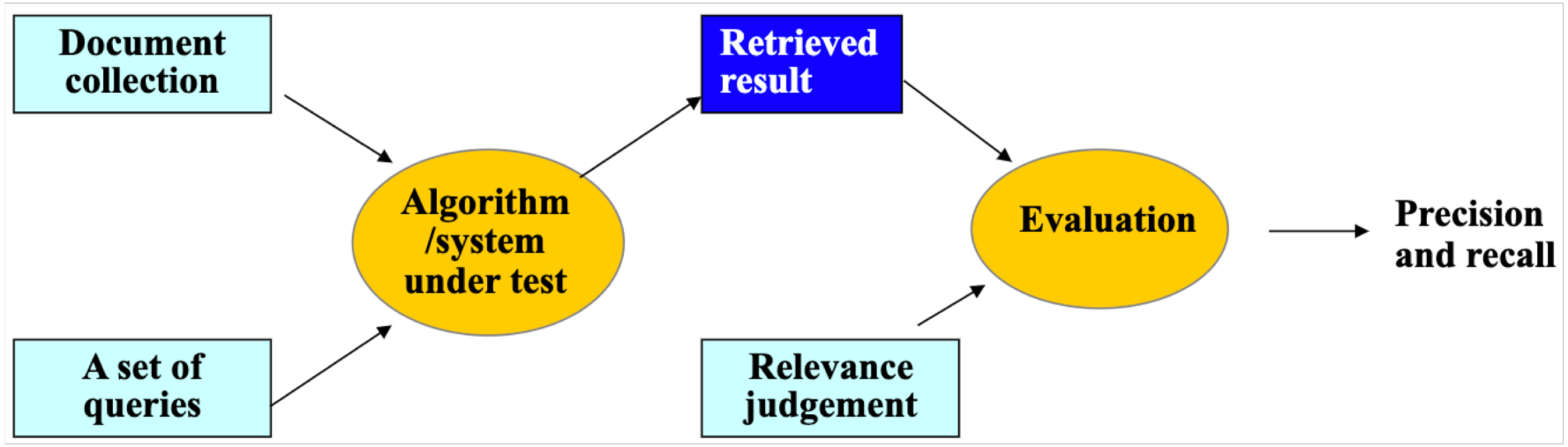
Evaluating Interactive Information Retrieval

Dorota Glowacka

glowacka@cs.helsinki.fi

Traditional IR Evaluation

- **Test collection**: a collection of documents, a set of queries, the relevance judgement
- **Process**: input the documents, put each query to the system, collect the output
- Measurement: usually precision and recall



Problems with system-oriented experiments

- Pros:
  - Advanced the system development

- Cons:
  - System is an input-output device, while most real searches involve interaction.
  - Relevance is binary and judged independently of context, while relevance is:
    - ***Subjective***: Depends upon a specific user's judgment.
    - ***Situational***: Relates to user's current needs.
    - ***Cognitive***: Depends on human perception and behavior.
    - ***Dynamic***: Changes over time.

The TREC Benchmark

- Text Retrieval Conference - organized by NIST, started in 1992

- Purposes:
  - To encourage research in IR based on large text collections.
  - To provide a common ground/task evaluation that allows cross-site comparison.
  - To develop new evaluation techniques, particularly for new applications, e.g. filtering, cross-language retrieval, web retrieval, high precision, question answering

TREC Interactive Track

- Goal: to investigate searching as an interactive task by examining the process as well as the outcome.
- Interactive track tasks:
  - TREC3-4: finding relevant documents
  - TREC5-9: finding any *N* short answers to a question, to which there are multiple answers of the same type.
  - TREC10-11: finding any *N* short answers to a question and finding any N websites that meet the need specified in the task statement
  - TREC12: topic distillation

Experimental Step-by-Step

1. Establish the aims of the evaluation, the intended users and context of use for the system; obtain or construct scenarios illustrating how the application will be used.
2. Select evaluation methods – should be a combination of expert review and end-user testing
3. Carry out expert review
4. Plan user testing; use the results of the expert review to help focus this
5. Recruit users and organise testing venue and equipment
6. Carry out user testing
7. Analyse results, write up and report back to designers

How to measure outcome?

- Aspectual precision:
- The proportion of the documents identified by a subject that were deemed to contain topic aspects.

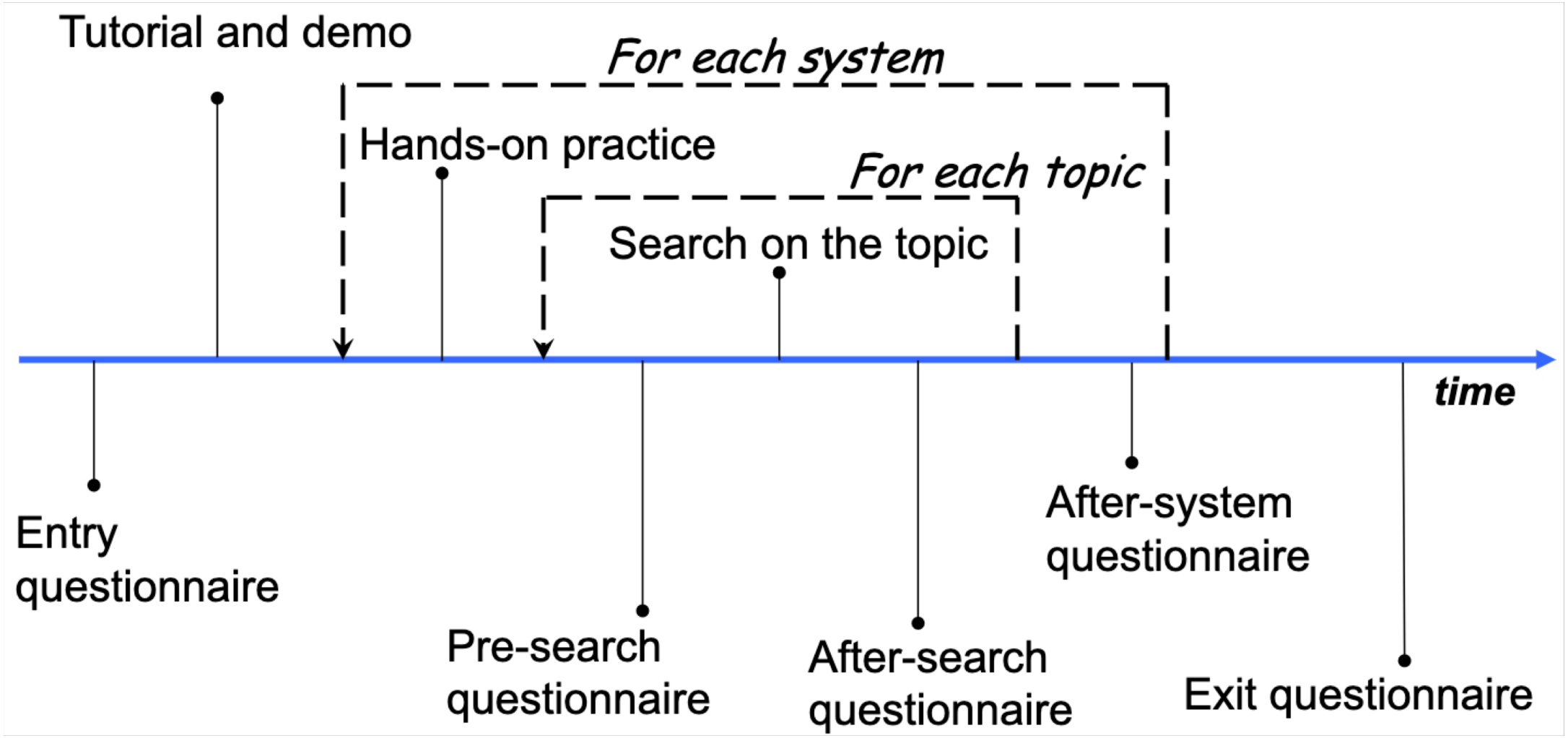
- Aspectual recall:
- The proportion of the known topic aspects contained in the documents identified by a subject.

How to measure process?

- Objective measures:
  - Number of query iterations
  - Number of document surrogates seen
  - Number of documents read
  - Number of documents saved
  - Actual time used with the system/searching

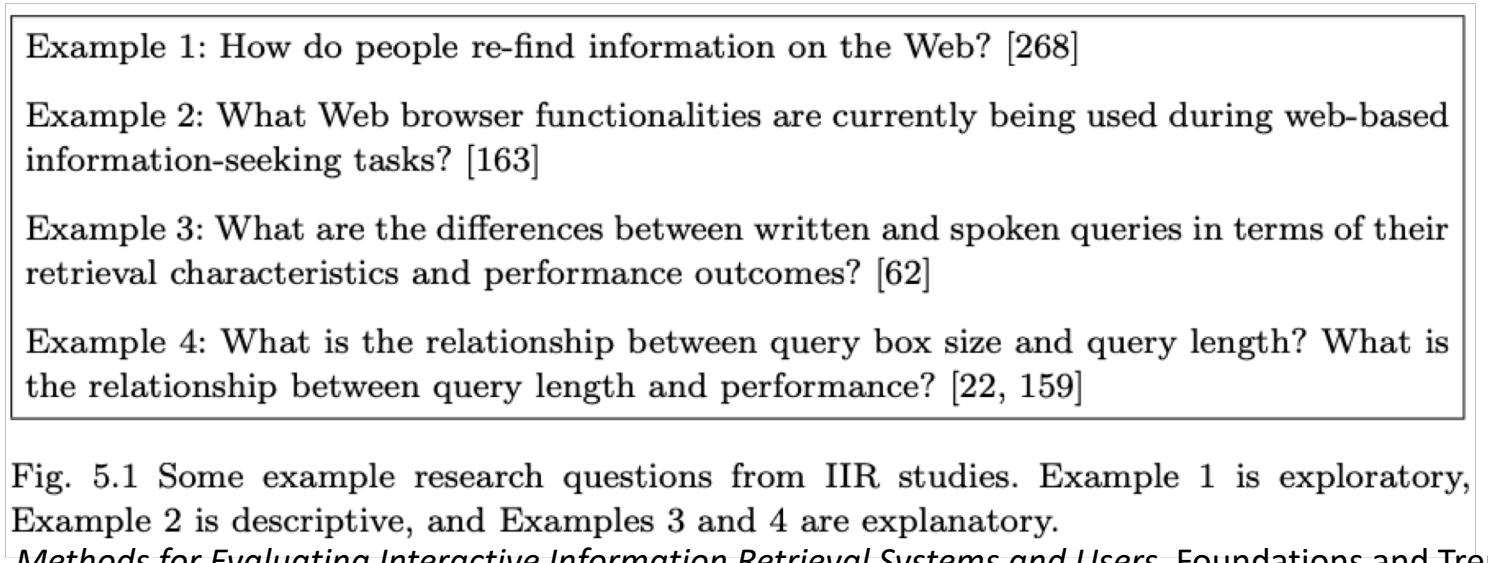
- Subjective measures:
  - Searchers' ***satisfaction*** with the interaction
  - Searchers' ***self-perception*** of their task completeness
  - Searchers' ***preference*** of an search system/interface

Experimental Procedure



Problems and Questions

- Identify and ***describe the problem*** – helps to draw attention to what is currently known about a particular issue
- ***Research question*** should be narrow and specific enough that it can be addressed in a study



Example 1: How do people re-find information on the Web? [268]

Example 2: What Web browser functionalities are currently being used during web-based information-seeking tasks? [163]

Example 3: What are the differences between written and spoken queries in terms of their retrieval characteristics and performance outcomes? [62]

Example 4: What is the relationship between query box size and query length? What is the relationship between query length and performance? [22, 159]

Fig. 5.1 Some example research questions from IIR studies. Example 1 is exploratory, Example 2 is descriptive, and Examples 3 and 4 are explanatory.

D. Kelly. 2009. *Methods for Evaluating Interactive Information Retrieval Systems and Users*. Foundations and Trends in Information Retrieval. Vol. 3.

Hypothesis

- Hypotheses follow from research questions and ***state expected relationships between the concepts identified in the questions*** (such concepts may be more or less definable, but they are eventually represented by variables).
- Hypothesis should be stated at the ***beginning of the study*** (rather than after analysing the data)
- Alternative vs. null hypothesis:
  - ***Alternative hypothesis*** (research hypothesis) is the researcher's statement about the expected relationship between the concepts under study, e.g. *system A is more usable than system B*.
  - ***Null hypothesis*** states that there is no relationship or difference between tested concepts. The null hypothesis is accepted by default.

Variables and Measurements

- ***Variables represent concepts***. Specifically they represent ways of defining, observing and measuring the concepts that researchers aim to study, e.g. relevance, performance or satisfaction.
- To investigate concepts, researchers must engage in two basic processes:
  - ***Conceptualization***: process by which researchers specify what they mean by particular terms, e.g. relevance. No claim is made about the universality of the definition.
  - ***Operationalization***: operational definitions, which state the precise way the concept will be measured. For instance, one might decide to measure topical relevance by asking subjects to indicate how useful they find documents and giving them a five-point scale to indicate this.

Direct and Indirect Observables

- ***Direct observables*** - byproducts of a user's behaviors and interactions, produced as the user searches: number of queries entered, number of documents opened, and the amount of time spent searching.
  - Direct observables are often easier to measure becuase they refer to easily measuarable ground-truth, e.g. number of clicks

- ***Indirect observables*** - cannot be observed and that essentially exist within the user's head, e.g. user satisfaction.
  - Instrumentation is more difficult
  - Researchers must ensure that indirect measures are good representations of particular concepts and that this information is properly captured, e.g., does a five-point Likert-type item adequately capture satisfaction?

Measurement Considerations

- ***Range of variation*** - extent to which a measure presents an adequate number of categories with which to respond, e.g. when creating an instrument for eliciting relevance judgments, is a binary scale, a tertiary scale, or a five-point scale provided?
- ***Exhaustiveness*** - extent to which a response set can be used to characterize all elements under study, e.g. with a binary relevance scale, a user might have a difficult time characterizing a document that is partially relevant.
- ***Exclusiveness*** - extent to which items in the response set overlap. When this property has been violated, there might be more than one response that can be used to characterize a single object, e.g. a user might be provided with the following options for indicating relevance: *not relevant, partially relevant, somewhat relevant* and *relevant*. Most subjects would have a difficult time distinguishing between the middle two options (unless provided clear definitions of each choice).

Measurement Considerations

- ***Equivalence*** - extent to which items in a response set are of the same type and same level of specificity. Consider a scale that is meant to assess a person's familiarity with a search topic and has at one end of the scale the label *very unfamiliar* and at the other end, *I know details*. It would be better to associate the first label with *very familiar*, and the second label with *I know nothing* since these are true opposites and at the same level.

- ***Appropriateness*** - extent to which the provided response set makes sense in relation to the question being asked, e.g. consider question "*How likely are you to recommend this system to others*?". If the researcher provided subjects with a five-point scale with *strongly agree* and *strongly disagree* as anchors, then this response set would be inappropriate because the scale anchors do not match the question

Levels of Measurement

- A critical concept that ultimately determines what types of statistical tests are possible.
- ***Discrete measures*** provide and elicit categorical responses:
- ***Nominal*** data types provide response choices that represent different kinds of things but not different degrees, e.g. ***independent variables***, such as interface type and task-type.
- ***Ordinal*** measures provide response choices that are ordered, where choices represent different degrees, e.g. ***rank-order measure*** (user asked to order a set of documents from most relevant to least) or ***Likert-type scales***, where numbers represents labels rather than numerical values, e.g. a 5-point Likert scale, where 1=not relevant and 5= relevant, indicate which documents are more relevant than others but not the amount of these differences. (document rated 4 is more relevant than a document rated 2, but not necessarily twice as relevant)

Levels of Measurement

Continuous measures - differences between consecutive points are equal:

- ***Interval scales*** – no true zero exists, e.g. Fahrenheit temperature scale and intelligence quotient (IQ) test scores. For both measures, a score of zero does not indicate the complete absence of heat or intelligence.

- ***Ratio level of measurement*** - represents the highest level of measurement, e.g. time and almost any measure that can be verbally described as the number of occurrences (the number of queries issued, the number of pages viewed, and the number of documents saved). It is possible for these values to be zero, e.g. it is possible for someone not to enter a single manual query or not open any documents during a search session.

Experimental Design – Baseline Selection

- Baselines are used in IIR evaluations, but in a way that differs slightly from the classic experimental model.
- In IIR evaluations, baselines are often introduced as an ***alternative to the experimental system***.
- Instead of taking a baseline measure before a user interacts with a stimulus, the baseline is more often represented by ***one level of the stimulus variable***, e.g. different values of a given parameter in an IIR system.
- Baselines in IIR evaluations are more similar to ***control groups***.
- Commercial search engine as a baseline not always possible (use of specific datasets, unknown optimization used by commercial search engines, etc.)
- However, users' experience with commercial search engines may affect the experimental results, e.g. users feeling more comfortable with a specific display of results

Experimental Design

- ***Within-subjects design*** (repeated measures design): each participant is tested under each condition, e.g. system A and system B
- ***Between-subjects design***: each participant is tested under one condition only. One group of participants is tested under condition A, a separate group is tested under condition B, and so on.
- within-subjects design is generally preferred:
  - **fewer participants are needed** since each participant is tested on all levels/systems. Although more testing is required for each participant, there is an advantage in having fewer participants overall, since recruiting, scheduling, briefing, demonstrating, practicing, and so on, are easier if there are fewer participants.
  - **less variance due to *participant disposition*** (since there are fewer participants), e.g. a participant who is predisposed to be meticulous (or reckless!) will likely exhibit such behaviour consistently across the experimental conditions.

Within-subject design: issues

- ***interference*** between experimental conditions, e.g. testing conflicting motor skills, such as typing on keyboards with different arrangements of keys, where the required skill to operate one keyboard tends to inhibit, or interfere with, the skill required for the other keyboard. Such an experiment should be assigned between-subjects.

- ***Learning effects*** due to the *order of presentation*. For example, if participants are tested under condition A first, then under condition B, they could potentially exhibit better performance under condition B simply due to prior practice under condition A.

Couterbalancing

- ***Counterbalancing*** -- placing participants in groups and presenting conditions to each group in a different order.
- ***Latin Square*** -- If experiment has two conditions (e.g., A and B), participants are randomly assigned to groups of equal size: Group 1 is given condition A followed by condition B, while Group 2 is given condition B followed by condition A:

A	B
B	A

→ time

A	B	C
B	C	A
C	A	B

3 × 3 Latin Square

A	B	C	D
B	C	D	A
C	D	A	B
D	A	B	C

4 × 4 Latin Square

A	B	C	D	E
B	C	D	E	A
C	D	E	A	B
D	E	A	B	C
E	A	B	C	D

5 × 5 Latin Square

Balanced Latin Squares

- Latin Square does not fully eliminate the learning effect, e.g. in 3 × 3 design condition B follows condition A for two of the three groups of participants, while in 4 × 4 design, condition B follows condition A for three of the four groups.
- Thus, there is a tendency for better performance on condition B simply because most participants benefited from practice on condition A prior to testing on condition B.
- ***Balanced Latin Square*** -- each condition appears before and after each other condition an equal number of times, e.g. B follows condition A two times and it also precedes condition A two times.

A	B	D	C
B	C	A	D
C	D	B	A
D	A	C	B

4 × 4 Balanced Latin Square

Balanced Latin Squares

- Balanced Latin Squares do not exist for odd-order squares, such as 3 × 3, 5 × 5, etc.
- Here's the rubric for any even number of conditions:
  - The 1st column is in order, starting at A.
  - The top row has the sequence, A, B, *n*, C, *n* - 1, D, *n* - 2, etc.
  - Entries in the 2nd and subsequent column are in order, with wrap around.

A	B	F	C	E	D
B	C	A	D	F	E
C	D	B	E	A	F
D	E	C	F	B	A
E	F	D	A	C	B
F	A	E	B	D	C

6 × 6 Balanced Latin Square

Protocals and Tutorials

- ***Study protocol*** - a step-by-step account of what will happen in a study:
  - ensures consistency in the administration of the study.
  - helps to maintain the integrity of the study
  - ensures that subjects experience the study in similar ways

- ***Tutorials*** – instructions for study participants how to use the system
  - Allows users to understand what to expect during the study
  - Cons: may bias the user during the experiment

- ***Pilot testing:***
  - help researchers identify problems with instruments, instructions, and protocols;
  - allow systems to be exercised in the same way they will be in the actual study;
  - provide researchers with an opportunity to get detailed feedback from test subjects about the method;
  - help researchers gain comfort with administering the study.

Simulated Work Tasks

- ***Simulated work task*** - a short cover story that describes the situation leading to the information need.
- Simulated work task describes the following to the user:
  - the source of the information need,
  - the environment of the situation,
  - the problem which has to be solved. This problem serves to make the test person understand the objective of the search.
- Such descriptions provide a basis against which situational relevance can be judged.

Simulated Work Tasks

Simulated Situation

Simulated work task situation: After your graduation you will be looking for a job in industry. You want information to help you focus your future job seeking. You know it pays to know the market. You would like to find some information about employment patterns in industry and what kind of qualifications employers will be looking for from future employees.

Indicative request: Find, for instance, something about future employment trends in industry, i.e., areas of growth and decline.

Criticism of simulated work tasks:

- tasks are artificial
- subjects may not have a context for executing the task and making relevance judgments
- subjects may simply be unmotivated to search for artificial tasks

Data Collection

- ***Think-aloud*** method asks subjects to articulate their thinking and decision-making as they engage in IIR.
- ***Stimulated recall*** - researcher records the screen of the computer as the subject completes a searching task. After the task is complete, the recording is played back to the subject who is asked to articulate thinking and decision-making as the recording is played.
- ***Prompted self-report*** - elicit feedback from subjects periodically while they search.
- ***Observation*** - researcher is seated near subjects and observes them as they search. During real-time observation, the subject is not interrupted, but can be asked follow-up questions about particular events later during post-search interviews.

Data Collection

- ***System logs*** are used to characterize the interaction and record both what the system does as well as how the subject reacts. Typical logs will record the subject's queries, the results shown to the subject and the results selected by the subject.

- ***Interviews*** - used as a delivery mode for a set of open-ended questions:
  - allows to obtain more individualized responses
  - allows some flexibility with respect to probing and follow-up
  - often conducted at the end of an IIR study

Questionnaires

Table 9.1 Commonly used questionnaires in IIR evaluations.

Questionnaire	Purposes	Administration
Demographic	This questionnaire is used to elicit background information about subjects. This information is typically used to characterize and describe subjects, but it can also be used to explore and test specific hypotheses. For instance, a researcher might be interested in investigating the difference between male and female behavior, or among people with different amounts of search experience.	This questionnaire is usually given at the start of the study, but it can be given at the end. The rationale for waiting until the end is that subjects are likely to be fatigued and it is better to get this "easy" information then.
Pre-task	This questionnaire can be used to assess subjects' knowledge of the search task and/or topic. Questionnaire items are usually directly related to the search task in which the subject is about to engage.	Subjects complete this questionnaire before searching occurs so that the search experience does not bias responses.
Post-task	This questionnaire is most often used to gather feedback about the subject's experiences using a particular system to complete a particular task. Thus, the primary goal of this questionnaire is to assess the system–task interaction.	This questionnaire is administered following each task.
Post-system	This questionnaire elicits feedback from subjects about their experiences using a particular experimental system. It is typical to administer this type of questionnaire during within-subjects studies where subjects use more than one system. The assessment is usually focused on the subjects' experiences using the system to complete a number of tasks and represents an overall assessment of a particular system.	The questionnaire is administered after subjects finish using a system. Subjects complete one questionnaire for each system.
Exit	If the study is a between-subjects study, then this questionnaire functions similarly to the post-system questionnaire. However, for within-subjects studies, this questionnaire can be used to elicit cross-system comparisons and ratings.	As its name implies, it is typically administered at the end of the study.

Questionnaires

- ***Questionnaire for User Interface Satisfaction (QUIS)*** - elicits evaluations of several aspects of the interface using a 10-point scale, including the subject's overall reactions to the software, the screen, the terminology and system information, and learning and system capabilities.
- ***Software Usability Measurement Inventory (SUMI)*** consists of 50 items and provides subjects with three coarse responses: agree, do not know and disagree.
- ***NASA-Task Load Index (NASA-TLX)*** consists of six component scales, which are weighted to reflect their contribution to the workload according to the subject: mental demand, physical demand, temporal demand, performance, frustration and effort.
- ***Recommender systems' Quality of user experience (ResQue)*** aims at measuring the qualities of the recommended items, the system's usability, usefulness, interface and interaction qualities, users' satisfaction with the systems, and the influence of these qualities on users' behavioral intentions.
- ***System Usability Scale (SUS)*** is a "quick and dirty", reliable tool for measuring the usability.

Usability (IOS Standard 9241)

- ***Effectiveness*** is the “accuracy and completeness with which users achieve specified goals.” In IIR, the most common way to measure effectiveness is precision and recall and to elicit self-report data from subjects about their perceptions of performance.
- ***Efficiency*** is the “resources expended in relation to the accuracy and completeness with which users achieve goals.” A tool is efficient if it helps users complete their tasks with minimum waste, expense or effort, e.g. time it takes a subject to complete a task
- ***Satisfaction*** is the “freedom from discomfort, and positive attitudes of the user to the product”. Satisfaction can be understood as the fulfillment of a specified desire or goal, e.g. system preference.

SUS

- developed by John Brooke in 1986,
- allows evaluation of a wide variety of products and services in terms of usability, including hardware, software, mobile devices and websites
- a simple, ten-item Likert scale with five response options for respondents; from *Strongly agree* to *Strongly disagree*
- often referred to as an “industry standard” in the business and technology industries
- SUS is particularly relevant to user experience when you comparing two versions of an application that are based around different technologies.
- Because SUS is pretty much technology-neutral, you can continue to use it in usability testing as technology evolves over the years, and you don’t have to continually reinvent questionnaires.

System Usability Scale

Instructions: For each of the following statements, mark one box that best describes your reactions to the website *today*.

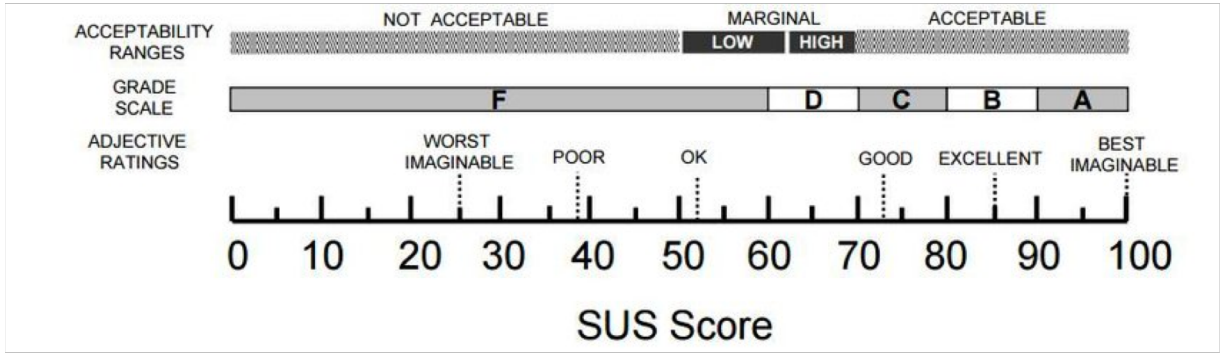
		Strongly Disagree				Strongly Agree
1.	I think that I would like to use this website frequently.	☐	☐	☐	☐	☐
2.	I found this website unnecessarily complex.	☐	☐	☐	☐	☐
3.	I thought this website was easy to use.	☐	☐	☐	☐	☐
4.	I think that I would need assistance to be able to use this website.	☐	☐	☐	☐	☐
5.	I found the various functions in this website were well integrated.	☐	☐	☐	☐	☐
6.	I thought there was too much inconsistency in this website.	☐	☐	☐	☐	☐
7.	I would imagine that most people would learn to use this website very quickly.	☐	☐	☐	☐	☐
8.	I found this website very cumbersome/awkward to use.	☐	☐	☐	☐	☐
9.	I felt very confident using this website.	☐	☐	☐	☐	☐
10.	I needed to learn a lot of things before I could get going with this website.	☐	☐	☐	☐	☐

Considerations when using SUS

- The scoring system is somewhat complex (more later)
- There is a temptation, when you look at the scores, since they are on a scale of 0-100, to interpret them as percentages, they are not
- The best way to interpret your results involves “normalizing” the scores to produce a percentile ranking
- SUS is not diagnostic - its use is in classifying the ease of use of the site, application or environment being tested; it differentiates usable and unusable sites
- Is a very easy scale to administer to participants
- Can be used on small sample sizes with reliable results

Interpreting SUS scores

- The participant's scores for each question are converted to a new number, added together and then multiplied by 2.5 to convert the original scores of 0 - 40 to 0 - 100.
- Though the scores are 0 - 100, these are not percentages and should be considered only in terms of their percentile ranking.
- Based on research, a SUS score above a 68 would be considered above average and anything below 68 is below average.

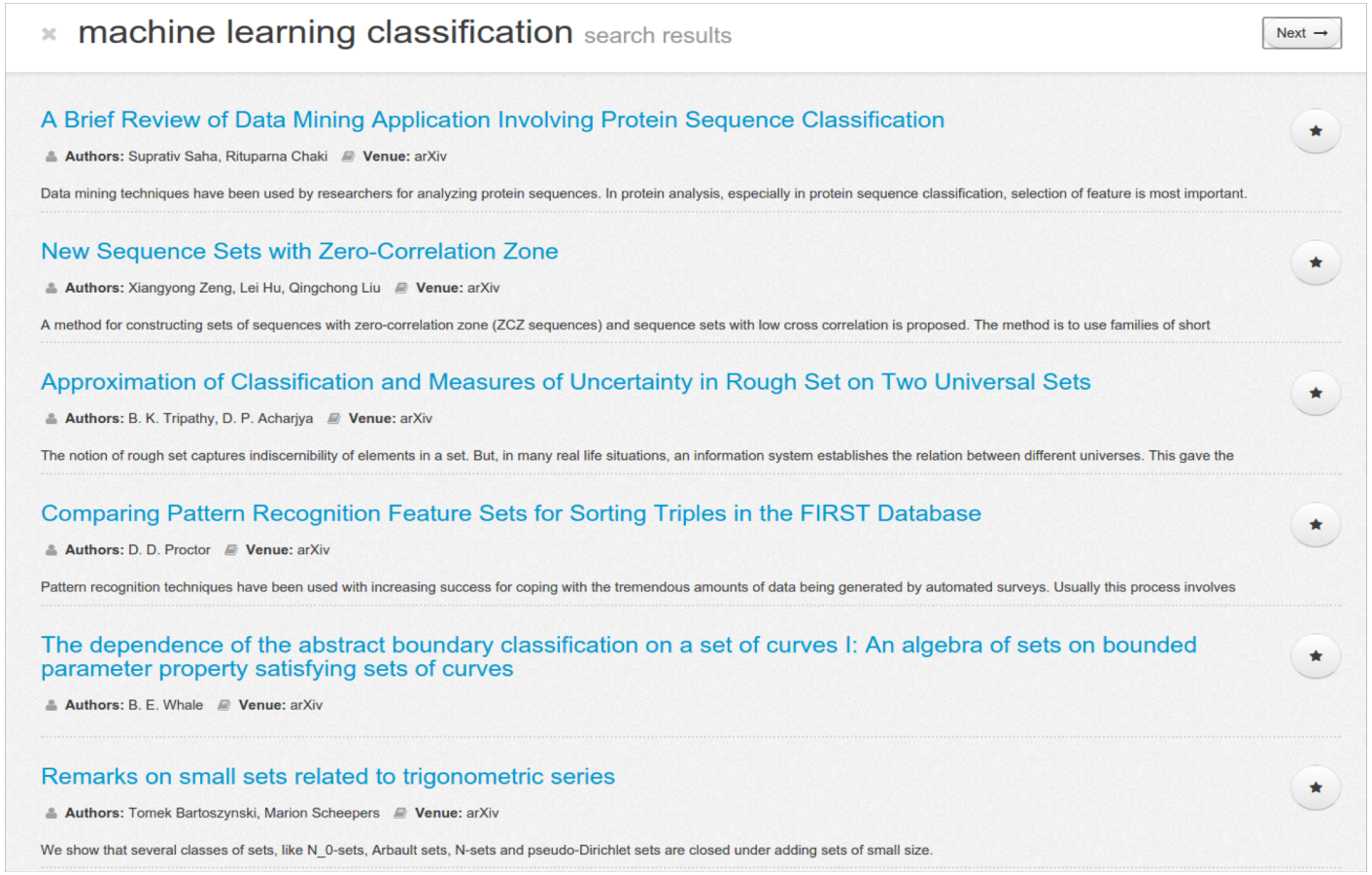


How to calcualte SUS score

- For every odd-numbered question, subtract 1 from the score (X - 1), e.g. if the response is 4, then the final calculation is 4 – 1 = 3
- For every even-numbered question, subtract the score from 5 (5 - X), e.g. if the response is 1, then the final calculation is 5 – 1 = 4
- Sum the scores from even and odd-numbered questions, then multiply the total by 2.5. The highest SUS score is now 100.

- Example:
  - Calculation values for each odd-numbered questions are 4, 2, 3, 4, 3 so that the accumulation of odd-numbered questions is 4+2+3+4+3 = 16.
  - Calculated values for even-numbered questions are 3, 3, 4, 2, 3, which makes the total accumulation of even-numbered questions to be 3+3+4+2+3 = 15.
  - Sum all scores of odd and even numbered questions. From this data, we got 16 (odd) + 15 (even) = 31.
  - If we **multiply the sum result with 2.5**, it would be **31 x 2.5 = 77.5**. The System Usability Scale (SUS) Score of this example is 77.5.

Personalizing Exploration-Exploitation



Medlar et al. *A System for Exploratory Search of Scientific Literature*. SIGIR 2016.

LinRel

In each iteration *t*, LinRel calculates:

$$a_i = x_i \cdot \left(X_t^T X_t + \mu I \right)^{-1} X_t^T$$

for each document *i* in dataset and selects for presentation top *n* documents that maximize:

$$\arg\max_X \{ a_i \cdot y_t + \frac{c}{2} \| a_i \| \}$$

for some constant $c > 0$

Study Design

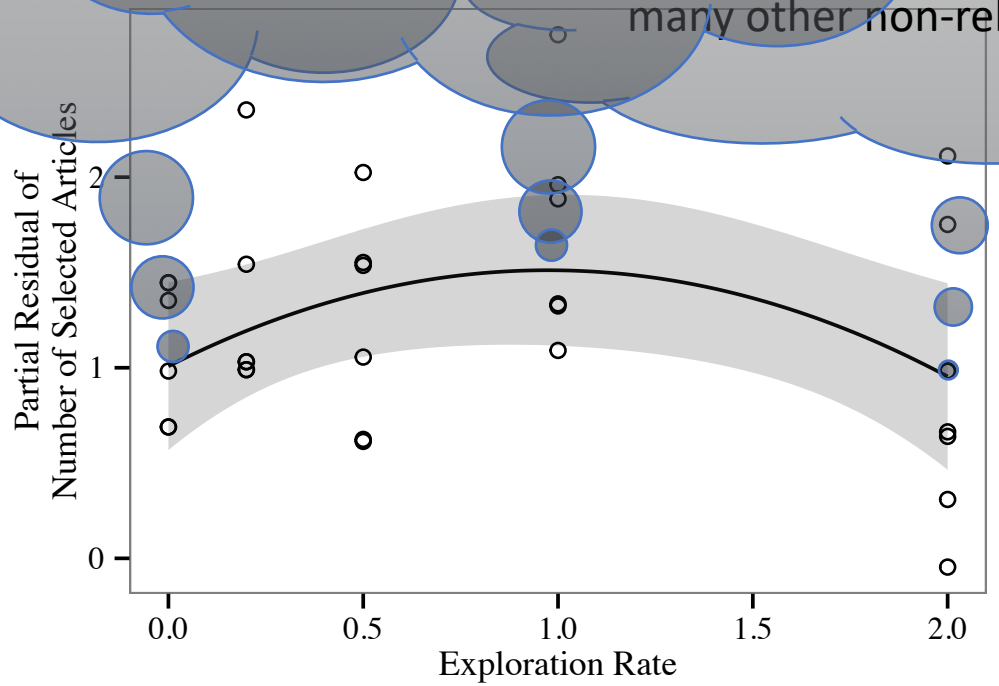
- Simulations: exploration rates to show different numbers of “exploratory” documents
- User study: MSc/PhD researchers in Machine Learning, 5 ML queries using different exploration rates
- Analysis: modelling combined with qualitative analysis of user performance data

Results

"Results are not quite satisfactory and about very specific definitions. I did not get [any] understanding of what topics are in this area"

"I went over several iterations. Results started getting way better [over iterations] and overall I am very satisfied"

"[results are] too scattered and many other non-related papers"



Further reading

D. Kelly. 2009. *Methods for Evaluating Interactive Information Retrieval Systems and Users*. Foundations and Trends in Information Retrieval. Vol. 3.

Assignment

A patent office is about to introduce a new patent search system. The system will be used by the in-house patent lawyers as well as general public. You are a consultant specialising in testing information retrieval and information search systems. You have been asked to evaluate this system before it is launched and make suggestions for any possible improvements. Your tasks are to:

1. Gather information about the requirements for this system.

2. Propose what aspects of the system need testing/evaluating based on the requirements.

3. Describe in detail how you would proceed with the evaluation (justify your proposed evaluation procedures).

4. Describe what results you would expect to obtain based on your selected evaluation methods.