Usability and Evaluation

Maneesh Agrawala

CS 558: Visualization
Winter 2005

[lecture adapted from Guimbretiere, Hearst, Tory and Stasko]

Reflection

So far we have:

- Learned about theories of perception/cognition
- Seen many visualization tools and techniques
- Asserted that visualizations are effective

How can we more rigorously evaluate the effectiveness of the visualizations we create?

Measuring effectiveness

Usability

- Does it help people?
- Does it convey the information?
  - Speed
  - Errors
- Is it better than other techniques?

Impact

- Is it an influential idea?
- Does it lead to a new way of thinking?

Topics

Connecting mental model with real world

User-centered design

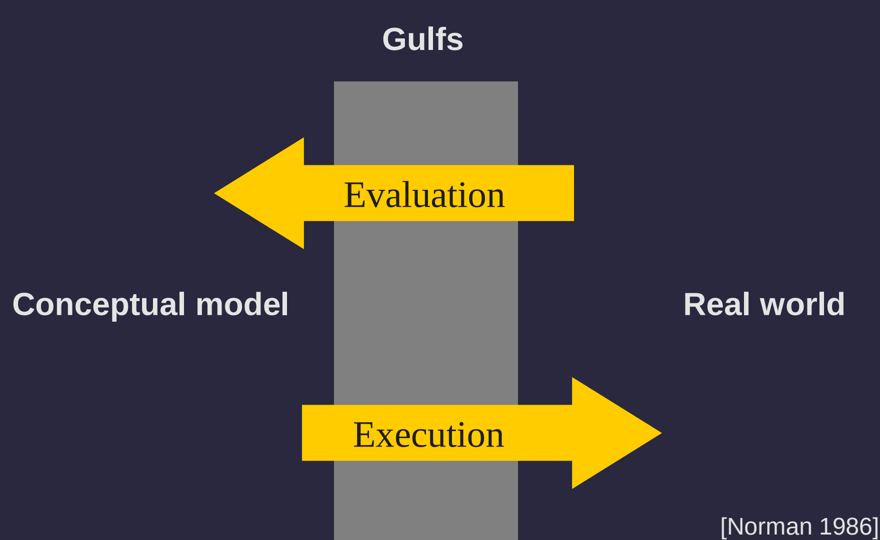
Evaluation techniques

Designing controlled experiments

Example: Identification and validation of design principles for assembly instructions

Mental Models and the Real World

Gulfs of execution & evaluation



[Norman 1986]

Gulf of evaluation

Gulf

Evaluation

Conceptual model:
x,y correlated?

Real world:

X	Y
0.67	0.79
0.32	0.63
0.39	0.72
0.27	0.85
0.71	0.43
0.63	0.09
0.03	0.03
0.20	0.54
0.51	0.38
0.11	0.33
0.46	0.46

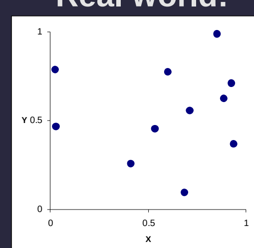
Gulf of evaluation

Gulf

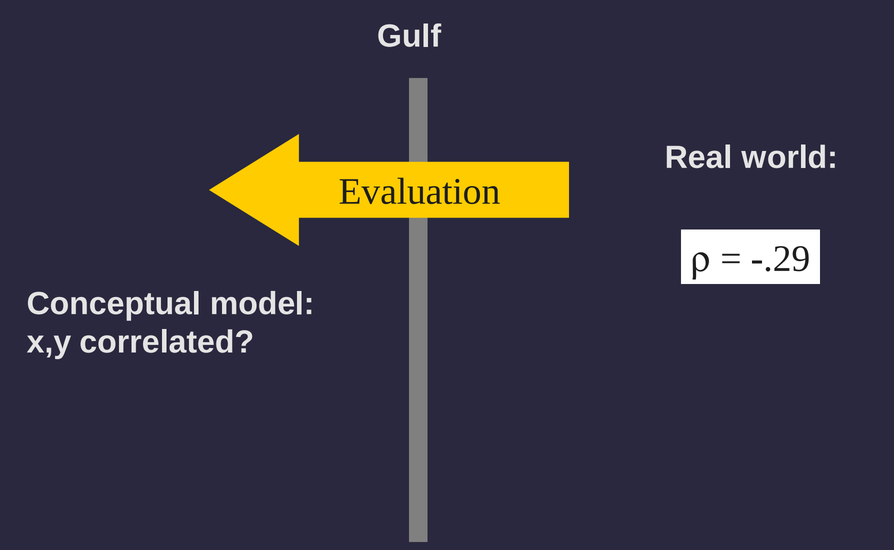
Evaluation

Conceptual model:
x,y correlated?

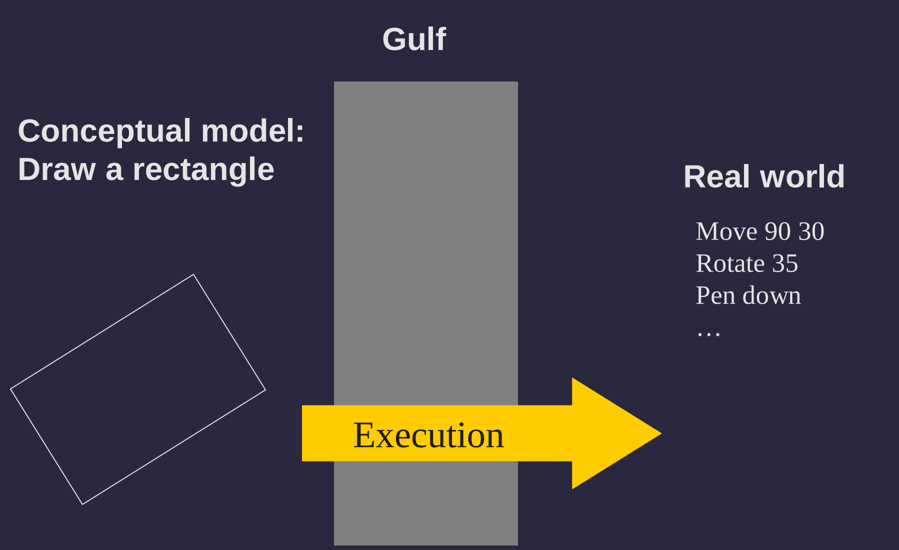
Real world:



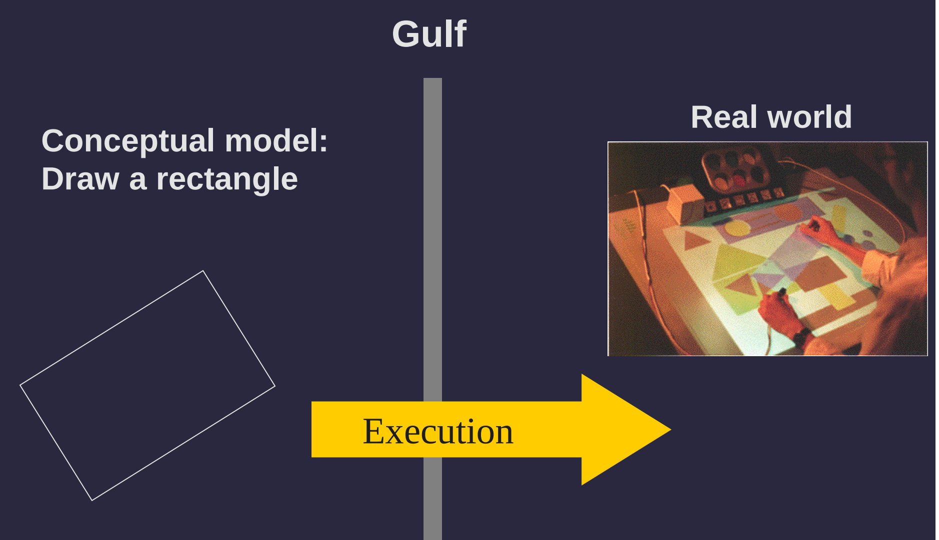
Gulf of evaluation



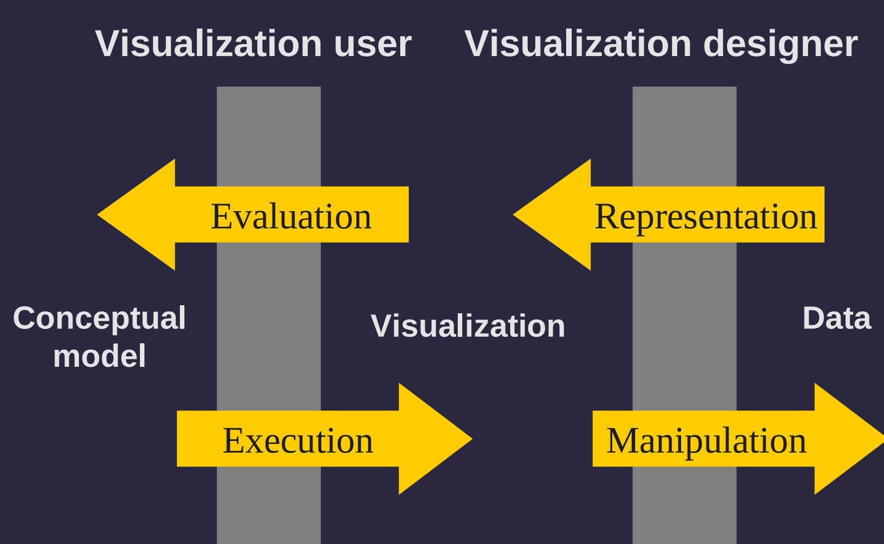
Gulf of execution



Gulf of execution



Visualization: A double gulf?



Bad visualization?

Visualization user | Visualization designer

Evaluation

Representation

x,y correlated?

X	Y
0.67	0.79
0.32	0.63
0.39	0.72
0.27	0.85
0.71	0.43
0.63	0.09
0.03	0.03
0.20	0.54
0.51	0.38
0.11	0.33
0.46	0.46

X	Y
0.67	0.79
0.32	0.63
0.39	0.72
0.27	0.85
0.71	0.43
0.63	0.09
0.03	0.03
0.20	0.54
0.51	0.38
0.11	0.33
0.46	0.46

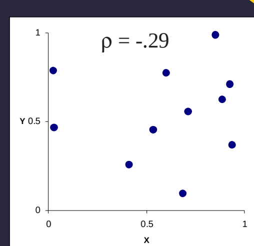
Good visualization?

Visualization user | Visualization designer

Evaluation

Representation

x,y correlated?



X	Y
0.67	0.79
0.32	0.63
0.39	0.72
0.27	0.85
0.71	0.43
0.63	0.09
0.03	0.03
0.20	0.54
0.51	0.38
0.11	0.33
0.46	0.46

User-Centered Design

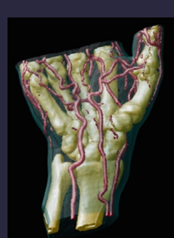
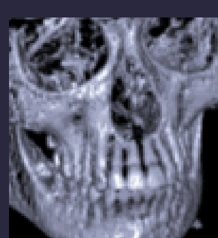
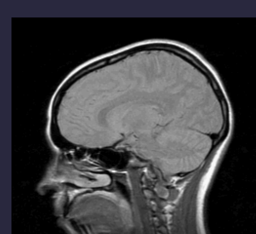
Why involve users?

Understand the users and their problems

- Visualization users are experts
- Our design intuition is not good enough

Expectation management

- Ensure users have realistic expectations
- Make users active stakeholders

[Slide adapted from Tory]

Philosophy

Focus on users and tasks

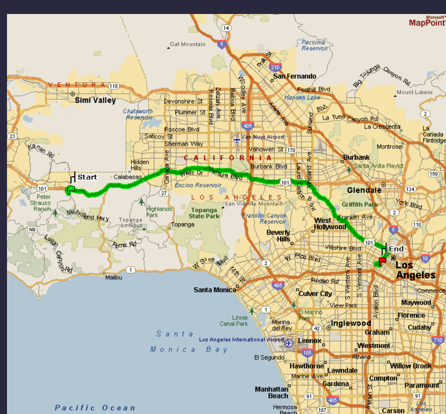
- Directly study information needs and tasks

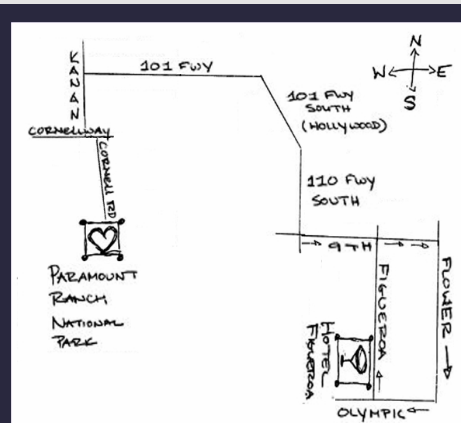
Empirical measurement

- Test reactions and performance with prototypes

Iterative design

Understanding the task





Which map is better?

Qualitative assessments

Observe users encounter problem and use visualization to solve it

- Is problem/task as expected?
- Does visualization address task?
- Observation may suggest new designs or improvements

Observing users

Ethnography

- Observer immersed in all aspects of users' life
- Long-term (weeks/months/years)

Structured observation / Contextual inquiry

- Watch user encounter problems in context
- Short-term (a few hours)

Think aloud method

- Users say what they are thinking as they encounter problem and use visualization
- Rich source of information

Subjective assessment

Ask users about their task and whether the visualization addresses it

- Is visualization enjoyable, confusing, fun, … ?
- Personal judgments can influence adoption and use

- Common assessment techniques
  - Meetings/collaborations: Interact and design with users
  - Surveys: Users fill out questionnaire about their experience

LineDrive: Understanding the task

Online map usage survey: (122 respondents, Apr 2000)

How often do you print online directions?

Always: 77.9% Most of the time: 17.2% Half the time: 4.9%

How often do you use directions in your own area?

In-town use: 76.3% Out-of-town use: 24.7%

How do you use the text versus maps?

Text: 15.6% Text mostly: 54.9% Equally: 14.8% Maps mostly: 12.3% Maps: 2.4%

Would you say online maps suffer from problems?

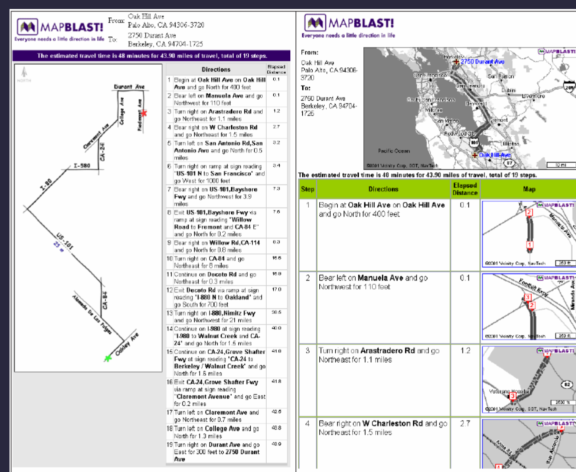
Print-outs too long:	50.0%
Difficult to recover from wrong turn:	42.6%
Overview map not useful:	50.1%
Focus maps not useful:	64.8%
Directions not reliable:	39.3%

LineDrive: Prototype

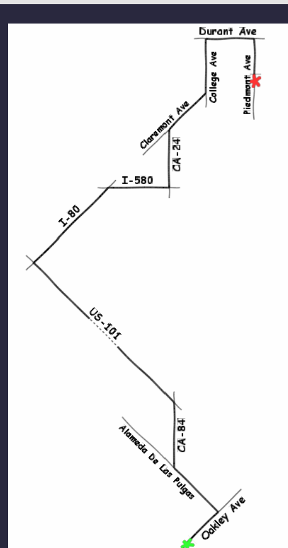
Survey #2: (90 respondents, July 2000)

Which map preferred?

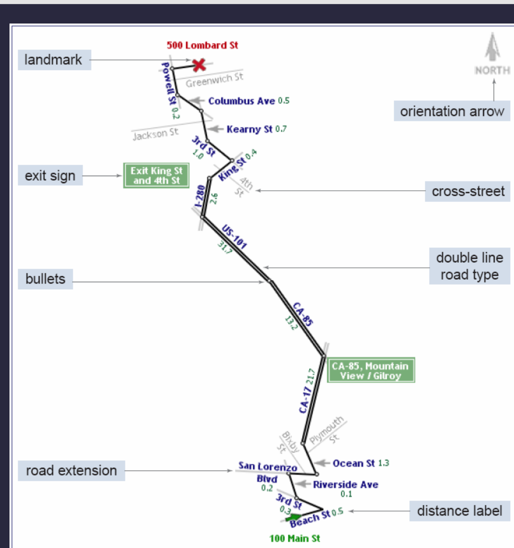
LineDrive: 87.8%
Standard: 12.2%



LineDrive: Design iterations



Prototype



Final Design

Evaluation Techniques

Techniques

With users

- Qualitative assessments
- Subjective assessments
- Controlled experiments

Without users

- Cognitive walkthroughs
- GOMS analysis
- Heuristic evaluation

Cognitive walkthrough http://hcibib.org/tcuid/

Formalized technique for imagining user's thoughts and actions when using a interface

- Given detailed design description of interface
- Select task
- Tell story motivating user actions required to do task
- Interface should give motivations via prompts/feedback
- Breakdown in motivations imply problem with interface

Walkthroughs are difficult to do when tasks are ill defined and can be accomplished in many ways

GOMS analysis [Card, Moran and Newell 83]

GOMS: Goals, operators, methods and selection

- Break task into simple physical/mental actions
- Predict time required for each action
- Identify bottlenecks and other problems

```
- take light meter out of pocket
- make sure film speed is set correctly on light meter
     -- remember: what speed of film is in the camera
- aim light meter at scene
- read light value from needle on light meter
- adjust calculator dial on light meter to light value
- the calculator shows several possible combinations of
  f-stop and shutter speed, all of which give the
  correct exposure.  So...
   -- remember: big f-stop number means small lens opening
   -- remember: small lens opening means more depth of field
   -- remember: big shutter speed number means short exposure
   -- remember: short exposure means moving things aren't blurred
   -- decide: which combination of f-stop and speed to use
- set the f-stop on the camera lens
- set the shutter speed on the camera speed dial
- remove the lens cap
- cock the shutter
- aim the camera
- focus the camera
   -- remember: if lens aperture is small, depth of field is shallow
- press the shutter release
- advance the film
```

Difficulty with GOMS

PHYSICAL MOVEMENTS		
Enter one keystroke on a standard keyboard:	.28 second	Ranges from .07 second for highly skilled typists doing transcription, to .2 second for an average 60-wpm typist, to over 1 second for a bad typist. Random sequences, formulas, and commands take longer than plain text.
Use mouse to point at object on screen	1.5 second	May be slightly lower -- but still at least 1 second -- for a small screen and a menu. Increases with larger screens, smaller objects.
Move hand to pointing device or function key	.3 second	Ranges from .21 second for cursor keys to .36 second for a mouse.
VISUAL PERCEPTION		
Respond to a brief light	.1 second	Varies with intensity, from .05 second for a bright light to .2 second for a dim one.
Recognize a 6-letter word	.34 second	
Move eyes to new location on screen (saccade)	.23 second	
MENTAL ACTIONS		
Retrieve a simple item from long-term memory	1.2 second	A typical item might be a command abbreviation ("dir"). Time is roughly halved if the same item needs to be retrieved again immediately.
Learn a single "step" in a procedure	25 seconds	May be less under some circumstances, but most research shows 10 to 15 seconds as a minimum. None of these figures include the time needed to get started in a training situation.
Execute a mental "step"	.075 second	Ranges from .05 to .1 second, depending on what kind of mental step is being performed.
Choose among methods	1.2 second	Ranges from .06 to at least 1.8 seconds, depending on complexity of factors influencing the decision.

Detailed GOMS analysis is daunting, especially if many action sequences lead to same results

Heuristic evaluation [Nielsen and Molich 90, 92]

Challenge: Identify general interface design guidelines

- Simple and natural dialog
- Speak user's language
- Minimize user memory load
- Be consistent
- Provide feedback
- Provide clearly marked exits
- Provide shortcuts
- Good error messages
- Prevent errors

Designing Controlled Experiments

Designing the experiment

Response variables (aka dependent variable(s))

- **Outcome of experiment**
- **Usually measure user performance**
  - **Time**
  - **Errors**

Factors (aka independent variables))

- **Attributes we manipulate/vary in each condition**

Levels (aka values for independent variables)

Replication

- **How often to repeat each combination of choices**

Example: Configuring a computer

Want to determine how to configure hardware for a personal workstation

- Hardware choices
  - Which CPU (three types)
  - How much memory (four amounts)
  - How many disk drives (from 1 to 3)
- Workload characteristics
  - Administration, management, scientific

We have four independent variables

Number of conditions

To isolate effect of each independent variable we consider all combinations (factorial design)

- WL1 CPU1 Mem1 Disk1
- WL1 CPU1 Mem1 Disk2
- WL1 CPU1 Mem1 Disk3
- WL1 CPU1 Mem2 Disk1
- WL1 CPU1 Mem2 Disk2
- …

(3 CPUs) * (4 memory sizes) * (3 disk sizes) * (3 workloads) = 108 combinations!

Goals

Internal validity

- Manipulation of independent variable is cause of change in dependent variable
  - Requires control of all independent variables
  - Required eliminating confounding variables
  - Requires that experiment is replicable

External validity

- Results are generalizable to real world situtations

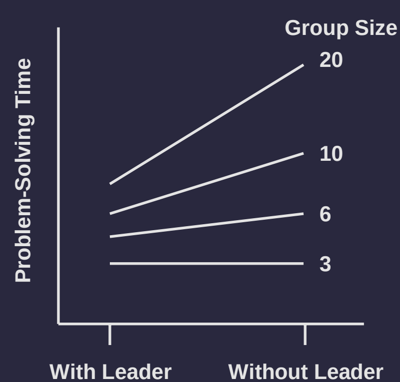
Confidence in results

- Statistics

Reducing number of conditions

Vary only one independent variable leaving others fixed

- Will miss effects of interactions



[from Martin 04]

Other reduction strategies

Compare a few independent variables at a time

- If strong effect, include variable in future studies
- Otherwise pick fixed control value for it

Fractional factorial design

- Procedures for choosing subset of independent variables to vary in each experiment

…

Ordering effects

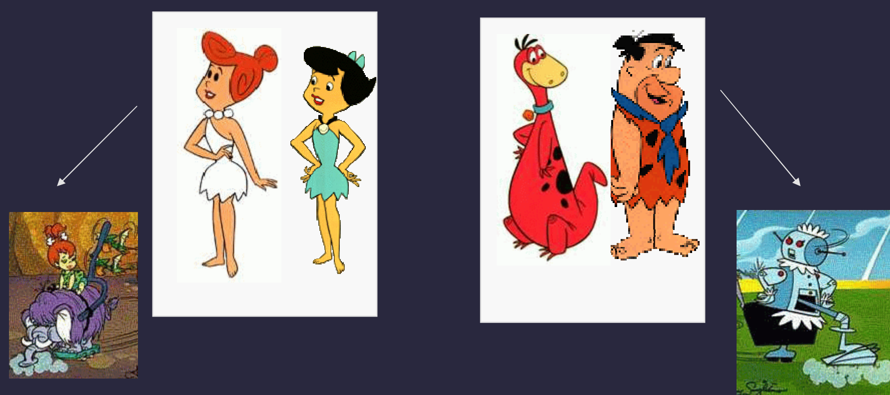
Ordering of conditions is a variable that can confound the results

- Randomization
- Counterbalancing
- Latin square (partial counterbalancing)
- …

Between subjects design

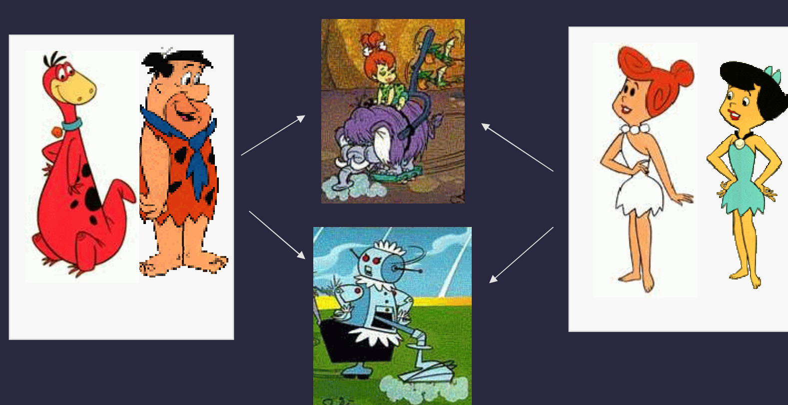
Wilma and Betty use one interface

Dino and Fred use the other



Within subjects design

Everyone uses both interfaces



Between vs. within subjects

Between subjects

- Each participant tries one condition
  - No ordering effects
  - Participants cannot compare conditions
  - Need more participants

Within subjects

- All participants try all conditions
  - Compare one person across conditions to isolate effects of individual differences (Statistically more powerful)
  - Requires fewer participants
  - Learning and fatigue effects

Statistical analysis

Compute aggregate statistics for each condition

- Usually mean and standard deviation

Compute significance (p value)

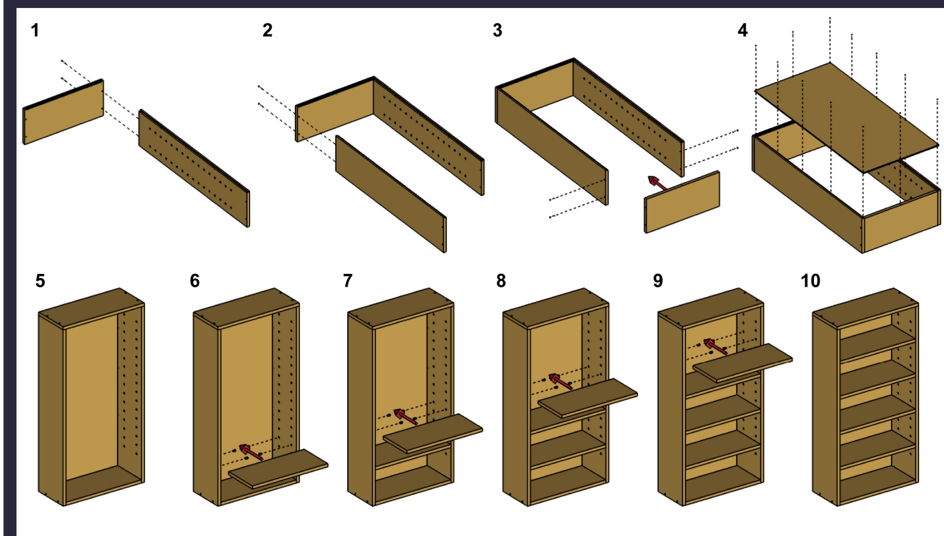
- Likelihood that results are due to chance variation
- $p = 0.05$ usually considered significant

Statistical tests

T-tests (compare 2 conditions)
ANOVA (compare >2 conditions)
Correlation and regression
Many others

Example

Designing assembly instructions

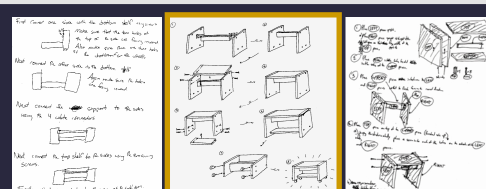


[Agrawala et al. 03]

Identification and validation [Heiser et al. 04]

Identification

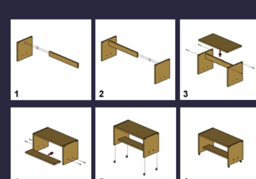
- Production
- Preference
- Comprehension



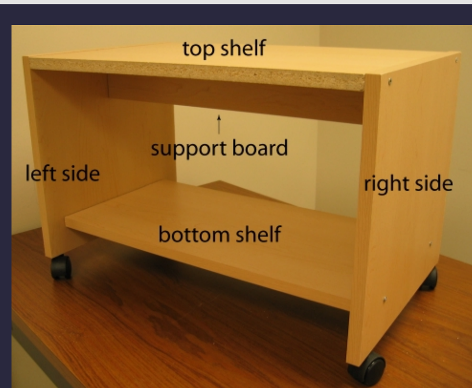
- Step-by-step, one diagram for each major step
- Clear and explicit order
- Parts added in each step should be visible
- Mode of attachment should be visible

Validation

- Instantiation
- Usability

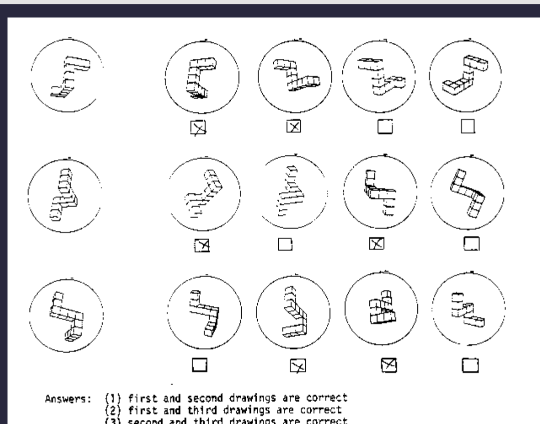


TV stand

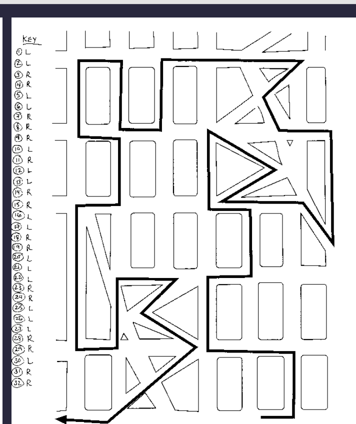


- Contains several parts and actions
- Ordering constraints
- One person can assemble
- Representative of other furniture

Spatial ability tests



Mental Rotation [Vandenburg 78]



Navigation [Money 78]

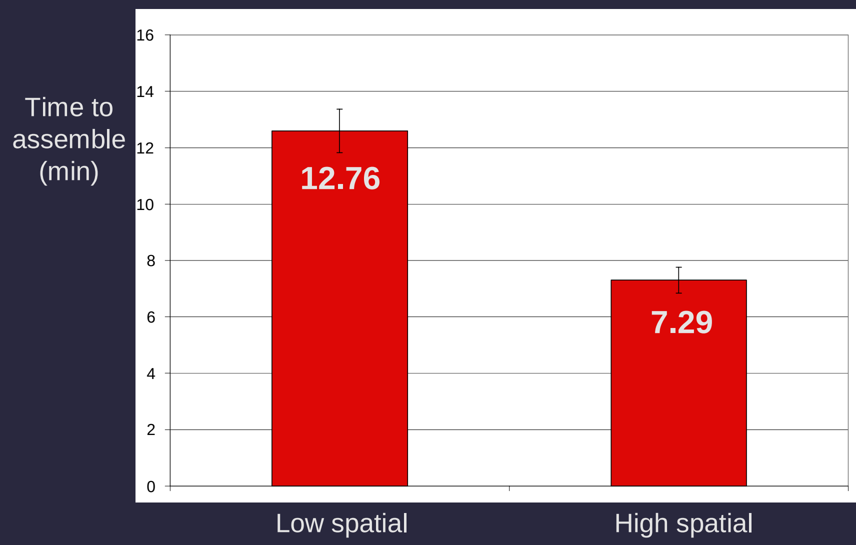
Separate high and low spatial ability

Stage 1: Production

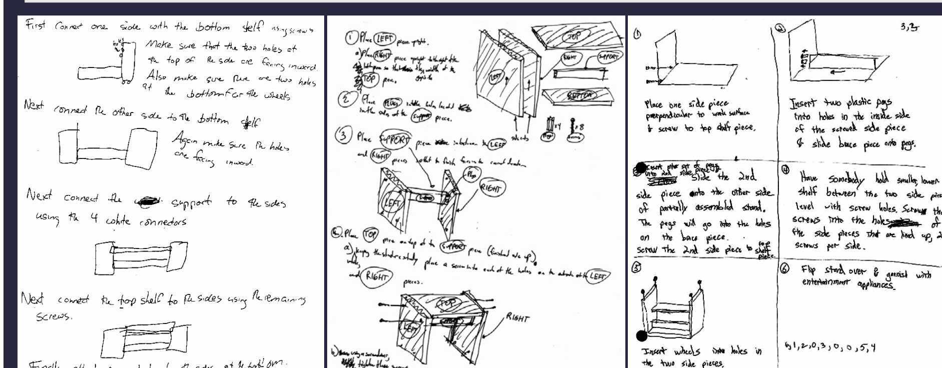
Stage 1: Production

- 43 Participants
- Assemble TV Stand without instructions
- Write instructions for novice assembler

Stage 1: Mean completion time

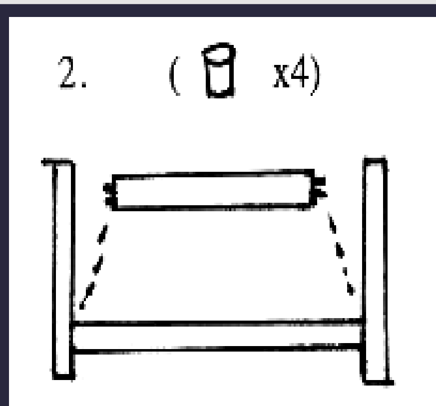


Stage 1: Instructions produced



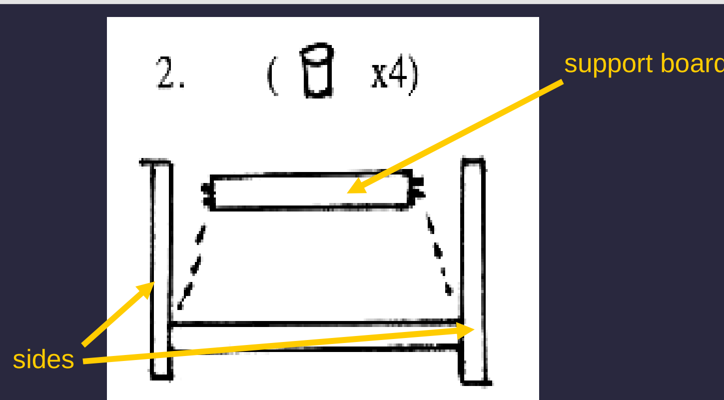
- Almost all contained diagrams 98%
- Text redundant with diagrams 62%

Stage 1: Errors in instructions



- Errors in low spatial instructions 86%
- Errors in high spatial instructions 12%

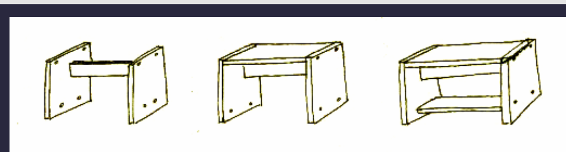
Stage 1: Errors in instructions



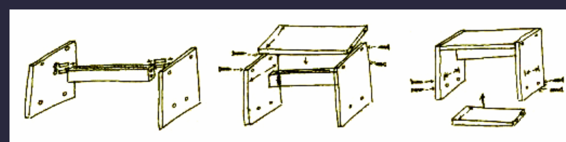
- Errors in low spatial instructions 86%
- Errors in high spatial instructions 12%

Stage 1: Classes of Diagrams

Parts menu



Structural diagrams

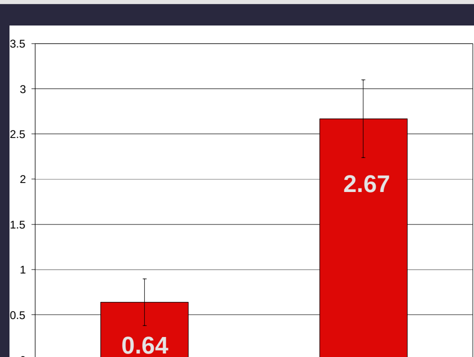


Action diagrams

- Parts menu to differentiate parts
- Structural diagrams depict completed step
- Action diagrams show assembly action/operation

Stage 1: Action diagrams

Mean number per set

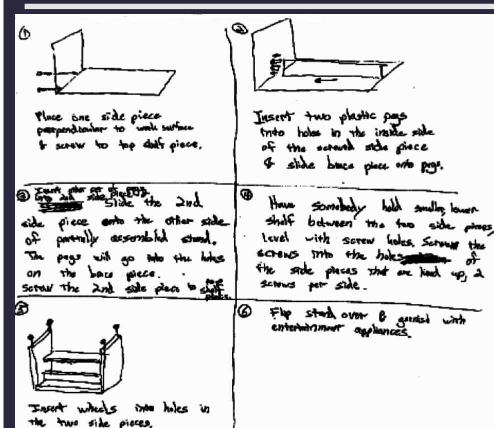


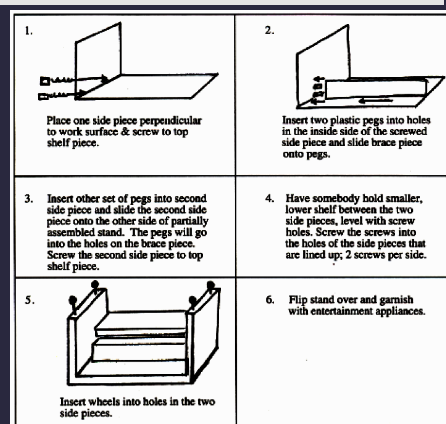
Low spatial High spatial

- High spatial
  - More action diagrams
  - More 3D diagrams
  - Less text

Stage 2: Preference

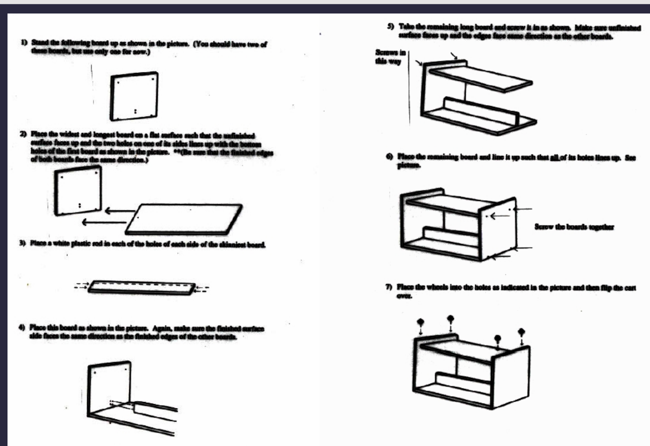
Stage 2: Preference





- 21 Participants
- Assemble TV Stand without instructions
- Rated 39 sets of redrawn instructions

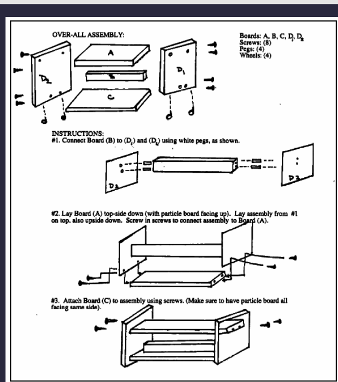
Stage 2: Highest Rated



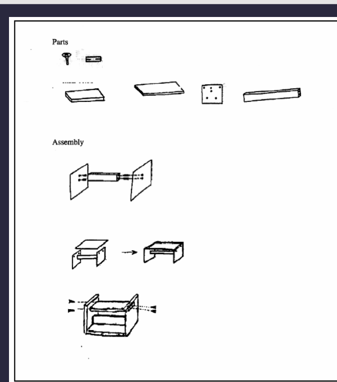
- Ratings similar across all participants
- Spatial ability does not affect preference

Stage 3: Comprehension

Stage 3: Comprehension



Set 1: Text + Action



Set 3: Parts menu + Structural + Action

- 44 Participants
- Given 1 of 4 instruction sets from Stage 2
- Assemble TV stand using instructions

Stage 3: Results

- No difference in assembly time by condition
- Instruction consultations: Low 8.9 High 7.1
- Box picture consultations: Low 9.1 High 3.4

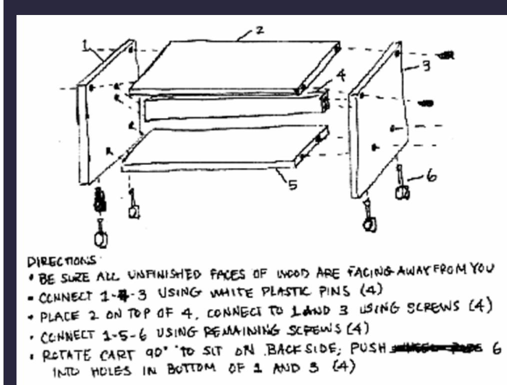
Comments

- Should show relevant parts and attachments
- Structural diagrams and exploded view hard to use
- Text not very useful

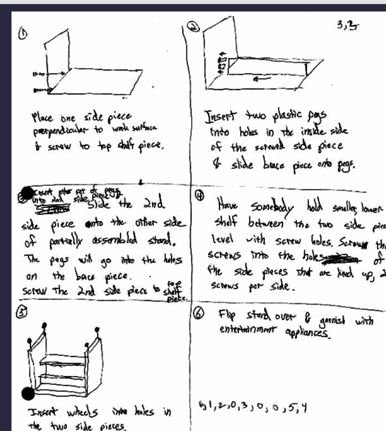
Cognitive design principles

- Sequence assembly operations
- Ensure visibility of parts
- Illustrate assembly operations

Sequence assembly operations



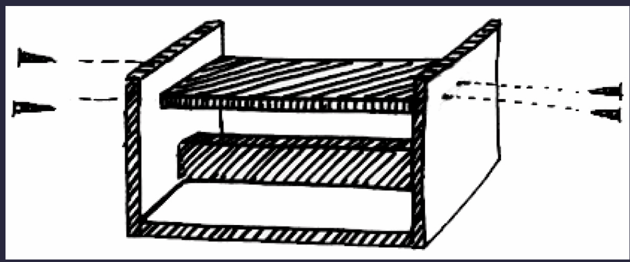
Single exploded view diagram



Step-by-step diagrams

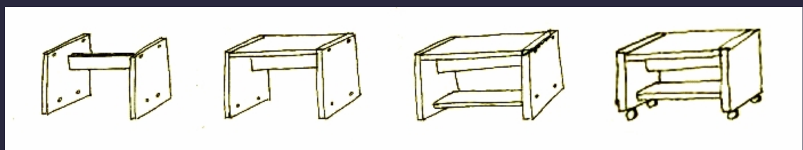
- Step-by-step, one diagram per major step
- Clear and explicit order

Ensure visibility of parts

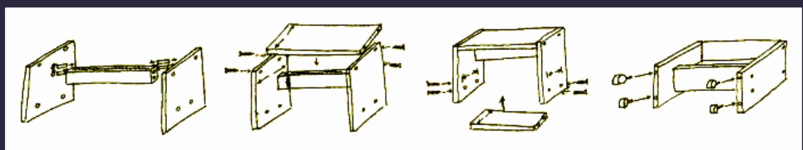


- Show parts added in each step
- Show mode and location of attachment
- Avoid changing viewpoint
- Use physically stable orientation

Illustrate assembly operations



Structural diagrams

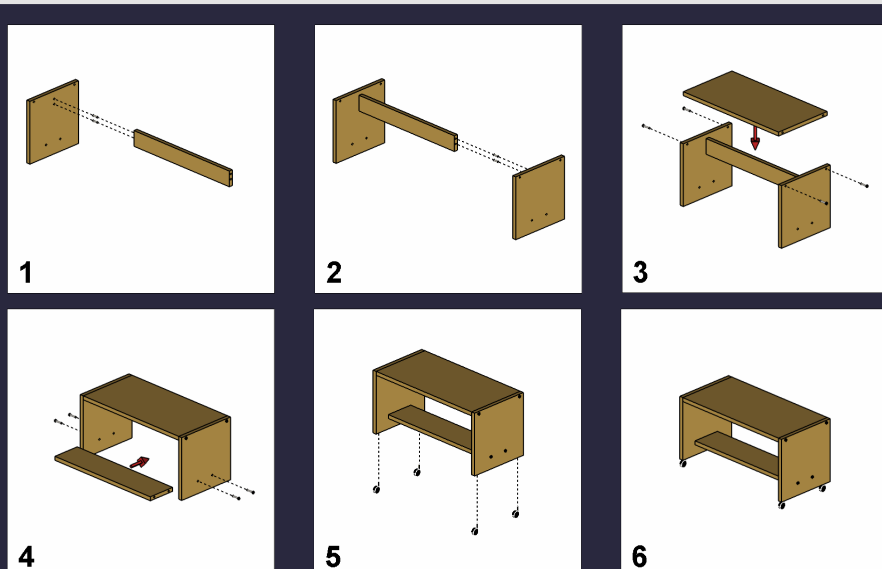


Action diagrams

- Use action diagrams rather than structural
- Use arrows and guidelines to indicate attachment

Stage 4: Instantiation

Stage 4: Instantiation [Agrawala et al. 03]

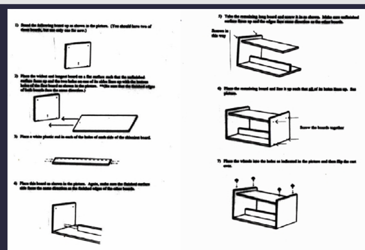


Stage 5: Usability

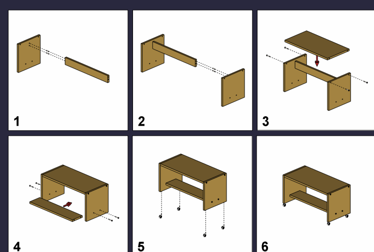
Stage 5: Usability

- 30 Participants
- Given 1 of 3 instruction sets: hand-drawn, factory, computer
- Assemble TV stand using instructions

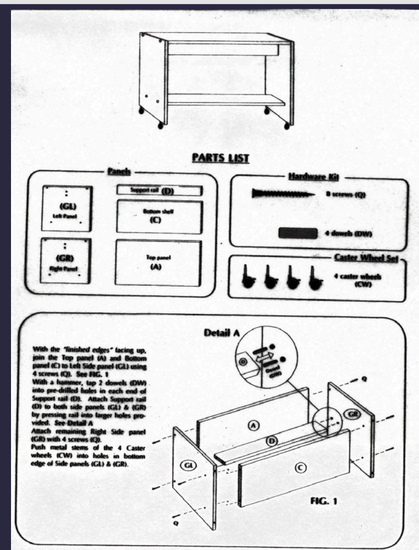
Stage 5: Instructions Tested



Hand-drawn

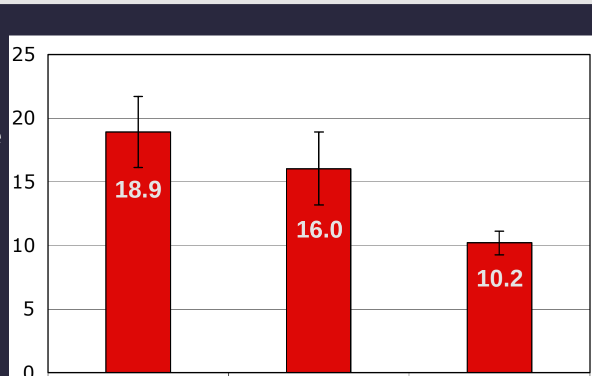


Computer generated



Factory

Stage 5: Results



- Errors: Hand-drawn 1.6 Factory 0.6 Computer 0.5
- Users rated task as easiest in computer condition

Summary

- Visualizations must support specific users doing specific tasks
- “Showing the data” is not enough!
- Evaluation techniques can tell us whether our visualization techniques are effective & usable