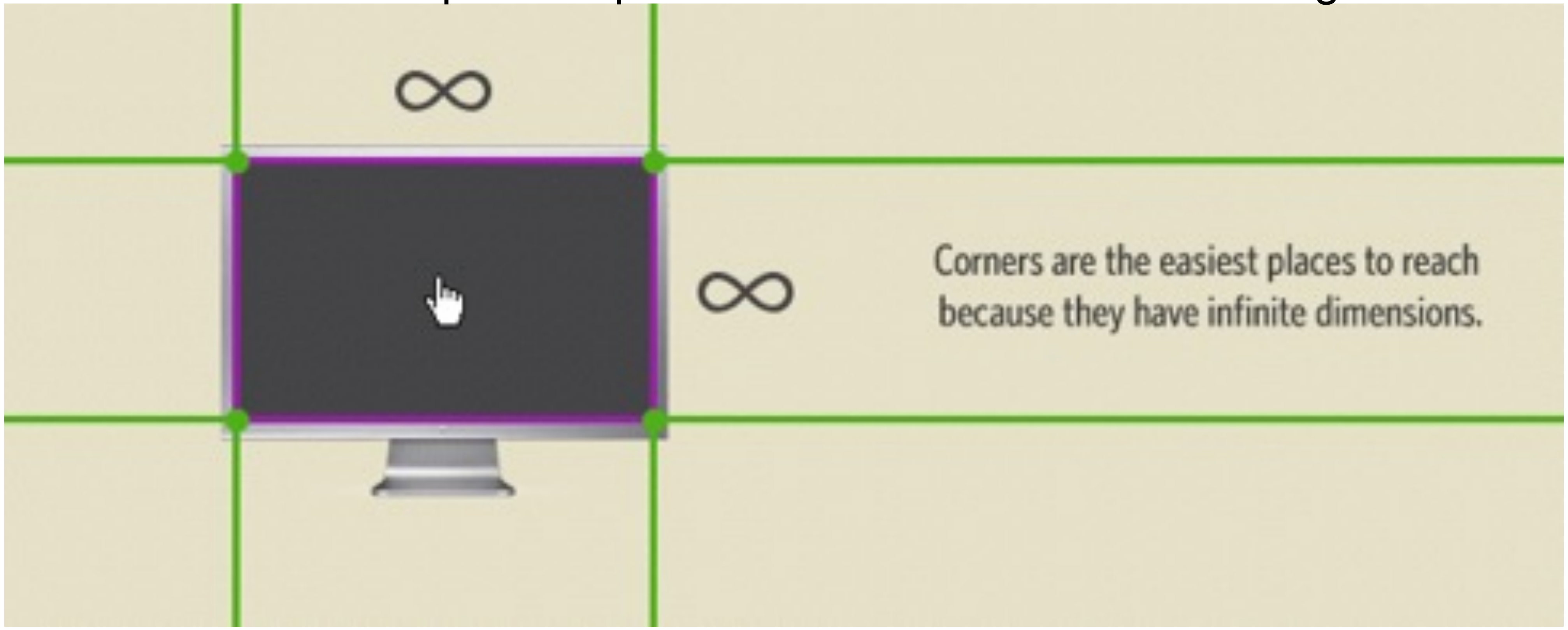


http://www.particletree.com/features/visualizing-fittss-law/

Not All Pixels Are Equal (before Fitts' Law)



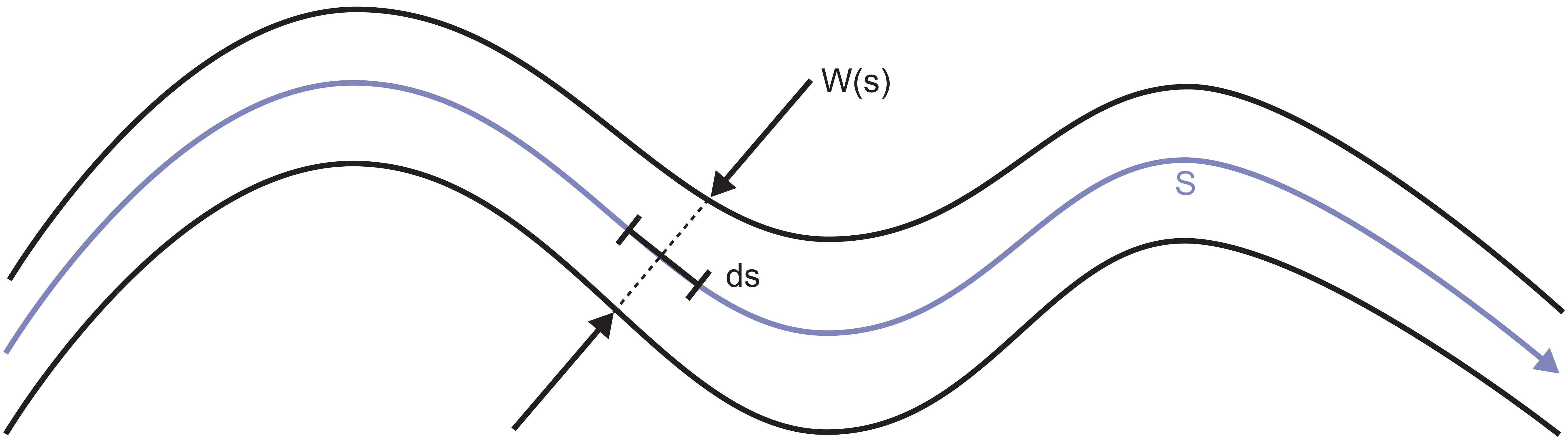
http://www.particletree.com/features/visualizing-fittss-law/


∞
∞
Corners are the easiest places to reach
because they have infinite dimensions.

Steering Law ???

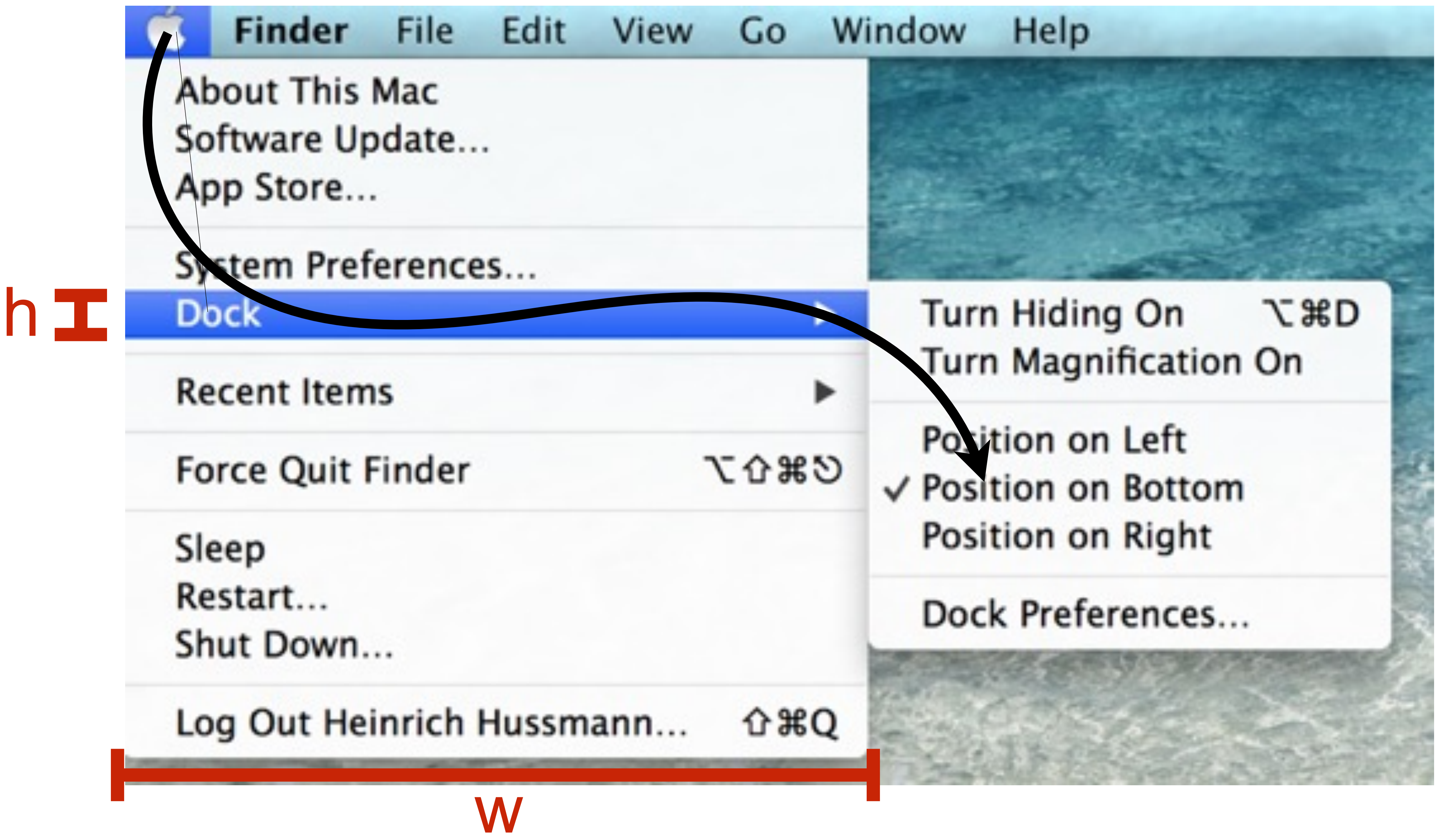
http://www.all-wallpapers.net/wp-content/uploads/2012/12/Winding-Road-Nature.jpeg

Time for Driving Along a Narrow Road



$$T = a + b * \int_S \frac{1}{W(s)} \mathrm{d}s$$

Narrow Roads on Screens

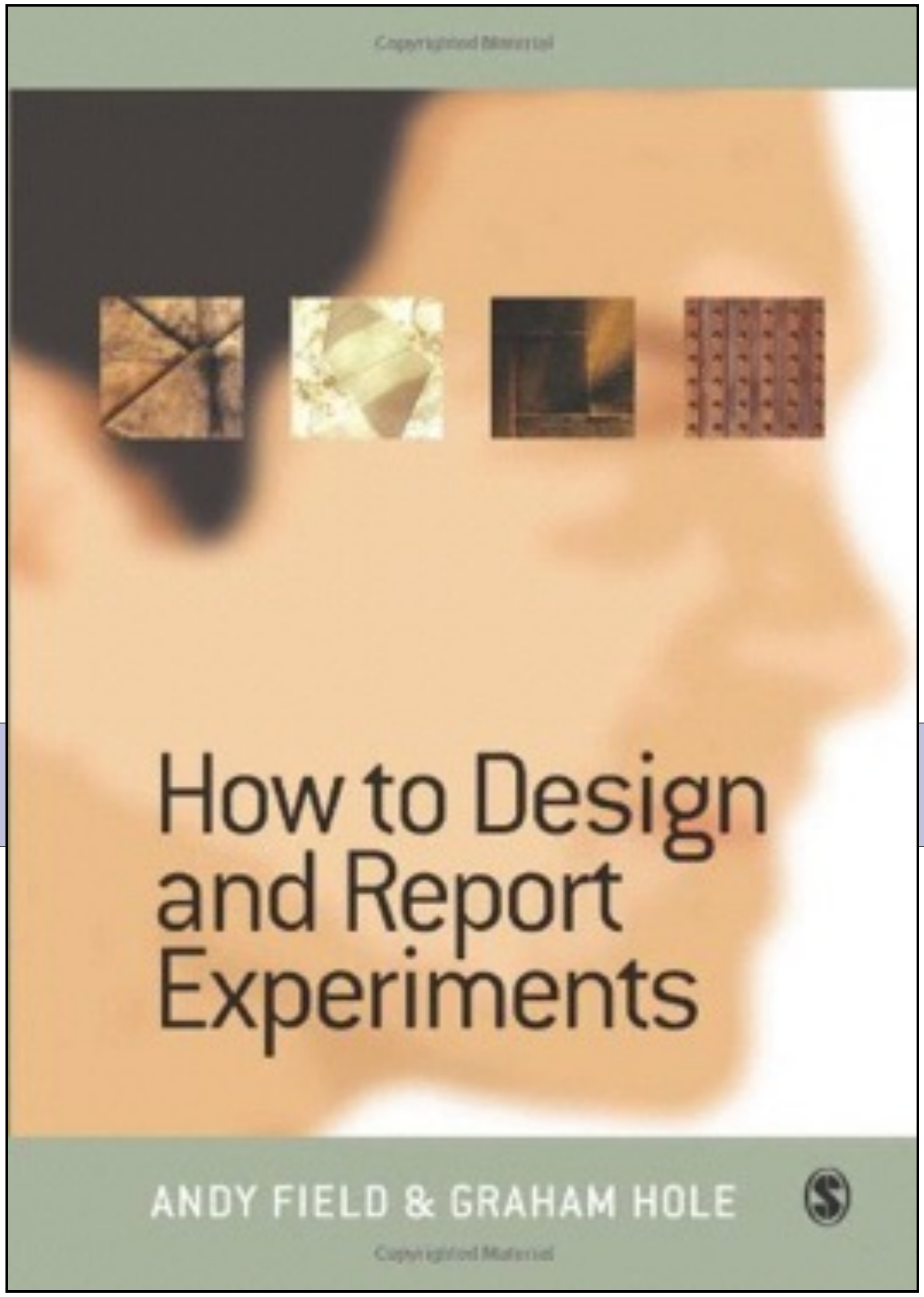


$$T = \boxed{a_1 + b_1 * \log_2(\frac{nh}{h} + 1)} + \boxed{a_2 + b_2 * \frac{w}{h}} + \ldots$$

vertical: Fitts' law horizontal: steering law

Chapter 5 - Evaluation

- Types of Evaluation
  - Formative vs. Summative
  - Quantitative vs. Qualitative
  - Analytic vs. Empirical
- Analytic Methods
  - Cognitive Walkthrough
  - Heuristic Evaluation
  - GOMS and KLM
  - Motor Functions: Fitt's Law, Steering Law
- Empirical Methods
  - Field Studies und Lab Studies
  - Longitudinal und Diary Studies
  - Usability Scales

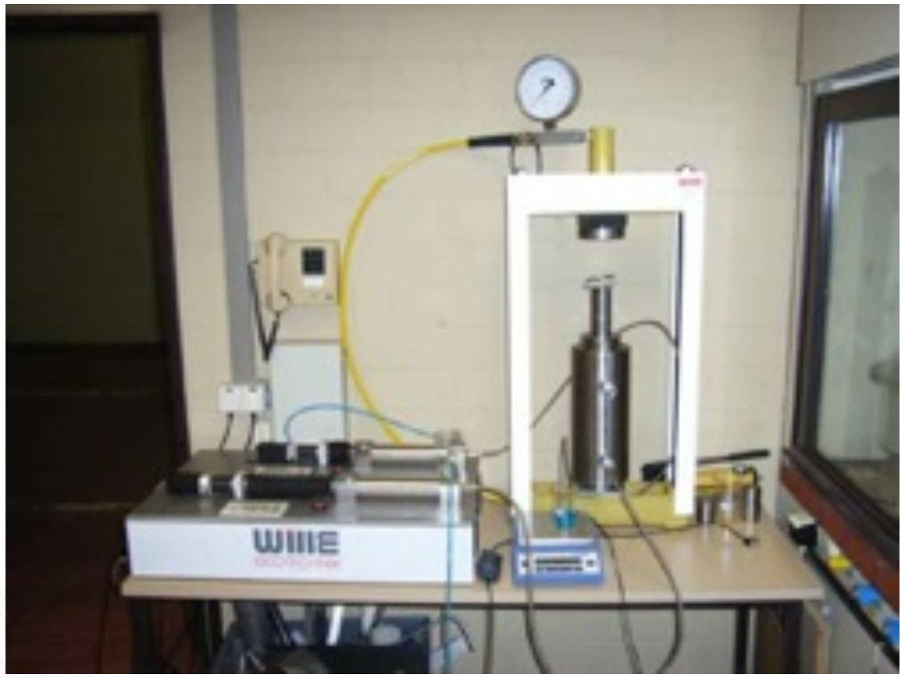


http://www.amazon.de/dp/0857028294

Quality Properties of Empirical Methods

- Objectivity
- Reproducibility
- Validity
  - internal
  - external
- Relevance

http://www.schwimmvereinapolda.de/images/Webelemente/Stoppuhr.jpg



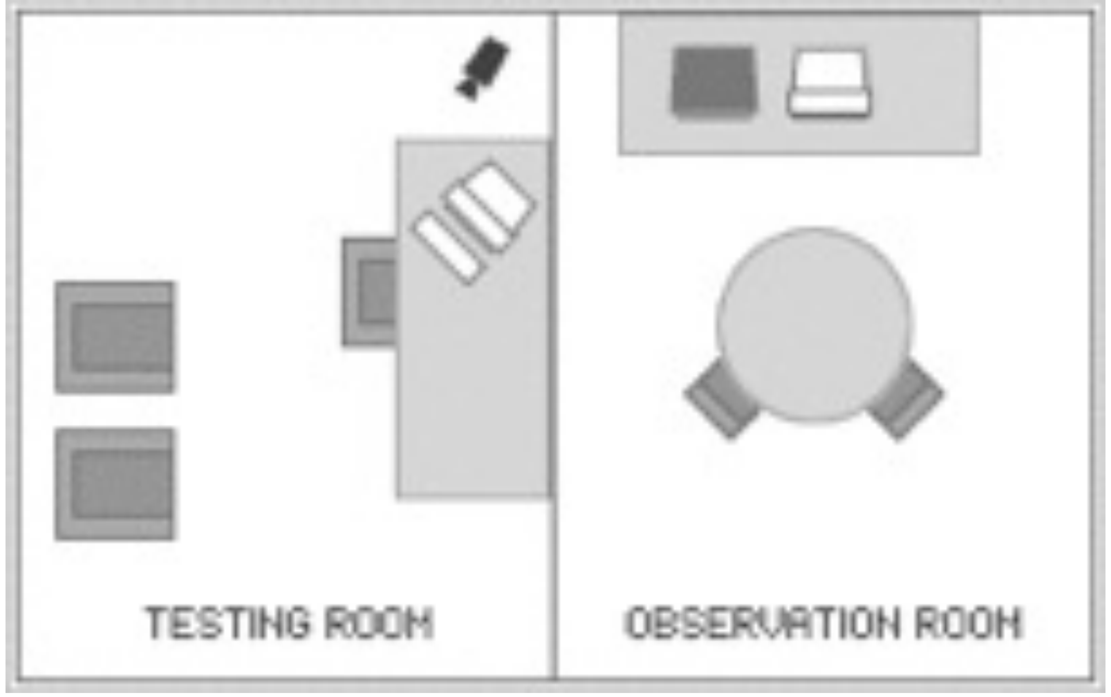
http://www.bgr.bund.de/DE/Themen/Endlagerung/Bilder/end_nfpro_hyperf_g.jpg?__blob=normal&v=2

http://wl15www815.webland.ch/travelinfos/images/mensch/gehirn4.jpg

http://bilder.n3po.com/cache/Photos/Bach-Fliessend-Bergab_w475_h230_cw475_ch230_thumb.jpg

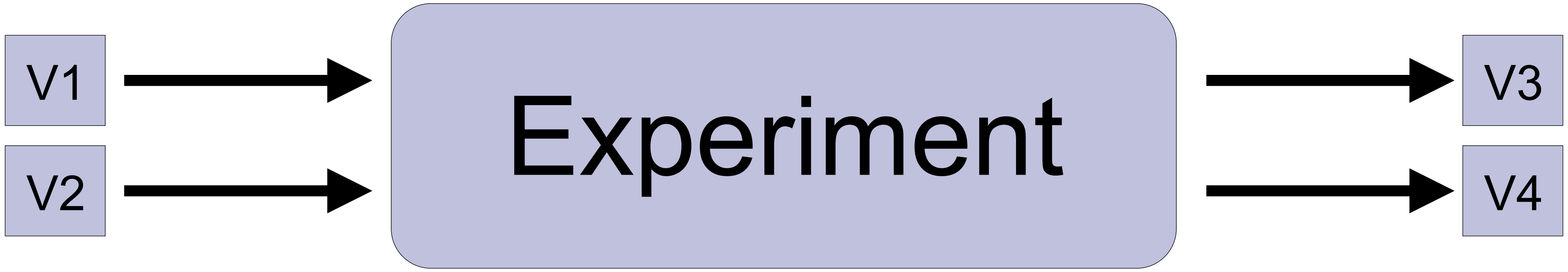
Field Study vs Lab Study

- External Validity
- Internal Validity
- Effort



Source: www.xperienceconsulting.com

Variables and Values

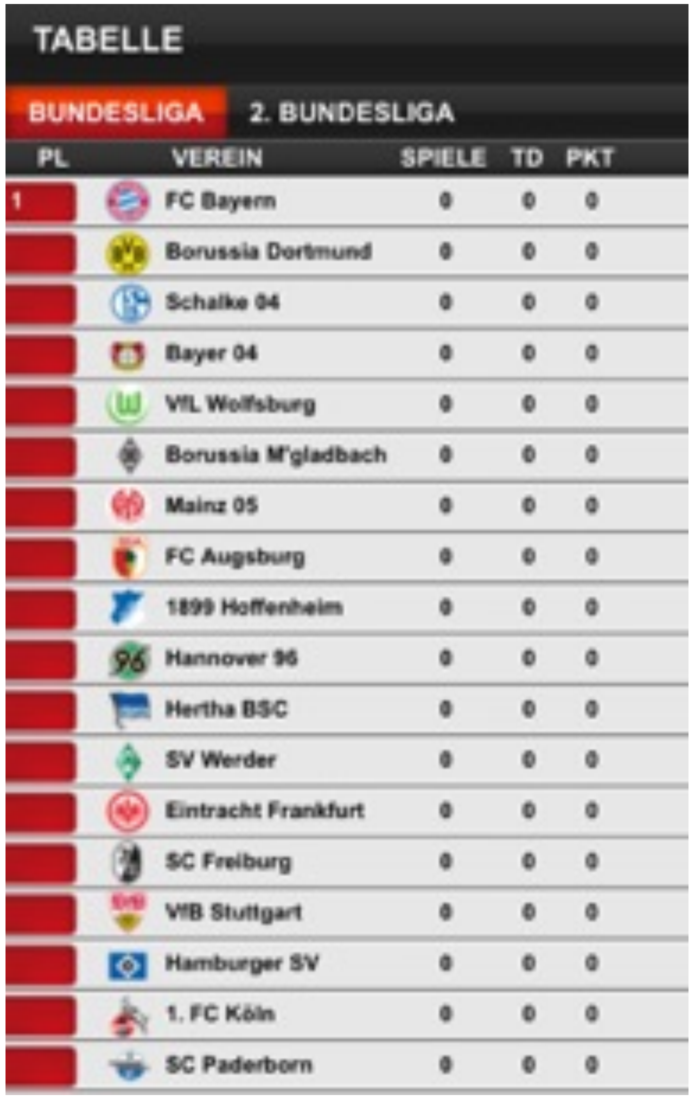


independent

dependent

- Nominal
- Ordinal
- Cardinal

http://www.gebr-clasen.de/-g400/weltkarte_KB.png



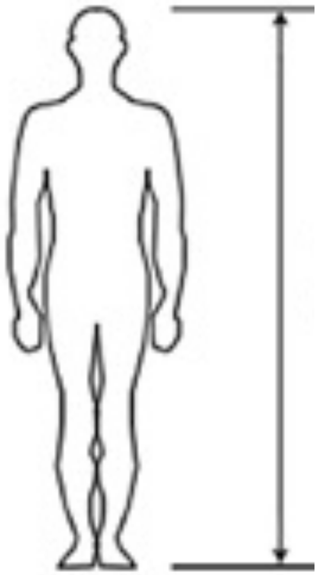
TABELLE

BUNDESLIGA | 2. BUNDESLIGA

PL	VEREIN	SPIELE	TD	PKT
1	FC Bayern	0	0	0
	Borussia Dortmund	0	0	0
	Schalke 04	0	0	0
	Bayer 04	0	0	0
	VfL Wolfsburg	0	0	0
	Borussia M'gladbach	0	0	0
	Mainz 05	0	0	0
	FC Augsburg	0	0	0
	1899 Hoffenheim	0	0	0
	Hannover 96	0	0	0
	Hertha BSC	0	0	0
	SV Werder	0	0	0
	Eintracht Frankfurt	0	0	0
	SC Freiburg	0	0	0
	VfB Stuttgart	0	0	0
	Hamburger SV	0	0	0
	1. FC Köln	0	0	0
	SC Paderborn	0	0	0

http://www.bundesliga.de/

http://www.zukunftskinder.org/wp-content/uploads/2013/01/großfamilie.jpg



http://www.kreativrad.de/img/parts/fahrrad-massanfertigung-koerpergroesse.png

Observation Study (Example)

http://cdn3.spiegel.de/images/image-109402-panoV9free-nsqt.jpg

http://www.experto.de/software-fuer-studenten-800px-534px0.jpg

- One independent variable: Participation in tutorials (Yes / No)
  - Assuming participation is voluntary
- One dependent variable: Achieved grade in test
- 108 subjects, 54 “yes”, 54 “no” (to participation question)
- Measurement shows: Grade positively ***correlated*** with tutorial participation
- Beware of ***confounding variables***!

Controlled Experiment

http://cdn3.spiegel.de/images/image-109402-panoV9free-nsqt.jpg

http://www.experto.de/software-fuer-studenten-800px-534px0.jpg

- One independent variable: Participation in tutorials (Yes / No)
  - assigned randomly to subjects !!!
- One dependent variable: Achieved grade in test
- 108 subjects, 54 “participating” condition, 54 “not-participating” condition
- Measurement: Grade positively ***correlated*** with participation
- Causal relationship established: Participation in tutorials leads to better grade

Experiment Design

	HCI1	Analysis	Algebra
Yes	Condition 1	Condition 2	Condition 3
No	Condition 4	Condition 5	Condition 6

- 2 Variables with 2 resp. 3 values: 2x3 = 6 Conditions
- **within-subjects**: everybody does everything
- **between-groups**: groups, each group does one condition
- Vary the order to avoid **learning** and **fatigue effects**
  - Randomisation
  - Permutation
  - Latin square

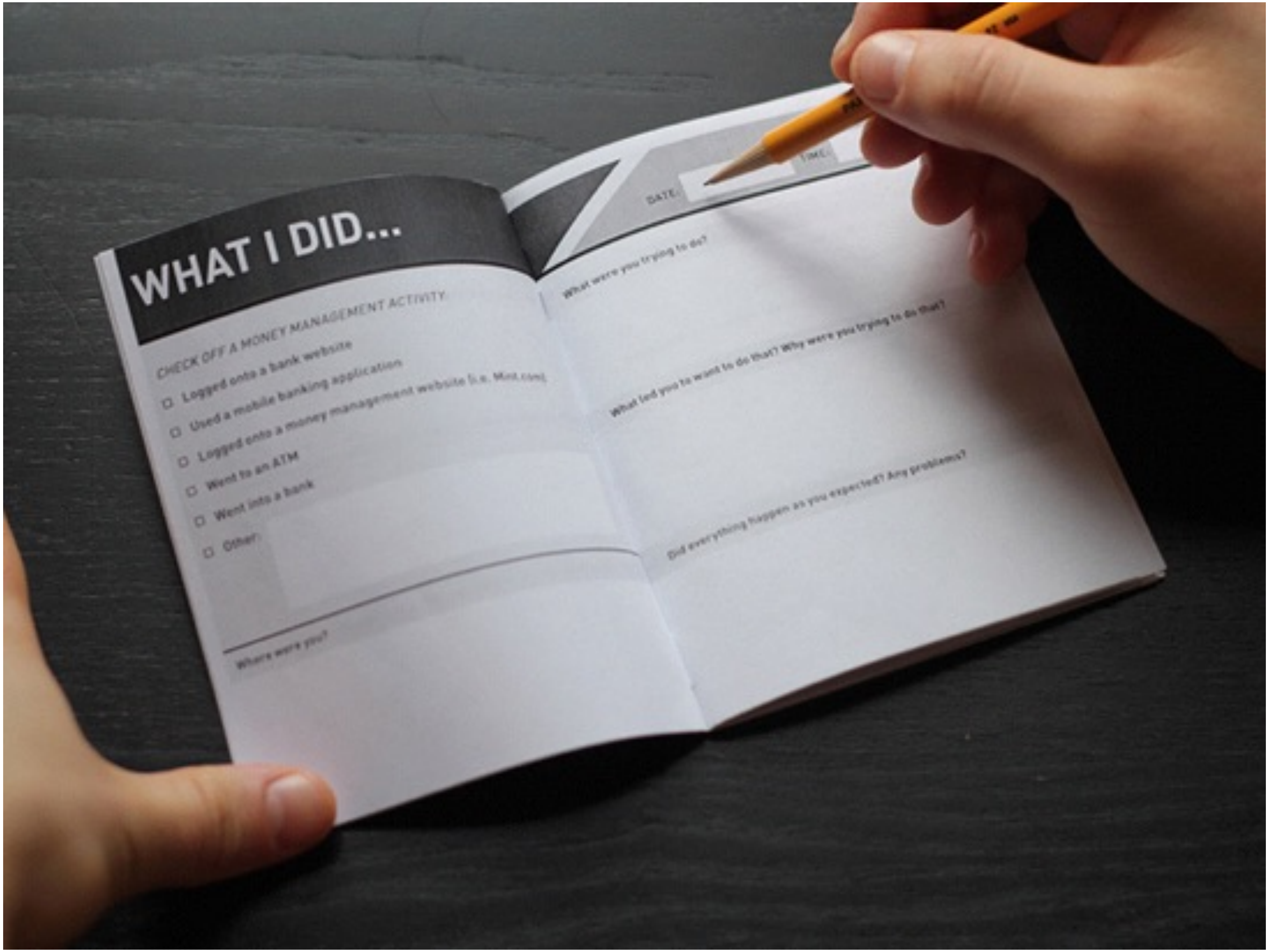
Cond. 6	Cond. 1	Cond. 5	Cond. 2	Cond. 4	Cond. 3
Cond. 5	Cond. 6	Cond. 4	Cond. 1	Cond. 3	Cond. 2
Cond. 2	Cond. 3	Cond. 1	Cond. 4	Cond. 6	Cond. 5
Cond. 1	Cond. 2	Cond. 6	Cond. 3	Cond. 5	Cond. 4
Cond. 4	Cond. 5	Cond. 3	Cond. 6	Cond. 2	Cond. 1
Cond. 3	Cond. 4	Cond. 2	Cond. 5	Cond. 1	Cond. 6

Hypotheses and Significance

- H: Tutorial participants achieve better grades in test.
- H_0: Tutorial participants and non-participants achieve in average the same grades in test. *(null hypothesis)*
- Effect size = difference of mean values (unknown in advance)
- Trick: Instead of proving H, dis-prove H_0.
  Then H is implicitly proven – independent of effect size.

- Significance:
  - *p-value: probability* of obtaining the observed results when null hypothesis is true
  - *statistical significance:* p-value less than *significance level*
    Often 0,05 (= 5%)
  - obtaining p-values: *tests* dependent on experiment design

Longitudinal and Diary Studies



WHAT I DID...

CHECK OFF A MONEY MANAGEMENT ACTIVITY:

- □ Logged onto a bank website
- □ Used a mobile banking application
- □ Logged onto a money management website (i.e. Mint.com)
- □ Went to an ATM
- □ Went into a bank
- □ Other:

Where were you?

DATE:

TIME:

What were you trying to do?

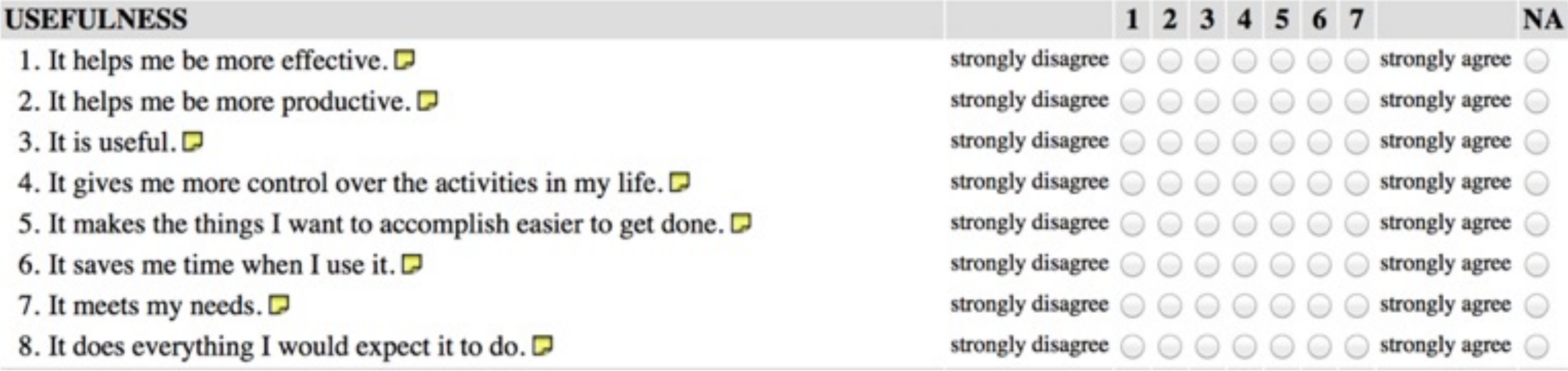
What led you to want to do that? Why were you trying to do that?

Did everything happen as you expected? Any problems?

http://www.hcii.cmu.edu/M-HCI/2011/BOA-PlanningTools/images/diary_study.jpg

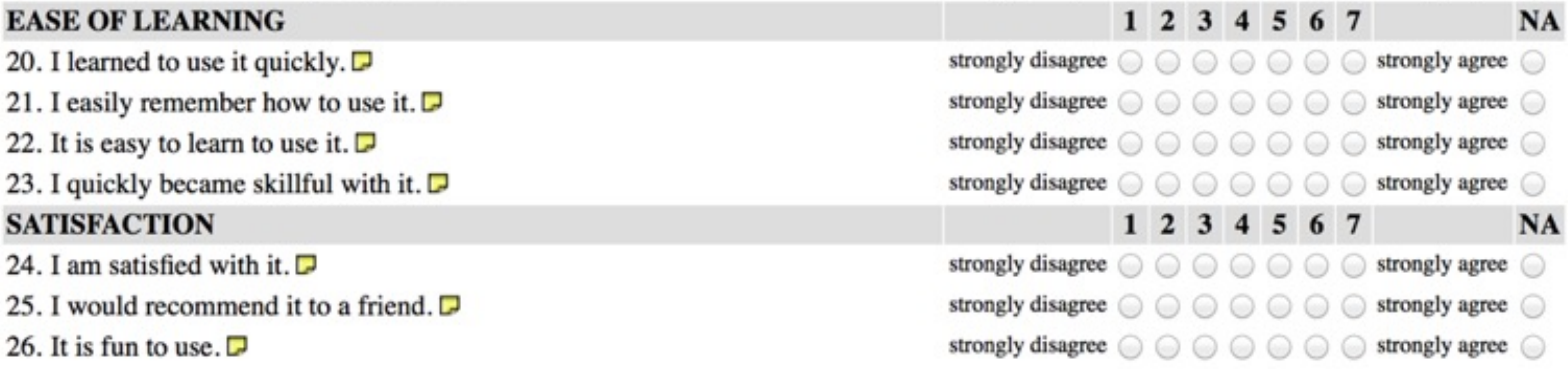
USE:
Usefulness, Satisfaction and Ease of Use

- Lund 2001: 30 questions with 7-point Likert scales



USEFULNESS		1	2	3	4	5	6	7		NA
1. It helps me be more effective.	strongly disagree	○	○	○	○	○	○	○	strongly agree	○
2. It helps me be more productive.	strongly disagree	○	○	○	○	○	○	○	strongly agree	○
3. It is useful.	strongly disagree	○	○	○	○	○	○	○	strongly agree	○
4. It gives me more control over the activities in my life.	strongly disagree	○	○	○	○	○	○	○	strongly agree	○
5. It makes the things I want to accomplish easier to get done.	strongly disagree	○	○	○	○	○	○	○	strongly agree	○
6. It saves me time when I use it.	strongly disagree	○	○	○	○	○	○	○	strongly agree	○
7. It meets my needs.	strongly disagree	○	○	○	○	○	○	○	strongly agree	○
8. It does everything I would expect it to do.	strongly disagree	○	○	○	○	○	○	○	strongly agree	○

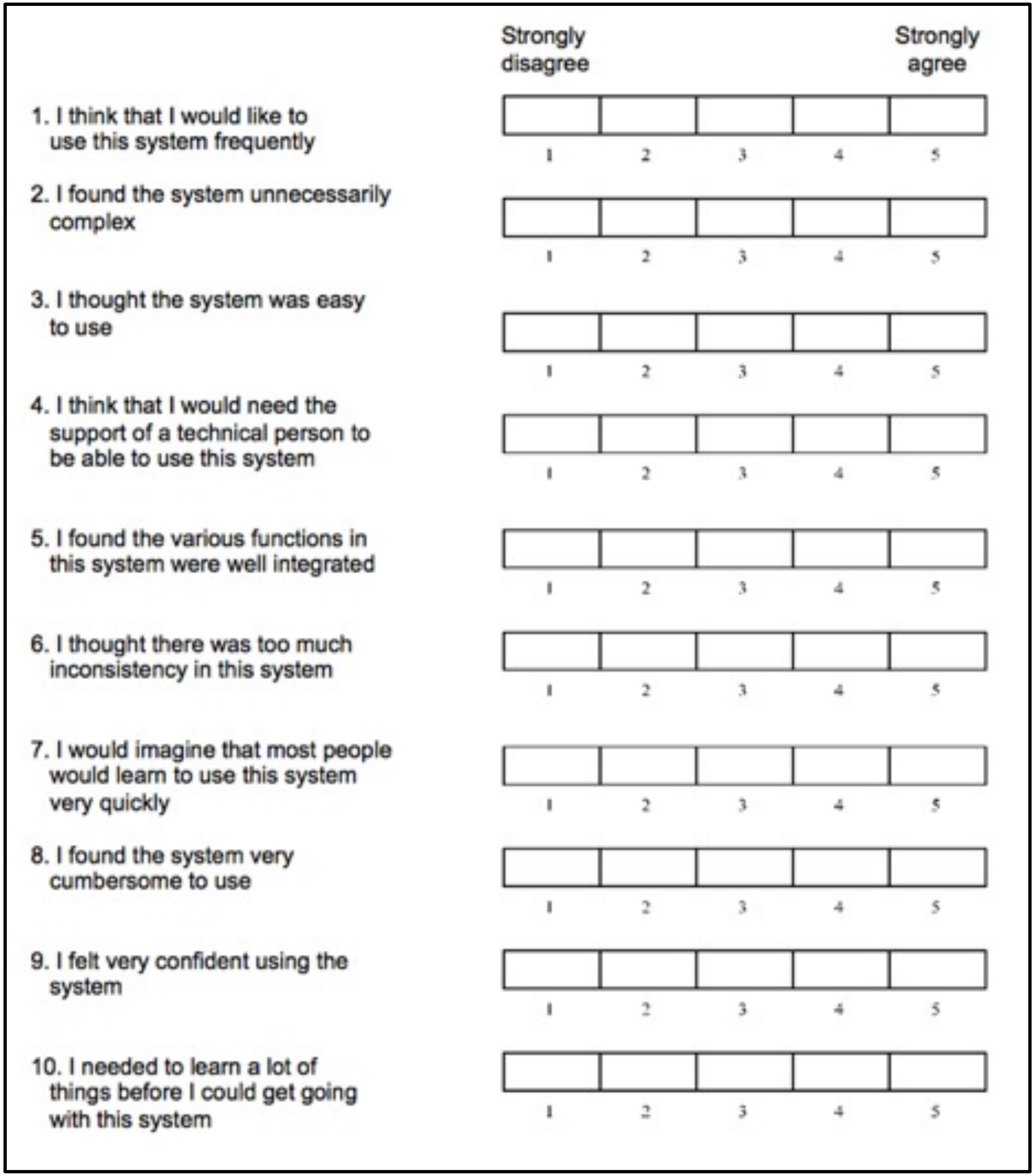
...



EASE OF LEARNING		1	2	3	4	5	6	7		NA
20. I learned to use it quickly.	strongly disagree	○	○	○	○	○	○	○	strongly agree	○
21. I easily remember how to use it.	strongly disagree	○	○	○	○	○	○	○	strongly agree	○
22. It is easy to learn to use it.	strongly disagree	○	○	○	○	○	○	○	strongly agree	○
23. I quickly became skillful with it.	strongly disagree	○	○	○	○	○	○	○	strongly agree	○
SATISFACTION		**1**	**2**	**3**	**4**	**5**	**6**	**7**		**NA**
24. I am satisfied with it.	strongly disagree	○	○	○	○	○	○	○	strongly agree	○
25. I would recommend it to a friend.	strongly disagree	○	○	○	○	○	○	○	strongly agree	○
26. It is fun to use.	strongly disagree	○	○	○	○	○	○	○	strongly agree	○

SUS: System Usability Scale

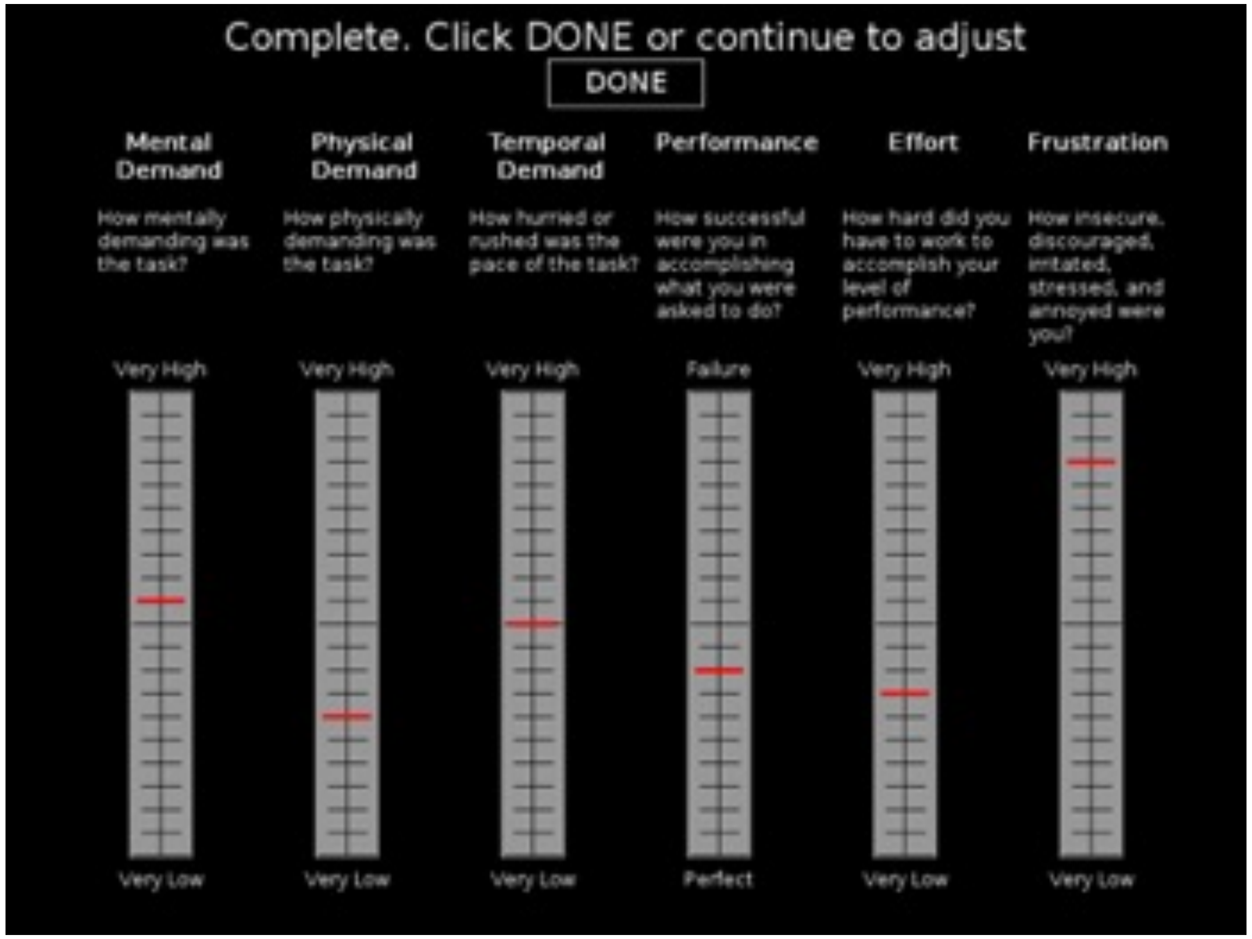
- Brooke (DEC) 1986
  - "Quick and dirty", very popular
  - 10 questions
  - 5-point Likert scale
  - Adapted for Web sites: Tullis / Stetson (Fidelity Investments) 2004



	Strongly disagree				Strongly agree
1. I think that I would like to use this system frequently	□ 1	□ 2	□ 3	□ 4	□ 5
2. I found the system unnecessarily complex	□ 1	□ 2	□ 3	□ 4	□ 5
3. I thought the system was easy to use	□ 1	□ 2	□ 3	□ 4	□ 5
4. I think that I would need the support of a technical person to be able to use this system	□ 1	□ 2	□ 3	□ 4	□ 5
5. I found the various functions in this system were well integrated	□ 1	□ 2	□ 3	□ 4	□ 5
6. I thought there was too much inconsistency in this system	□ 1	□ 2	□ 3	□ 4	□ 5
7. I would imagine that most people would learn to use this system very quickly	□ 1	□ 2	□ 3	□ 4	□ 5
8. I found the system very cumbersome to use	□ 1	□ 2	□ 3	□ 4	□ 5
9. I felt very confident using the system	□ 1	□ 2	□ 3	□ 4	□ 5
10. I needed to learn a lot of things before I could get going with this system	□ 1	□ 2	□ 3	□ 4	□ 5

NASA TLX

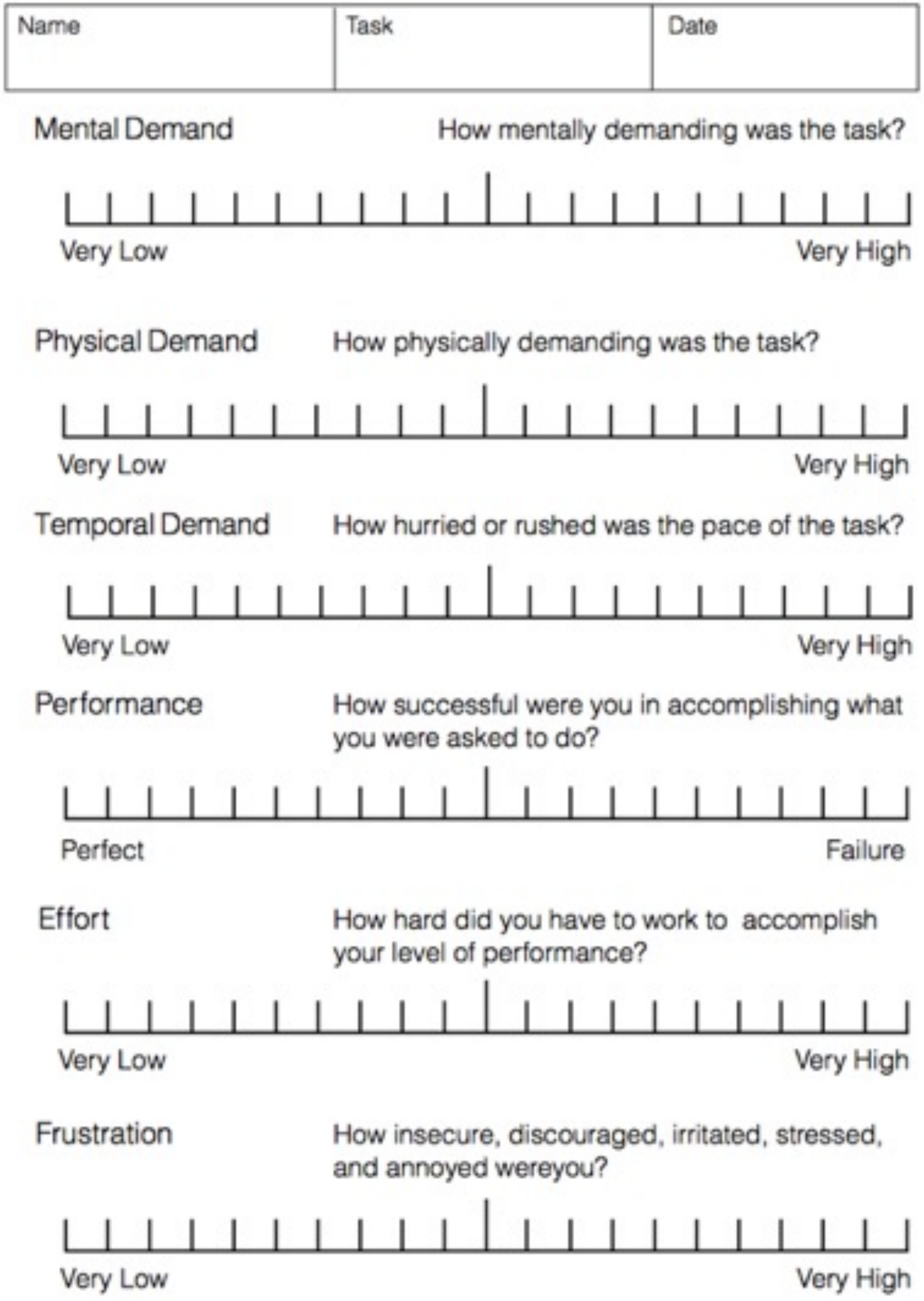
- Measurement for perceived workload
  - NASA AMES Research 1986
  - 100 points per subscale, 5-point steps (i.e. neutral plus 10 values in each direction)



Complete. Click DONE or continue to adjust

Mental Demand	Physical Demand	Temporal Demand	Performance	Effort	Frustration
How mentally demanding was the task?	How physically demanding was the task?	How hurried or rushed was the pace of the task?	How successful were you in accomplishing what you were asked to do?	How hard did you have to work to accomplish your level of performance?	How insecure, discouraged, irritated, stressed, and annoyed were you?
Very High	Very High	Very High	Failure	Very High	Very High
Very Low	Very Low	Very Low	Perfect	Very Low	Very Low

http://humansystems.arc.nasa.gov/groups/TLX/



Name	Task	Date

Mental Demand — How mentally demanding was the task?

Very Low — Very High

Physical Demand — How physically demanding was the task?

Very Low — Very High

Temporal Demand — How hurried or rushed was the pace of the task?

Very Low — Very High

Performance — How successful were you in accomplishing what you were asked to do?

Perfect — Failure

Effort — How hard did you have to work to accomplish your level of performance?

Very Low — Very High

Frustration — How insecure, discouraged, irritated, stressed, and annoyed wereyou?

Very Low — Very High

PANAS

Positive and Negative Affect Scale

Watson, D., Clark, L. A. & Tellegen, A. (1998). Development and validation of brief measures of positive and negative affect: The PANAS Scales. Journal of Personality and Social Psychology, 54, 1063–1070.

attentive

interested

alert

excited

enthusiastic

inspired

proud

determined

strong

active

distressed

upset

hostile

irritable

scared

afraid

ashamed

guilty

nervous

jittery

User Experience (UX) Design

- Marc Hassenzahl
- “Good UX is the consequence of fulfilling the human needs for ***autonomy***, ***competency***, stimulation (self-oriented), ***relatedness***, and popularity (others-oriented) through interacting with the product or service (i.e. hedonic quality). Pragmatic quality facilitates the potential fulfillment of be-goals.”
- Goal types:
  - *Do-goals:* Want to send a message through a digital medium
  - *Be-goals:* Send a message to feel related to another person
- Criteria for usability:
  change from technical aspects
  to aspects of human personality

http://hassenzahl.wordpress.com

M&C MORGAN&CLAYPOOL PUBLISHERS

Experience Design

Technology for All the Right Reasons

Marc Hassenzahl

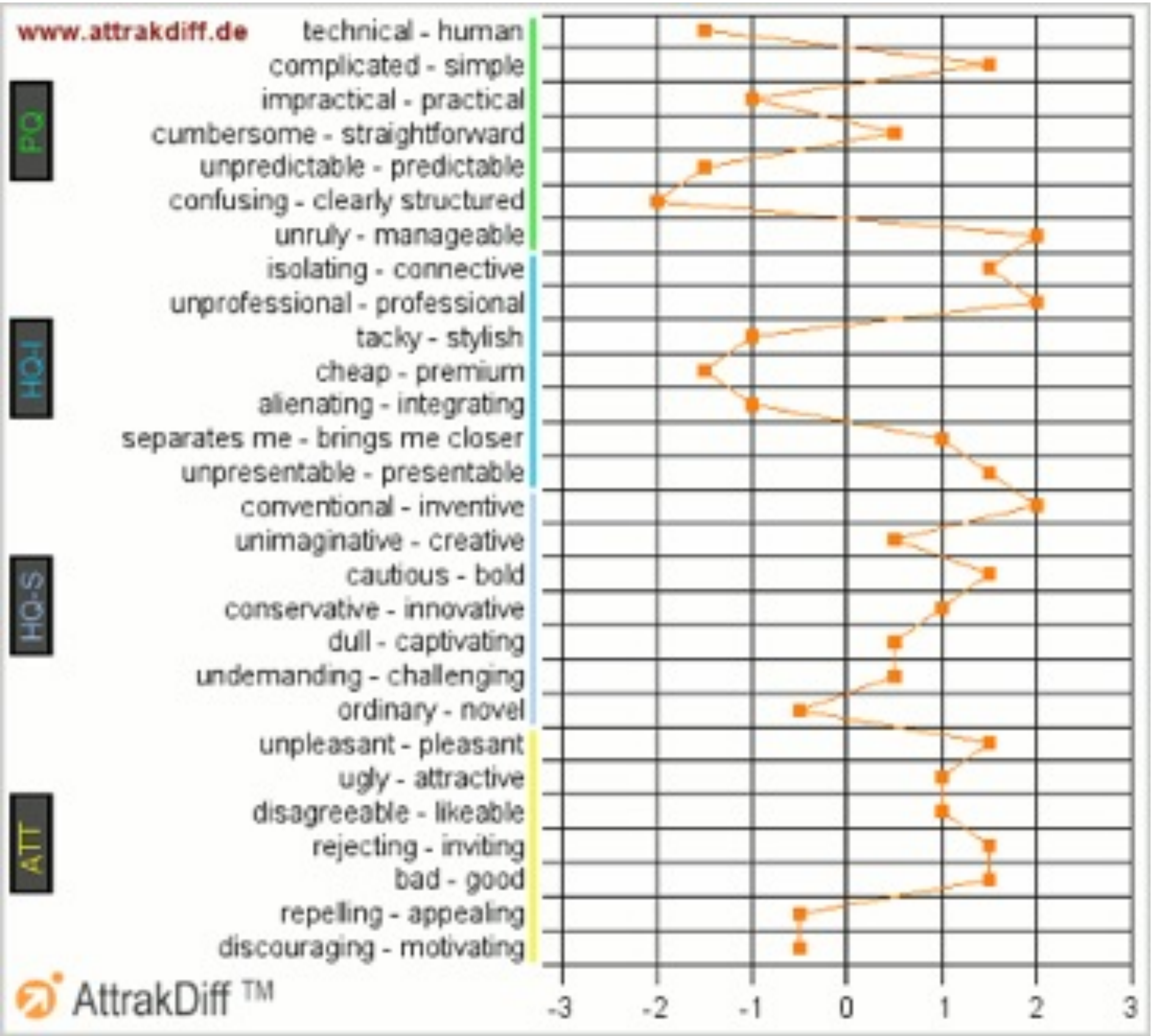
Synthesis Lectures on Human-Centered Informatics

John M. Carroll, *Series Editor*

AttrakDiff

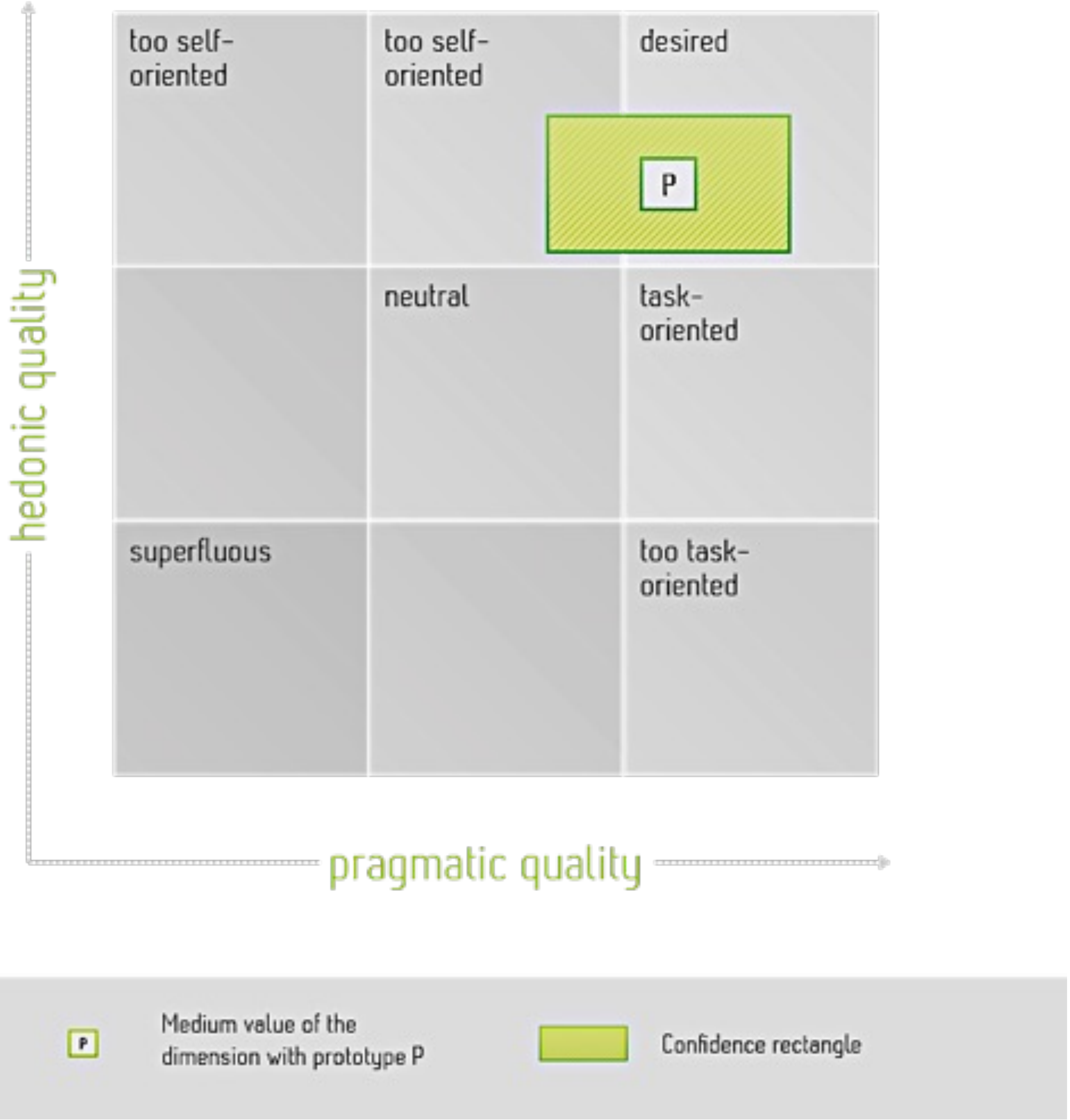
Four dimensions:

- pragmatic quality (PQ)
- hedonic quality - identity (HQ-I)
- hedonic quality - stimulation (HQ-S)
- attractiveness (ATT).



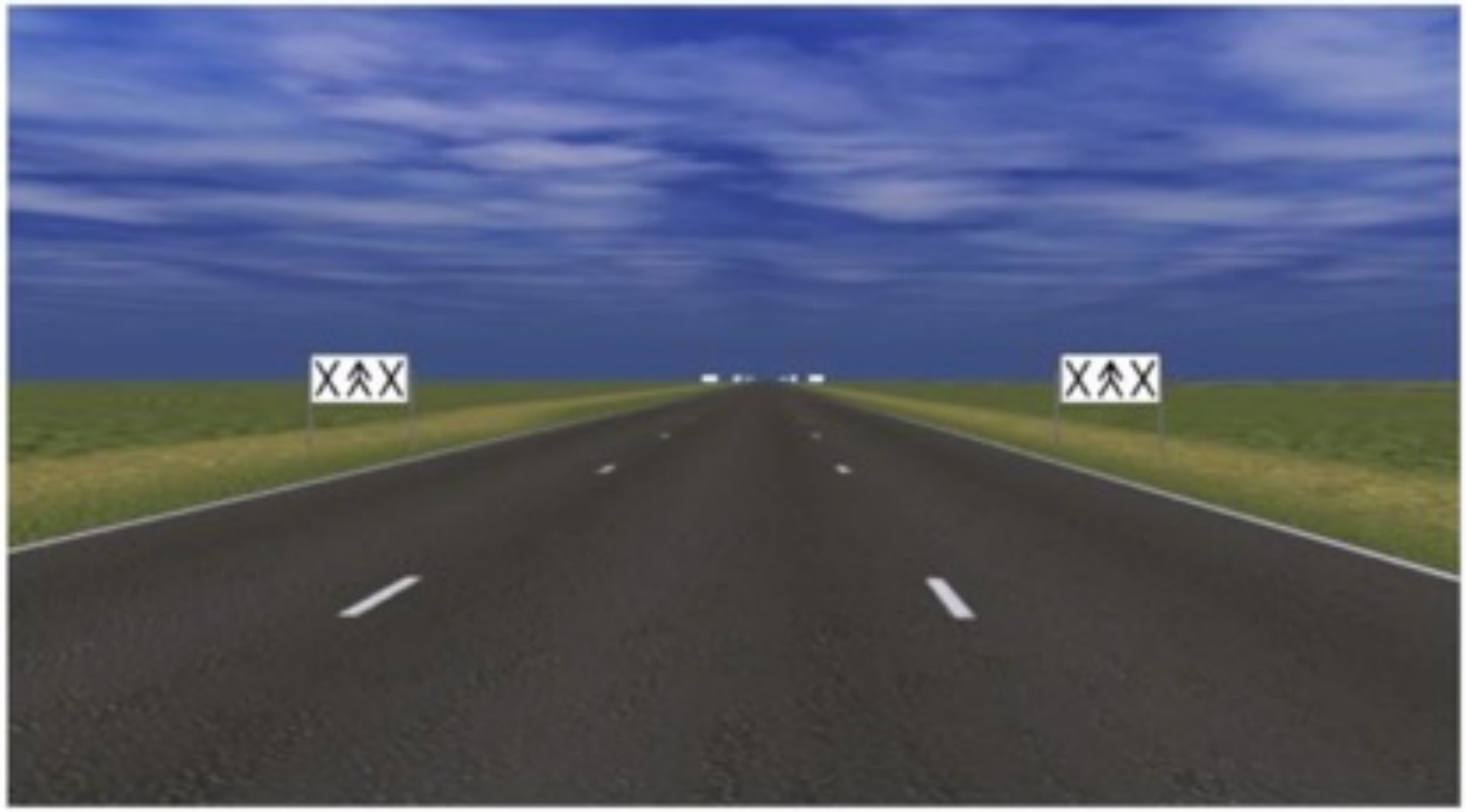
www.attrakdiff.de

AttrakDiff Visualization



http://attrakdiff.de

Domain-Specific Tests: Automotive Example Lane Change Task



- Standardized test (ISO 26022)
- Driving situation (primary task)
  - Demands for lane changes at non-predictable times
- Accompanied by secondary task
- Measures attention split primary/secondary task