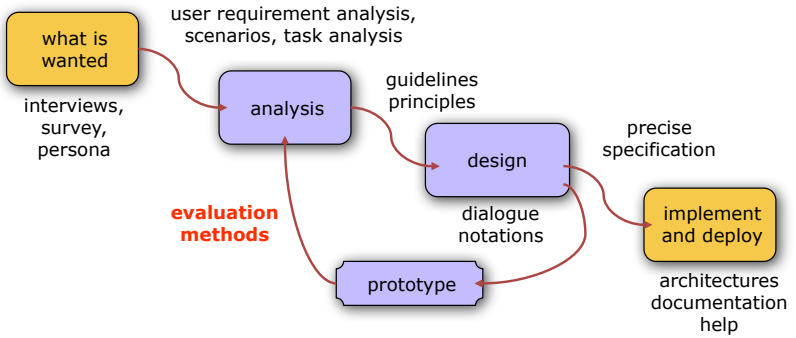
Human-Computer Interaction

Session 8:
User Interface Evaluation

Reading:
- Dix et al., Human-Computer Interaction, chapter 9
- Shneiderman, Designing the User Interface, chapter 4

User-centered design process



Process to develop interactive systems such that usability will be maximized.

Evaluation methods

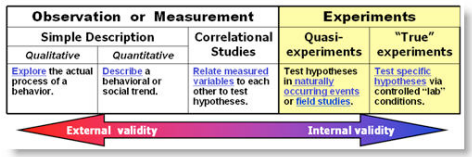
Usability inspection (expert reviews)
- Guidelines review & consistency inspection
- Cognitive walkthrough
- Heuristic evaluation
- Focus group

User studies
- Usability testing -> statistical analysis
- Thinking-Aloud
- Field studies
- Interviews & questionnaires

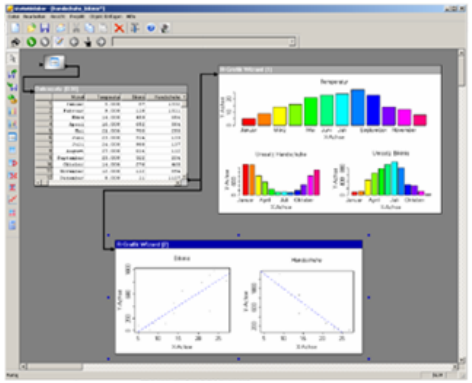
Model-based evaluation

Usability Testing

1. get representative users
2. define criteria for evaluation
3. develop test scenario: setup + context + task
4. consider ethical issues
4. run pilot tests & refine design
5. actual testing & data gathering
6. data analysis
   - statistics over data, must match kind of study/experiment
   - post task user interview
7. report & recommendations



Data & statistics



Statistik & Daten

- Durchführung eines Experimentes und Beobachtung von empirischen Sachverhalten
- Beobachtungen bzw. Messungen ergeben Daten
  - im Hinblick auf das jeweilige Erkenntnisinteresse informativ
  - durch den Bezug auf eine Skala mit spezifischen mathematischen Eigenschaften statistisch auswertbar
- Zwei Auswertungsarten:
  - Deskriptive Statistik: math. Beschreibung der Daten
  - Inferenzstatistik: Überprüfung einer Hypothese

Messen & Werte - Skalen

- Skalen für Messwerte in Typenhierarchie einteilbar
- Typ einer Skala wird auch als Skalenniveau bezeichnet
- man unterscheidet die Niveaus
  1. Nominalskala*: gleiche Merkmalsausprägung, gleiche Werte:* Blutgruppen A, B, AB, Null; Geschlecht m., w.
  2. Ordinalskala*: größerer Merkmalsausprägung, größere Werte:* Kleidungsgrößen S, M, L, XL, XXL; Urteile --, -, 0, +, ++
  3. Intervallskala*: Größenordnung von Wertdifferenzen entspricht Merkmalsunterschieden:* Temperatur in °C oder Fahrenheit
  4. Verhältnisskala*: Größenverhältnisse der Werte entspricht Merkmalsausprägungen:* Länge in mm
  5. Absolutskala*: absolute Werte entsprechen Merkmalsausprägungen:* Anzahl der Kinder

Skalenhierarchie

	Skalenniveau	*Mögl. Aussagen über Messwerte*	*Zulässige Transformation*
metrisch (Wertabstände sinnvoll)	Absolutskala	Absoluter Messwert	Identität $T(x) = x$
	Verhältnisskala	Gleichheit von Verhältnissen	Ähnlichkeit $T(x) = ux$
	Intervallskala	Gleichheit von Abständen	Positiv linear $T(x) = ux + v$
nichtmetrisch	Ordinalskala	Größer-kleiner-Relationen	Monot. steigend $x > y \Leftrightarrow T(x) > T(y)$
	Nominalskala	Gleichheit, Verschiedenheit	Eineindeutigkeit $x = y \Leftrightarrow T(x) = T(y)$

Messwerte aussagekräftiger

Deskriptive Statistik

Angenommen, zu Evaluationszwecken wurde reichlich Datenmaterial gesammelt bzw. erhoben.

Zentrale Frage der deskriptiven Statistik:

Wie können die empirischen Daten aufbereitet, dargestellt, zusammengefasst und strukturiert werden, so dass zentrale Merkmale sichtbar werden?

Hier Fokus auf

- Darstellung von Häufigkeitsverteilungen durch Kennwerte
- Darstellung durch graphische Visualisierung

Empirische Häufigkeitsverteilungen

- einfache Art der Zusammenfassung von Datenmatrizen
- enthält beobachtete *Häufigkeiten* der Werte einer Variablen
  - absolute Häufigkeit *f:* Gesamtzahl der Auftretens eines Wertes
  - relative Häufigkeit f_r*:* abs. Häufigkeit/n , mit *n*=Anzahl Werte
  - kumulierte Häufigkeit f_K*:* sukzessives Aufsummieren der Häufigkeiten; in wie vielen Fällen größerer bzw. kleinerer Wert gemessen (für mind. ordinalskalierte Daten)

Alter	f	f_r	f_K	f_{rK}
17	1	0.25	1	0.25
29	2	0.50	3	0.75
36	1	0.25	4	1.00

Kennwerte der „zentralen Tendenz"

Gesucht ist ein Wert, der die ganze Messreihe, also die Gesamtheit aller Ausprägungen einer Variablen in einer Stichprobe, möglichst gut repräsentiert.

Arithmetisches Mittel (Mittelwert, Durchschnitt)

Summe der Messwerte x_i einer Messreihe, dividiert durch Anzahl i

$$\bar{x} = \frac{\sum_i x_i}{n}$$

- geeignet bei Intervalldaten (oder "höheren" Skalen), großen Stichproben und symmetrischer Verteilung
- sensibel gegenüber Extremwerten („outliers"), daher bei schiefer Verteilung nicht angebracht!

Kennwerte der zentralen Tendenz

Median (Zentralwert) $\tilde{x}$

der Wert, der die nach Größe geordnete Messreihe halbiert (daher auch 2-Quantil genannt)

1. Meßwerte der Größe nach ordnen $x_1 \le x_2 \le \ldots \le x_n$
2. Halbierungsstelle H identifizieren
   bei geradem n: $H = n/2$
   bei ungeradem n: $H = (n+1)/2$
3. Median berechen
   bei geradem n: $\tilde{x} = (x_H + x_{H+1})/2$
   bei ungeradem n: $\tilde{x} = x_H$

- geeignet bei Ordinaldaten oder "höheren" Skalen, kleinen Stichproben oder asymmetrischer Verteilung
- unsensibel gegenüber Extremwerten

Kennwerte der zentralen Tendenz

Modus (Modalwert) $\ddot{x}$

der am häufigsten vorkommende Wert

1. Häufigkeit $f(x)$ für alle vorkommenden Kategorien bestimmen
2. Modus = Wert der meistbelegten Kategorie

bei mehreren benachbarten Spitzenwerten, Modus des arithmetischen Mittels

bei mehreren nicht-benachbarten Spitzenwerten, alle angeben

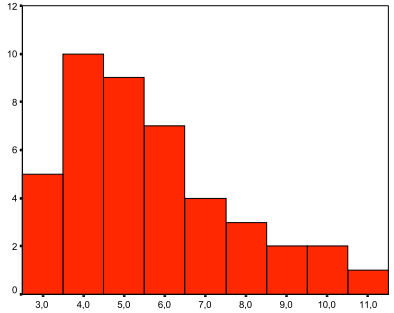
- ☐ geeignet bei Nominaldaten und bei mehrgipfliger Verteilung
- ☐ insgesamt weniger gebräuchlich

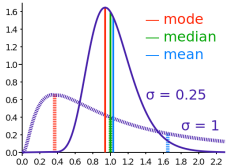
Zentrale Tendenz am Beispiel

Häufigkeitsverteilung für eine Variable:

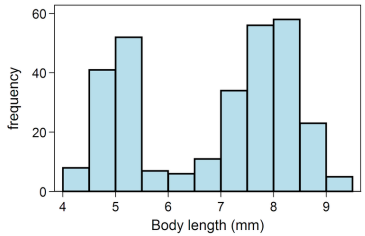
Wert	3	4	5	6	7	8	9	10	11
Häufigkeit	5	10	9	7	4	3	2	2	1

arithm. Mittel	5,65
Median	5
Modus	4

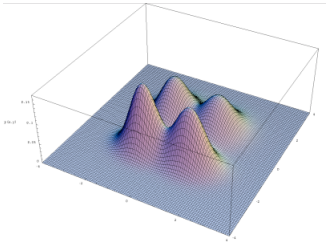




Bimodal distribution



Multi-modal distribution



Kennwerte der „Dispersion" (Streuung)

Dispersion bezeichnet das Ausmaß, in dem sich die Verteilungen der Werte unterscheiden (wie "weit" oder "eng" sie sich um die zentrale Tendenz gruppieren).

Varianz und Standardabweichung s^2 s

Summe der quadrierten Abweichungen vom arithm. Mittel bezogen auf ihre Anzahl (Standardabweichung: Wurzel davon)

$$s^2 = \frac{\sum_i (x_i - \bar{x})^2}{n-1} \qquad s = \sqrt{\frac{\sum_i (x_i - \bar{x})^2}{n-1}}$$

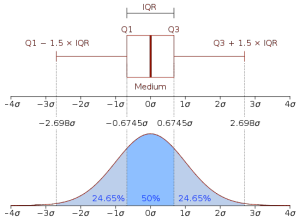
- ☐ geeignet bei Intervalldaten oder "höheren" Skalen, großen Stichproben und symmetrischer Verteilung

Kennwerte der Dispersion

Quartilsabstand q (Quartil = 4-Quantil)

absolute Differenz der Quartile, d.h. der Mediane der am Median geteilten Messreihe (0.25-Quantil bis 0.75-Quantil)

1. Meßwerte der Größe nach ordnen $x_1 \le x_2 \le \ldots \le x_n$
2. Median $\tilde{x}$ bestimmen
3. unteres Quartil $q_{0.25}$ (Median aller $x_i \le \tilde{x}$) bestimmen
4. oberes Quartil $q_{0.75}$ (Median aller $x_i \ge \tilde{x}$) bestimmen
5. absoluten Quartilsabstand $q = (|\, q_{0.75} - q_{0.25} \,|)$ berechnen



Kennwerte der Dispersion

Quartilsabstand q (Quartil = 4-Quantil)

absolute Differenz der Quartile, d.h. der Mediane der am Median geteilten Messreihe (0.25-Quantil bis 0.75-Quantil)

1. Meßwerte der Größe nach ordnen $x_1 \le x_2 \le \ldots \le x_n$
2. Median $\tilde{x}$ bestimmen
3. unteres Quartil $q_{0.25}$ (Median aller $x_i \le \tilde{x}$) bestimmen
4. oberes Quartil $q_{0.75}$ (Median aller $x_i \ge \tilde{x}$) bestimmen
5. absoluten Quartilsabstand $q = (|\, q_{0.75} - q_{0.25} \,|)$ berechnen

- geeignet bei Ordinaldaten oder "höheren" Skalen, kleinen Stichproben oder asymmetrischer Verteilung
- ebenso werden manchmal Abstände zwischen anderen Quantilen betrachtet, z.B. der Interdezilbereich $d = (|\, q_{0.9} - q_{0.1} \,|)$

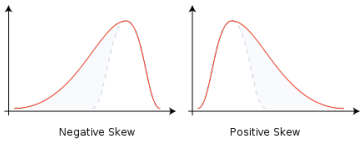
Kennwerte der „Schiefe"

Schiefe bezeichnet, wie stark rechts-/linkslastig eine Verteilung ist:
Schiefe < 0: Verteilung ist linksschief (rechtssteil)
Schiefe > 0: Verteilung ist rechtsschief (linkssteil)
Schiefe = 0: Verteilung ist symmetrisch

Pearsonsches Schiefemaß

Verhältnis der Differenz von Mittelwert und Modus zur Standardabweichung

$$S_p = \frac{\bar{x} - M}{s}$$



Bowley-Fishersches Schiefemaß

$$S_b = \frac{\sum_i (x_i - \bar{x})^3 / n}{s^3}$$

Inferenzstatistik

Zentrale Frage der Inferenzstatistik:

Wie können auf Basis empirischer Daten (Wahrscheinlichkeits-)Aussagen über die Allgemeinheit getroffen werden?

Fokus auf

- Prinzip inferenzstatistischer Analysen
- ein einfaches Beispiel
- Überblick über statistischer Testverfahren

Experimental design

- subjects/participants
  - representative, sufficient number *N*
- variables
  - entities being altered (controlled) or measured, resp.
- conditions
  - experimental conditions, differing only with respect to controlled variable
- hypotheses
  - what you want to show
  - derived from theory/literature, not the same set of experimental data

Variables

- independent variable (IV)
  - characteristics changed to produce different conditions, e.g., interface style, number of menu items, colors, sizes, etc.
  - also called *controlled* variables
- dependent variable (DV)
  - characteristics measured in the experiment, e.g., time taken, number of errors, reaction time, etc.

Hypotheses

- formualte as *if-then* or *the-the* („je..desto") statement
- formulate in three steps
  1. in terms of the underlying theory
     "Alle Computerbenutzer können das Programm xy schneller mit der Maus als mit der Tastatur bedienen."
  2. in terms of the variables
     "*Wenn bei Programm xy Interaktionsgerät = Maus, dann Interaktionszeit < als wenn Interaktionsgerät = Tastatur."*
  3. in terms of statistical measures („Kennwerte")
- need to frame theoretical concepts in statistical terms (operationalization)

Hypotheses

- statistical formulation calls for comparison of data gathered in test series under different conditions
- formulate and test possible explanations
- working hypothesis or alternative hypothesis H_1
  - differences in test series are systematic and due to changes in controlled variables (IVs)
  - H_1 states expected outcome (how IVs influence DVs) $\bar{x}_A \neq \bar{x}_B$
- null hypothesis H_0 :
  - there is no difference between conditions other than random variation
  - contraposition to working hypothesis
  - aim is to disprove this $\bar{x}_A = \bar{x}_B$
    e.g. null hypothesis = "no change with font size"

Principle of statistical tests

Disprove the null hypothesis, i.e. prove that differences between the conditions did *not* happen by chance.

Note:
Statistical conclusions are always generalizations from a sample to an overall population, where the sample will *always* be affected by random variation! There are thus no absolute decisions againts the null hypothesis, but only probabilities of their (in)validity!

Do not reject the null hypothesis before the results disprove it with a sufficient probability (significance)!

Signifikanz & Fehler

Signifikanz

Ergebnis der Analyse ist Wahrscheinlichkeitsaussage über Bedeutsamkeit (Signifikanz) des Unterschieds bzw. des Zusammenhangs

Fehler: zwei Arten von Fehlentscheidungen sind möglich:

- *α – Fehler (1. Art):* Entscheidung für H_1 obwohl H_0 zutrifft
- *β – Fehler (2. Art):* Entscheidung für H_0 obwohl H_1 zutrifft

In der Praxis v.a.: Wahrscheinlichkeit für *α-Fehler* minimieren

		In der Grundgesamtheit gilt die	
		H_0	H_1
Entscheidung auf Grund der Stichprobe	H_0	richtige Entscheidung, kein Fehler	β-Fehler
	H_1	α-Fehler	richtige Entscheidung, kein Fehler

Signifikanzniveaus

Die Irrtumswahrscheinlichkeit, d.h. die WK einer Fehlentscheidung unter der Annahme, die H_0 treffe zu, ist berechenbar.

Als vertretbar gilt: *α – Fehlerrisiko* von *max. 5 %*

Schreibweise: $p(\alpha) \leq 0.05$

Lesart: Meßreihen-Unterschied bzw. –Zusammenhang ist mit einer Wahrscheinlichkeit von 0.95 oder mehr systematisch bedingt

- Aussage auf 95%-Niveau $p(\alpha) \leq 0.05$ *"signifikant"*
- Aussage auf 99%-Niveau $p(\alpha) \leq 0.01$ *"sehr signifikant"*
- Aussage auf 99.9%-Niveau $p(\alpha) \leq 0.001$ *"hochsignifikant"*

Experimental design

Goal: controlled evaluation of aspects of interactive behavior

1. define appropriate task (must encourage cooperation)
2. define variables (IV, DV)
3. formulate hypothesis to be tested in terms of variables
4. choose conditions to test; changes in measure are attributed to different conditions; *control* condition without variable manipulation
5. choose how to and gather data
6. choose statistical technique to test the hypotheses

Before you start to do any statistics

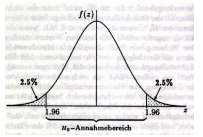
- look at data
- check for *outliers*
- save original data for later re-analyses

Statistical test – choice depends on

- type of data/variables
  - discrete - can take finite number of values (*levels*)
  - continuous - can take any value
  - ranking scale – interval, nominal, etc.
- type of experimental variation
  - are dependent variables subject to random errors?
  - do they follow a known probability distribution?
- type of information required
  - is there a difference between…
    - distributions? Anpassungstests
    - frequencies? Häufigkeitstests
    - means? Unterschiedstests
    - dispersions? Homogenitätstest
  - correlation of test series? Zusammenhangstests
  - influential factors? Varianzanalyse
  - how accurate is the estimate?

Statistical test - basic types

- parametric
  - powerful
  - assume particular distribution of DV
  - robust (give reasonable results also when data not exactly normal)
  - *Example*: completion time of complex task, depending on independent subtasks



- non-parametric (distribution-free)
  - less powerful, more reliable
  - do not assume normal distribution
  - *Example*: subjective usability rating
- contingency
  - classify data by discrete attributes
  - count number of data items in each group

		Relevant R	Not Relevant $\bar{R}$
Retrieved	G	$G \cap R$	$G \cap \bar{R}$
Not Retrieved	$\bar{G}$	$\bar{G} \cap R$	$\bar{G} \cap \bar{R}$

Statistical test by form of IV and DV

	IV	DV	Test
Parametric	Two-valued	Normal	Student's t-test on difference of means
	Discrete	Normal	ANOVA (ANalysis Of VAriance)
	Continuous	Normal	(Non-)linear regression factor analysis
Non-parametric	Two-valued	Cont.	Wilcoxon/Mann-Whitney rank-sum test
	Discrete	Cont.	Rank-sum versions of ANOVA
	Continuous	Cont.s	Spearman's rank correlation
Contingency test	Two-valued	Discrete	No special test, see next entry
	Discrete	Discrete	Contingency table and Chi-squared test
	Continuous	Discrete	Group indep. Variable and then as above

Dix et al., Human-Computer Interaction, p. 334

Summary

- statistics helps to describe experimental data and to test hypotheses on them
- statistics can be (coarsely) divided into descriptive statistics and inferential statistics
- methods are standardized – in science, everybody knows what you want to say
- methods, especially of inferential statistics, are not easily applied; experience needed, refer text books!
- make sure the statistical test you are using is applicable, check the requirements!
- use software for analyzing the data
  - *Example: R (www.r-project.org/)*

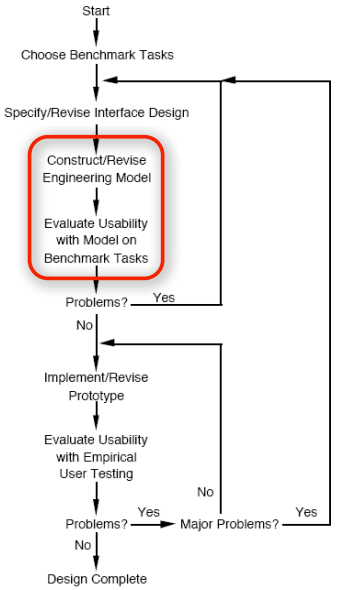
Model-based evaluation

Model-based evaluation

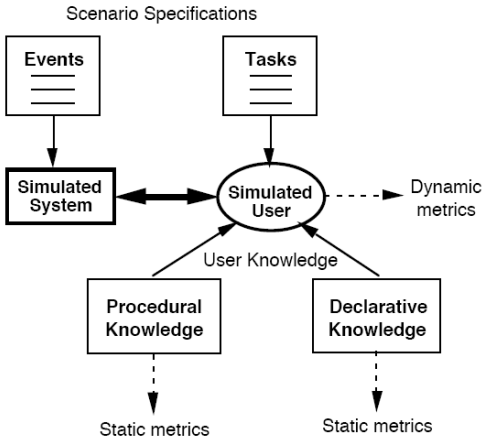
Idea: Provide usability test before building prototype or user testing, allow for more design iterations

Four steps:

1. describe interface design in detail
2. build simulation model of system & user doing a task
3. run the model to predict execution or learning time
4. revise or choose design depending on results



Model-based evaluation



Models = simulations of human users and system in interaction

Procedural knowledge
how-to procedures
→ executable

Declarative knowledge
facts, beliefs
→ reportable

Model-based evaluation

Current models can rather predict few aspects:

- perceptual processing, motor control
- time required to execute specific (low-level) tasks
- ease of learning procedures, consistency effects

Actual user testing is indispensible!

Modeling approaches

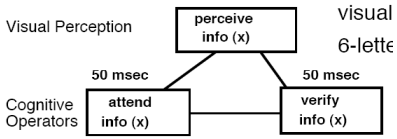
Three basic approaches:

1. Task network models – before detailed design
2. Cognitive Architecture Models – test constraints
3. GOMS models – simple & effective, rapid testing

Differ with respect to...

- human constraints modeled (cognitive/psychological vs. perceptual vs. motoric)
- level of detail
- when to use it in the design process

Task Network Models



Tasks = mixture of human and machine tasks

Each task characterized by a distribution of completion times, and arbitrary dependencies and effects

Connected network of tasks:

- Connection: one task is a prerequisite of the other
- Both serial and parallel execution of tasks
- Final completion time computed from chain of serial and parallel tasks
- Key: determine critical path with largest execution time

Cognitive architectures

"Programmed" with a strategy to perform specific tasks

- provides constraints on form and content of the strategy
- architecture + specific strategy = model of a specific task

To model a specific task...

- do task analysis to arrive at human's task strategy
- "program" architecture with representation of strategy
- run the model using task scenarios

Result: predicted behavior and time course for that scenario and task strategy

Needs comprehensive psychological theory, quite complex; used mostly in a research settings

EPIC Architecture

(Kieras & Meyer, mid-1990s)

Developed to represent executive processes that control other processes during multiple task performance

Executive-**P**rocess **I**nteractive ***C***ontrol
http://www.eecs.umich.edu/~kieras/epic.html

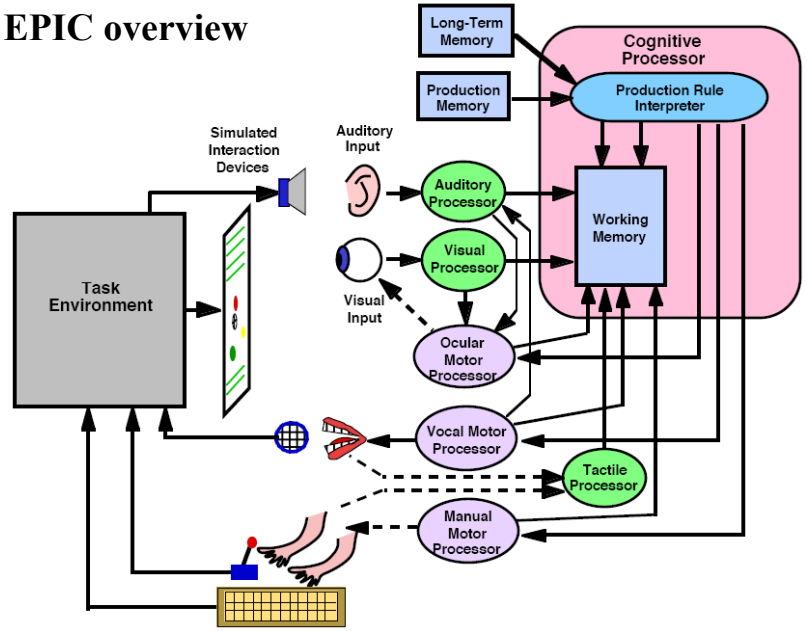
General properties

- production-rule cognitive processor
- parallel perceptual and motor processors
- components, pathways, and time parameters

Task-dependent properties

- cognitive processor production rules (strategy)
- perceptual recoding
- response requirements and styles

EPIC overview



GOMS (Card, Moran, & Newell, 1983)

Model-based methodology based on simplified cognitive architectures

An approach to describing the knowledge of *procedures* that a user must have in order to operate a system

- **G**oals - what goals can be accomplished with the system
- **O**perators - what basic actions can be performed
- **M**ethods - what sequences of operators can be used
- **S**election Rules - which method should be used

Well worked out, practical, but limited due to simplifications

Often in the "sweet spot" - lots of value for modest modeling effort

GOMS - Keystroke-level model

1. Choose one or more representative task scenarios
2. Have design specified to the point that keystroke-level actions can be listed.
3. List the keystroke-level actions (operators) involved in doing the task.
4. Insert mental operators for when user has to stop and think.
5. Look up the standard execution time to each operator.
6. Add up the execution times for the operators.
7. The total is the estimated time to complete the task (sum of times for tasks t_i multiplied by frequency n_i)

$$T_{execute} = \sum_i t_i * n_i$$

KLM – operators and times

K - Keystroke (0.12 - 1.2 sec; 0.28 for ordinary user)

- Pressing a key or button on the keyboard
- Different experience levels have different times
- Pressing SHIFT or CONTROL key is a separate keystroke
- Use type operator T(n) for series of *n* Ks done as a unit

P - Point with mouse to a target on the display

- Follows Fitts' law if possible: 0.1 * log2 (D/S + 0.5)
- Typically ranges from .8 to 1.5 sec, average (text editing) is 1.1 sec.

B - Press/release mouse button (.1 sec; click is .2).

- Highly practiced, simple reaction

KLM – operators and times

H - Home hands to keyboard or mouse (.4 sec)

W - Wait for system response
- Only when user is idle because can not continue
- Have to estimate from system behavior
- Often essentially zero in modern systems

M - Mental act of thinking
- Represents pauses for routine activity
- New users often pause to remember or verify each step
- Experienced users pause and think only when logically necessary
- Estimates ranges from .6 to 1.35 sec; 1.2 sec is good single value

Example: File deletion in MacOS

Find file icon and drag into trash can

Assumptions:
- user thinks of selecting+dragging as a single operation
- Finding to-be-deleted icon is still required
- Moving icons to the trash can is highly practiced

Operator sequence:
initiate the deletion **M,** find the file icon **M,** point to file icon **P,** press and hold mouse button **B,** drag file icon to trash can icon **P,** release mouse button **B,** point to original window **P**

- **Total time = 3P + 2B + 2M = 5.9 sec**

Example: Command key file deletion

Select file icon and hit a command key

Assumptions:
- User operates both mouse + key with right hand
- Right hand starts and ends on the mouse

Operator sequence:
initiate the deletion **M,** find the icon for the to-be-deleted file **M,** point to file icon **P,** click mouse button **BB,** move hand to keyboard **H,** hit command key **KK,** move hand back to mouse **H**

- Total **time = P + 2B + 2H + 2K + 2M = 5.06 sec**
- Only slightly faster, due to the need to move the hand

Other models in GOMS family

Critical-Path Method GOMS (CPM-GOMS)
- Express activities in terms of Model Human Processor → task network → analyze for critical path

Natural GOMS Language (NGOMSL)/
Cognitive Complexity Theory (CCL)
- basic GOMS concept as simple production system
- hierarchical actions as sequential/hierarchical rules, eventually keystroke level operators

Executable GOMS Language (GOMSL)/GLEAN
- Formalized and executable version of NGOMSL.
- *GLEAN* **-** a simplified version of the EPIC simulation system (***G***OMS ***L***anguage ***E***valuation and ***An***alysis)

Summary: model-based vs. inspection evaluation

	Cognitive walkthrough	*Heuristic evaluation*	*Model-based*
Stage	Throughout	Throughout	Design
Style	Lab	Lab	Lab
Objective?	No	No	Somewhat
Measure	Qualitative	Qualitative	Qual. & Quan.
Information	Low level	High level	Low level
Intrusive?	No	No	No
Time demand	Medium	Low	Medium
Equipment	Low	Low	Low
Expertise	High	Medium	High

Outlook - next sessions

Year	**Paradigm**	**Implementation**
1950s		*Switches, punched cards*
1970s	*Typewriter*	*Command-line interface*
1980s	*Desktop*	*Graphical user interface, direct manipulation*
1980s+	*Spoken Language*	*Speech recognition/synthesis, natural language processing, dialogue systems*
1990s+	*Natural interaction*	*Perceptual, multimodal, interactive, conversational, tangible, adaptive*
2000+	*Social interaction*	*Agent-based, anthropomorphic, social, emotional, affective, collaborative*