HUMAN-COMPUTER INTERACTION
THIRD EDITION
DIX
FINLAY
ABOWD
BEALE

chapter 9

Evaluation techniques

Evaluation Techniques

- **Evaluation**
  - Tests **usability** and **functionality** of system
  - Occurs in **laboratory**, **field** and/or in **collaboration** with users
  - Evaluates both **design** and **implementation**
  - Should be considered at **all stages** in the design life cycle:

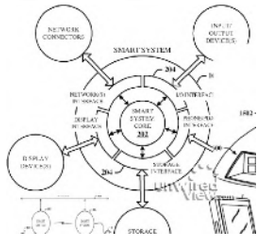
Design → Implementation

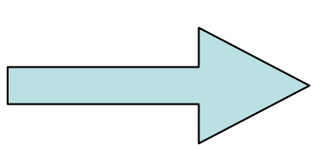
$ $$$ yet +real

Goals of Evaluation

- **Assess extent of system functionality**

- **Assess effect of interface on user**





ex: overload areas

- **Identify specific problems**

unexpected results!

Goals of Evaluation

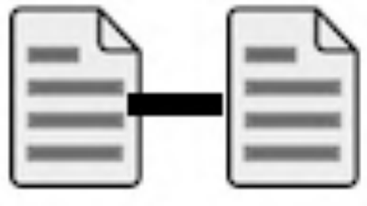
- **Assess effect of interface on user**

Abstract versus **Concrete** icons.

Copy

Save

Delete



Evaluating Designs

Cognitive Walkthrough
Heuristic Evaluation
Review-based evaluation

Cognitive Walkthrough

Proposed by Polson *et al.*

- Evaluates design on how well it supports user in **learning** task
- Usually <u>performed by expert</u> in cognitive psychology
- Expert '**walks through**' design to identify potential problems using psychological principles
- Forms used to guide analysis

A kind of virtual experience!

Cognitive Walkthrough (ctd)

- **What is needed?**
  - Specification or prototype of the system.
  - A description of the task the user is to perform (not the system).
  - Complete written list of the actions needed to complete te task.
  - Indications of who the users are, in terms of evaluation one can assume from them.

Cognitive Walkthrough (ctd)

- **For each task walkthrough considers**
  - –what impact will interaction have on user?
  - –what cognitive processes are required?
  - –what learning problems may occur?

- **Analysis focuses on goals and knowledge:** does the design lead the user to generate the correct goals?

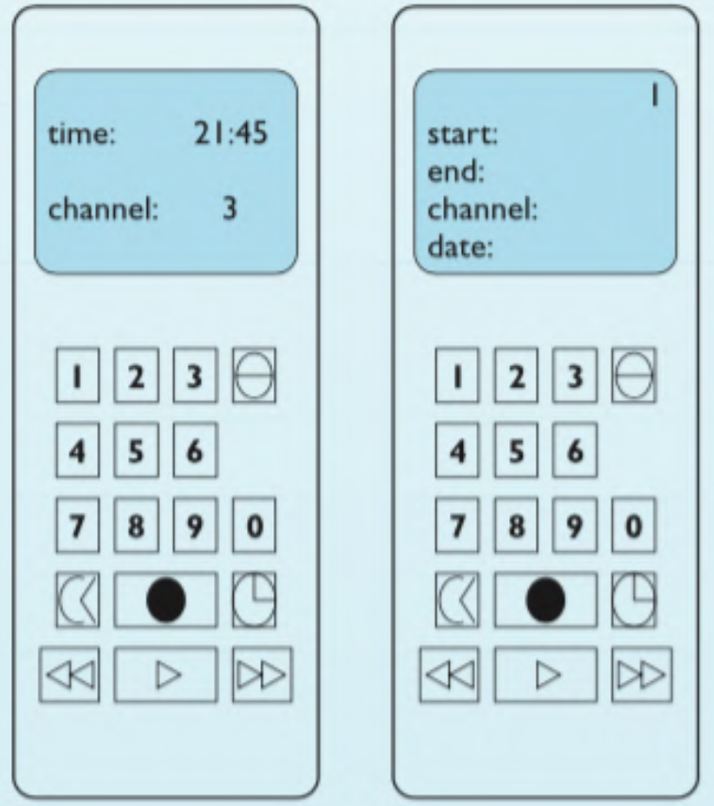
Cognitive Walkthrough (ctd)

Example:

We will assume that the user is familiar with VCRs but not with this particular design.

The next step in the walkthrough is to identify the action sequence for this task. We specify this in terms of the user's action (UA) and the system's display or response (SD). The initial display is as the left-hand picture in Figure 9.1.

UA 1: Press the 'timed record' button
SD 1: Display moves to timer mode. Flashing cursor appears after 'start:'
UA 2: Press digits 1 8 0 0
SD 2: Each digit is displayed as typed and flashing cursor moves to next position
UA 3: Press the 'timed record' button
SD 3: Flashing cursor moves to 'end:'
UA 4: Press digits 1 9 1 5
SD 4: Each digit is displayed as typed and flashing cursor moves to next position
UA 5: Press the 'timed record' button
SD 5: Flashing cursor moves to 'channel:'
UA 6: Press digit 4
SD 6: Digit is displayed as typed and flashing cursor moves to next position
UA 7: Press the 'timed record' button
SD 7: Flashing cursor moves to 'date:'
UA 8: Press digits 2 4 0 2 0 5
SD 8: Each digit is displayed as typed and flashing cursor moves to next position
UA 9: Press the 'timed record' button
SD 9: Stream number in top right-hand corner of display flashes
UA 10: Press the 'transmit' button
SD 10: Details are transmitted to video player and display returns to normal mode



An initial remote control design

Cognitive Walkthrough (ctd)

Example:

Having determined our action list we are in a position to proceed with the walkthrough. For each action (1–10) we must answer the four questions and tell a story about the usability of the system. Beginning with UA 1:

UA 1: Press the 'timed record' button

Question 1: Is the effect of the action the same as the user's goal at that point?

The timed record button initiates timer programming. It is reasonable to assume that a user familiar with VCRs would be trying to do this as his first goal.

Question 2: Will users see that the action is available?

The 'timed record' button is visible on the remote control.

Question 3: Once users have found the correct action, will they know it is the one they need?

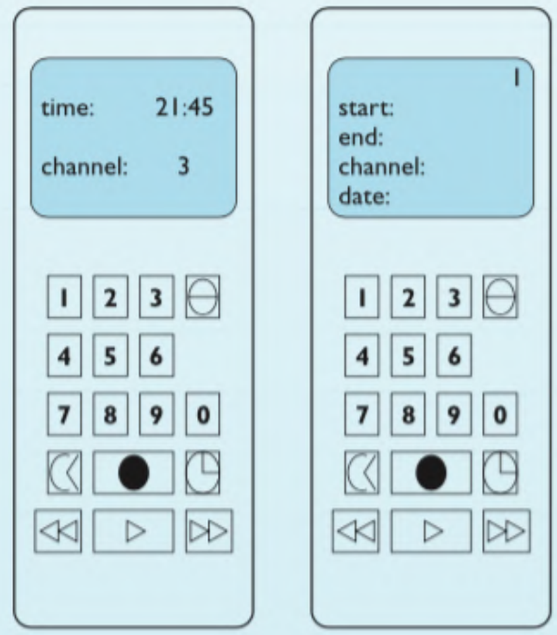
It is not clear which button is the 'timed record' button. The icon of a clock (fourth button down on the right) is a possible candidate but this could be interpreted as a button to change the time. Other possible candidates might be the fourth button down on the left or the filled circle (associated with record). In fact, the icon of the clock is the correct choice but it is quite possible that the user would fail at this point. This identifies a potential usability problem.

Question 4: After the action is taken, will users understand the feedback they get?

Once the action is taken the display changes to the timed record mode and shows familiar headings (start, end, channel, date). It is reasonable to assume that the user would recognize these as indicating successful completion of the first action.

So we find we have a potential usability problem relating to the icon used on the 'timed record' button. We would now have to establish whether our target user group could correctly distinguish this icon from others on the remote.

The analysis proceeds in this fashion, with a walkthrough form completed for each action. We will leave the rest of the walkthrough for you to complete as an exercise. What other usability problems can you identify with this design?



An initial remote control design

Heuristic Evaluation

- **Proposed by Nielsen and Molich.**

- **Usability criteria (heuristics) are identified**
- **Design examined by experts to see if these are violated**

- **Example heuristics**
  - System behavior is predictable
  - System behavior is consistent
  - Feedback is provided

- **Heuristic evaluation "debugs" design.**
  - Cheap approach
  - Nielsen participants interval: **{3,…,5} 3 => 75%**

Heuristic Evaluation

- **Nielsen's 10 Heuristics**

1. Visibility of system status
2. Match between system and the real world
3. User control and freedom
4. Consistency and standards
5. Error prevention
6. Recognition rather than recall
7. Flexibility and efficiency of use
8. Aesthetic and minimalist design
9. Help users recognize, diagnose, and recover from errors
10. Help and documentation

Heuristic Evaluation

- **Each evaluator assesses heuristics violations**
  - They identify usability problems and more detailed characterization and actions to be taken:
    - How common is the problem
    - Easiness of overcoming the problem
    - Problem persistency
    - Impact in user perception
    - Severity of the problem:
      - 0: this isn’t a problem at all :-)
      - 1: cosmetic problem only: no need to fix, unless extra dt.
      - 2: minor usability problem: low priority to fix
      - 3: major usability problem: important to fix
      - 4: usability catastrophe: imperative to fix

Evaluating through user Participation

Evaluating through user Participation

- **These includes:**
  - Empirical or experimental methods
  - Observational methods.
  - Query techniques
  - Physiological monitoring

Laboratory studies

- **Advantages:**
  - specialist equipment available
  - uninterrupted environment

- **Disadvantages:**
  - lack of context
  - difficult to observe several users cooperating

- **Appropriate**
  - if system location is dangerous or impractical for constrained single user systems to allow controlled manipulation of use

Field Studies

- **Advantages:**
  - natural environment
  - context retained (though observation may alter it)
  - longitudinal studies possible

- **Disadvantages:**
  - distractions
  - noise
  - interruption (ex: save/resume)

- **Appropriate**
  - where context is crucial for longitudinal studies

Heisenberg Uncertainty Principle

Evaluating Implementations

Requires an artifact:

simulation, prototype,

full implementation

Experimental evaluation

- Controlled evaluation of specific aspects of interactive behavior

- Evaluator chooses **hypothesis** to be tested

- A number of experimental conditions are considered which differ only in the value of some controlled variable.

- Changes in behavioral measure are attributed to different conditions

Experimental factors

- **Subjects**
  - who – representative, sufficient sample
  - Nielsen: 1 user => identify about 1/3 problems, and >5 not very much to gain.
- **Variables**
  - things to modify and measure
- **Hypothesis**
  - what you'd like to show
- **Experimental design**
  - how you are going to do it

Variables

- **independent variable (IV)**

  characteristic changed to produce different conditions

  e.g. interface style, number of menu items

- **dependent variable (DV)**

  characteristics measured in the experiment

  e.g. time taken, number of errors.

$$y = F(x)$$

Hypothesis

- **prediction of outcome**
  - framed in terms of IV and DV
    e.g. "error rate will increase as font size decreases"

- **null hypothesis (H_0):**
  - states no difference between conditions
  - aim is to disprove this

    e.g. **H**0 = "no change with font size"

Hypothesis Testing

- **assess effect of interface on user**

Abstract versus Concrete icons.

Analysis of data

- **Before you start to do any statistics:**
  - look at data
  - save original data

- **Choice of statistical technique depends on**
  - type of data
  - information required

- **Type of data**
  - **discrete** —> finite number of values
  - **continuous** —> any value

Analysis - types of test

- **parametric**
  - assume normal distribution
  - robust
  - powerful
- **non-parametric**
  - do not assume normal distribution
  - less powerful
- **contingency table**
  - classify data by discrete attributes
  - count number of data items in each group

Analysis - types of test

Independent variable	Dependent variable	
Parametric		
Two valued	Normal	Student's *t* test on difference of means
Discrete	Normal	ANOVA (ANalysis Of VAriance)
Continuous	Normal	Linear (or non-linear) regression factor analysis
Non-parametric		
Two valued	Continuous	Wilcoxon (or Mann–Whitney) rank-sum test
Discrete	Continuous	Rank-sum versions of ANOVA
Continuous	Continuous	Spearman's rank correlation
Contingency tests		
Two valued	Discrete	No special test, see next entry
Discrete	Discrete	Contingency table and chi-squared test
Continuous	Discrete	(Rare) Group independent variable and then as above

Analysis - Non-parametric example

Experiment:

Condition A: 33, 42, 25, 79, 52

Condition B: 87, 65, 92, 93, 91, 55

Rank:

Condition A: 2, 3, 1, 7, 4 Sum=17 (LVS=1+2+3+4+5=15)

Condition B: 8, 6, 10, 11, 9, 5 Sum=49 (LVS=1+…+6=21)

Statistic U:

Condition A: U(A) = Sum - LVS = 2

Condition B: U(B) = Sum - LVS = 28

H_0 rejected

min{ U(A), U(B) } = 2 < 3 (critical value at critical level of 5%)

Experimental design

- **Between-Subjects design**
  - each subject performs under only one condition
  - no transfer of learning
  - more users required
  - variation can bias results.

- **Within-Subjects design**
  - each subject performs experiment under each condition.
  - transfer of learning possible
  - less costly and less likely to suffer from user variation

Observational Methods

Think Aloud

Cooperative evaluation

Protocol analysis

Automated analysis

Post-task walkthroughs

Think Aloud

- **user observed performing task**
- **user asked to describe what he is doing and why, what he thinks is happening etc.**

- **Advantages**
  - simplicity - requires little expertise
  - can provide useful insight
  - can show how system is actually use
- **Disadvantages**
  - subjective
  - selective
  - act of describing may alter task performance

Cooperative evaluation

- **Variation on think aloud**
- **User collaborates in evaluation**
- **Both user and evaluator can ask each other questions throughout**

- **Additional advantages**
  - less constrained and easier to use
  - user is encouraged to criticize system
  - user clarification possible, when necessary

Protocol analysis

- **paper and pencil** – cheap, limited to writing speed
- **audio** – good for think aloud, difficult to match with other protocols

- **video** – accurate and realistic, needs special equipment, obtrusive, normally 2 cameras

- **computer logging** – automatic and unobtrusive, large amounts of data difficult to analyze
- **user notebooks** – coarse and subjective, useful insights, good for longitudinal studies

---

- **Mixed use in practice.**
- **audio/video** transcription difficult and requires skill.
  - Difficult to combine - ex: audio + 2x video + key log
  - Some automatic support tools available

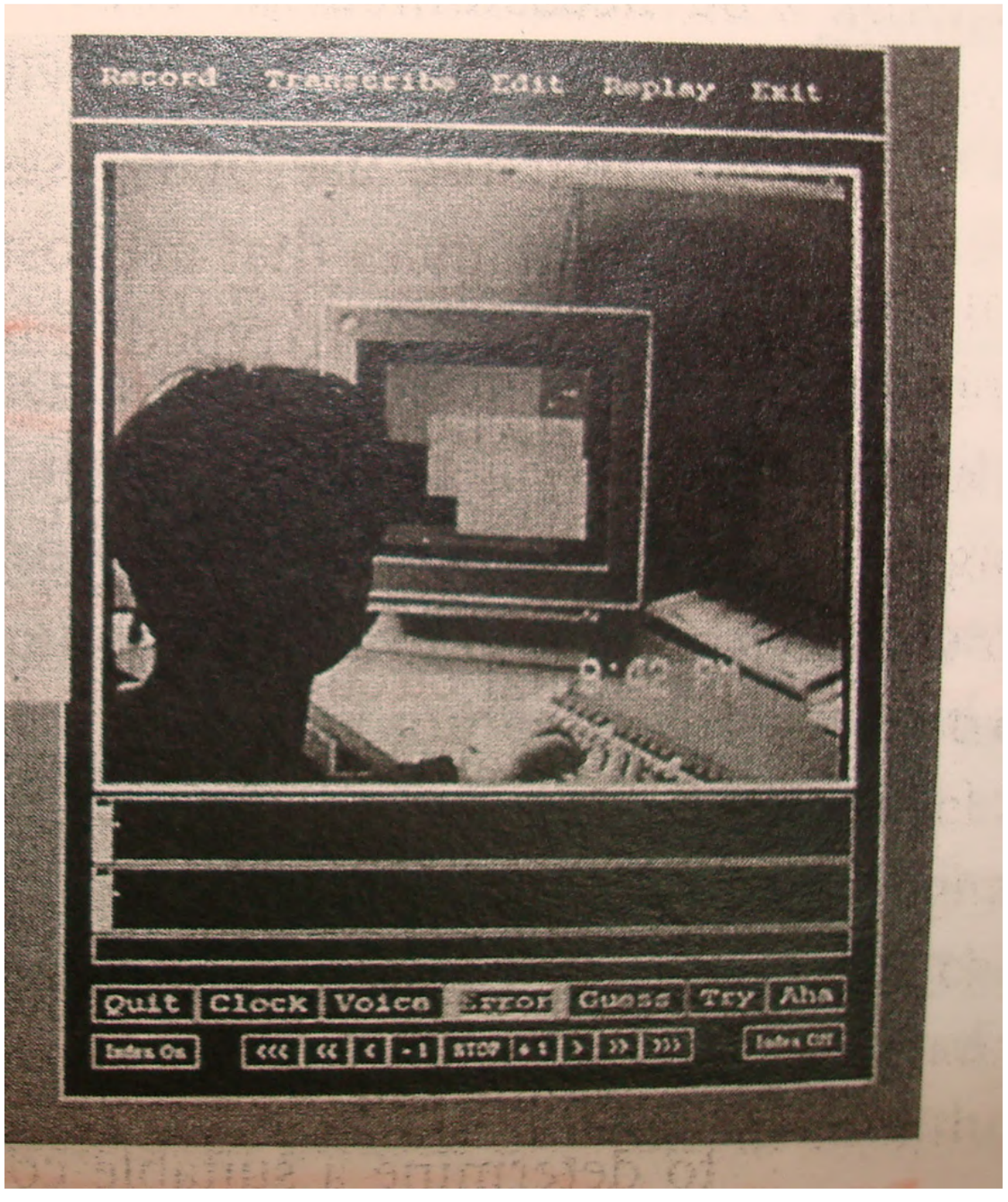
Automated Analysis Tools

- **Software running in a multimedia workstation**

– Several buttons to make specific annotations during recording session.
– System may be logging as well.
– **Advantage**: Alleviates burden of video analysis.
– **Disadvantage**: evaluator loose focus!

More recently:

– Existing comercial tools, for example:
– "Observer Pro" (www.noldus.com)
– DRUM



EVA - Experimental Video Annotator

post-task walkthroughs

- **Recorded data => lack of interpretation!**
- **Workplace project**
- **Post task walkthrough**
  - user reacts on action after the event
  - used to fill in intention
- **Advantages**
  - analyst has time to focus on relevant incidents
  - avoid excessive interruption of task
- **Disadvantages**
  - lack of freshness
  - may be post-hoc interpretation of events

post-task walkthroughs

- **transcript played back to participant for comment**
  - immediately → fresh in mind
  - delayed → evaluator has time to identify questions
- **useful to identify reasons for actions and alternatives considered**
- **necessary in cases where think aloud is not possible**

Query Techniques

Interviews
Questionnaires

Interviews

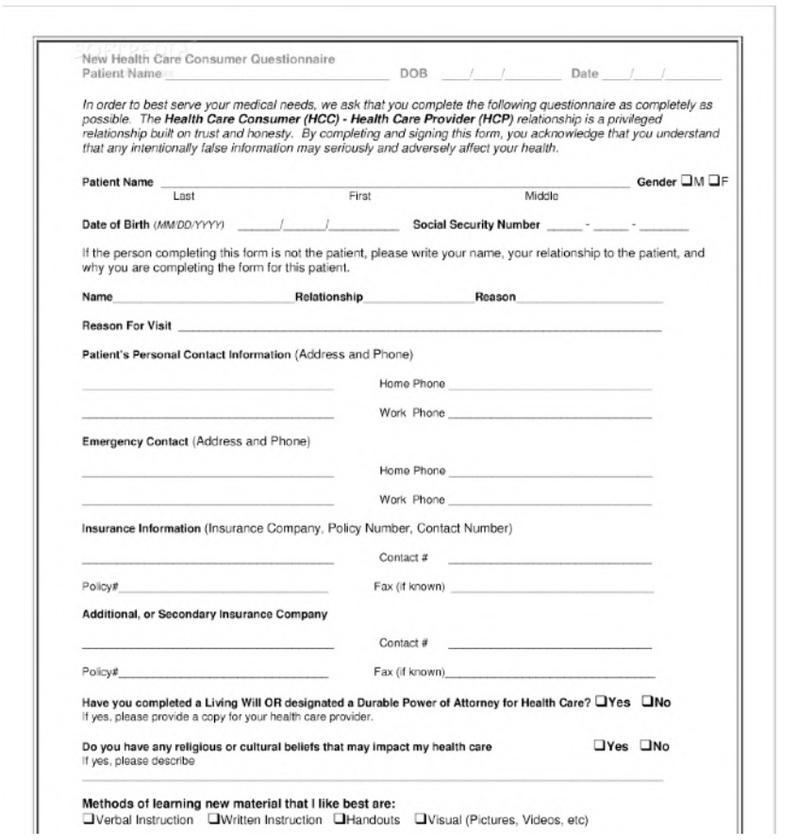
- **Analyst questions user on one-to-one basis usually based on prepared questions**
- **Informal, subjective and relatively cheap**

- **Advantages**
  - can be varied to suit context
  - issues can be explored more fully
  - can elicit user views and identify unanticipated problems

- **Disadvantages**
  - very subjective
  - time consuming

Questionnaires (ctd)

- **Need careful design**
  - what information is required?
  - how are answers to be analyzed?



New Health Care Consumer Questionnaire
Patient Name ____ DOB ___/___/____ Date ___/___/____

In order to best serve your medical needs, we ask that you complete the following questionnaire as completely as possible. The ***Health Care Consumer (HCC) - Health Care Provider (HCP)*** *relationship is a privileged relationship built on trust and honesty. By completing and signing this form, you acknowledge that you understand that any intentionally false information may seriously and adversely affect your health.*

Patient Name _____ **Gender** ❑M ❑F
Last First Middle

Date of Birth *(MM/DD/YYYY)* _____/_____/_____ **Social Security Number** _____ - _____ - _____

If the person completing this form is not the patient, please write your name, your relationship to the patient, and why you are completing the form for this patient.

Name_____ **Relationship**_____ **Reason**_____

Reason For Visit _____

Patient's Personal Contact Information (Address and Phone)

_____ Home Phone _____

_____ Work Phone _____

Emergency Contact (Address and Phone)

_____ Home Phone _____

_____ Work Phone _____

Insurance Information (Insurance Company, Policy Number, Contact Number)

_____ Contact # _____

Policy#_____ Fax (if known) _____

Additional, or Secondary Insurance Company

_____ Contact # _____

Policy#_____ Fax (if known)_____

Have you completed a Living Will OR designated a Durable Power of Attorney for Health Care? ❑**Yes** ❑**No**
If yes, please provide a copy for your health care provider.

Do you have any religious or cultural beliefs that may impact my health care ❑**Yes** ❑**No**
If yes, please describe

Methods of learning new material that I like best are:
❑Verbal Instruction ❑Written Instruction ❑Handouts ❑Visual (Pictures, Videos, etc)

- **Styles of question**
  - general
  - open-ended
  - scalar
  - multi-choice
  - ranked

Physiological methods

Eye tracking
Physiological measurement

eye tracking

- **head or desk mounted equipment tracks the position of the eye**
- **eye movement reflects the amount of cognitive processing a display requires**
- **measurements include**

  - **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
  - **scan paths**: moving straight to a target with a short fixation at the target is optimal

eye tracking

DANS, KÖN OCH JAGPROJEKT

På jakt efter ungdomars kroppsspråk och den "synkretiska dansen", en sammansmältning av olika kulturers dans, har jag i mitt fältarbete under hösten rört mig på olika arenor inom skolans värld. Nordiska, afrikanska, syd- och östeuropeiska ungdomar gör sina röster hörda genom sång, musik, skrik, skratt och gestaltar känslor och uttryck med hjälp av kroppsspråk och dans.

Den individuella estetiken framträder i kläder, frisyrer och symboliska tecken som förstärker ungdomarnas "jagprojekt" där också den egna stilen i kroppsrörelserna spelar en betydande roll i identitetsprövningen. Uppehållsrummet fungerar som offentlig arena där ungdomarna spelar upp sina performanceliknande kroppsshower

physiological measurements

- **emotional response linked to physical changes**
- **these may help determine a user's reaction to an interface**
- **measurements include:**
  - 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)
- **some difficulty in interpreting these physiological responses - more research needed**

physiological measurements

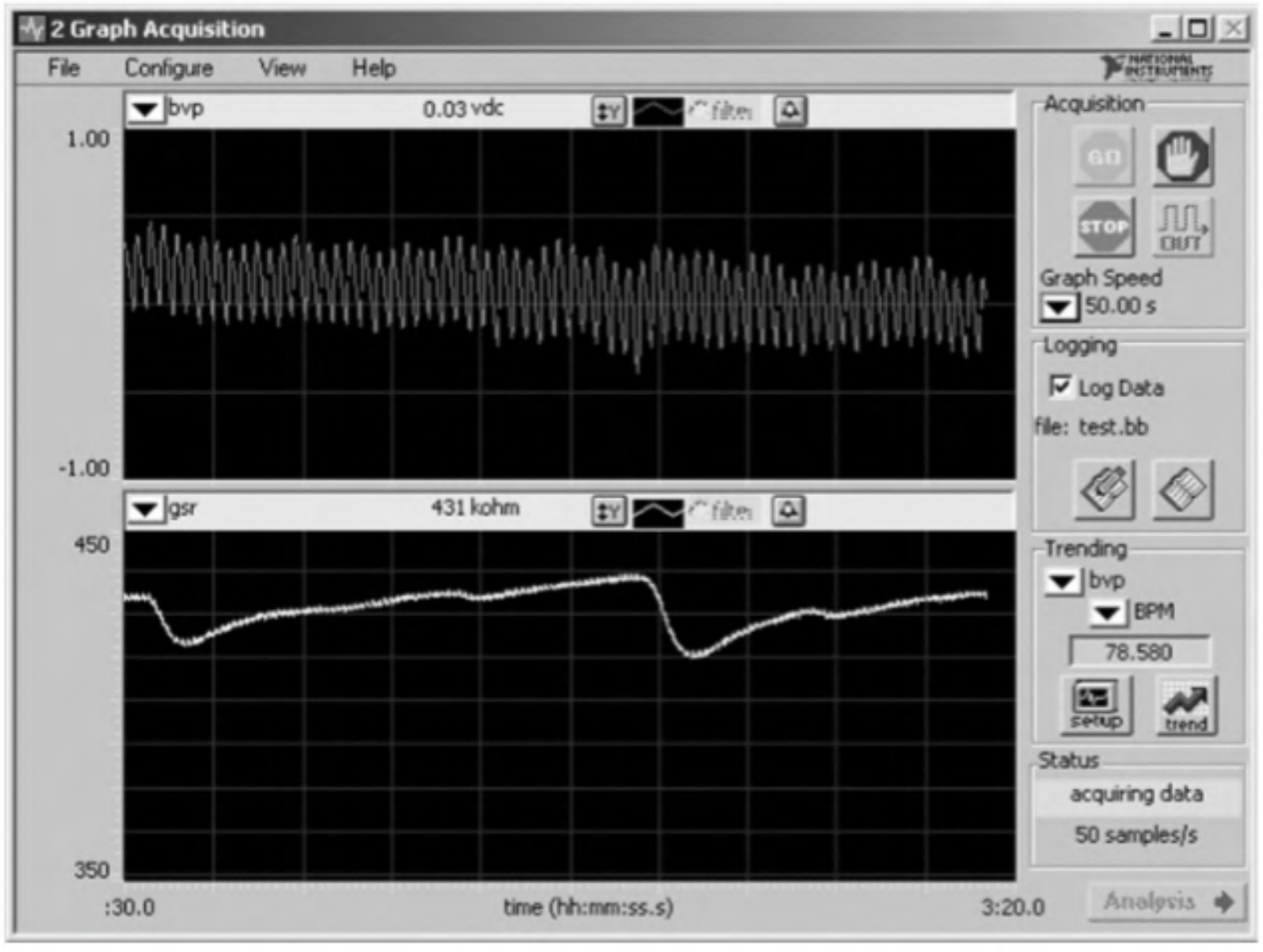


Figure 9.8 Output of monitoring pulse rate (above) and skin conductivity (below). Source: Screen shot courtesy of Dr R. D. Ward; frame source: National Instruments BioBench software

physiological measurements

Heart activity, indicated by blood pressure, volume and pulse. These may respond to stress or anger

Activity of the sweat glands, indicated by skin resistance or galvanic skin response (GSR). These are thought to indicate levels of arousal and mental effort.

Electrical activity in muscle, measured by the electromyogram (EMG). These appear to reflect involvement in a task.

Electrical activity in the brain, measured by the electroencephalogram (EEG). These are associated with decision making, attention and motivation.

Choosing an Evaluation Method

When in process: design vs. implementation

Style of evaluation: laboratory vs. field

How objective: subjective vs. objective

Type of measures: qualitative vs. quantitative

Level of information: high level vs. low level

Level of interference: obtrusive vs. unobtrusive

Resources available: time, subjects, equipment, expertise