An Introduction to Usability Testing

HCIL Open House
May 2018

Bill Killam, MA CHFP
President, User-Centered Design. Inc.
Adjunct Professor, University of Maryland
Adjunct Professor, George Mason University

www.user-centereddesign.com

Caveat

The title of this presentation is correct:

"An Introduction to Usability Testing"…

…as long as we replace the word "Usability'…

…and the word "Testing"

The origin of the species

- “Usability testing” is the common name for multiple forms both user and non-user based system evaluation focused on a specific aspect of the design
- Done for many, many years prior, but popularized in the media by Jakob Neilson in the 1990’s

What does “usability” mean?

- ISO 9126
  - “A set of attributes that bear on the effort needed for use, and on the individual assessment of such use, by a stated or implied set of users”

- ISO 9241
  - “Extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use.”

ISO 9241

- Effectiveness - The most common data collected are success rate.
- Efficiency - The most common data collected are time-on-task, number of keystrokes, number of mouse clicks, or number of screens. This is flawed.
- Satisfaction - The most common data collected use subjective rating scales like the SUS.
- Some people have tried to combine these three elements into an über measure of usability with no real success. They have concluded that there is “a complex relationship” between these measures.

Effectiveness, Efficiency, and Satisfaction

- Logic suggests the following:
  - We need to achieve success (effectiveness) first
  - For two designs with equal effectiveness, or if a desired level of effectiveness is reached, consider increasing efficiency as a secondary goal as long as you don't lose effectiveness…
  - …but true efficiency is the measure of cognitive effort required regardless of task-times, number of mouse clicks, number of keystrokes, or number of screens.
- Satisfaction is a separate measure of the user experience that may show an impact on effectiveness at extreme levels. (Jeff Sauro's analysis of SUS scores compared to completion rate showed a correlation of only 0.24. He concluded that "approximately 6% of the satisfaction scores can be accounted for by performance differences". This is coincidence, not correlation.)
- Satisfaction is, generally, independent of usability. It is, at best, the user's *perception* of the product usability.

Elements of the User Experience

- Accessibility
  - A precursor to usability: if users cannot gain access to the product, all other elements of the experience are moot points.
- Functional Suitability
  - Does the product contain the functionality required by the user? This is the product's utility, but related to usability in terms of desirability to try to use it.
- Functional Discoverability
  - Can the user “discover” the functions of a product?
- Ease-of-learning
  - Can the user figure out how to exercise the functionality provided once it has been discovered?
- Ease-of-use
  - Can the user exercise the functionality accurately and efficiently once it’s learned?
- Ease-of-recall
  - Can the knowledge of operation be easily maintained over time?
- Safety
  - Can the user operate the system in relative safety, and recover from errors?
- Subjective Preference
  - Do user’s like it and like using it?

What are we testing?

Conscious and Unconscious Processing

- The car versus elephant analogy. - Daniel Gilbert
- System 1: Automated and Unconscious Processes
  - Fast, multi threaded (massive parallel processing)
    - 95% or more of our daily decision making
- System 2: Conscious Processes
  - Slower, one thread

Limits of Attention

Test Your Attention

Expert Behavior

- Expert behavior is exhibited as we transfer processing from slow, single threaded conscious processing to faster, multi threaded unconscious processing
- The development of expert behavior is inherent in all human learning
- Example: Do you know how to drive a car?

Goal of Interaction Design

- The goal of interaction design is to allow product interaction (*how* we do what we are doing) to occur (ideally) as all non conscious (System 1) processing, thus allowing our limited, single threaded conscious attention to focus on the goal (*what* we trying to accomplish). Ideally, this would be to the point we don't even notice (consciously) the device, product, or interface we used to get the job done.

- The less often we have to redirect our attention from our task to attend to how we accomplish the task, the more *transparent* the product design, the easier it is to use, the less errors we make, the faster we work (to a point), and the happier we are.

USER-CENTERED DESIGN

Loudoun County Public Schools / Overview

www.lcps.org/loudoun

Public Surpl...lus Auctions | INFM605-Fall 2013 | Entrepreneurs | 401(k) | We are hirin...ility People | Toyota Financial Services | 2016-12?ym...mini=2016

DISTRICT | Select A School | Translate this page from... | SIGN IN | REGISTER | Search This Site...

"A Climate for Success"

LOUDOUN COUNTY PUBLIC SCHOOLS

SCHOOL MENU | SCHOOL ATTENDANCE BOUNDARIES | CONTACT US! | CALENDAR OF EVENTS

PARENTS | STUDENTS | STAFF

Home | Our District | Academics | Administration | Employment | Information | Purchasing | School Board | Services

National African-American History Month Activities

National African-American History Month is always observed the entire month of February.

- National African-American History Month Activities
- Douglass History Presentation February 9th at Balch Library
- Briar Woods Presents 'Guys and Dolls'
- Stone Hill Middle School Leo Club Induction
- Monroe Technology Center Open House
- Trailside Middle School Mascot Revealed

view full ARTICLE | view all HEADLINES | view NEWS ARCHIVES

WHAT DO YOU WANT TO DO?

UPCOMING EVENTS

Today

5:30 PM - 7:00 PM Health, Heath, Safety, Wellness & Transportation Committee

6:30 PM - 8:30 PM Community Public Meeting on Middleburg Charter School

Tomorrow

5:30 PM - 7:00 PM Adhoc Committee on AOS/MATA facility

5:30 PM - 7:30 PM Legislative & Policy Committee Meeting

6:30 PM - 10:30 PM School Board Work Session & Adoption - FY15 Operating Budget

View Calendar

ANNOUNCEMENTS

Featured Video

Raiders Girls Varsity Volleyball Ranked 18th in Nation

Watch

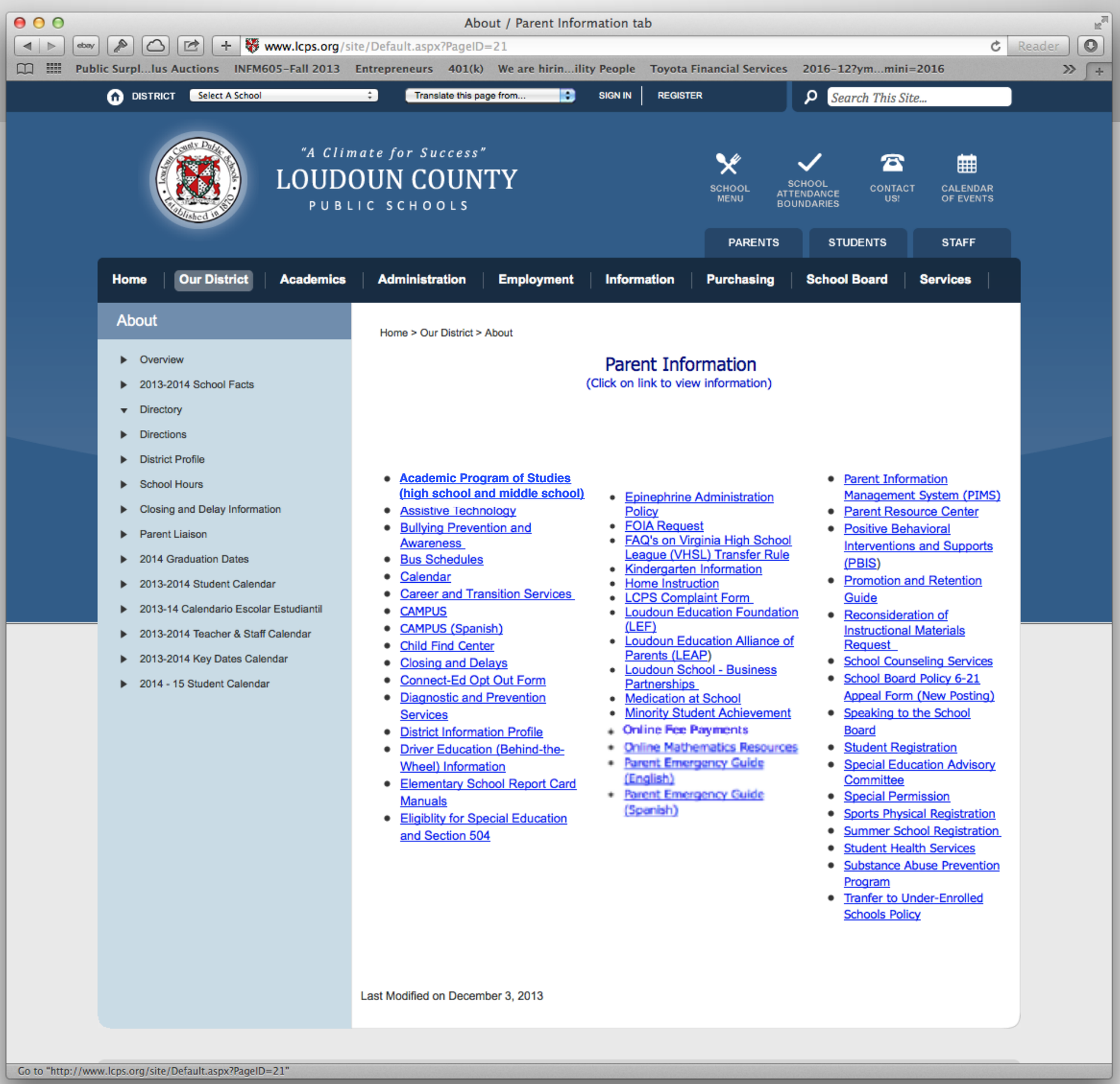
YourVoice
Powered by K12 *Insight*

Loudoun County Public Schools listens to your views on critical issues. The results of the Community Budget Priorities Survey were presented during the School Board meeting on January 14th and are viewable here. Thank you for your participation.

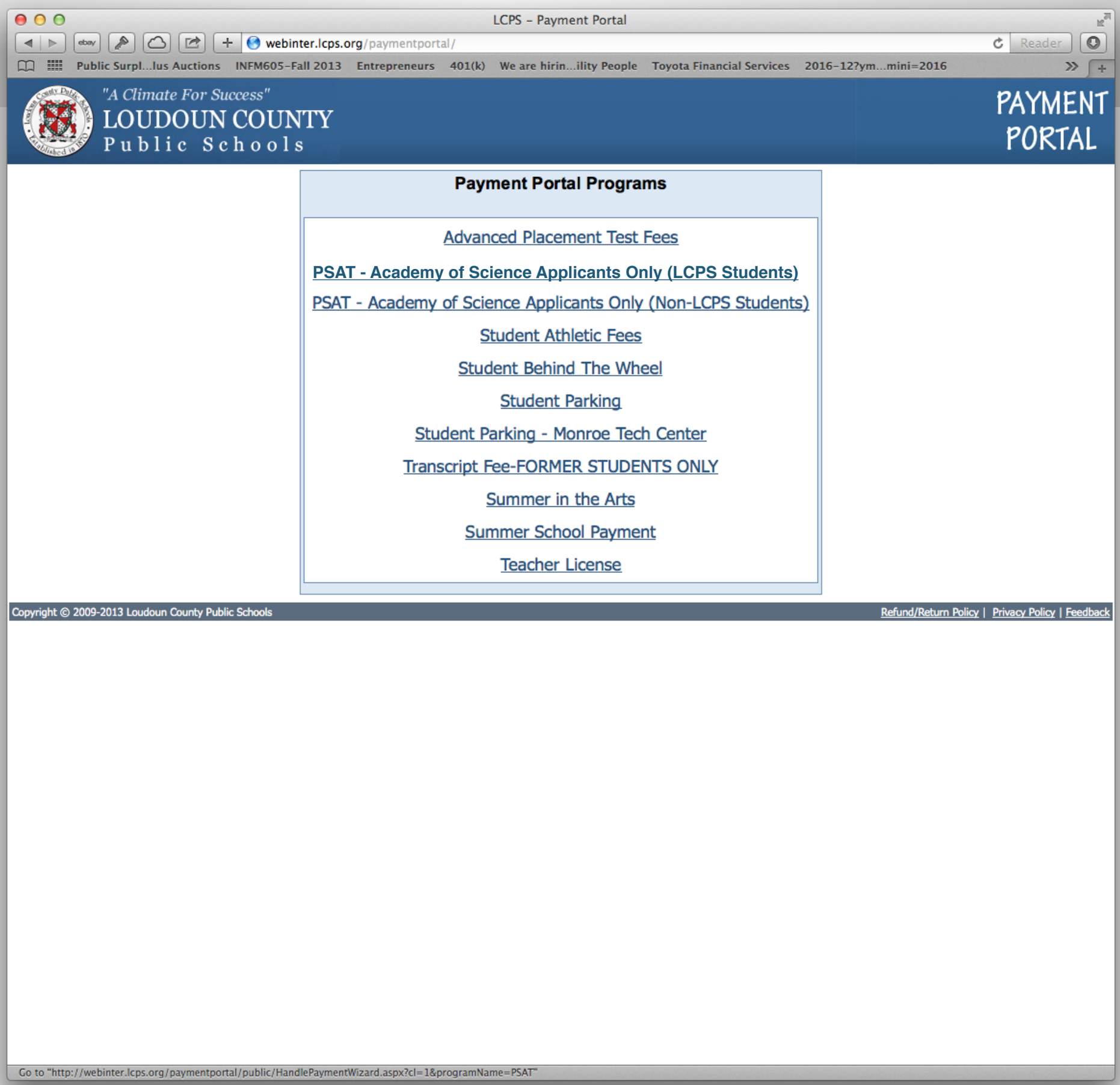
QUICK LINKS

- Apply Online Now
- Bus Routes
- Budget (Operating) Information
- BoardDocs - School Board Information
- Closing and Delay Information
- Connect-Ed Phone Messaging System Opt-Out Form
- CTE Non-discrimination policy
- District Profile
- Energy & Environment
- Fine Arts
- FOIA Requests
- FY 14 Salary Scale
- High School Athletics
- Kindergarten
- LCPS Comment and Complaint Form

USER-CENTERED DESIGN



USER-CENTERED DESIGN



USER-CENTERED DESIGN

LCPS – Payment Portal

webinter.lcps.org/paymentportal/public/HandlePaymentWizard.aspx?cl=1&programName=ap_fees

"A Climate For Success"
LOUDOUN COUNTY
Public Schools

PAYMENT PORTAL

Loudoun County Public Schools - Payment Portal

Step 1: Select the items you would like to pay for

Available Items

To add an item to the shopping cart: click on the item then click on the 'Add selected item' button or drag the item to the shopping cart.

Item	Description	Cost
AOS AB Calculus - AP w/ Statistics (Academy of Science Students Only)	Advanced Placement Course	$83.41
AOS BC Calculus - AP w/Statistics (Academy of Science Students Only)	Advanced Placement Course	$83.41
AOS Biology - AP (Academy of Science Students Only)	Advanced Placement Course	$83.41
AOS Chemistry - AP (Academy of Science Students Only)	Advanced Placement Course	$83.41
AOS Environmental Science - AP (Academy of Science Students Only)	Advanced Placement Course	$83.41
AOS Physics - AP (Academy of Science Students Only)	Advanced Placement Course	$83.41
Biology - AP	Advanced Placement Course	$83.41
Calculus AB - AP	Advanced Placement Course	$83.41
Calculus BC - AP	Advanced Placement Course	$83.41
Chemistry - AP	Advanced Placement Course	$83.41
Comparative US Government - AP	Advanced Placement Course	$83.41
Computer Science A - AP	Advanced Placement Course	$83.41
Economics - Macro - AP	Advanced Placement Course	$83.41
Economics - Micro - AP	Advanced Placement Course	$83.41
Environmental Science - AP	Advanced Placement Course	$83.41
European History - AP	Advanced Placement Course	$83.41
French - AP	Advanced Placement Course	$83.41
German - AP	Advanced Placement Course	$83.41
Human Geography - AP	Advanced Placement Course	$83.41
Language & Composition - AP	Advanced Placement Course	$83.41
Latin - AP	Advanced Placement Course	$83.41
Literature & Composition - AP	Advanced Placement Course	$83.41
Music Theory - AP	Advanced Placement Course	$83.41
Physics C - AP	Advanced Placement Course	$83.41
Psychology - AP	Advanced Placement Course	$83.41
Spanish - AP	Advanced Placement Course	$83.41
Statistics - AP	Advanced Placement Course	$83.41
Studio Art - AP	Advanced Placement Course	$83.41
U.S. Government & Politics - AP	Advanced Placement Course	$83.41
U.S. History - AP	Advanced Placement Course	$83.41
World History - AP	Advanced Placement Course	$83.41

A $81.00 Advance Placement Test fee (AP) will be administered for each AP test the student chooses to take. Fees are to be paid in advance before the test is taken. Payment of the fee does not guarantee a testing slot as all requirements must be met and applications accepted by your school in advance. In addition, payment of the testing fee does not guarantee a passing grade. Payment of the AP fee will be required for all AP tests taken. Although the Loudoun County School Board strongly believes students benefit from the Advance Placement program, this program is voluntary. On line payment of the AP testing fees is $83.41 per test. Parents and students who wish to receive a discount of $2.41 may elect to pay at the school site with cash, check, or money order. All refunds will be issued after the student fills out a refund request form and has it approved by the high schools director of guidance. Refunds will be submitted via check.

USER-CENTERED DESIGN

LCPS – Payment Portal

webinter.lcps.org/paymentportal/public/HandlePaymentWizard.aspx?cl=1&programName=ap_fees

"A Climate For Success"
LOUDOUN COUNTY
Public Schools

PAYMENT PORTAL

Loudoun County Public Schools - Payment Portal

Step 1: Select the items you would like to pay for

Available Items

To add an item to the shopping cart: click on the item then click on the 'Add selected item' button or drag the item to the shopping cart.

Item	Description	Cost
AOS AB Calculus - AP w/ Statistics (Academy of Science Students Only)	Advanced Placement Course	$83.41
AOS BC Calculus - AP w/Statistics (Academy of Science Students Only)	Advanced Placement Course	$83.41
AOS Biology - AP (Academy of Science Students Only)	Advanced Placement Course	$83.41
AOS Chemistry - AP (Academy of Science Students Only)	Advanced Placement Course	$83.41
AOS Environmental Science - AP (Academy of Science Students Only)	Advanced Placement Course	$83.41
AOS Physics - AP (Academy of Science Students Only)	Advanced Placement Course	$83.41
Biology - AP	Advanced Placement Course	$83.41
Calculus AB - AP	Advanced Placement Course	$83.41
Calculus BC - AP	Advanced Placement Course	$83.41
Chemistry - AP	Advanced Placement Course	$83.41
Comparative US Government - AP	Advanced Placement Course	$83.41
Computer Science A - AP	Advanced Placement Course	$83.41
Economics - Macro - AP	Advanced Placement Course	$83.41
Economics - Micro - AP	Advanced Placement Course	$83.41
Environmental Science - AP	Advanced Placement Course	$83.41
European History - AP	Advanced Placement Course	$83.41
French - AP	Advanced Placement Course	$83.41
German - AP	Advanced Placement Course	$83.41
Human Geography - AP	Advanced Placement Course	$83.41
Language & Composition - AP	Advanced Placement Course	$83.41
Latin - AP	Advanced Placement Course	$83.41
Literature & Composition - AP	Advanced Placement Course	$83.41
Music Theory - AP	Advanced Placement Course	$83.41
Physics C - AP	Advanced Placement Course	$83.41
Psychology - AP	Advanced Placement Course	$83.41
Spanish - AP	Advanced Placement Course	$83.41
Statistics - AP	Advanced Placement Course	$83.41
Studio Art - AP	Advanced Placement Course	$83.41
U.S. Government & Politics - AP	Advanced Placement Course	$83.41
U.S. History - AP	Advanced Placement Course	$83.41
World History - AP	Advanced Placement Course	$83.41

A $81.00 Advance Placement Test fee (AP) will be administered for each AP test the student chooses to take. Fees are to be paid in advance before the test is taken. Payment of the fee does not guarantee a testing slot as all requirements must be met and applications accepted by your school in advance. In addition, payment of the testing fee does not guarantee a passing grade. Payment of the AP fee will be required for all AP tests taken. Although the Loudoun County School Board strongly believes students benefit from the Advance Placement program, this program is voluntary. On line payment of the AP testing fees is $83.41 per test. Parents and students who wish to receive a discount of $2.41 may elect to pay at the school site with cash, check, or money order. All refunds will be issued after the student fills out a refund request form and has it approved by the high schools director of guidance. Refunds will be submitted via check.

LCPS – Payment Portal

webinter.lcps.org/paymentportal/public/HandlePaymentWizard.aspx?cl=1&programName=ap_fees

Public Surpl...lus Auctions | INFM605-Fall 2013 | Entrepreneurs | 401(k) | We are hirin...ility People | Toyota Financial Services | 2016-12?ym...mini=2016

AOS AB Calculus - AP w/ Statistics (Academy of Science Students Only)	Advanced Placement Course	$83.41
AOS BC Calculus - AP w/Statistics (Academy of Science Students Only)	Advanced Placement Course	$83.41
AOS Biology - AP (Academy of Science Students Only)	Advanced Placement Course	$83.41
AOS Chemistry - AP (Academy of Science Students Only)	Advanced Placement Course	$83.41
AOS Environmental Science - AP (Academy of Science Students Only)	Advanced Placement Course	$83.41
AOS Physics - AP (Academy of Science Students Only)	Advanced Placement Course	$83.41
Biology - AP	Advanced Placement Course	$83.41
Calculus AB - AP	Advanced Placement Course	$83.41
Calculus BC - AP	Advanced Placement Course	$83.41
Chemistry - AP	Advanced Placement Course	$83.41
Comparative US Government - AP	Advanced Placement Course	$83.41
Computer Science A - AP	Advanced Placement Course	$83.41
Economics - Macro - AP	Advanced Placement Course	$83.41
Economics - Micro - AP	Advanced Placement Course	$83.41
Environmental Science - AP	Advanced Placement Course	$83.41
European History - AP	Advanced Placement Course	$83.41
French - AP	Advanced Placement Course	$83.41
German - AP	Advanced Placement Course	$83.41
Human Geography - AP	Advanced Placement Course	$83.41
Language & Composition - AP	Advanced Placement Course	$83.41
Latin - AP	Advanced Placement Course	$83.41
Literature & Composition - AP	Advanced Placement Course	$83.41
Music Theory - AP	Advanced Placement Course	$83.41
Physics C - AP	Advanced Placement Course	$83.41
Psychology - AP	Advanced Placement Course	$83.41
Spanish - AP	Advanced Placement Course	$83.41
Statistics - AP	Advanced Placement Course	$83.41
Studio Art - AP	Advanced Placement Course	$83.41
U.S. Government & Politics - AP	Advanced Placement Course	$83.41
U.S. History - AP	Advanced Placement Course	$83.41
World History - AP	Advanced Placement Course	$83.41

A $81.00 Advance Placement Test fee (AP) will be administered for each AP test the student chooses to take. Fees are to be paid in advance before the test is taken. Payment of the fee does not guarantee a testing slot as all requirements must be met and applications accepted by your school in advance. In addition, payment of the testing fee does not guarantee a passing grade. Payment of the AP fee will be required for all AP tests taken. Although the Loudoun County School Board strongly believes students benefit from the Advance Placement program, this program is voluntary. On line payment of the AP testing fees is $83.41 per test. Parents and students who wish to receive a discount of $2.41 may elect to pay at the school site with cash, check, or money order. All refunds will be issued after the student fills out a refund request form and has it approved by the high schools director of guidance. Refunds will be submitted via check.

Add selected item(s) to shopping cart

Shopping Cart

To remove an item from the shopping cart: click on the item then click on the 'Remove selected item' button or drag the item out of the shopping cart.

Item	Description
The shopping cart is empty	
	Total:

Next

Refund/Return Policy | Privacy Policy | Feedback

Types of Testing

"Know the rules well so you can break them effectively."

- Fourteenth Dalai Lama

Bad Science

- Ten users representing the 10 user profiles of HHS participated in a usability evaluation. Later, this work was published in 2 Federal magazines as: “According to research conducted, 40% of people can’t find what they’re looking for on the HHS website.”
- The NYT reported the following finding from a research project: Nearly 42% of consumers surveyed judged the credibility of health Web sites on their visual appeal, while just 7.6% of health experts mentioned a site's design when assessing its credibility. Also consumers pick sites based on “superficial aspects of the site–the graphics and visual cues” whereas health–care experts pick sites based on “sourcing and credentials” over the site’s “attractiveness and ease-of-use.”
- A Federal agency released an RFP for a “performance-based” contract (payment tied to actual performance measured). Vendors were told they would have to redesign a site and provide evidence of improvements through an 8 person usability evaluation.

Formative Testing

Formative evaluation is a type of usability evaluation that helps to "form" the design for a product or service. Formative evaluations involve evaluating a product or service during development, often iteratively, with the goal of detecting and eliminating usability problems.

One important aspect of formative evaluation is that the audience for the observations and recommendations is the project team itself, used to immediately improve the design of the product or service and refine the development specifications. Results can be less formal than in summative evaluation, as suits the needs of designers, developers, project managers, and other project participants.

- Usability Book of Knowledge

Qualitative Research

The examination, analysis and interpretation of observations for the purpose of discovering underlying meanings and patterns of relationships, including classifications of types of phenomena and entities, in a manner that does not involve mathematical models.

The goal of qualitative testing is to understand the problem and possibly generate potential solutions.

Summative Testing

Summative usability testing is used to obtain measures to establish a usability benchmark or to compare results with usability requirements. The usability requirements should be task-based, and should tie directly to product requirements, including results from analytic tools such as personas, scenarios, and task analysis. Testing may validate a number of objective and subjective characteristics, including task completion, time on task, error rates, and user satisfaction.

The main purpose of a summative test is to evaluate a product through defined measures, rather than diagnosis and correction of specific design problems, as in formative evaluation. The procedure is similar to a controlled experiment, testing the product in a controlled environment. However, it is common to note usability problems that occur during testing, and to interview the participant after the task to obtain an understanding of the problems.

- Usability Book of Knowledge

Quantitative Research

Your goal in conducting a quantitative research study is to determine the relationship between one thing [an independent variable] and another [a dependent or outcome variable] within a population.

Quantitative research focuses on numeric and unchanging data and detailed, convergent reasoning rather than divergent reasoning.

In quantitative testing, samples are compared to test for a difference or a sample population is used to predict the outcome for the larger (target) population.

Experimental Design

Your Hypothesis

- “A supposition or proposed explanation made on the basis of limited evidence as a starting point for further investigation” - Dictionary.com
- A hypothesis a never proved. The antithesis or null hypothesis is rejected, allowing you to claim the hypothesis is likely true
- The hypothesis is “usability testing”:
  - one product is better than this other
  - this design is better then the last
  - this design works (defined how?)
- Variables
  - Independent and Dependent Variables
  - Constants
  - Random Variables
  - Confounding Variables

Validity

- “Validity is the degree to which the results of a research study provide trustworthy information about the truth or falsity of the hypothesis.”*
- Construct validity is the degree to which a variable actually measures what it purports to measure.
  - e.g., Is someone's opinion about a product’s ease of use an accurate measurement of a product actual ease of use?
- Content validity is the extent to which a measure represents all facets of a construct.
  - e.g., Is time on task a complete measure of efficiency?
- Criterion validity is the extent to which a measure is consistent with other measures taken (concurrent validity) or if measured in the future (predictive validity).
  - e.g., Are observed or measured levels of difficulty consistent with reported levels of difficulty (concurrent validity)?
  - e.g., Would you perform a task with errors today, but perform differently on the same task if attempted later?

*Cherulnik, P.D. 2001. Methods for Behavioural Research: A Systematic Approach

Construct Validity of Usability Measures

- Pupil dilation, eye tracking, galvanic skin response, “smile meter”, and secondary task time, for measuring user reaction or workload vary in construct validity and usually only in isolated settings
- Some subjective measures of difficulty (cognitive workload) such as the Cooper-Harper have limited construct validity show reasonable correlation
- Self preference or assessment of usability (e.g., SUS, SUMI, QUIZ) do not have construct validity with performance
- You can compare a design against some type of standard or benchmark, but that standard or benchmark needs be meaningful (have construct validity) For example:
  - Accomplishing a task within 3 or 5 or whatever minutes is not meaningful (has no construct validity) unless there is an actual external time constraint.
  - Finding all content within 3 clicks is not meaningful (has no construct validity)
- Task pass/fail has contract validity if an objective measures is defined (e.g., no partial credit like “completed with help”)
  - Relative success rate is the best approach (A:B Testing)
  - Single sample with benchmark is possible (e.g., 95% of the users can accomplish the task has construct validity), but tends to be against an arbitrarily defined value

Validity (continued)

- Internal validity refers to the situation where the “experimental treatments make a difference in this specific experimental instance.”**
- How you set up and run a study determines if you have internal validity. IOW, don’t screw up your data collection if you want your results to be valid.
- Typical threats to internal validity include:
  - Recruitment/selection bias
  - Any interference with the participants, including the mere presence of a moderator and/or observer
  - Differences in how tasks are administered
  - Lack of objective measures

**Cambell, D.T. & Stanley, J.C. (1963) Experimental and Quasi-experimental Designs for Research

Validity (concluded)

- External validity asks the question of “generalizability” – can the result from the experiment correctly predict the behavior of the larger audience they represent.
- You must have construct validity and internal validity before even attempting to address external validity.
- Typical threats to external validity include:
  - Lack of a representative sample
  - Differences in environment (e.g., lab versus real life)

**Cambell, D.T. & Stanley, J.C. (1963) Experimental and Quasi-experimental Designs for Research

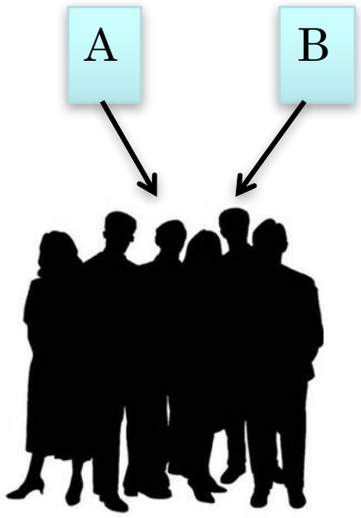
Use of Confidence Intervals

- When working with samples, a confidence interval provides a way to represent the inherent uncertainty in any test results.
- Since each sample and each test is different, the confidence level tells the reader the likelihood that another sample will provide the same results. (In other words, if you ran the test again, what value are you likely to get the next time?).
- Typical confidence intervals desired in research include the 90% or 95% confidence interval.
- Behavioral research often uses an 80% confidence interval.

Internal Validity and the Experimental Design Protocol

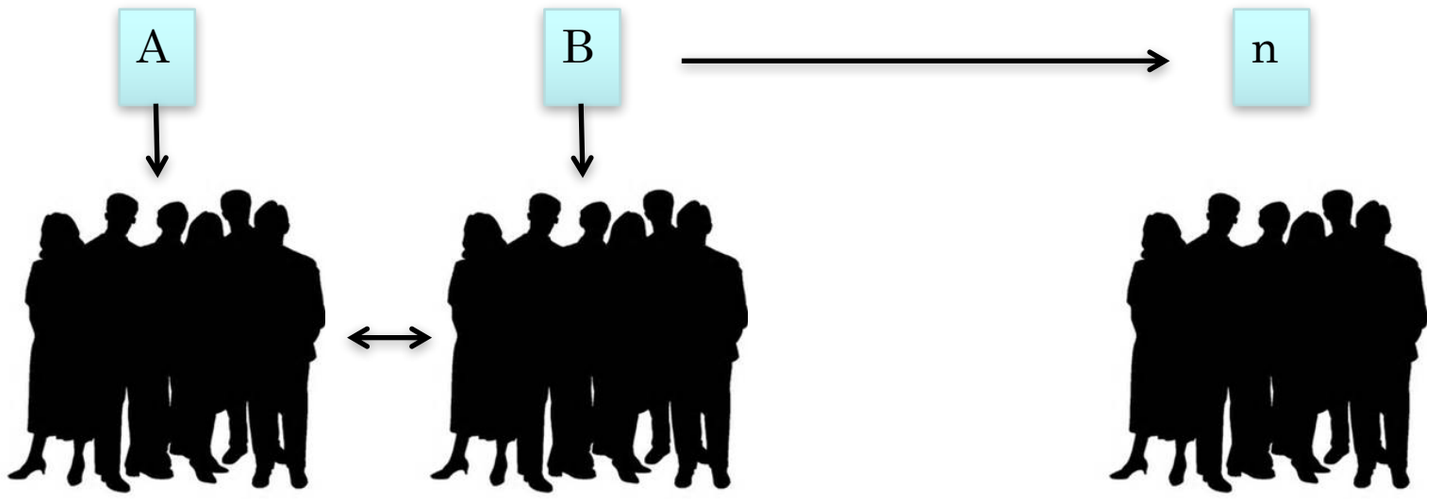
Within Subject Design

- A single group of participants are exposed to multiple (generally 2) independent variable (i.e., try out 2 designs).
- The data of interest is a comparison of the difference in the dependent variable(s) measured within the group itself (i.e., all of your data is within the group).
- Best restricted to two independent variables.
- Since group is consistent with itself, matching participants across groups is not an issue. However, exposure to the independent variables have to be counterbalanced to address possible order effect.
- Participants get to make a direct comparison between the designs, which is valuable with small groups.
- The amount of time with each product is limited compared to a between subject design.



Between Subject (Group) Design

- Having a group of participants to test each independent variable (i.e., each design).
  - The data of interest is a comparison of the difference in the dependent variable(s) measured or observed in each groups (i.e., you are comparing the difference between the groups).
- Can be used with any number of products to be evaluated.
  - Each group has to be identical to avoid introducing possible confounding variables, something that is difficult to do with small groups of participants.



Internal Validity & Sampling Methods

- Random
  - The sample frame (the part of the population you have access to) is homogeneous
  - All members of the sample frame have equal chance of being selected
  - The sample population is selected by some random selection methodology
- Stratified
  - The population is divided into separate strata (i.e., user profiles or personas), each of which is homogenous
  - All members of the each strata have equal chance of being selected
  - The sample population is selected by some random selection methodology from each strata
- Convenience
  - The sample frame is whatever part of the total population can be effectively and efficiently selected, often preselected
  - The sample population is polled to solicit participation

External Validity & Sample Size

- You can test with any sample size, but your confidence will be effected
- There is a relationship between the sample population and the larger population they are supposed to represent
- The research rule of thumb is that you need a minimum of 25-30 people to detect a medium effect size

Population size	Needed sample
1000	150
10,000	300
100,000	800

Required samples size for 5% error, 95% confidence interval, assuming a true random sample from a normally distributed, homogeneous population

Quantitative Research (Summative Testing) Examples

Quantitative Testing: Example 1

- You own a company that sells products on the web. You have always required people to register to purchase from the site. It has been suggested that sales would increase if you allowed people to purchase as a “guest.” (Your hypothesis.)
- You operationalize your dependent variable as sales or number of abandoned shopping carts.
- You create a new design for your site that allows people to purchase without registering. You make NO OTHER CHANGES.
- You set up two servers – one hosts the current design and one hosts the new design (your independent variables.)
- For one month (or more), you run both servers. Every other person who comes to the site is routed to the the alternate server. If you have large traffic, you will have near equivalent groups in both populations.
- If there is a difference in sales in favor of the new site design, you can conclude that adding the ability to check out as a guest is the likely cause of higher sales. Knowing the values of N (your population sizes), statistics can provide a confidence interval for this finding.

Quantitative Testing: Example 2

- You have 3 potential designs for brake lights - standard (your control), redundant, and redundant centered (CHMSL).
- You operationalize your dependent variables as the reduction in rear end collisions and the reduction in costs of rear end collisions.
- You obtain 3 sets of equivalent drivers operating in a real world environment (cab drivers in NYC)
- After the end of the experiment trial, you measure differences in the number of rear-end collisions and the cost of repair to test for statistically significant differences.
- Results from the original study showed an approx. 50% reduction in rear end collisions and $100 less in repair costs.
- This testing has good construct validity and good internal validity but limited external validity (generalizability) since you don't know how other drivers in other environments will compare.
- Results for the general public the first year of incorporation was 8.5%, 95% CI 6.1-10.9.

Limitations of Quantitative Testing

- In addition to the rigorous and often impractical sampling requirements, quantitative testing assumes you have a completed, fielded product or a fully operational prototype to evaluate
- Therefore, it is fiscally impossible to do valid quantitative testing during the design phase of a product

Qualitative Research (Formative Testing)

Pick one and only one

- You can either do quantitative (summative) testing or qualitative (formative) testing, but not at the same time (i.e., You can't have your cake and eat it, too)
- Formative testing requires us to observe and interact with our participants. This is done at a cost, not only in the loss of internal validity, but other factors associated with human interaction and performance
- Same size is almost always too small to get significant differences if you use the necessary confidence interval
- Sample size is too small to generalize to a broader population

Typical Usability Sample Sizes

- Usability is typically done with very few people per round
  - Neilson says you only need 5 people (but not for the right reason)
  - Krug says you only need 2 or 3 people (also not for the right reason)
  - The IUSR and the related ISO standard says 3 per user group, profile, or persona
- From a practical standpoint, a single day of testing can test with, at most, 8-9 people

Confidence Intervals for 8 Users

Success Rate	Success Rate	95% Confidence Interval
1 of 8	12.5%	1% - 52%
2 of 8	25%	3% - 65%
3 of 8	37.5%	8% - 75.5%
4 of 8	50%	16% - 84%
5 of 8	62.5%	24.5% - 91.5%
6 of 8	75%	35% - 97%
7 of 8	82.5%	47% - 99%
8 of 8	100%	63% - 100%

Potential issues when we interact with participants

Observer Effect/Social Facilitation

- Any of a general class of changes to a user's behavior as the result of being observed (or thinking they are being observed).
- Most well known is the Hawthorne Effect. This effect causes a divergence in performance - the good do better, the poor do worse.
- The effect of observers is powerful and unconscious. And they don't even need to be real observers. In research on stealing and lying, children are less likely to cheat and lie if there is a mirror in the room.
- Melissa Bateson (Newcastle University) ran a field experiment with her own (psychology) department. Coffee was paid for on a faith basis. She alternated images above the donation box - even weeks had a poster with flowers on it, odd weeks had a poster with eyes on it. On odd weeks, contributions were 3x what was received on even weeks.

Confabulation

- If System 2 does not have access to the information of system 1, it will use logic to answer the question even if it's incorrect. We cannot think about our own thought processes.

- In a split brain study, people were shown a picture of a chicken's leg and a picture of a car covered in snow and then asked to point to a related picture in a set. People pointed to either a picture of a chicken or a picture of a snow shovel. If the image was shown to the left hemisphere, they could describe the reason why they pointed to this picture. If the picture was shown to the right hemisphere, they pointed to the picture but could not explain why.

- When participants were shown the picture of a chicken's leg to the left hemisphere and a picture of the car in snow to the right hemisphere at the same time, they would point to the same 2 pictures. When asked why they pointed to the picture of the shovel, participants reported that chickens produce a lot of chicken poop, so you need a shovel to clean it up.

Answer Substitution

- You are shown a picture of a person running for office and asked if you think they will win. There are far too many variables for you to make a good prediction, so the task is too hard for system 2 to work out.
- System 1 substitutes the hard question for an easier one – does the person look like a person who will win?
- System 1 provides an answer to that new question, but System 2 reports it as the answer to the first question without realizing the substitution.

Projected Responding

- Respondents believe they understand the goal of the project and attempt to provide the information they think is being asked for.
- Encouraged by subtle differences in responses (correctly or incorrectly) perceived. Why it's extremely hard to test your own designs. And why you NEVER take notes in the participants' presence.
- Almost unavoidable.

Observer Differences/Bias

Some people see a monster. We see improper metering, poor lens selection, and a total lack of composition.

Style Guide, Heuristics Evaluation, and Expert Reviews

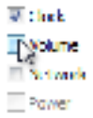
Style Guide Reviews

- The Spelling and Grammar checker of usability testing
- Possible (within limits) to be performed by anyone
- Can remove the low level usability issues that often mask more significant usability issues
- Available Standards
  - Commercially GUI & Web Standards and Style Guides
  - Domain Specific GUI & Web Standards and Style Guides
  - Internal Standards and Style Guides

Style Guide Reviews (concluded)

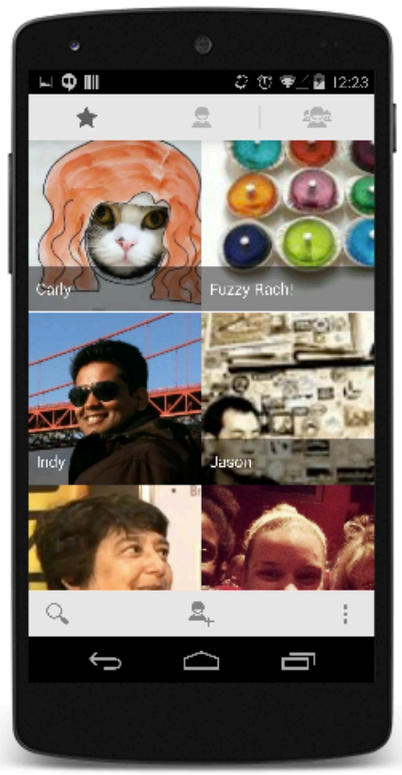
Check Boxes

With a check box, users make a decision between two clearly opposite choices. The check box label indicates the selected state, whereas the meaning of the cleared state must be the unambiguous opposite of the selected state. Consequently, check boxes should be used only to toggle an option on or off or to select or deselect an item.



A typical group of check boxes.

Swiping Between Tabs



People app with swipe gesture navigation between top-level screens.

Click device screen to replay movie.

If your app uses action bar tabs, use swipe to navigate between the different views.

Checklist

- Use swipe to quickly navigate between detail views or tabs.
- Transition between the views as the user performs the swipe gesture. Do not wait for the gesture to complete and then transition between views.
- If you used buttons in the past for previous/next navigation, replace them with the swipe gesture.
- Consider adding contextual information in your detail view that informs the user about the relative list position of the currently visible item.
- For more details on how to build swipe views, read the developer documentation on Implementing Lateral Navigation.

Alerts & Action Sheets

Alerts and action sheets are full-screen system interfaces that you use to convey information and request feedback. Action sheets let you prompt the user to choose from one of several possible options. Alerts let you display errors or other important information related to the state of your app and its activities. Alerts and action sheets are modal interfaces, and you can present them from any of your app's screens.

Alerts and action sheets come in three different styles, and each has a specific use:

Action sheets ask the user to select from a set of possible options. An action sheet displays a title, an optional message, and one or more buttons from which to select. One button is always designated as a Cancel button and displayed in the upper-left corner of the sheet. You can customize the title of the Cancel button as needed.

Alerts communicate errors or unusual conditions to the user. An alert displays a title, an optional message, and a button to dismiss the sheet. Use the title and message to communicate precisely what happened. Do not display more than one button.

Side-by-side alerts communicate errors or unusual conditions where you need to give the user a choice between two options. A side-by-side alert displays a title, an optional message, and exactly two buttons. Use the title and message to describe precisely what happened and the choice that the user must make. The left button should always let the user dismiss the alert without taking any action.

Each button in an alert or action sheet has an associated style that conveys information about the button's purpose. Most buttons use the default style, which indicates that the button has no special meaning. The destructive style indicates that the button destroys user data or performs a destructive action on the app. The cancel style indicates that the button dismisses the sheet without taking any action. Button styles affect the appearance and placement of buttons in each sheet.

Use action sheets to prompt the user to choose an option. An action sheet displays a list of options and conveys how the user's choice is to be applied. Use action sheets to request information from the user about how to proceed. For example, a messaging app might offer options to respond to an incoming message or ignore it.

Use alerts sparingly. Alerts and side-by-side alerts inform the user about errors or unusual conditions in your app. As a result, you should use them rarely, if at all.

Always include a Cancel button. Always give the user a way to dismiss the sheet without taking further action. The button you include may have a title other than "Cancel" but it should always use the cancel button style.

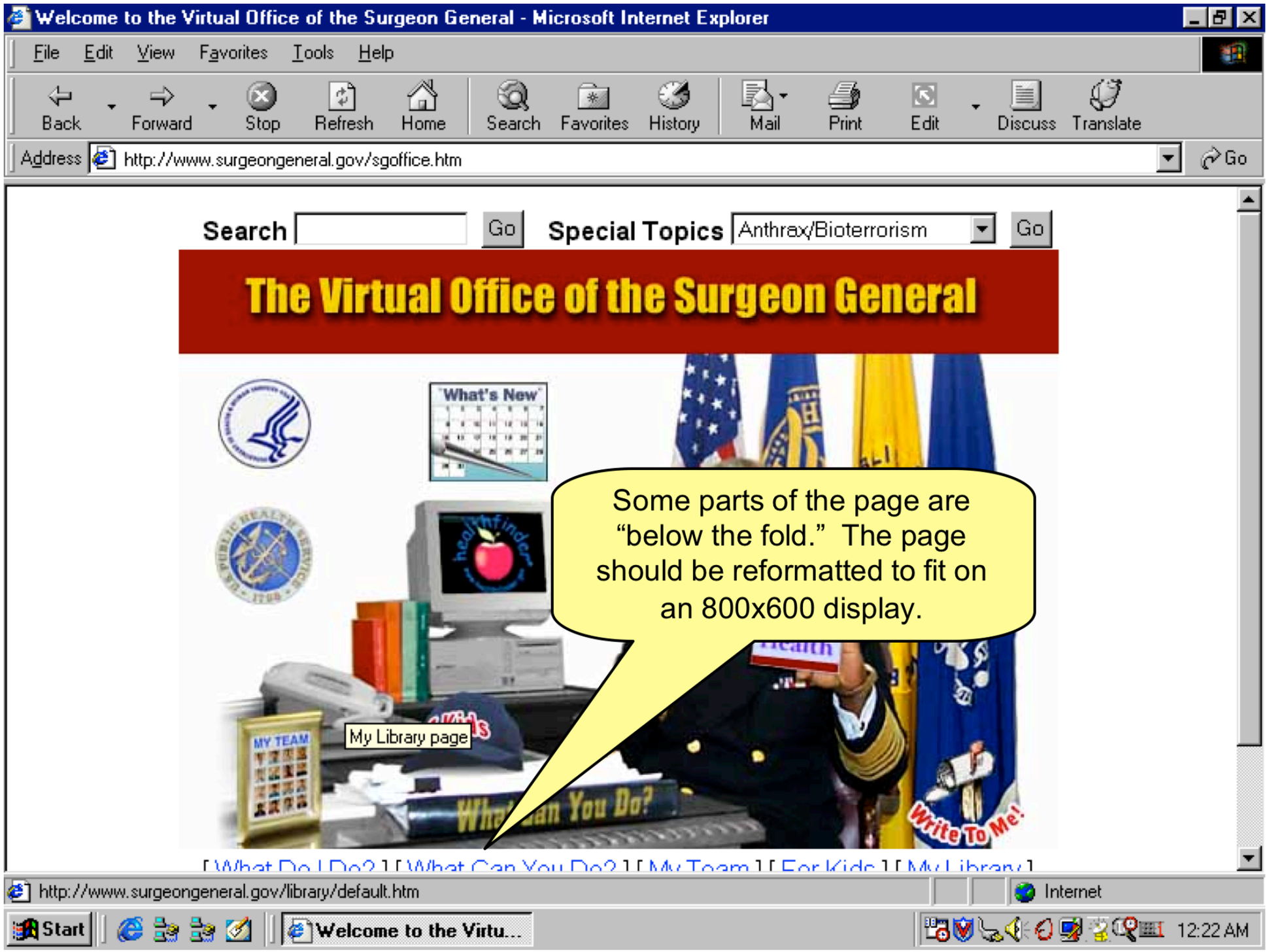
Heuristic Evaluations

- A semi structured method of reviewing a product
- Based on established usability principles (heuristics)
  - Hix and Hartson Design Principles
  - Shneiderman's 8 Golden Rules
  - Tognazinni's First Principles of Interaction Design
  - Neilson or Molich/Neilson
  - Norman Principles (Derived)

1st Heuristic

Functional discoverability through obvious interactive elements and adequate feedback

J.K.Rowling
official site


Welcome to the Virtual Office of the Surgeon General - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Forward Stop Refresh Home Search Favorites History Mail Print Edit Discuss Translate
Address http://www.surgeongeneral.gov/sgoffice.htm
Go
Search Go
Special Topics Anthrax/Bioterrorism Go
The Virtual Office of the Surgeon General
What's New
My Library page
Some parts of the page are "below the fold." The page should be reformatted to fit on an 800x600 display.
Write To Me!
http://www.surgeongeneral.gov/library/default.htm
Internet
Start
Welcome to the Virtu...
12:22 AM

2nd Heuristic

Design for the Intended User (Not Yourself)


CD-RW

The Game of 15

Let's play the game of "15." The pieces of the game are the numbers 1, 2, 3, 4, 5, 6, 7, 8, and 9. Each player takes a digit in turn. Once a digit is taken, the other player cannot use it. The first player to get three digits that sum to 15 wins.

Here's a sample game: Player A takes 8. Player B takes 2. Then A takes 4, and B takes 3. A takes 5. What digit should B take?

The Game of 15 (continued)

4	3	8
9	5	1
2	7	6

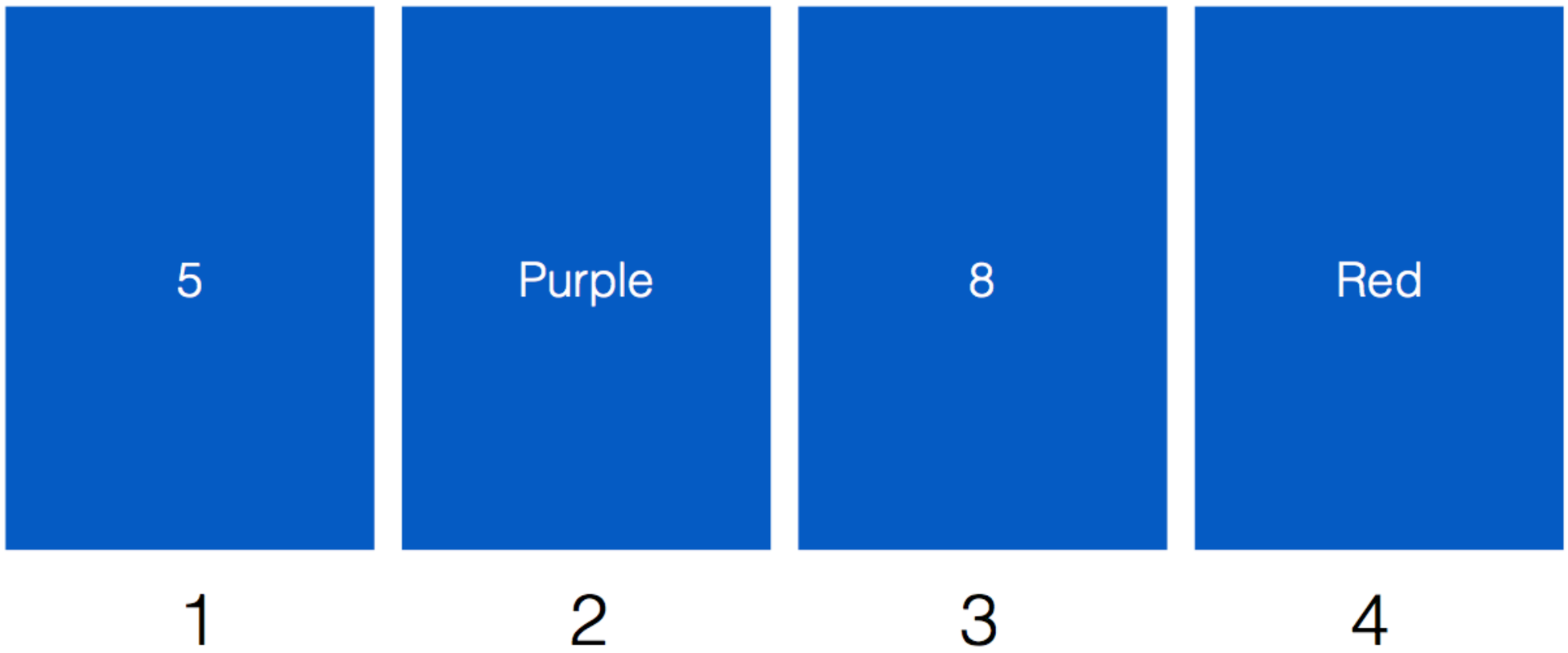
The Game of 15 (concluded)

~~4~~	~~3~~	~~8~~
9	~~5~~	1
~~2~~	7	6

Designing in the User's Domain

- The rule to test: If a card has an even number on its face, its has a primary color on its opposite face.
- How many cards in the next slide do you need to look at to confirm this rule is being followed?

Designing in the User's Domain



Designing in the User's Domain

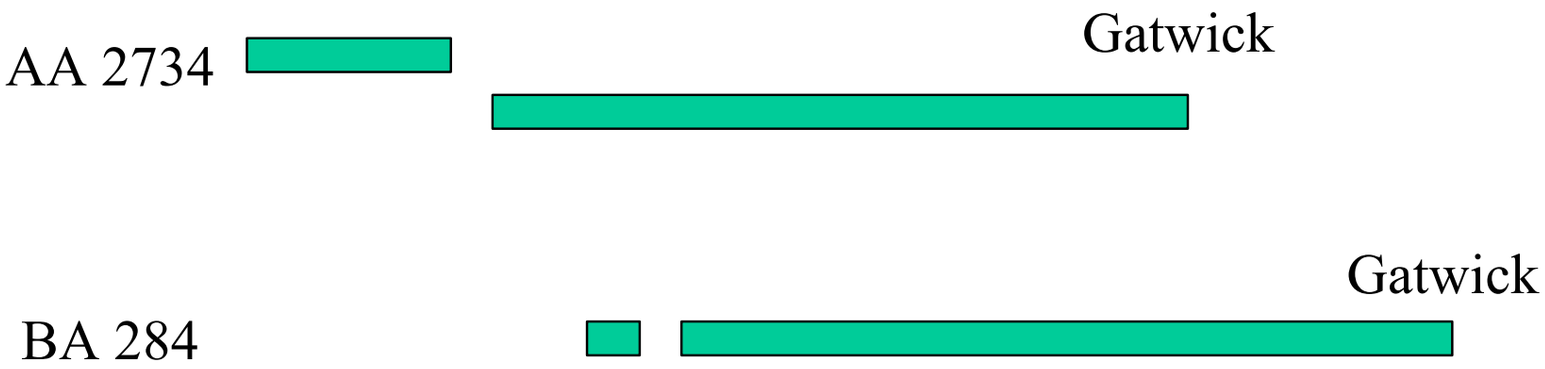
- The rule to test: You cannot drink alcohol if you are under 21.
- How many cards in the next slide do you need to look at to confirm this rule is being followed?

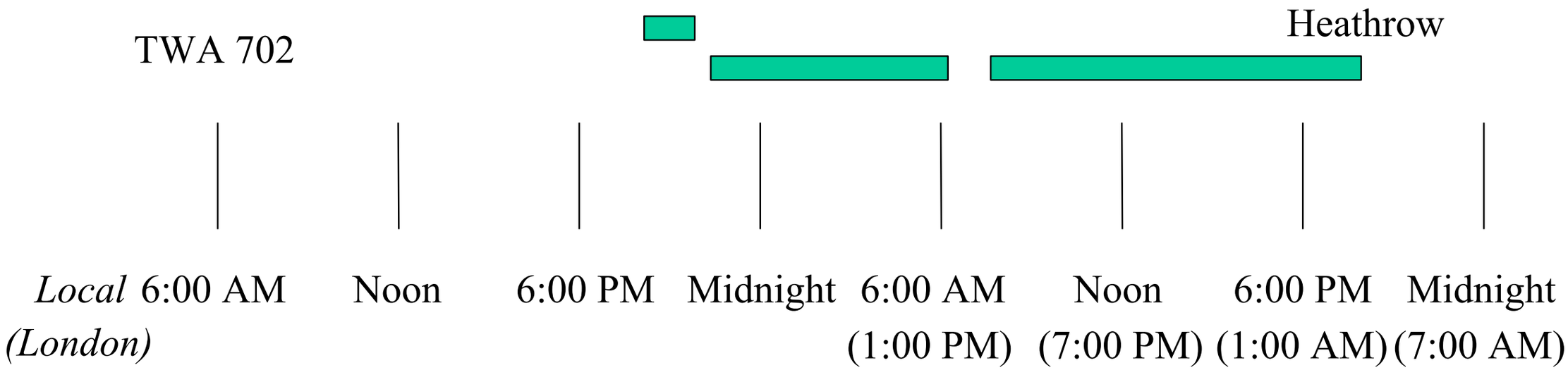
Designing in the User's Domain

Designing for Experience

	1131	SAN	0820+1	LGW	AA	2734	FCYBM	D10	1
		AA	2734	CHG PLANE AT DFW					
X12	1805	SAN	1425+1	LGW	BA	284	FJMSB	D10	1
	2100	SAN	2030+1	LHR	TW	702	FCYBQ	*	2
		TW	702	EQUIPMENT 767 LAX L-10					

Designing for Experience





3rd Heuristic

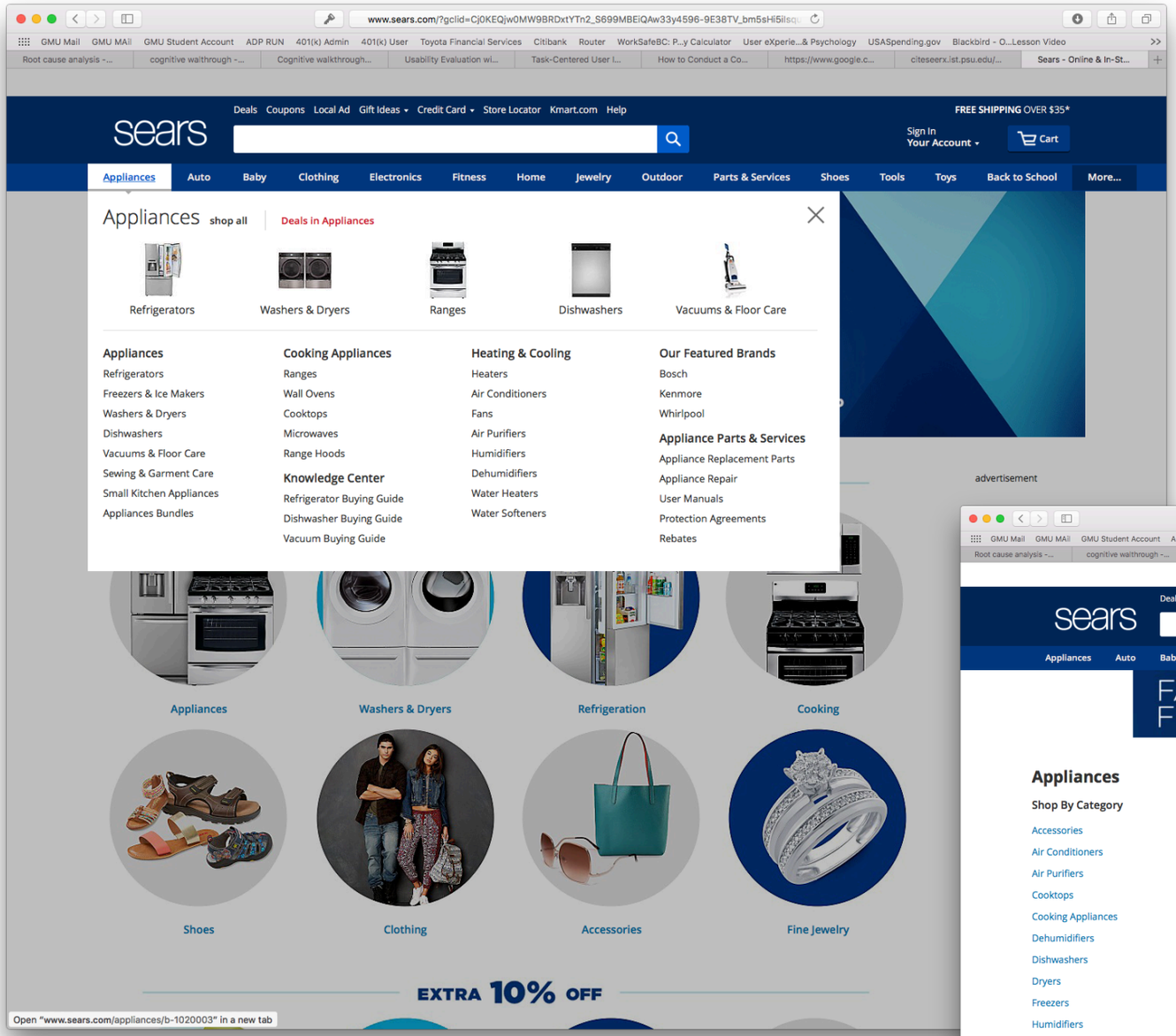
The user must be able to develop a good, complete, and unambiguous cognitive (or conceptual) model of the product to predict the effects of our actions

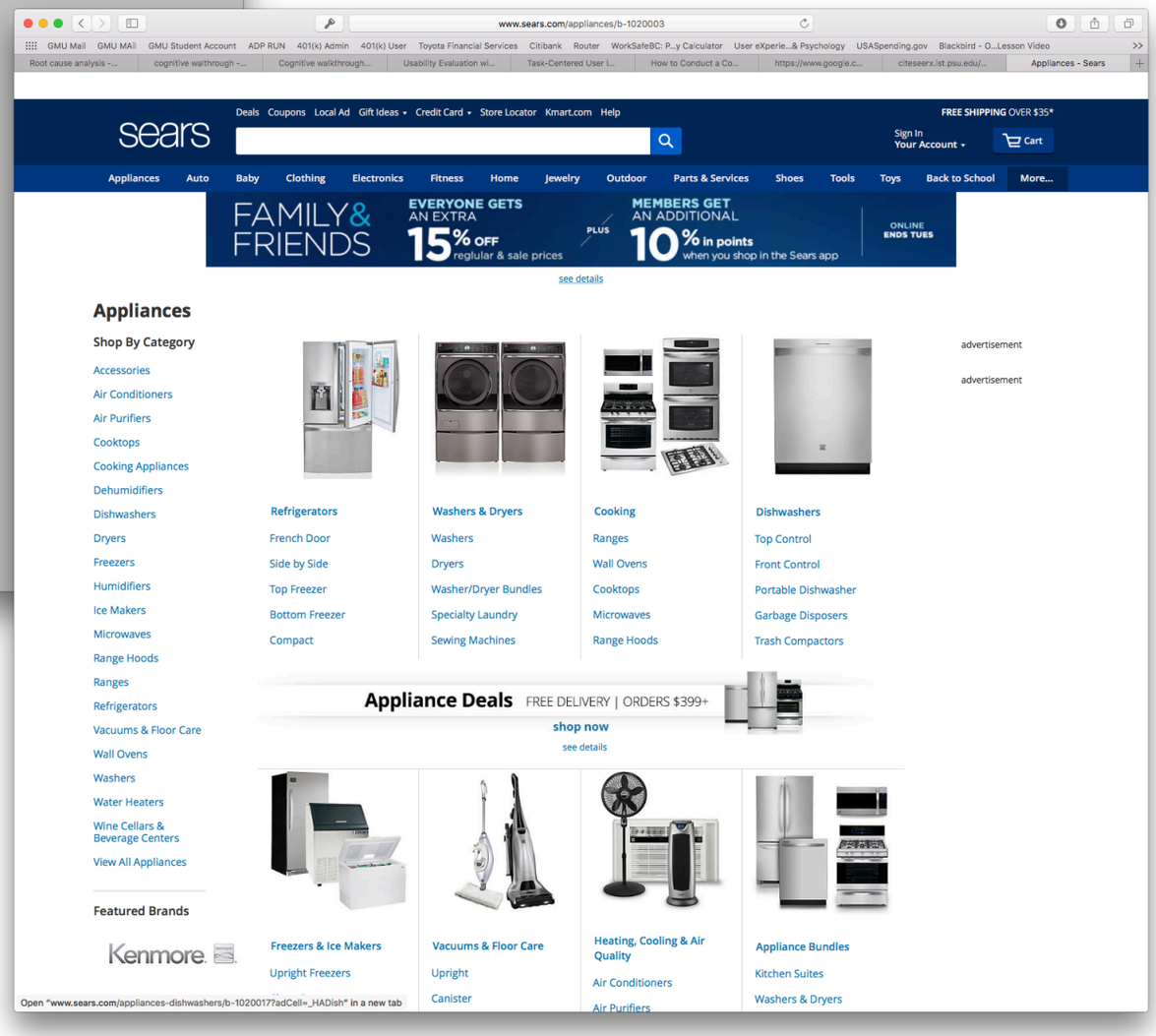
Simple Cognitive Models

More Difficult Cognitive Model

Another Cognitive Model with Limits

Cognitive Model of Information





4th Heuristic

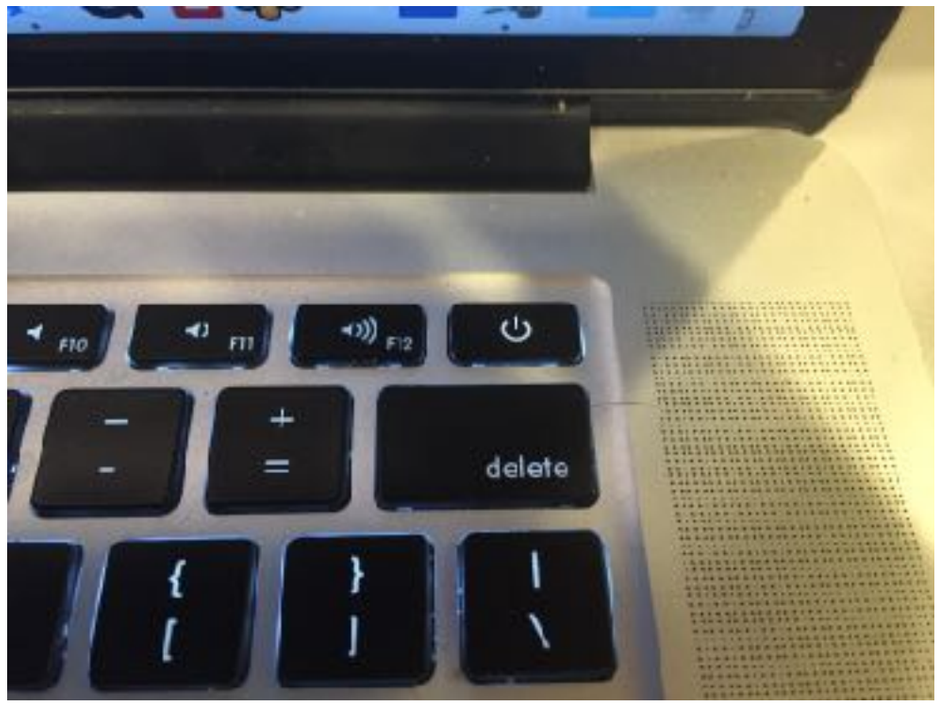
Design for Errors

(Expect, eliminate, limit the impact of, or compensate for errors)

)

Error Type 1: Slips

- Slips are common users issues
- Hand/eye coordination or basic control of our psychomotor systems
- Exacerbated by distraction, speed, attention overload
- Unavoidable by design but need to be anticipated and addressed by the designer
- “To err is human. To forgive: Design”

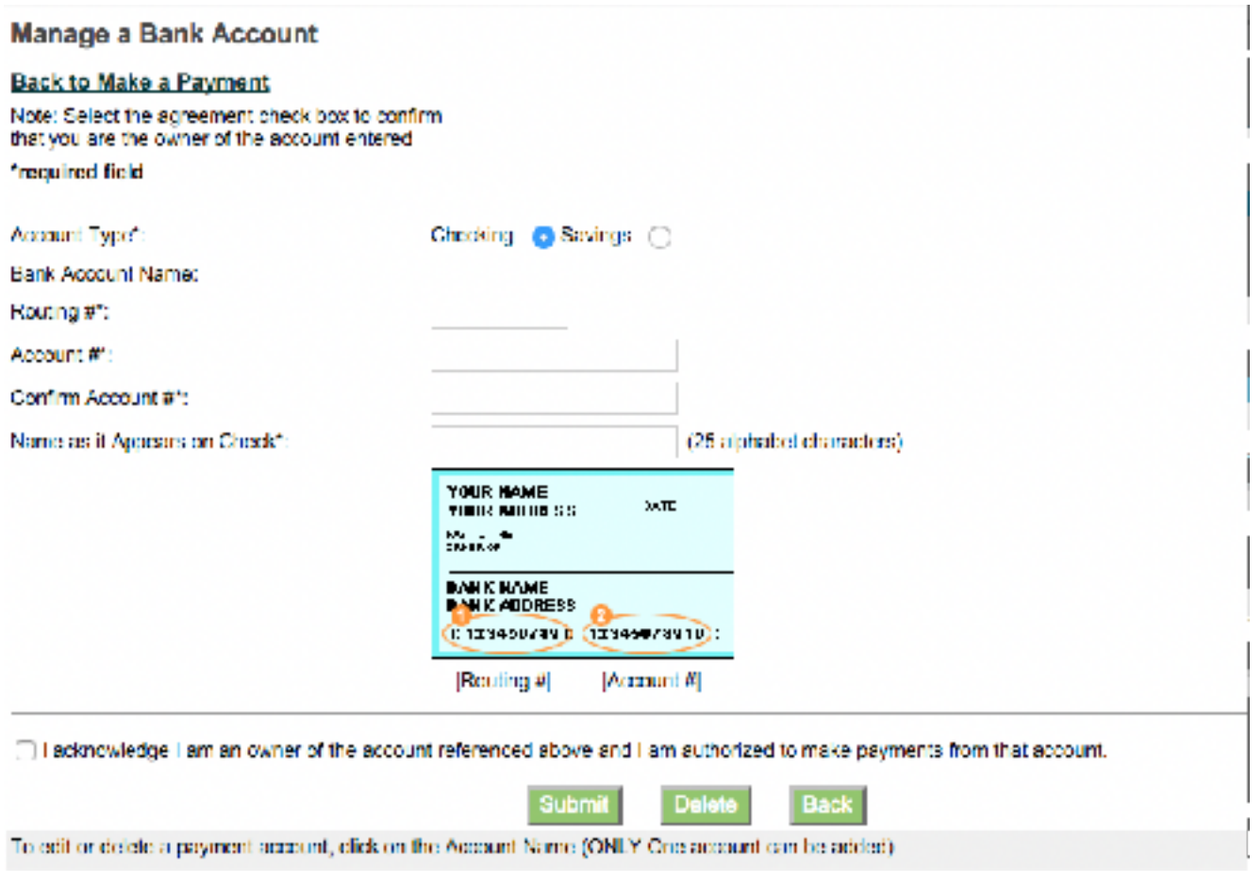


Error Type 2: Lapse

- Lapses are induced by inconsistencies or lack of good ease of recall
- Can be caused by retro and proactive interference in memories
- Common Example: “I forgot your name”
- Design Example: NIST’s need to “chop” a form

Error Type 3: Mistake

- Mistakes are generated by a lack of understanding or a lack of sufficient or correct information
- Lack of sufficient or correct information is the responsibility of the designer in the presentation layer of an interface
- Mistake are often undetectable by the end user



Expert Review

- An expert review, is a review by an expert (duh) – experts in interaction design and/or human factors
- An expert review includes a style guide review and a heuristic evaluation since an expert should have their own set of heuristics or may have adopted one of a set of published heuristics
- But an expert review also includes
  - A review of the design against industry standards and best practices
  - Experience from prior evaluations
- The domain and user population knowledge is the limiting factor for a good expert review

Qualitative User-Based Testing Protocols 1: The Think Aloud

Think Aloud Protocol

A direct observation method of user testing that involves asking users to think out loud as they are performing a task. Users are asked to say whatever they are looking at, thinking, doing, and feeling at each moment. *This method is especially helpful for determining users' expectations and identifying what aspects of a system are confusing. [italics added]*

- Usability Book of Knowledge

Issues

- Limit of introspection leading to confabulation
- Split attention effecting performance
- Increased anxiety (based on a shift in focus from product evaluation to participant evaluation)
- Incorrect focus on user's understanding of the design and not their ability to perform the task

Proper Application

- Think aloud protocol studies are appropriate when you have a non interactive product or concept to evaluate
  - Storyboard Evaluations
  - Static Mockup Evaluations
- Since it expects a participant's accurate introspection, results are unlikely to be inaccurate

Qualitative User-Based Testing Protocols 2: Interrupted Task-based Protocol

Interrupted Task-based Testing

- The goal of design is allow allow the user to complete a task with minimal or no attention to the means required to perform the task. We need to detect when htis rule is violated.
- Participants are asked to perform a task
  - If, during task performance, a participant is able to accomplish the task without distraction by the means of accomplishing the task, there is no reason to interrupt them (or distract them with a secondary task)
- If, during task performance, a participant shows some sign of a potential usability issue, the participant is interrupted to explore the potential issue

Interrupted Task-based Testing

- Allows for semi-accurate performance data collection (within the limits of the evaluation environment) or the exploration of potential usability issues
- Highly dependent on the skill set of the facilitator for both accurate administration of the session and evaluation/generalization of the results

Proper Application

- Interrupted task-based protocols are appropriate when you have an interactive product to evaluate and you want to learn about why they work or don't work
  - Partial Interactive Mockup
  - Full Interactive Mockup
  - Prototype
  - Fielded Product

Satisfaction Data

Satisfaction Data

- Satisfaction data does not correlate with performance except in extreme situations
- Satisfaction data can be operationalized in a number of ways, but is always opinion data
  - Standardized survey instrument (e.g. SUS, QUIS, SUMI)
  - Simple Likert scale assessments
- Questionnaires suffer from numerous issues that threaten their validity
  - Halo effect, leniency bias, strictness bias, central tendency bias
  - As a result, the data is not normally distributed.
  - In addition, the data is ordinal or on an interval or ratio scale
- Proper analysis of satisfaction data requires non-parametric statistics

Satisfaction Survey Data

- Take a poll on participants comparing the new product against the old version of the product. People might be asked to comment on the statement ‘The new design is an improvement over the old design.’ and given a choice of answers from “Definitely, it’s the tops.” to “No definitely not, it’s awful.” The data would be a collection of opinions. Assume the following scale and results…

Person Number	*Definitely an improvement, it’s the tops.*	*It’s a good improvement*	*it’s OK*	*I have no opinion*	*Not much of an improvement.*	*I don’t think its an improvement*	*No, definitely not an improvement, it’s awful*
1					X		
2		X					
3	X						
4				X			
5					X		
6				X			
7	X						
8						X	
9			X				
10		X					
11			X				
12	X						

Satisfaction Survey Data

- Take a poll on participants comparing the new product against the old version of the product. People are asked to comment on the statement 'The new design is an improvement over the old design.' and given a choice of answers from "Definitely, it's the tops." to "No definitely not, it's awful." Assume the following results…

Person Number	*Definitely an improvement, it's the tops.*	*It's a good improvement*	*it's OK*	*I have no opinion*	*Not much of an improvement.*	*I don't think its an improvement*	*No, definitely not an improvement, it's awful*
1					X		
2		X					
3	X						
4				X			
5					X		
6				X			
7	X						
8						X	
9			X				
10		X					
11			X				
12	X						

Wilcoxin Rank Sum Analysis - Single Population

- The hypothesis you would want to test would be: “The participants consider the new product an improvement.”
- A quick ‘eyeball’ test shows that none of those questioned thought it was awful and only one person thought it not very good, so a first impression is that people generally approve. If you start by assuming that in the population there is no opinion one way or the other, and that people’s responses are symmetrically distributed about ‘no opinion’, you can test the hypothesis that people think the new design is an improvement, with the null hypothesis that people have no opinion about it. The median value is 4.

Wilcoxin Rank Sum Analysis - Single Population

- You need to be careful to choose the appropriate test statistic for the problem you are tackling
  - For a one tailed test, where the alternative hypothesis is that the median is greater than a given value, the test statistic is W- . For a one tailed test, where the alternative hypothesis is that the median is less than a given value, the test statistic is W+ .
  - For a two tailed test the test statistic is the smaller of W+ and W
- As people who think it an improvement will give a rating of less than 4, the null and alternative hypotheses can be stated as follows.
  - H0 : the median response is 4
  - H1 : the median response is less than 4
  - 1 tail test, Significance level 5%

Wilcoxin Rank Sum Analysis - Single Population

- List the value
- Find the difference between each value and the median.
- Ignore the zeros and rank the absolute values of the remaining scores.
- Ignore the signs, start with the smallest difference and give it rank 1. Where two or more differences have the same value find their mean rank, and use this.
- Now check that W+ + W- are the same as ½ n(n+1), where n is the number in the sample (having ignored the zeros). In this case n = 10.
  - ½ n(n+1)= ½ x 10 x 11 = 55
  - W+ + W- = 9·5 + 45·5 = 55

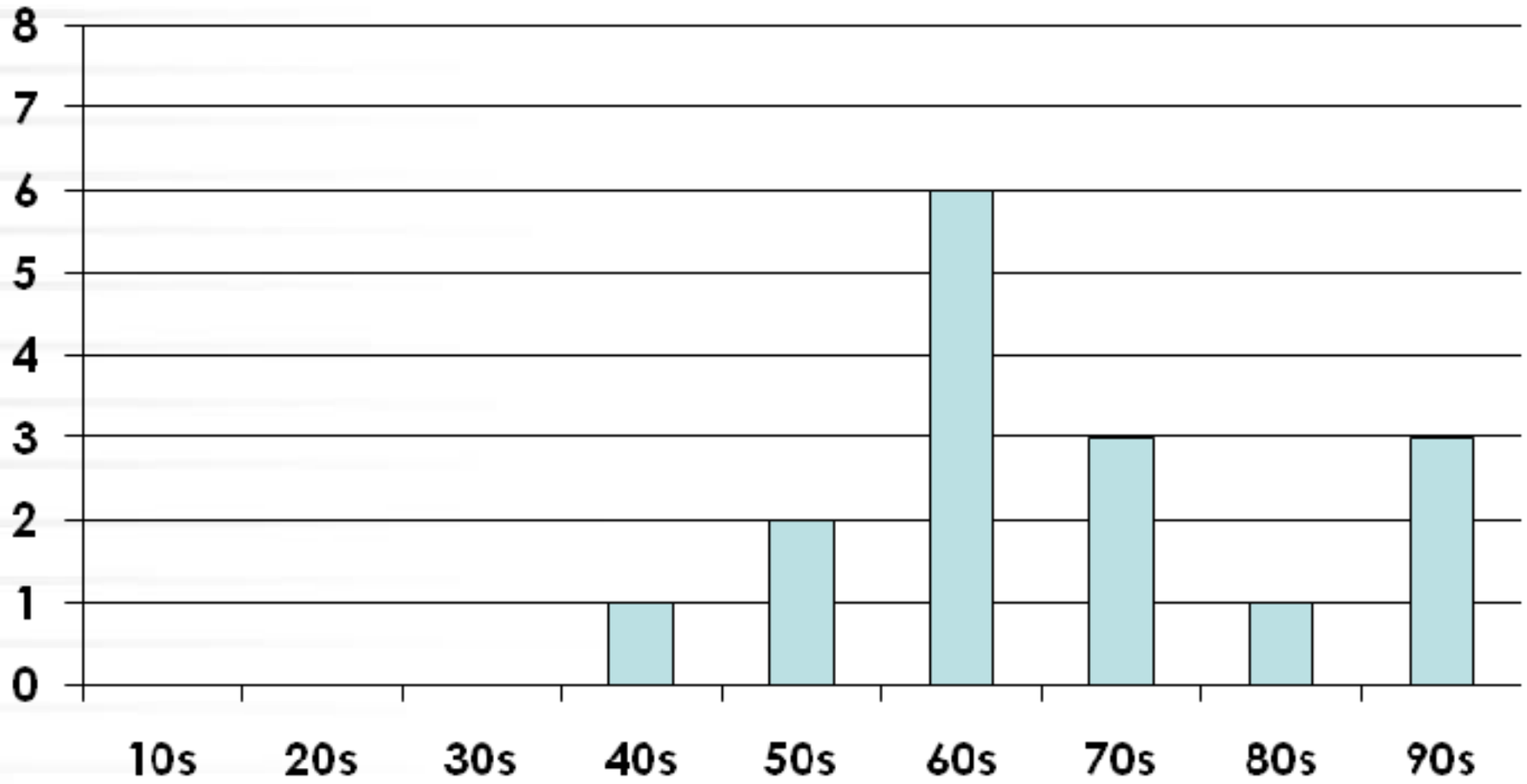
rating	rating – median (4)	absolute value	ranking	+	_
5	5 – 4 = 1	1	2	2	
2	2 – 4 = -2	2	5.5		5.5
1	1 – 4 = -3	3	9		9
4	4 – 4 = 0	0	Ignore		
5	5 – 4 = 1	1	2	2	
2	2 – 4 = -2	2	5.5		5.5
4	4 – 4 = 0	0	Ignore		
1	1 – 4 = -3	3	9		9
6	6 – 4 = 2	2	5.5	5.5	
3	3 – 4 = -1	1	2		2
2	2 – 4 = -2	2	5.5		5.5
1	1 – 4 = -3	3	9		9
			Total	9.6	45.5

Wilcoxin Rank Sum Analysis - Single Population

- Compare the test statistic with the critical value in the tables. If the null hypothesis were true, and the median is 4, you would expect W+ and W- to have roughly the same value. There are two possible test statistics here, W+ = 9·5 and W- = 45·5, and you have to decide which one to use. We are interested in the sum of the ranks of ratings greater than 4. W+ is much less than W- which suggests that more people felt the shopping center was an asset. It could also suggest that those who expressed a negative view expressed a very strong one, with lots of high numbers in the ratings.
- Now you need to compare the value of W+ , the test statistic, with the critical value from the table. Given that W+ is small, the key question becomes “Is W+ significantly smaller than would happen by chance?” The table helps you decide this by supplying the critical value. For a sample of 10, at the 5% significance level for a 1 tailed test, the value is 10. As W+ is 9·5, which is less than this, the evidence suggests that we can reject the null hypothesis.
- Your conclusion is that the evidence shows, at the 5% significance level, that the public thinks the design is better than the old design

1-tail	5%	2½%	1%	½%
2-tail	10%	5%	2%	1%
n				
2	-	-	-	-
3	-	-	-	-
4	-	-	-	-
5	0	-	-	-
6	2	0	-	-
7	3	2	0	-
8	5	3	1	0
9	8	5	3	1
10	10	8	5	3
11	13	10	7	5
12	17	13	9	7
13	21	17	12	9
14	25	21	15	12
15	30	25	19	15

Correlated User Ratings



SUS

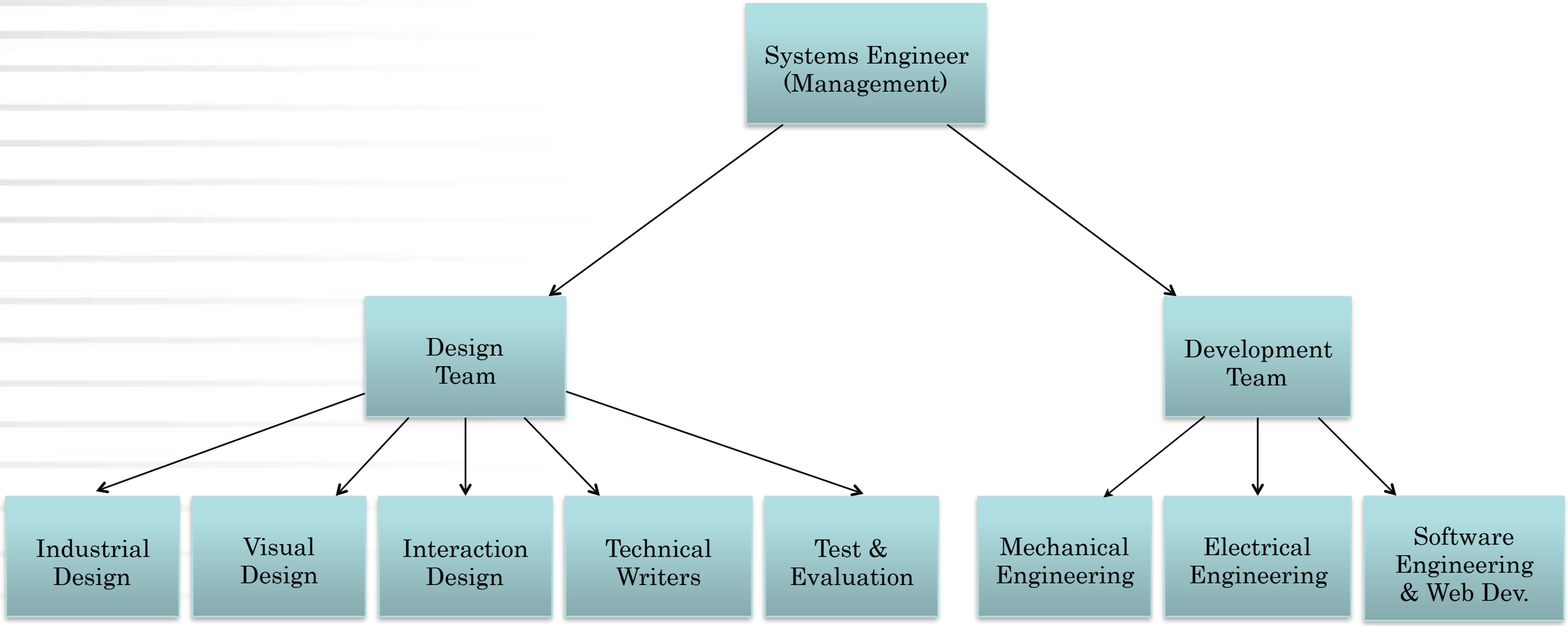
USER-CENTERED DESIGN

Usability, Organization, and Processes

Thought From CHI '92

- The 1970s, when Hardware is King
  - 1950s – its an art
  - 1960s – there are degrees
  - 1970s – they're in management
- The 1980s, when Software is King
  - 1960s – its an art
  - 1970s – there are degrees
  - 1980s – they're in management
- 1990s, when "Interaction" should be King
  - 1970s – its an art
  - 1980s – there are degrees (?)
  - 1990s – they should be in management

Product Design & Development



Processes

- System Development Models
  - Waterfall
  - Spiral
  - V-Model
- Software Development Models
  - Dynamic System Development Process (DSDP)
  - Joint Application Development Process (JAD) (circa 1970)
  - Structured Systems Analysis and Design Methodology (SSADM) (circa 1980)
  - Information Requirement Analysis/Soft System (circa 1980)
  - Object Oriented Programming (origins in 1960, but a common methodology in the 1990s)
  - Rapid Application Development (circa 1991)*
  - Agile*
    - Extreme Programming (circa 1990)
    - SCRUM

Processes (concluded)

- Interface Design Models
  - Star (Hartson & Hix, 1989)
  - LUCID (Cognetics, 2008)
  - ISO 13407/ISO 9241
  - Human Centered Design (IDEO)
  - User-Centered Design (the common term)
- Characteristics of a User-Centered Design Process
  - Design is a separate activity, distinct from development
  - Design should occur, completely, before development begins
  - Feedback is needed at many steps in the design process to…
    - Confirm the direction of design
    - Evaluate alternatives