Method Bias? The Effects of Performance Feedback on Users' Evaluations of an Interactive IR System

or

What we did last semester in the IIR Seminar

Diane Kelly, Chirag Shah, Cassidy R. Sugimoto, Earl W. Bailey, Rachael A. Clemens, Ann K. Irvine, Nicholas A. Johnson, Weimao Ke, Sanghee Oh, Anezka Poljakova, Marcos A. Rodriguez, Megan G. van Noord, & Yan Zhang

CRADLE | 04 April 2008

UNC SCHOOL OF INFORMATION AND LIBRARY SCIENCE

Motivation

- Let's start with the traditional IIR evaluation model (a lá TREC)
- What's wrong with this model?
  - Relevance Assessments
  - Assigned Topics
  - Not much variability in Tasks
  - …

Motivation

- How can users be expected to evaluate a system when they have no idea how well they performed?
  - Some problems with ‘perceptions’
  - Users like everything

Research Question

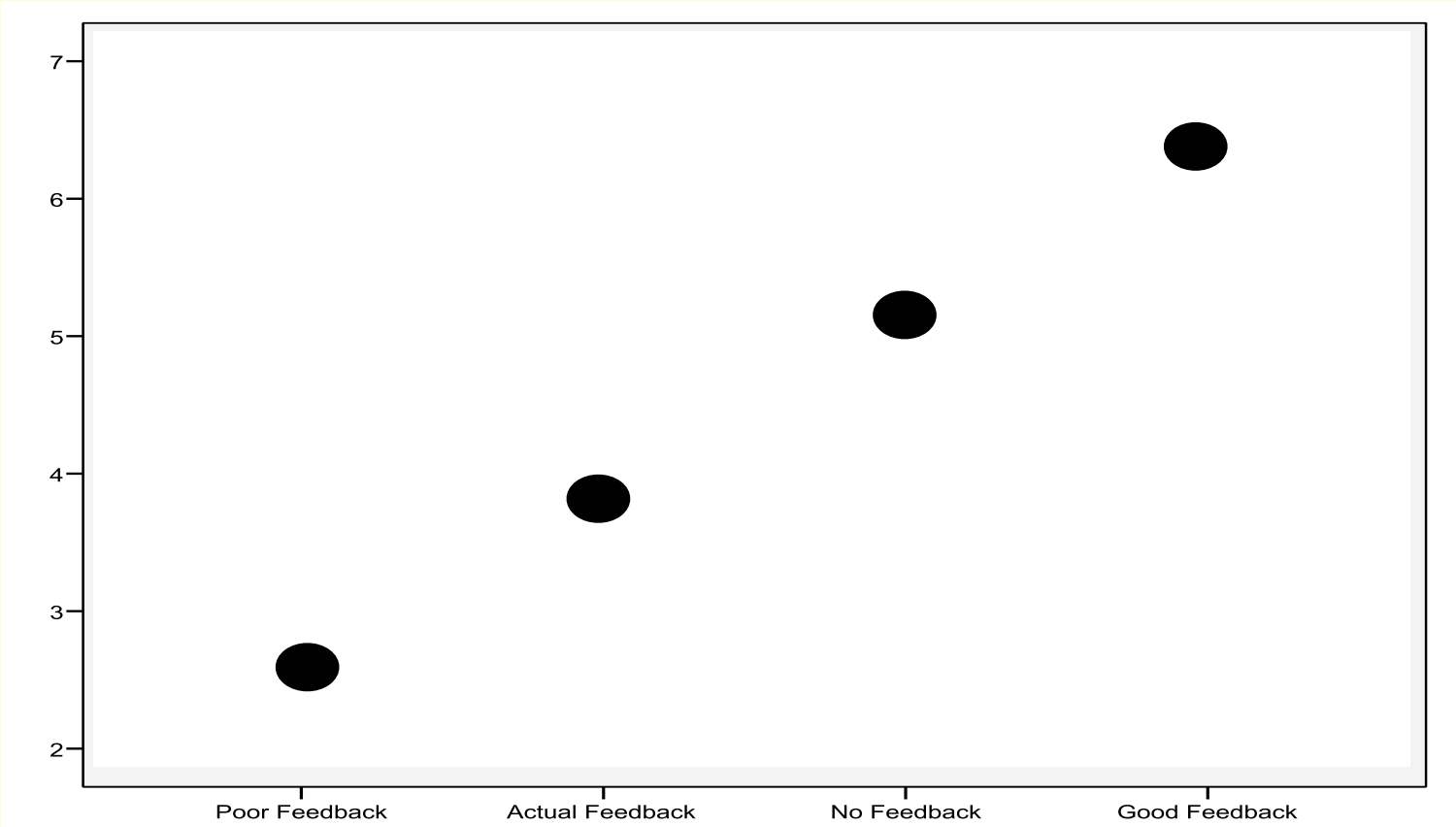
- How does providing feedback to users about their performances for high-recall tasks affect their evaluations of an experimental IIR system?

- "New" Evaluation Method
- Feedback: Recall (relevant documents found/relevant documents in corpus)

Feedback Conditions

- **No Feedback (NF):** users are not provided with any feedback about their performances (baseline)
- **Actual Feedback (AF):** users are told their real performances
- **Good Feedback (GF):** users are deceived and told they performed very well (92%)
- **Poor Feedback (PF):** users are deceived and told they performed very poorly (12%)

Expectations



Poor Feedback < Actual Feedback < No Feedback < Good Feedback

Method

- XRF experimental IR system + TREC collection + Lemur (BM25)
- 60 undergraduate subjects assigned randomly to condition (between subjects)
- Three recall-oriented search topics followed by Post-Task Questionnaires
- For three conditions (PF, AF, and GF) *performance feedback* is provided before completing Exit Questionnaire

Example Topic

[simulated work task] + Your goal is to identify as many different developments in robotic technology and their uses as possible, and to find as much information about each development and its uses as possible.

“Flavor” of Questionnaires

- Post-Task Questionnaire (6 items)
  - Familiarity, Easy to search, Satisfaction with search results, Confidence, Enough time, Satisfaction with performance
  - 7-pt scale: 1=not at all, 4=somewhat, 7=extremely

“Flavor” of Questionnaires

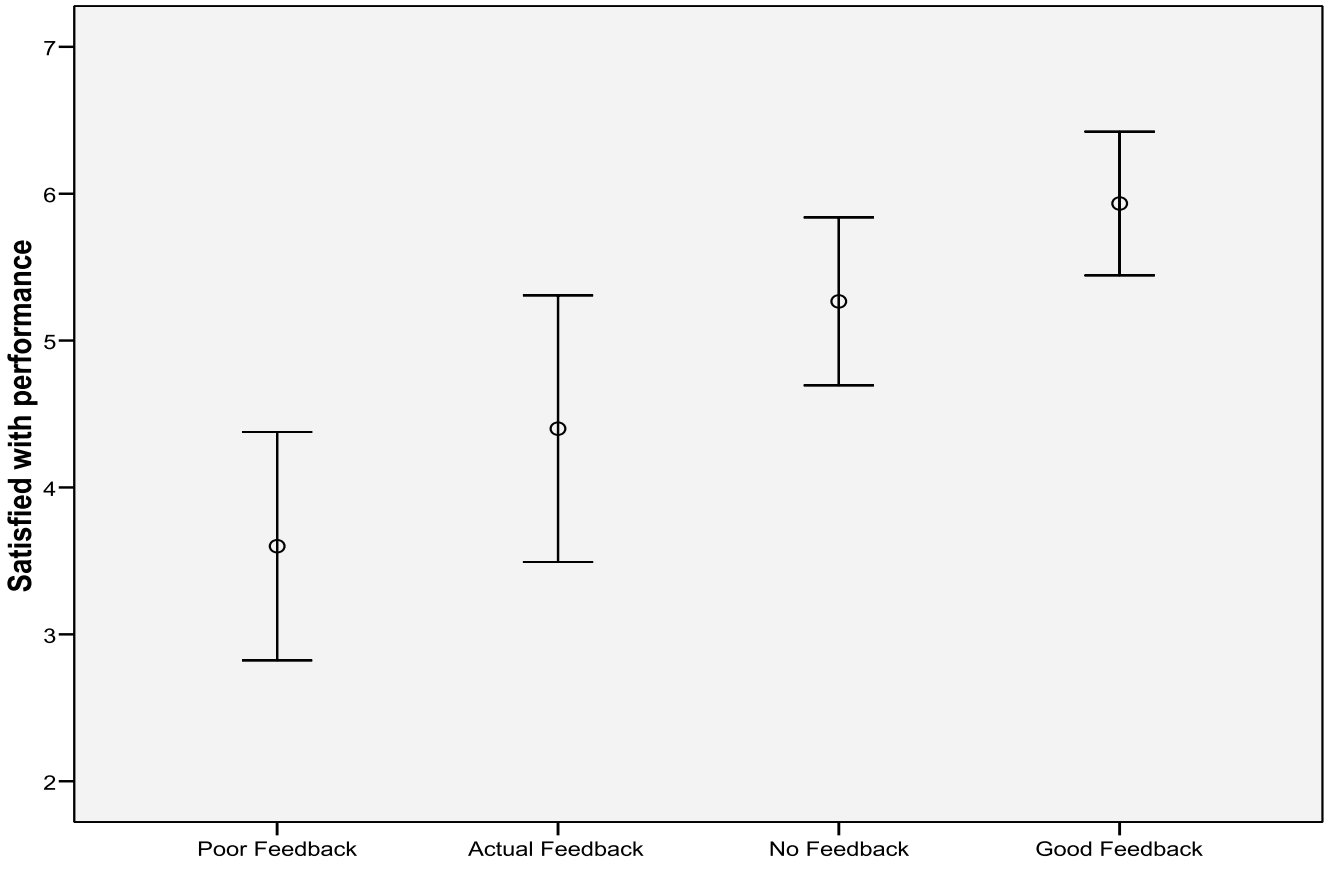
- Exit Questionnaire (13 items)
  - Easy to learn, Inconsistencies, Easy to query, Easy to navigate, Similarity of search methods, Color-coding, Easy to use, Accomplish, Easy to find relevant documents, Easy to understand why documents were retrieved, Various functions well-integrated, Overall effectiveness, Overall satisfaction
  - 7-pt scale: 1=strongly disagree, 7=strongly agree

Results

- Post-Task Questionnaire Responses
  - For most items (except familiarity), ratings were higher than the scale mid-point
  - For all items, there were no significant differences in responses according to condition
  - Overall, things seems to be going okay …

	Feedback Condition			
	Poor	**Actual**	**No**	**Good**
Satisfaction	4.69 (1.20)	5.22 (1.19)	4.89 (1.01)	4.82 (1.28)

Exit Questionnaire #13: Satisfaction

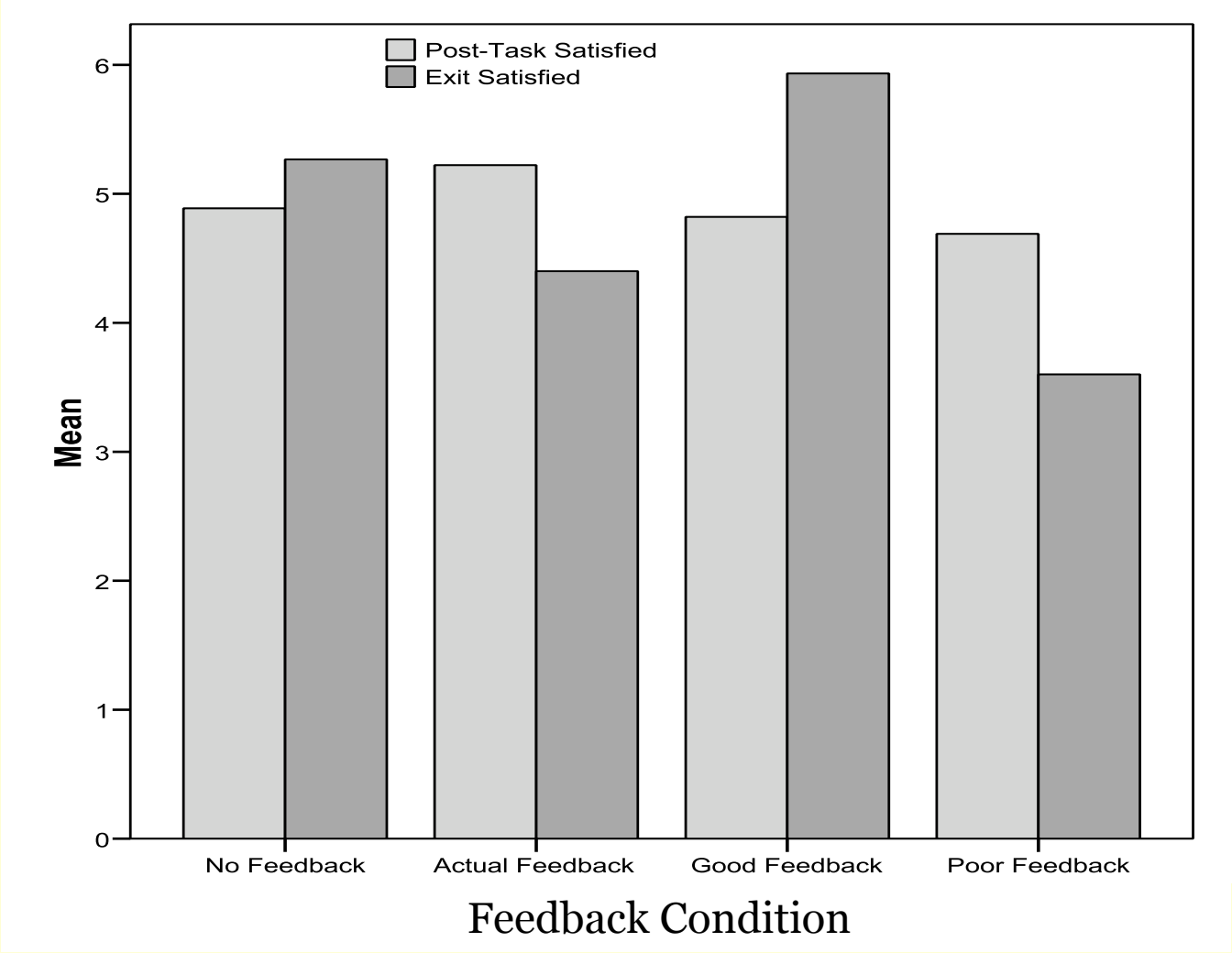


$F(3,59)=9.54$, $p<.01$

PF < NF, GF
AF < GF

Effect Size = .34

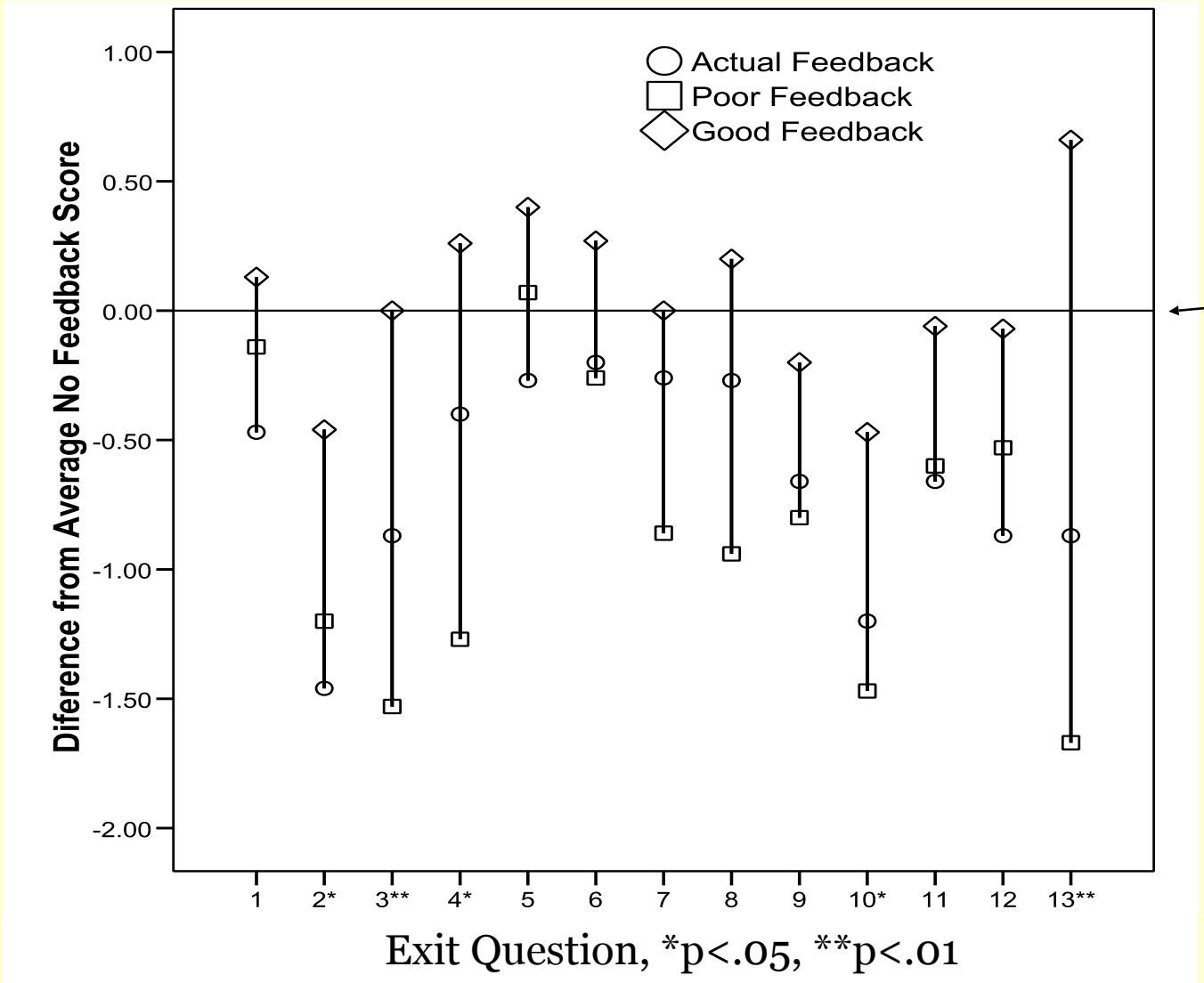
Pre- & Post-Feedback Satisfaction



NF: $t(14) = 1.89^{*}$
AF: $t(14) = -3.14^{**}$
GF: $t(14) = 6.29^{**}$
PF: $t(14) = 3.58^{**}$

$^{*}ns$
$^{**}p<.01$

Exit Questionnaire Responses



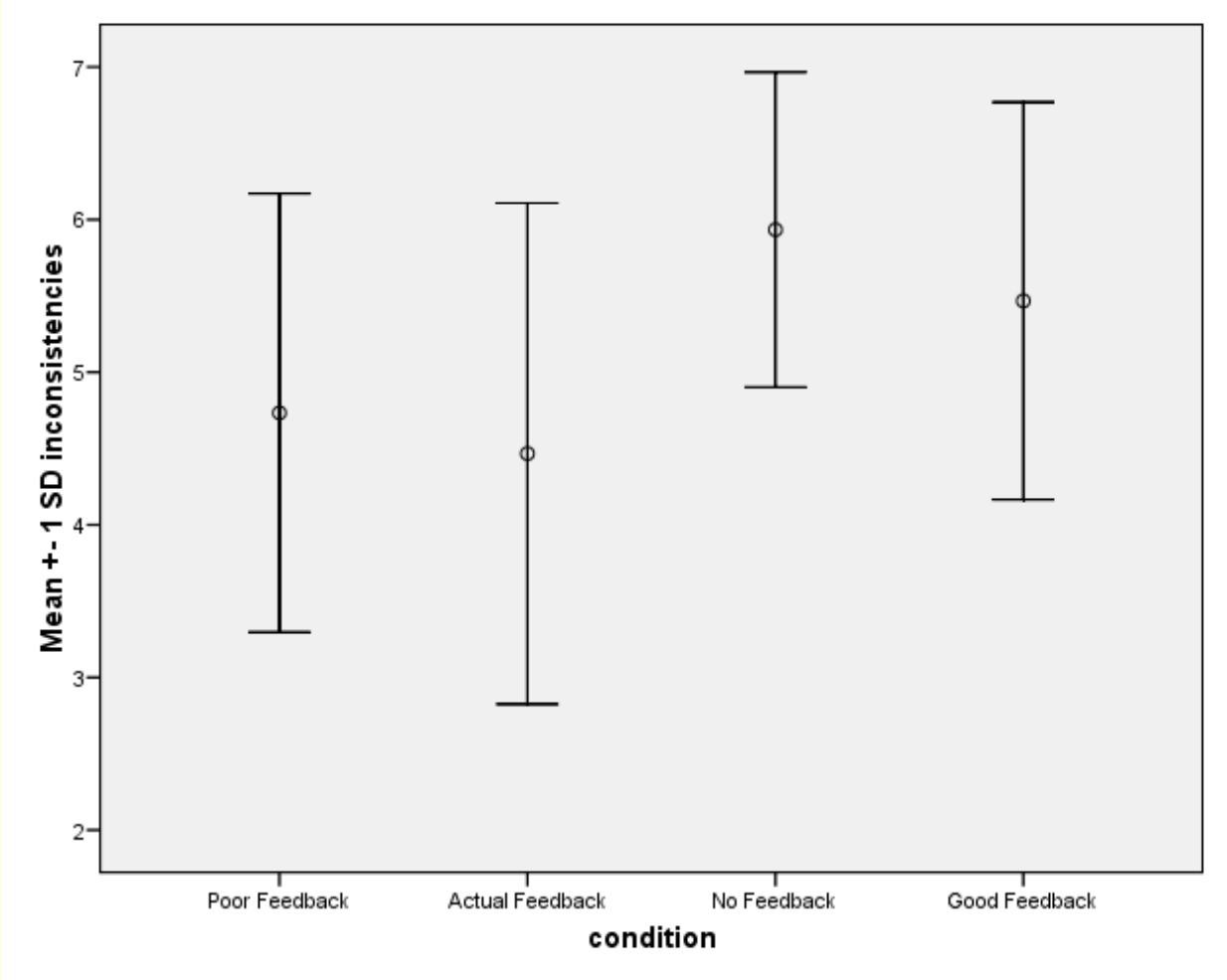
Average Score of subjects in the No Feedback Condition for each Question is set to 0.

A Closer Look at Three Other Questions

2. I didn't notice any inconsistencies when I used the system.

3. It was easy to pose queries to the system.
4. It was easy to navigate the search results.

Exit Questionnaire #2

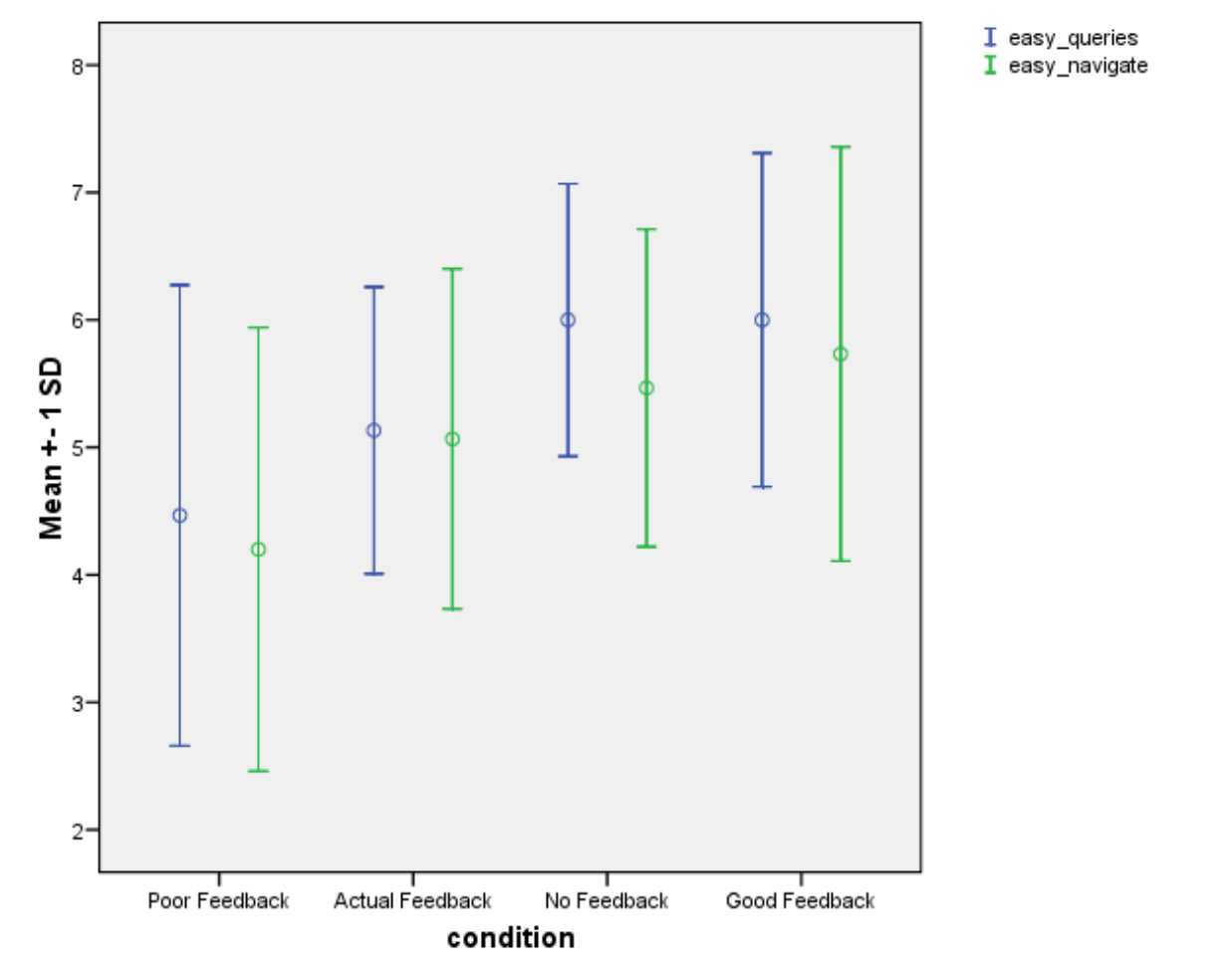


$F(3,59)=3.60$, $p=.02$

AF < NF

Effect Size = .16

Exit Questionnaire #3 and #4



$F(3,59)=4.50$, $p=.01$

PF < NF, GF

Effect Size = .19

$F(3,59)=2.99$, $p=.04$

ns

Effect Size = .14

What do we make of these results?

- We consider several explanations:
  - Inaccurate perceptions of performance
  - Inflation bias
  - Attribution bias

Inaccurate Perceptions

- Subjects' perceptions of how well they performed were inaccurate and providing them with information about this helped them to make more *accurate* evaluations
- No strong relationship between objective and subject measures
- Subjects in the No Feedback condition gave the system some of the highest ratings despite being some of the worst performers
- Pre- and post-test changes in 'inconsistencies' question

Inflation Bias

- Subjects' initial ratings were inflated and providing them with information about their performances motivated them to make more critical evaluations
- Several items affected by feedback condition were about interface features
- Regardless of feedback condition, average ratings were still above the scale mid-point

Attribution Bias

- Subjects in the Poor and Actual Feedback conditions blamed the computer for their 'failures'
- Attribute theory is often used to explain how people perceive the causes of success and failure and how people attribute blame
- Does IIR create an equal-opportunity 'blame' scenario? Who is responsible for the outcome?
- Our study was not set-up to really support exploration of this explanation

Some Limitations

- Subjects' interpretations of performance measures (i.e., 30% = F)
- Validity and reliability of questionnaire items
- Expectations about the relationship between objective and subjective evaluation measures

Conclusions

- Lots of undergraduates are no-shows
- Subjects' evaluation behaviors are brittle and susceptible to slight changes in evaluation method
- Researchers should provide users with feedback about their performances when this information is available (?)
- Researchers should *at least* start to question and evaluate method variance

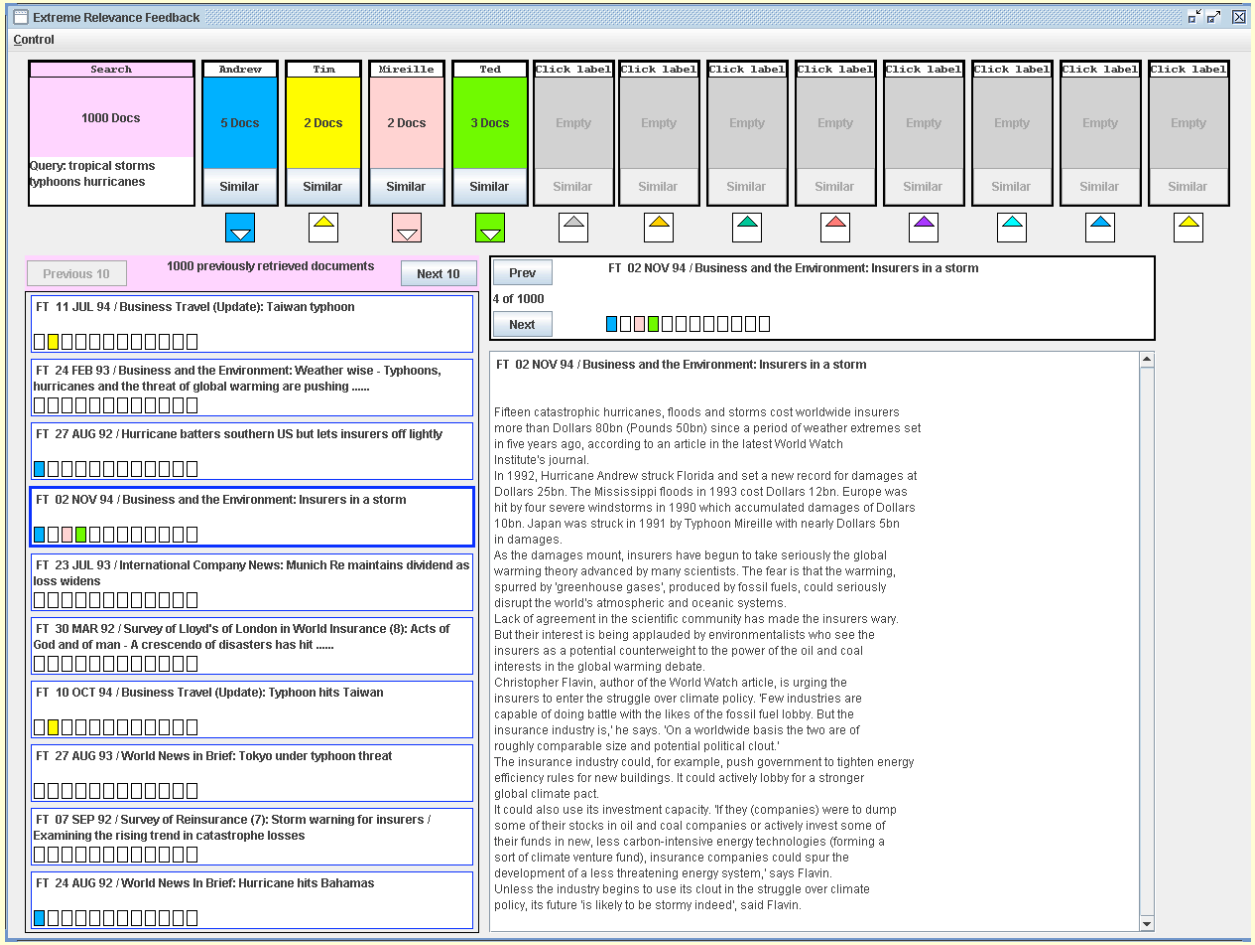
Teaching + Research

- Start with something you *already* have available and a *simple* research question.
- Ideally you have published a paper about this topic. Ask students to read this and discuss it in class.
- Spend class time discussing and refining instruments and method
- Provide lots of training & structure
- (Where'd I get the $600?)

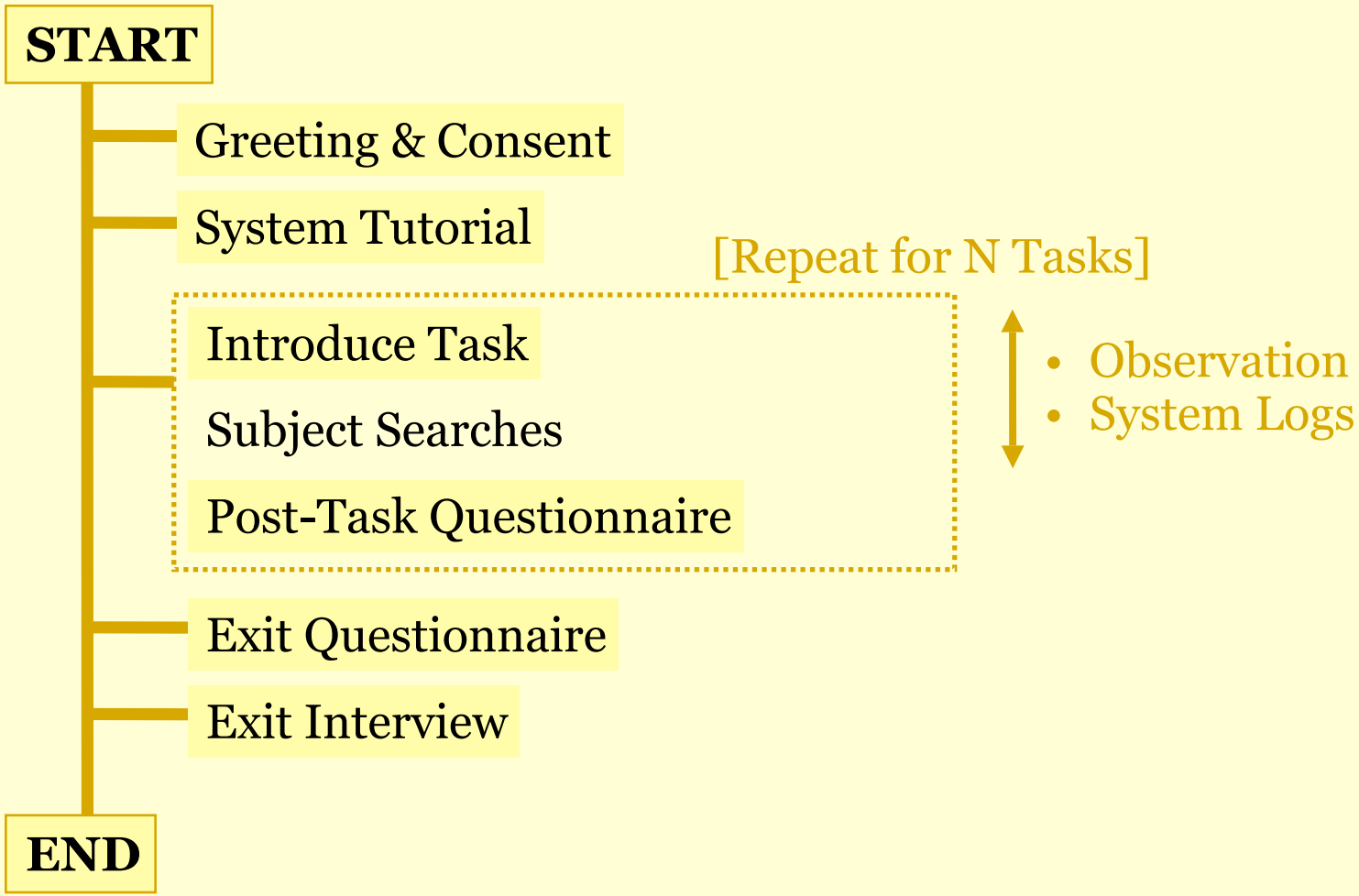
Teaching + Research

- As data are being collected, discuss experiences (research therapy)
- Provide a template for writing report describing data *each student* collected
- Discuss individual findings and present overall key findings
- If possible, integrate some of material from the reports into the final manuscript
- Master's papers

XRF Interface

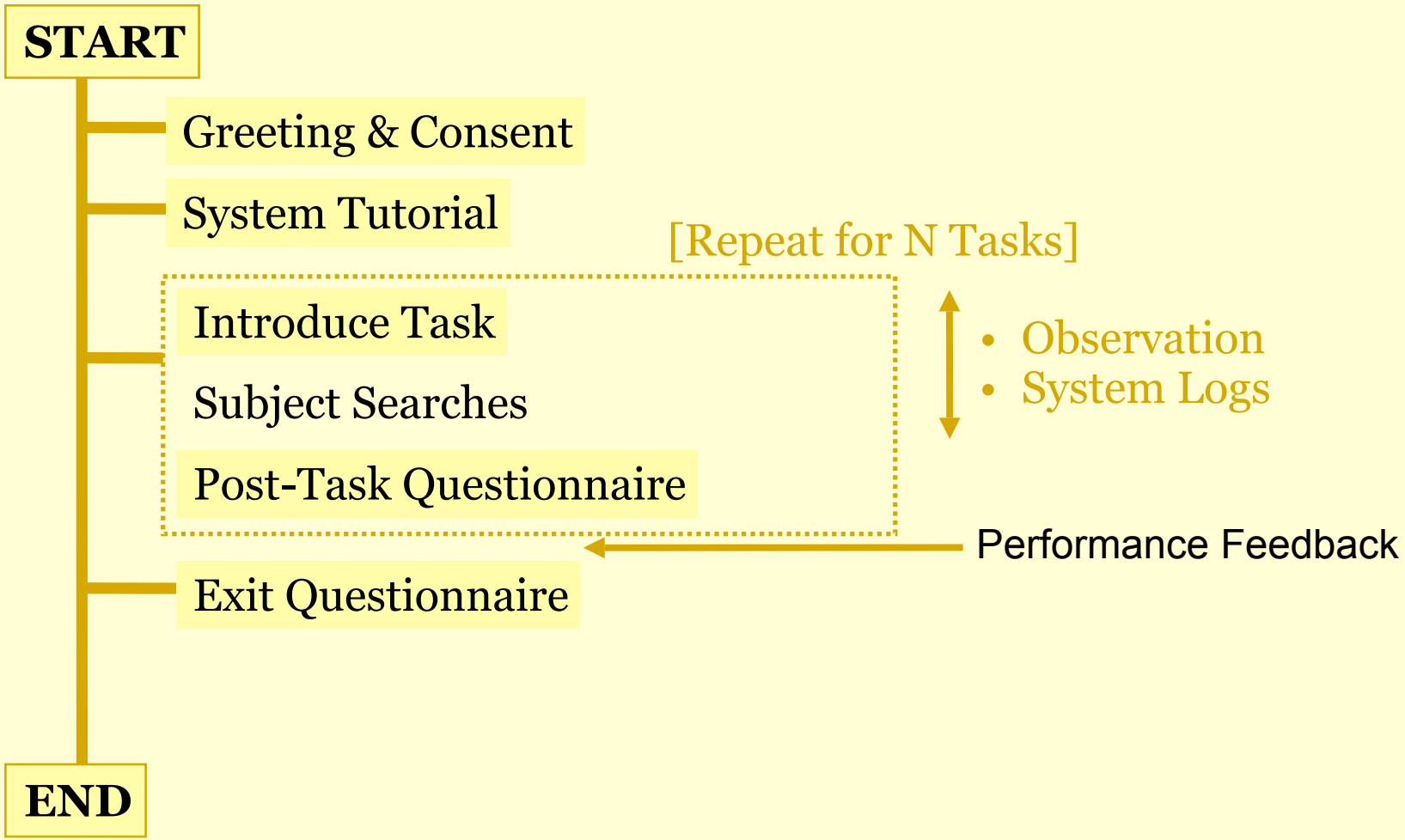


Basic Protocol



BAC

Basic Protocol



BAC

TREC
[Text REtrieval Conference]

It's not this …

What is TREC?

- TREC is a workshop series sponsored by the National Institute of Standards and Technology (NIST) and the US Department of Defense.
- It's purpose is to build infrastructure for large-scale evaluation of text retrieval technology.
- TREC collections and evaluation measures are the *de facto* standard for evaluation in IR.
- TREC is comprised of different tracks each of which focuses on different issues (e.g., question answering, filtering).

Table 1.1
Number of participants per track and total number of distinct participants in each TREC

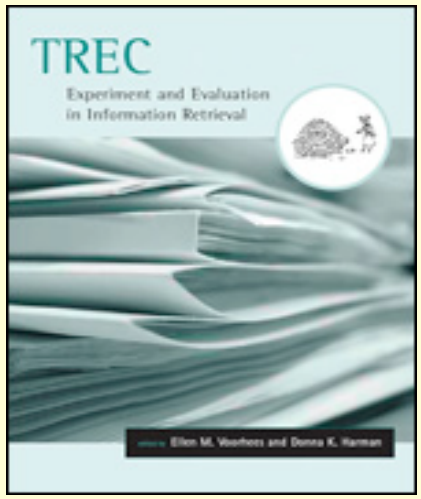
	TREC											
Track	92	93	94	95	96	97	98	99	00	01	02	03
Ad hoc	18	24	26	23	28	31	42	41	—	—	—	—
Routing	16	25	25	15	16	21	—	—	—	—	—	—
Interactive	—	—	3	11	2	9	8	7	6	6	6	—
Spanish	—	—	4	10	7	—	—	—	—	—	—	—
Confusion	—	—	—	4	5	—	—	—	—	—	—	—
Database merging	—	—	—	3	3	—	—	—	—	—	—	—
filtering	—	—	—	4	7	10	12	14	15	19	21	—
Chinese	—	—	—	—	9	12	—	—	—	—	—	—
NLP	—	—	—	—	4	2	—	—	—	—	—	—
Speech	—	—	—	—	—	13	10	10	3	—	—	—
Cross-language	—	—	—	—	—	13	9	13	16	10	9	—
High precision	—	—	—	—	—	5	4	—	—	—	—	—
Very large corpus	—	—	—	—	—	—	7	6	—	—	—	—
Query	—	—	—	—	—	—	2	5	6	—	—	—
Question answering	—	—	—	—	—	—	—	20	28	36	34	33
Web	—	—	—	—	—	—	—	17	23	30	23	27
Video	—	—	—	—	—	—	—	—	—	12	19	—
Novelty	—	—	—	—	—	—	—	—	—	—	13	14
Genome	—	—	—	—	—	—	—	—	—	—	—	29
HARD	—	—	—	—	—	—	—	—	—	—	—	14
Robust	—	—	—	—	—	—	—	—	—	—	—	16
Total participants	25	31	33	36	38	51	56	66	69	87	93	93

TREC Collections

- Central to each TREC Track is a collection, which consists of three major components:
  1. A corpus of documents (typically newswire)
  2. A set of information needs (called *topics*)
  3. A set of relevance judgments.
- Each Track also adopts particular evaluation measures
  - Precision and Recall; F-measure
  - Average Precision (AP) and Mean AP (MAP)

Learn more about TREC

- http://trec.nist.gov
- Voorhees, E. M., & Harman, D. K. (2005). *TREC: Experiment and Evaluation in Information Retrieval*, Cambridge, MA: MIT Press.



BACK