Research Methods

Evan Strasnick
CS 347
Your paper is an argument*

Your methods provide evidence
Different arguments require different evidence
Figure 2: The strategy circumplex (adapted from Runkel & McGrath).

Maxima for each criterion:
- A = Generalizability
- B = Precision
- C = Realism

Quadrant I: Field Strategies
- Field Experiment
- Field Study
- Computer Simulation
- Formal Theory

Quadrant II: Experimental Strategies
- Laboratory Experiment
- Experimental Simulation
- Judgment Study
- Sample Survey

Quadrant III: Respondent Strategies
- Formulation
- Analysis
- Observation

Quadrant IV: Theoretical Strategies
- Conceptual
- Propositional
- Analytical

Abstract

Unobtrusive

Concrete
Method triangulation

All methods are flawed
...but multiple methods can support each other!

E.g. complement your statistics with semi-structured interviews
E.g. complement qualitative work with primary source evidence or log data
How do we decide which methods to use?
## Common claims – Systems

<table>
<thead>
<tr>
<th>I built a system that...</th>
<th>Likely questions</th>
<th>Possible methods</th>
</tr>
</thead>
<tbody>
<tr>
<td>...solves an entirely new problem</td>
<td>Is the problem important? (How well) does it work?</td>
<td>Field study, lab experiment, technical evaluation</td>
</tr>
<tr>
<td>...solves an old problem more effectively</td>
<td>(How well) does it work? How much better is it?</td>
<td>Technical evaluation, lab experiment, field experiment</td>
</tr>
<tr>
<td>...improves task performance</td>
<td>By how much? Under what circumstances?</td>
<td>Lab experiment, formal theory, judgement study</td>
</tr>
<tr>
<td>...lowers the threshold/raises the ceiling/widens the walls</td>
<td>What can it now make? Who can now make it?</td>
<td>Interviews, Demonstrative applications, long-term deployment</td>
</tr>
<tr>
<td>...is more accessible</td>
<td>Who can now use it? How much better is it?</td>
<td>Interviews, field study, field experiment, sample survey</td>
</tr>
<tr>
<td>I hypothesize that...</td>
<td>Likely questions</td>
<td>Possible methods</td>
</tr>
<tr>
<td>-----------------------</td>
<td>------------------</td>
<td>------------------</td>
</tr>
<tr>
<td>...people behave in accordance with model X</td>
<td>How do you know?</td>
<td>Field study, formal theory, experimental simulation, field experiment</td>
</tr>
<tr>
<td>...we can get better outcomes using mechanism Y</td>
<td>What other factors might be at play?</td>
<td>Lab experiment, field experiment, sample survey, experimental simulation</td>
</tr>
<tr>
<td>...dimension X plays a significant role in how people interact with system Y</td>
<td>How can you be sure?</td>
<td>Field study, field experiment, sample survey</td>
</tr>
<tr>
<td>...understanding system X can inform us about broader problem Y</td>
<td>How much better?</td>
<td>Field study, formal theory, field experiment</td>
</tr>
<tr>
<td></td>
<td>Why do you think the two are sufficiently similar?</td>
<td></td>
</tr>
</tbody>
</table>
Determining your methods

Your methods = Your claims + Standards of evidence in your area
Standards of evidence

Every field has an accepted standard of evidence — a set of methods that are agreed upon for proving a point:

- Medicine: Double-blind randomized controlled trial
- Philosophy: Rhetoric
- Math: Formal proof
- Applied Physics: Measurement
Standards of evidence

In computing, because areas use different methods, the standard of evidence differs based on the area.

Your goal: convince an expert in your area
So, use the methods that those experts expect.
Don’t reinvent the wheel

There’s no need to start from scratch on this.

Your nearest neighbor paper, and the rest of your literature search, has likely already introduced evaluation methods into this literature that can be adapted to your purpose.

Start here: figure out what the norms are, and tweak them. Talk to your TA if helpful.
Designing an evaluation
Problematic point of view

“But how would we evaluate this?”

Why is this point of view problematic?

Implication: “I believe the idea is right, but I don’t believe that we can prove it.”

Implication: “The thread of designing the evaluation is separate from the process of claiming the idea.”

Neither implication is correct. If you can precisely articulate your idea and your claim, then you can design an appropriate evaluation. If you can’t design an appropriate evaluation, then you haven’t precisely articulated your idea and your claim.
A better way: derive evaluation from your thesis
Step 1: Articulate your thesis

**Bit**
Labeling images is a tedious task, so the only way to get hand-labeled data is by paying workers.

The best gestural interactions result from the careful planning of an expert designer.

**Flip**
If we create an entertaining game that produces image labels, players will voluntarily label lots of images.

Elicitation from non-expert users can produce better gesture sets.
Step 2: Map your thesis onto a claim

There are only a small number of claim structures implicit in most theses:

- $x > y$: approach $x$ is better than approach $y$ at solving the problem
- $\exists x$: it is possible to construct an $x$ that satisfies some criteria, whereas it was not known to be possible before
- bounding $x$: approach $x$ only works given certain assumptions (i.e. has limitations)
Bit
Labeling images is a tedious task, so the only way to get hand-labeled data is by paying workers.

The best gestural interactions result from the careful planning of an expert designer.

Flip
If we create an entertaining game that produces image labels, players will voluntarily label lots of images.

Elicitation from non-expert users can produce better gesture sets.

Claim
∃ x: games can both yield high-quality image labels and be sufficiently fun that users will play voluntarily.

x > y: gestures elicited from non-technical users will have better coverage and agreement than those designed by experts.
Step 3: claims imply an evaluation design

Each claim structure implies an evaluation design:

- $x > y$: given a representative task or set of tasks, test whether $x$ in fact outperforms $y$ at the problem

- $\exists x$: demonstrate that your approach achieves $x$

- bounding $x$: demonstrate bounds inside or outside of which approach $x$ fails
<table>
<thead>
<tr>
<th>Flip</th>
<th>Claim</th>
<th>Implied evaluation</th>
</tr>
</thead>
<tbody>
<tr>
<td>If we create an entertaining game that produces image labels, players will voluntarily label lots of images.</td>
<td>∃ x: games can both yield high-quality image labels and be sufficiently fun that users will play voluntarily.</td>
<td>Demonstrate a game that produces image labels judged as high quality, and that users voluntarily play.</td>
</tr>
<tr>
<td>Elicitation from non-expert users can produce better gesture sets.</td>
<td>x &gt; y: gestures elicited from non-technical users will have better coverage and agreement than those designed by experts.</td>
<td>Compare coverage and agreement scores of gesture sets elicited from non-technical users and those designed by experts.</td>
</tr>
</tbody>
</table>
Let’s play a game
Guess the evaluation

Flip

We can encourage users to lead more active lifestyles via an ambient interface which detects physical activity and displays progress through a calm narrative.

Implied evaluation

1) “Can accurately classify” – Validate classification of user activity by comparing it to a manually recorded activity log
2) “Feel encouraged to adopt healthier habits” – Survey users’ attitudes towards the interface and observe their exercise habits over a time period

Claim

∃ x: an activity-sensing wearable device can accurately classify and present an ambient summary of users’ recent activity levels, such that users feel encouraged to adopt healthier habits.
Guess the evaluation

Flip
Instead of teaching a design cycle focused on repeatedly iterating on a given design, we might get better results by iterating less on more designs in parallel.

Implied evaluation
1) “more successful designs” – Measure the success of designs produced for their target function, in this case, by measuring the click-through rates of designed advertisements.

Claim
$x > y$: Designers will produce more successful designs by iterating on multiple in parallel, rather than by performing more iterations on a single design.
Architecture of an evaluation
Four constructs that matter

Dependent variable
Independent variable
Task
Threats
DV: dependent variable

In other words, what's the outcome you're measuring? Efficiency? Accuracy? Performance? Satisfaction? Trust? The choice of this quantity should be clearly implied by your thesis. Then, all that remains is to operationalize it. It’s often tempting to:

• ...measure many DVs. Instead, let one be your central outcome, and the others auxiliary.

• ...choose DVs that are easily quantifiable (clicks, time, completions). However, selecting DVs based on what we can easily measure often misses the point. Is your claim about clicks?
In other words, what determines what x and y are? What are you manipulating in order to cause the change in the dependent variable?

The IV leads to *conditions* in your evaluation. Examples might include:

- Algorithm
- Dataset size or quality
- Interface
Task

What, specifically, is the routine being followed in order to manipulate the independent variable and measure the dependent variable?

E.g. “Participants will have thirty seconds to identify each article as disinformation or not, within-subjects, randomizing across interfaces”
Threats

What are your threats to validity? Internal validity? External validity?

- Might your participants feel experimenter demand?
- Are your participants biased toward healthy young technophiles?
- Do your participants always see the best interface first?
- Is there some other variable (confound) responsible for differences you see (e.g. one interface is easier to use)?
Threats

Ways to handle these kinds of issues:

1) Manipulate – turn it into an IV
2) Control – equalize across groups through stratification or randomization
3) Measure – record the confound to later account for it statistically
4) Argue as irrelevant – yes, that bias might exist, but it’s not conceptually important to the phenomenon you’re studying and is unlikely to strongly effect the outcome or make the results less generalizable
Reminder: Stats review!

Feeling less than confident about statistical analyses? Don’t know when to use a nonparametric test or how to correct for family-wise error rate? Come to Littlefield 103 at 6 p.m. tonight for a crash course!
Discussion!