TUTORIAL 2019
- EVALUATION OF ARTIFACTS IN DESIGN SCIENCE - PART 1

GONDY LEROY, PH.D.
MANAGEMENT INFORMATION SYSTEMS
UNIVERSITY OF ARIZONA
GONDYLEROY@EMAIL.ARIZONA.EDU
PRESENTER BACKGROUND

Education

• B.S. and M.S. in Cognitive, Experimental Psychology, University of Leuven, Leuven, Belgium
• M.S. and Ph.D. in Management Information Systems, University of Arizona, Tucson, AZ

Relevant Experience

• Principal Investigator $2.3M of funded research, by NIH, NSF, AHRQ, Microsoft Research

Current Position and Contact Information

• Gondy Leroy, PhD
• Professor, Management Information Systems
• Director - Tomorrow’s Leaders Equipped for Diversity
• Eller College of Management, University of Arizona
• http://nlp.lab.arizona.edu/
ADDITIONAL RESOURCES

Content based on

*Designing User Studies in Informatics*

Gondy Leroy, Ph.D.

*Springer, 2011*

- Freely available in most academic libraries (chapters can be downloaded electronically)
- Book information
  - bridges gap between informatics, design science, and behavioral sciences.
  - explains what an experimenter should pay attention to and why
  - Practical, to-the-point, hands-on
  - Contains a ‘cookbook’ with step-by-step instructions for different types of artifact evaluations
DESIGN SCIENCE CONTEXT
DESIGN SCIENCE

Two paradigms in IS: behavioral and design science (Hevner et al. 2004)

- Behavioral paradigm
  - Develop theories, test theories
  - Theories to understand, predict organizational and human phenomena relevant to information systems

- Design science
  - Roots in engineering and AI (Simon 1996)
  - Problem solving paradigm: create artifact to solve a problem

Evaluation

- Can use many approaches to evaluate artifacts
  - Simulation, mathematical description, impact studies
  - Focus on correctness, effectiveness, efficiency …

- Studies to evaluate impact
  - Work with users in some cases
  - Work with experts in some cases
  - Work with gold standards in some cases
DEVELOPMENT LIFE CYCLE

Cyclical nature of development in informatics

- Digital libraries, online communities, mobile apps, ...
- Education, business, medical informatics, biology, commerce, ...

Good fit of cycle (regardless of which one is adopted) with different types of user studies

Example:

Importance of testing EARLY and FREQUENTLY:
Requirements Analysis - most errors originate here, 60% undiscovered until user acceptance testing (Gartner Group, 2009)

Remodeling, renewing, improving: test new features against previous version

System evaluation: test additional features, test synergy of all components together

Paper prototyping of interfaces: easy to accept changes

Algorithm development and evaluation:
1) possibility to pinpoint location of strengths and weaknesses
2) Possibility of batch process evaluations
FOCUSBNG THE STUDY

Increase your chances of getting published and funded by focusing the study according to these three principles:

- Define the goal of the system
  - Will help choose comparison points (=define independent variable) and pinpoint measures (=dependent variables)
- Keep stakeholders in mind
  - Why is the study conducted (to improve?), who is interested in it (what are they interested in?)
  - Will help relate to people evaluating your artifact and study
- Timeline and Development Cycle
  - What is available for testing, what can be tested in design phase?
  - Will help design appropriate (series of) studies

Example: reducing the number of no-shows at a clinic

- Goal: have more patients show up for appointment, don’t increase workload of staff
- Stakeholders: patients (will need to be easy, no effort), staff at clinic (low training, easy to manage), purchaser at clinic (demonstrated effect on no-shows)
- Timeline and Development Cycle: paper prototyping for interface, comparison after implementation with existing system
EXISTING STUDY TYPES
Naturalistic observation

- Individuals in their natural setting, no intrusion, ideally, people not aware
- Passive form of research, observation:
  - Observe in person
  - Use technology to observe (tracking, video, alerts, …)

Case Studies, Field Studies and Descriptive Studies

- Several types of each exist
- Help explain and answer difficult questions, e.g., “Why was the system not accepted?”
- Can consider characteristics of work environment, culture, lifestyle and personal preferences when searching for explanations systematically controlled in experiments
- can be combined with action research

Action Research

- Case studies + direct involvement of the researcher
- Goal is to solve a problem or improve an existing situation
- Less role of observer but iterative and error-correcting approach
DIFFERENT STUDY TYPES (2/3)

Surveys

- Surveys are useful to measure opinions, intentions, feelings and beliefs, ...
- Dangers:
  1) often hastily constructed, misconception about easiness of constructing a survey and validity,
  2) Not a measure of behaviors or actions
  3) Few people in IS properly trained to design surveys

Correlation Studies

- About changes in variables, to find where change in one variable coincides with change in another.
- Involve many variables and many data points (surveys and large population samples)
- Do not attempt to discover what causes a change
DIFFERENT STUDY TYPES (3/3)

Quasi-experiment

• No randomization (main difference with experiments)
• Often useful when groups pre-exist because
  • Geographical constraints, e.g., population across a country, in different cities
  • social constraints, e.g., siblings in families
  • time constraints, e.g., comparable group already studied in past

Experiments (focus of tutorial)

• Goal is to evaluate hypotheses about causal relations between variables
• In informatics:
  • evaluate the impact, benefit, advantages, disadvantages or other effects of information systems, algorithms, or interfaces, ...
  • New/improved system is compared to other systems or under different conditions and evaluated for its impact
TWO DIFFERENT KINDS OF EXPERIMENTS RELATED TO STAGES OF DEVELOPMENT

Early Stages: focusing on algorithm/system development

- Indirect involvement of users
- Batch-process approach to experiments
- Example Advantages: large scale, efficient, quick turn around
- Example Dangers: confusing development and evaluation, forgetting design considerations (randomization, double blind evaluation, …)

Later Stages: focusing on human-computer interaction, longitudinal evaluations of impact

- Direct involvement of users
- Example Advantages: rich data, useful information for product improvement
- Example Dangers: users not representative, IRB slows down process
EXPERIMENTAL DESIGN IN A NUTSHELL

STEPS TO TAKE
THREE STEPS TO DESIGN A STUDY

STEP 1: What is goal
  • The answer helps define *independent variables*
  • Independent variables: what will be manipulated

STEP 2: How will we know that goal is reached
  • The answer helps define *dependent variables*
  • Dependent variables: what will be measured

STEP 3: What can affect the system, users, use, actions, opinions, ...
  • The answer helps define *confounded and nuisance variables*
  • Confounded variables: 2 variables that change from one treatment to another
  • Nuisance variables: variables that add errors and variance but are of no interest to the researcher (they should be controlled)
STEP 1: CHOOSE THE INDEPENDENT VARIABLES

Define the goal of the study

- Examples:
  - Evaluate a new system where there was no system before?
  - Evaluate whether a system is better than another system
  - Evaluate whether different types of users get better use with system

Independent Variable (IV)

- Other terms: treatment, intervention
- Manipulated by researcher
- Goal of a user study is to compare the results for different treatments

Types of Independent Variable (IV)

- Qualitative Independent Variables
  - Describe different kinds of treatments
- Quantitative Independent Variables
  - Describes different amounts of a given treatment.
STEP 1: CHOOSE THE INDEPENDENT VARIABLES

There can be multiple independent variables

- In informatics, often only 1 or 2.
- Seldom more.

Critical to make this a TRUE experiment

- Assignment to a condition/treatment has to be done randomly
Abstract—As more patients use the Internet to answer health-related queries, simplifying medical information is becoming increasingly important. To simplify medical terms when synonyms are unavailable, we must add multi-word explanations. Following a data-driven approach, we conducted two user studies to determine the best formulation for adding explanatory content as parenthetical expressions. Study 1 focused on text with a single difficult term (N=260). We examined the effects of different types of text, types of content in parentheses, difficulty of the explanatory content, and position of the term in the sentence on actual difficulty, perceived difficulty, and reading time. We found significant support that enclosing the difficult term in parentheses is best for difficult text and enclosing the explanation in parentheses is best for simple text. Study 2 (N=116) focused on lists with multiple difficult terms. The same interaction is present although statistically insignificant, but parenthetical insertion can still significantly simplify text.
Y. Gu, G. Leroy, D. Kauchak, "When synonyms are not enough: Optimal parenthetical insertion for text simplification," Accepted for the AMIA Fall Symposium, November 2017, Washington DC.

Abstract—As more patients use the Internet to answer health-related queries, simplifying medical information is becoming increasingly important. To simplify medical terms when synonyms are unavailable, we must add multi-word explanations. Following a data-driven approach, we conducted two user studies to determine the best formulation for adding explanatory content as parenthetical expressions. Study 1 focused on text with a single difficult term (N=260). We examined the effects of different types of text, types of content in parentheses, difficulty of the explanatory content, and position of the term in the sentence on actual difficulty, perceived difficulty, and reading time. We found significant support that enclosing the difficult term in parentheses is best for difficult text and enclosing the explanation in parentheses is best for simple text. Study 2 (N=116) focused on lists with multiple difficult terms. The same interaction is present although statistically insignificant, but parenthetical insertion can still significantly simplify text.
STEP 1: FIND THE INDEPENDENT VARIABLES
EXERCISE


Abstract—Today’s search engines provide a single textbox for searching. This input method has not changed in decades and, as a result, consumer search behaviour has not changed either: few and imprecise keywords are used. Especially with health information, where incorrect information may lead to unwise decisions, it would be beneficial if consumers could search more precisely. We evaluated a new user interface that supports more precise searching by using query diagrams. In a controlled user study, using paper based prototypes, we compared searching with a Google interface with drawing new or modifying template diagrams. We evaluated consumer willingness and ability to use diagrams and the impact on query formulation. Users had no trouble understanding the new search method. Moreover, they used more keywords and relationships between keywords with search diagrams. In comparison to drawing their own diagrams, modifying existing templates led to more searches being conducted and higher creativity in searching.
STEP 1: FIND THE INDEPENDENT VARIABLES
EXERCISE


Abstract—Today’s search engines provide a single textbox for searching. This input method has not changed in decades and, as a result, consumer search behaviour has not changed either: few and imprecise keywords are used. Especially with health information, where incorrect information may lead to unwise decisions, it would be beneficial if consumers could search more precisely. We evaluated a new user interface that supports more precise searching by using query diagrams. In a controlled user study, using paper based prototypes, we compared searching with a Google interface with drawing new or modifying template diagrams. We evaluated consumer willingness and ability to use diagrams and the impact on query formulation. Users had no trouble understanding the new search method. Moreover, they used more keywords and relationships between keywords with search diagrams. In comparison to drawing their own diagrams, modifying existing templates led to more searches being conducted and higher creativity in searching.
Abstract—To solve crimes, investigators often rely on interviews with witnesses, victims, or criminals themselves. The interviews are transcribed and the pertinent data is contained in narrative form. To solve one crime, investigators may need to interview multiple people and then analyze the narrative reports. There are several difficulties with this process: interviewing people is time consuming, the interviews — sometimes conducted by multiple officers — need to be combined, and the resulting information may still be incomplete. For example, victims or witnesses are often too scared or embarrassed to report or prefer to remain anonymous. We are developing an online reporting system that combines natural language processing with insights from the cognitive interview approach to obtain more information from witnesses and victims. We report here on information extraction from police and witness narratives. We achieved high precision, 94% and 96%, and recall, 85% and 90%, for both narrative types.
Abstract—To solve crimes, investigators often rely on interviews with witnesses, victims, or criminals themselves. The interviews are transcribed and the pertinent data is contained in narrative form. To solve one crime, investigators may need to interview multiple people and then analyze the narrative reports. There are several difficulties with this process: interviewing people is time consuming, the interviews – sometimes conducted by multiple officers – need to be combined, and the resulting information may still be incomplete. For example, victims or witnesses are often too scared or embarrassed to report or prefer to remain anonymous. We are developing an online reporting system that combines natural language processing with insights from the cognitive interview approach to obtain more information from witnesses and victims. We report here on information extraction from police and witness narratives. We achieved high precision, 94% and 96%, and recall, 85% and 90%, for both narrative types.
Abstract—Research has shown that most users’ online information searches are suboptimal. Query optimization based on a relevance feedback or genetic algorithm using dynamic query contexts can help casual users search the Internet. These algorithms can draw on implicit user feedback based on the surrounding links and text in a search engine result set to expand user queries with a variable number of keywords in two manners. Positive expansion adds terms to a user’s keywords with a Boolean “and,” negative expansion adds terms to the user’s keywords with a Boolean “not.” Each algorithm was examined for three user groups, high, middle, and low achievers, who were classified according to their overall performance. The interactions of users with different levels of expertise with different expansion types or algorithms were evaluated. The genetic algorithm with negative expansion tripled recall and doubled precision for low achievers, but high achievers displayed an opposed trend and seemed to be hindered in this condition. The effect of other conditions was less substantial.
Abstract—Research has shown that most users’ online information searches are suboptimal. Query optimization based on a relevance feedback or genetic algorithm using dynamic query contexts can help casual users search the Internet. These algorithms can draw on implicit user feedback based on the surrounding links and text in a search engine result set to expand user queries with a variable number of keywords in two manners. Positive expansion adds terms to a user’s keywords with a Boolean “and,” negative expansion adds terms to the user’s keywords with a Boolean “not.” Each algorithm was examined for three user groups, high, middle, and low achievers, who were classified according to their overall performance. The interactions of users with different levels of expertise with different expansion types or algorithms were evaluated. The genetic algorithm with negative expansion tripled recall and doubled precision for low achievers, but high achievers displayed an opposed trend and seemed to be hindered in this condition. The effect of other conditions was less substantial.
STEP 2: CHOOSE THE DEPENDENT VARIABLES

What needs to be measured to know if goal was reached

Examples:

• Were enough relevant articles found? Did people like using the system (or did they get frustrated)?

Dependent Variable (DV)

• Other terms: *outcome* or *response variable*
• Metrics chosen by researcher
• Goal is to use metrics to compare different treatments, preferably use complementary measures to assess the impact of a treatment

Choose a relevant dependent variable

• Keep the stakeholders in mind, what do they care about?
• Development phase affects choice
  • Early on: do usability in formative phases of development
  • Then: relevance, completeness of results, …
  • Later: cost savings, risk taking, improved decision making, …
• Historically used metrics are good to include
  • Because they are probably already well understood
  • Because it will be expected by stakeholders

• Good starting point: effectiveness, efficiency and satisfaction
  (aka: outcome measures, performance measures, satisfaction measures)
STEP 2: CHOOSE THE DEPENDENT VARIABLES

Commonly used metrics for outcome measures

- Precision, recall, F-measure
- Accuracy, true positive, true negative, false positive, false negative, specificity, sensitivity
- Counts

Commonly used metrics for performance measures

- Time (to completion)
- Errors
- Usability when measured in an objective manner: counting events, errors or measuring task completion.
- Interesting measures compare different users on their training or task completion times.
  - E.g., comparison between novice and expert users.

Satisfaction and Acceptance

- Usually measured with one survey.
- NOTE: users need to be satisfied in short term before they will accept in the long term

Abstract - Text simplification often relies on dated, unproven readability formulas. As an alternative and motivated by the success of term familiarity, we test a complementary measure: grammar familiarity. Grammar familiarity is measured as the frequency of the 3rd level sentence parse tree and is useful for evaluating individual sentences. We created a database of 140K unique 3rd level parse structures by parsing and binning all 5.4M sentences in English Wikipedia. We then calculated the grammar frequencies across the corpus and created 11 frequency bins. We evaluate the measure with a user study and corpus analysis. For the user study, we selected 20 sentences randomly from each bin, controlling for sentence length and term frequency, and recruited 30 readers per sentence (N=6,600) on Amazon Mechanical Turk. We measured actual difficulty (comprehension) using a Cloze test, perceived difficulty using a 5-point Likert scale, and time taken. Sentences with more frequent grammatical structures, even with very different surface presentations, were easier to understand, perceived as easier and took less time to read. Outcomes from readability formulas correlated with perceived but not with actual difficulty. Our corpus analysis shows how the metric can be used to understand grammar regularity in a broad range of corpora.
Abstract - Text simplification often relies on dated, unproven readability formulas. As an alternative and motivated by the success of term familiarity, we test a complementary measure: grammar familiarity. Grammar familiarity is measured as the frequency of the 3rd level sentence parse tree and is useful for evaluating individual sentences. We created a database of 140K unique 3rd level parse structures by parsing and binning all 5.4M sentences in English Wikipedia. We then calculated the grammar frequencies across the corpus and created 11 frequency bins. We evaluate the measure with a user study and corpus analysis. For the user study, we selected 20 sentences randomly from each bin, controlling for sentence length and term frequency, and recruited 30 readers per sentence (N=6,600) on Amazon Mechanical Turk. We measured actual difficulty (comprehension) using a Cloze test, perceived difficulty using a 5-point Likert scale, and time taken. Sentences with more frequent grammatical structures, even with very different surface presentations, were easier to understand, perceived as easier and took less time to read. Outcomes from readability formulas correlated with perceived but not with actual difficulty. Our corpus analysis shows how the metric can be used to understand grammar regularity in a broad range of corpora.
STEP 1: FIND THE DEPENDENT VARIABLES
EXERCISE


Abstract - Current approaches to word sense disambiguation use and combine various machine-learning techniques. Most refer to characteristics of the ambiguous word and surrounding words and are based on hundreds of examples. Unfortunately, developing large training sets is time-consuming. We investigate the use of symbolic knowledge to augment machine learning techniques for small datasets. UMLS semantic types assigned to concepts found in the sentence and relationships between these semantic types form the knowledge base. A naïve Bayes classifier was trained for 15 words with 100 examples for each. The most frequent sense of a word served as the baseline. The effect of increasingly accurate symbolic knowledge was evaluated in eight experimental conditions. Performance was measured by accuracy based on 10-fold cross-validation. The best condition used only the semantic types of the words in the sentence. Accuracy was then on average 10% higher than the baseline; however, it varied from 8% deterioration to 29% improvement. In a follow-up evaluation, we noted a trend that the best disambiguation was found for words that were the least troublesome to the human evaluators.
STEP 1: FIND THE DEPENDENT VARIABLES

EXERCISE


Abstract - Current approaches to word sense disambiguation use and combine various machine-learning techniques. Most refer to characteristics of the ambiguous word and surrounding words and are based on hundreds of examples. Unfortunately, developing large training sets is time-consuming. We investigate the use of symbolic knowledge to augment machine learning techniques for small datasets. UMLS semantic types assigned to concepts found in the sentence and relationships between these semantic types form the knowledge base. A naïve Bayes classifier was trained for 15 words with 100 examples for each. The most frequent sense of a word served as the baseline. The effect of increasingly accurate symbolic knowledge was evaluated in eight experimental conditions. Performance was measured by accuracy based on 10-fold cross-validation. The best condition used only the semantic types of the words in the sentence. Accuracy was then on average 10% higher than the baseline; however, it varied from 8% deterioration to 29% improvement. In a follow-up evaluation, we noted a trend that the best disambiguation was found for words that were the least troublesome to the human evaluators.

Abstract - Purpose: Willingness and ability to learn from health information in text are crucial for people to be informed and make better medical decisions. These two user characteristics are influenced by the perceived and actual difficulty of text. Our goal is to find text features that are indicative of perceived and actual difficulty so that barriers to reading can be lowered and understanding of information increased.

Methods: We systematically manipulated three text characteristics, – overall sentence structure (active, passive, extraposed-subject, or sentential-subject), noun phrases complexity (simple or complex), and function word density (high or low), – which are more fine-grained metrics to evaluate text than the commonly used readability formulas. We measured perceived difficulty with individual sentences by asking consumers to choose the easiest and most difficult version of a sentence. We measured actual difficulty with entire paragraphs by posing multiple-choice questions to measure understanding and retention of information in easy and difficult versions of the paragraphs.

Results: Based on a study with 86 participants, we found that low noun phrase complexity and high function words density lead to sentences being perceived as simpler. In the sentences with passive, sentential-subject, or extraposed-subject sentences, both main and interaction effects were significant (all p < .05). In active sentences, only noun phrase complexity mattered (p < .001). For the same group of participants, simplification of entire paragraphs based on these three linguistic features had only a small effect on understanding (p = .99) and no effect on retention of information.

Conclusions: Using grammatical text features, we could measure and improve the perceived difficulty of text. In contrast to expectations based on readability formulas, these grammatical manipulations had limited effects on actual difficulty and so were insufficient to simplify the text and improve understanding. Future work will include semantic measures and overall text composition and their effects on perceived and actual difficulty.
STEP 1: FIND THE DEPENDENT VARIABLES
EXERCISE


Abstract - Purpose: Willingness and ability to learn from health information in text are crucial for people to be informed and make better medical decisions. These two user characteristics are influenced by the perceived and actual difficulty of text. Our goal is to find text features that are indicative of perceived and actual difficulty so that barriers to reading can be lowered and understanding of information increased.

Methods: We systematically manipulated three text characteristics, – overall sentence structure (active, passive, extraposed-subject, or sentential-subject), noun phrases complexity (simple or complex), and function word density (high or low), – which are more fine-grained metrics to evaluate text than the commonly used readability formulas. We measured perceived difficulty with individual sentences by asking consumers to choose the easiest and most difficult version of a sentence. We measured actual difficulty with entire paragraphs by posing multiple-choice questions to measure understanding and retention of information in easy and difficult versions of the paragraphs.

Results: Based on a study with 86 participants, we found that low noun phrase complexity and high function words density lead to sentences being perceived as simpler. In the sentences with passive, sentential-subject, or extraposed-subject sentences, both main and interaction effects were significant (all \( p < .05 \)). In active sentences, only noun phrase complexity mattered (\( p < .001 \)). For the same group of participants, simplification of entire paragraphs based on these three linguistic features had only a small effect on understanding (\( p = .99 \)) and no effect on retention of information.

Conclusions: Using grammatical text features, we could measure and improve the perceived difficulty of text. In contrast to expectations based on readability formulas, these grammatical manipulations had limited effects on actual difficulty and so were insufficient to simplify the text and improve understanding. Future work will include semantic measures and overall text composition and their effects on perceived and actual difficulty.
CONFOUNDED VARIABLES

Confounding = when the effects of 2 (or more) variables cannot be separated from each other

• A variable, other than the independent variable, may have caused the effect
• Reduces the internal validity: unsure whether the independent variable caused the effect

Random assignment to experimental condition is essential in avoiding confounding (but not always sufficient)

Example: An experiment with 2 conditions to test a DSS for managers: old vs. new system

Old system – tested with experienced managers (25+ yrs)
New system – tested with inexperienced managers (1 yr)
NUISANCE VARIABLES

Nuisance variables add variation to the study outcome

- not due to the independent variables
- of no interest to the experimenter.
- reduce the chance of detecting the systematic impact of the independent variable

Noise = if the variation is unsystematic

- E.g., conduct experiment at different times of day. At some times, the environment was noisy (train passes by, class changes and noise in hallway, people partying) \( \rightarrow \) affects performance of subjects

Bias = if the variation is systematic

- E.g., conduct experiment and for each level of IV have different graduate student conduct study. One student is constantly chatting with friends during the experiment \( \rightarrow \) affects performance of subjects (true story)
END OF PART 1