Designing Formative Studies
(and a little bit about pilot studies)

Todd Kulesza
Summative vs Formative

Summative studies examine outcomes.

Formative studies inform design.
## Where they differ

<table>
<thead>
<tr>
<th>Summative</th>
<th>Formative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Answers about outcomes</td>
<td>Clues about behavior/needs</td>
</tr>
</tbody>
</table>

**Wednesday, October 27, 2010**
Where they differ

<table>
<thead>
<tr>
<th>Summative</th>
<th>Formative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Answers about outcomes</td>
<td>Clues about behavior/ needs</td>
</tr>
<tr>
<td>Large[r] participant pool</td>
<td>Small[er] participant pool</td>
</tr>
</tbody>
</table>

Wednesday, October 27, 2010
<table>
<thead>
<tr>
<th>Summative</th>
<th>Formative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Answers about outcomes</td>
<td>Clues about behavior/needs</td>
</tr>
<tr>
<td>Large[r] participant pool</td>
<td>Small[er] participant pool</td>
</tr>
<tr>
<td>Often analyzes aggregates</td>
<td>Often focuses on specific participants</td>
</tr>
</tbody>
</table>
An example

**Eventual goal:** help people make intelligent assistants smarter
An example

**Eventual goal:** help people make intelligent assistants smarter

**Roadblock:** how do people communicate with intelligent assistants?
Todo list

What do we want to learn?
Todo list

What do we want to **learn**?

↓

How can we **collect** data?
Todo list

What do we want to **learn**?

How can we **collect** data?

How should we **analyze** participant data?
Todo list

What do we want to **learn**?

How can we **collect** data?

How should we **analyze** participant data?

What are the design **implications**?
What do we want to learn?

Clearly frame the...
What do we want to learn?

Clearly frame the problem...
What do we want to learn?

Clearly frame the... problem domain
What do we want to learn?

Clearly frame the... { problem domain population}
Framing example

We want to help people improve intelligent assistants.
Framing example

We want to help *people* improve intelligent *assistants*.
Framing example

We want to help *people* improve intelligent *assistants*.

What’s a popular type of assistant?
Framing example

We want to help *people* improve intelligent *assistants*.

What’s a popular type of assistant? → Why do they matter?
Framing example

We want to help *people* improve intelligent *assistants*.

Why do they matter?

Who uses these assistants?

What’s a popular type of assistant?
Framing example

We want to help people improve intelligent assistants.

What's a popular type of assistant?

How can they be improved?

Why do they matter?

Who uses these assistants?
## Qualitative analysis

<table>
<thead>
<tr>
<th>Time</th>
<th>Text</th>
<th>Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>So just to make sure for box E these are already, is there something to be fixed in box E because there are already grades? <em>Proctor: I can't tell you that.</em> Okay, Sum of AF12&gt;7 AF12?</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>[examines the grade table near the top] Okay, let's evaluate the letter grades [traces dependents for letter grades. Scrolls right to look at the labs attended. Evaluates formula] Z12=3.</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Why does that arrow stop at five and not at total labs? [referring to the trace dependents of the first letter grade G12. Traces dependents on GPA]</td>
<td></td>
</tr>
</tbody>
</table>
Our domain and population
Our domain and population

**What:** Naive Bayes text classifiers
Our domain and population

**What:** Naive Bayes text classifiers

**Why:** Save time & effort in qualitative analysis
Our domain and population

**What:** Naive Bayes text classifiers

**Why:** Save time & effort in qualitative analysis

**Who:** Social science researchers
Our domain and population

What: Naive Bayes text classifiers

Why: Save time & effort in qualitative analysis

Who: Social science researchers

How: Explaining specific reasons
Our research questions
Our research questions

How do end users think text-classifying intelligent assistants make their decisions?
Our research questions

1. How do end users think text-classifying intelligent assistants make their decisions?

2. Can such assistants [intelligibly] explain their reasoning to end users?
Our research questions

1. How do end users think text-classifying intelligent assistants make their decisions?

2. Can such assistants [intelligibly] explain their reasoning to end users?

3. How do end users “naturally” try to fix an assistant’s reasoning?
Natural programming

Human-centered approach
Natural programming

Human-centered approach

Study how people perform tasks
Natural programming

Human-centered approach

Study how people perform tasks

Build tools around existing behavior
Collecting data

Surveys

Interviews

Low-fi prototypes
Collecting data

Surveys

Interviews

Low-fi prototypes

Only answers #1
Collecting data

- Surveys
  - Only answers #1

- Interviews
  - Answers #1 & maybe part of #3

- Low-fi prototypes
Collecting data

Surveys

Interviews

Low-fi prototypes (often paired with think-aloud study)

Only answers #1

Answers #1 & maybe part of #3

Ding! Ding! Ding! Ding!
Design considerations

Need to measure effects of explanations

Small sample size

Avoid influencing participant reasoning

Open-ended feedback
Design considerations

Two treatments: one with explanations of assistant’s reasoning, one without

Small sample size

Avoid influencing participant reasoning

Open-ended feedback
Design considerations

Two treatments: one with explanations of assistant’s reasoning, one without

Within-subjects design

Avoid influencing participant reasoning

Open-ended feedback
Design considerations

Two treatments: one with explanations of assistant’s reasoning, one without

Within-subjects design

Treatment lacking explanations needs to be first

Open-ended feedback
Design considerations

Two treatments: one with explanations of assistant’s reasoning, one without

Within-subjects design

Treatment lacking explanations needs to be first

Paper prototype, audio/video recording
Our paper prototype

How do end users explain reasoning to an intelligent assistant?
Our paper prototype

How do end users explain reasoning to an intelligent assistant?

Can an assistant explain its logic to end users?
Our paper prototype

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Okay, Sum of AF12&gt;7 AF12?</td>
</tr>
<tr>
<td></td>
<td>They didn't answer whether AF12&gt;AF12, so no info was gained or lost.</td>
</tr>
<tr>
<td></td>
<td>None</td>
</tr>
<tr>
<td>2</td>
<td>lets look at the formula [clicks the evaluates formula for the total points column.</td>
</tr>
<tr>
<td></td>
<td>looking for info.</td>
</tr>
<tr>
<td></td>
<td>None</td>
</tr>
<tr>
<td></td>
<td>None</td>
</tr>
<tr>
<td></td>
<td>Seeking info.</td>
</tr>
</tbody>
</table>
Our paper prototype

80

[looks at waive instructions]

This was coded as None because:
- Actions following "Info Gained" are almost certainly "None"
- Segments beginning with "looks at" are probably "None"

This was not coded as Seeking Info because:
- Segments following "Info Gained" are probably not "Seeking Info"
- Segments containing "instruction" are probably not "Seeking Info"

This was not coded as Info Gained because:
- Actions are probably not "Info Gained"
- Segments lacking "OK" are probably not "Info Gained"

This was not coded as Info Lost because:
- Since one of the next 3 segments is "Info Gained", this is probably not "Info Lost"
- Segments beginning with a verb are almost certainly not "Info Lost"

None

ok!
Pilots

**Pilot studies** debug your experiment

Materials and procedures may change

Software parameters may require tuning
### Our pilot results

<table>
<thead>
<tr>
<th>Problem</th>
<th>Adaptation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Written feedback was unclear</td>
<td>Proctor needs to ask questions during the study</td>
</tr>
<tr>
<td>Verbal feedback tapered off</td>
<td>Proctor needs to prod participants</td>
</tr>
<tr>
<td>Participants appeared annoyed after experiment</td>
<td>Shorten study duration</td>
</tr>
</tbody>
</table>
Our pilot results

<table>
<thead>
<tr>
<th>Problem</th>
<th>Adaptation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Written feedback was unclear</td>
<td>Proctor needs to ask questions during the study</td>
</tr>
</tbody>
</table>
Our pilot results

<table>
<thead>
<tr>
<th>Problem</th>
<th>Adaptation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Written feedback was unclear</td>
<td>Proctor needs to ask questions during the study</td>
</tr>
<tr>
<td>Verbal feedback tapered off</td>
<td>Proctor needs to prod participants</td>
</tr>
</tbody>
</table>
## Our pilot results

<table>
<thead>
<tr>
<th>Problem</th>
<th>Adaptation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Written feedback was unclear</td>
<td>Proctor needs to ask questions during the study</td>
</tr>
<tr>
<td>Verbal feedback tapered off</td>
<td>Proctor needs to prod participants</td>
</tr>
<tr>
<td>Participants appeared annoyed after experiment</td>
<td>Shorten study duration</td>
</tr>
</tbody>
</table>
Analysis

More often qualitative than quantitative

Focus on individual participants instead of aggregates

Look for ties to existing theories
## Our analysis methods

<table>
<thead>
<tr>
<th>Subject</th>
<th>Word single</th>
<th>Word multiple</th>
<th>Word present</th>
<th>Punctuation</th>
<th>Relationships</th>
<th>KB-English</th>
<th>Probabilities</th>
</tr>
</thead>
<tbody>
<tr>
<td>2101</td>
<td><img src="#" alt="Blue" /></td>
<td><img src="#" alt="Orange" /></td>
<td><img src="#" alt="Blue" /></td>
<td><img src="#" alt="Orange" /></td>
<td><img src="#" alt="Orange" /></td>
<td><img src="#" alt="Blue" /></td>
<td><img src="#" alt="Orange" /></td>
</tr>
<tr>
<td>2201</td>
<td><img src="#" alt="Blue" /></td>
<td><img src="#" alt="Orange" /></td>
<td><img src="#" alt="Blue" /></td>
<td><img src="#" alt="Orange" /></td>
<td><img src="#" alt="Orange" /></td>
<td><img src="#" alt="Blue" /></td>
<td><img src="#" alt="Orange" /></td>
</tr>
<tr>
<td>2301</td>
<td><img src="#" alt="Blue" /></td>
<td><img src="#" alt="Orange" /></td>
<td><img src="#" alt="Blue" /></td>
<td><img src="#" alt="Orange" /></td>
<td><img src="#" alt="Orange" /></td>
<td><img src="#" alt="Blue" /></td>
<td><img src="#" alt="Orange" /></td>
</tr>
<tr>
<td>2701</td>
<td><img src="#" alt="Blue" /></td>
<td><img src="#" alt="Orange" /></td>
<td><img src="#" alt="Blue" /></td>
<td><img src="#" alt="Orange" /></td>
<td><img src="#" alt="Orange" /></td>
<td><img src="#" alt="Blue" /></td>
<td><img src="#" alt="Orange" /></td>
</tr>
<tr>
<td>2801</td>
<td><img src="#" alt="Blue" /></td>
<td><img src="#" alt="Orange" /></td>
<td><img src="#" alt="Blue" /></td>
<td><img src="#" alt="Orange" /></td>
<td><img src="#" alt="Orange" /></td>
<td><img src="#" alt="Blue" /></td>
<td><img src="#" alt="Orange" /></td>
</tr>
<tr>
<td>2901</td>
<td><img src="#" alt="Blue" /></td>
<td><img src="#" alt="Orange" /></td>
<td><img src="#" alt="Blue" /></td>
<td><img src="#" alt="Orange" /></td>
<td><img src="#" alt="Orange" /></td>
<td><img src="#" alt="Blue" /></td>
<td><img src="#" alt="Orange" /></td>
</tr>
<tr>
<td>2902</td>
<td><img src="#" alt="Blue" /></td>
<td><img src="#" alt="Orange" /></td>
<td><img src="#" alt="Blue" /></td>
<td><img src="#" alt="Orange" /></td>
<td><img src="#" alt="Orange" /></td>
<td><img src="#" alt="Blue" /></td>
<td><img src="#" alt="Orange" /></td>
</tr>
<tr>
<td>3001</td>
<td><img src="#" alt="Blue" /></td>
<td><img src="#" alt="Orange" /></td>
<td><img src="#" alt="Blue" /></td>
<td><img src="#" alt="Orange" /></td>
<td><img src="#" alt="Orange" /></td>
<td><img src="#" alt="Blue" /></td>
<td><img src="#" alt="Orange" /></td>
</tr>
<tr>
<td>3002</td>
<td><img src="#" alt="Blue" /></td>
<td><img src="#" alt="Orange" /></td>
<td><img src="#" alt="Blue" /></td>
<td><img src="#" alt="Orange" /></td>
<td><img src="#" alt="Orange" /></td>
<td><img src="#" alt="Blue" /></td>
<td><img src="#" alt="Orange" /></td>
</tr>
</tbody>
</table>
Our analysis methods
Our analysis methods
Our analysis methods
Our analysis methods
Design implications

How can we support observed behaviors?

Priorities?

Motivation for other research?
How would you craft a formative study for your term project?