Experimentation

goal: understand causal relationship between X and Y

first question: are X and Y measurable?

relationship between

X
programming language
new debugging tool
image recognition algorithm
new 3D shader

Y
clarity of code
correctness of code
accuracy of recognition
realism of objects
Given our interest in $X$ and $Y$, we have to find measurable $\hat{X}$ and $\hat{Y}$ and can then run an experiment on $\hat{X}$ and $\hat{Y}$.

Closeness between $X$ and $\hat{X}$ and $Y$ and $\hat{Y}$ affects our EXTERNAL VALIDITY.

$\hat{X}$ and $\hat{Y}$ must be measurable in a REPEATABLE way.
Indep. Var. Considerations

You need to define the treatment levels for your indep. var. $X$

- Presence or absence of a tool
- Different cache sizes
- Tool A vs. Tool B vs nothing
**Dependent Variable Considerations**
- measurable
- repeatable

You may have to prove your measurements are reliable.

→ Analyze your data multiple times by different people and then report correlation

→ Have multiple ways to measure dep. var. and show they correlate
More Dep. Var. Considerations

Could your dep. var. be confounded with other variables? If so, you must at least measure them all.

(Better still to control the confounding var. in some way).

\[
\text{indep. var.} \quad \rightarrow \quad \text{dep. var.}
\]

size of buttons \quad \rightarrow \quad \text{accuracy of pressing}

width of fingers

speed of pressing the buttons

learning effects

OK experiment?
Between Subjects vs. Within Subjects

Between subjects: each study participant only exposed to one treatment level of independent variable.

Pros:
- shorter study session for each participant
- no learning or history effects

Cons:
- groups assigned to each treatment level may not be the same
Matched Between Subjects Designs

(Java) Treatment Level A
      Subj 1 ←→ Subj 5
      Subj 2 ←→ Subj 6
      Subj 3 ←→ Subj 7
      Subj 4 ←→ Subj 8

(C) Treatment Level B

Match: equal score on variable closely related to dep. var.

4.0 gpa
3.5 gpa
2.5 gpa

4.0 gpa
3.5 gpa
2.5 gpa
Within-Subject Experiments

All participants see all levels.

Pros:
- Fewer participants needed.
- No worries about differences between groups at each treatment level

Cons:
- Learning, history, order effects → fatigue → boredom
- Range effects: Learning tends to improve performance on the treatment level most like the other levels
Counterbalancing

give some participants AB
give others BA

- this assumes that order effects are symmetric
  (often NOT TRUE)

Complete counterbalancing  \(4! = 24\)

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>B</td>
<td>C</td>
<td>D</td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>B</td>
<td>D</td>
<td>C</td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>C</td>
<td>D</td>
<td>B</td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>C</td>
<td>B</td>
<td>D</td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>D</td>
<td>B</td>
<td>C</td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>D</td>
<td>C</td>
<td>B</td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>D</td>
<td>C</td>
<td>B</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Balanced Latin Square Countercalancing

- each level at each position equally often,
  and
- each level precedes or follows equally often.

1. A B C D
2. C A D B
3. B D A C
4. D C B A
Range effects

Learning effects tend to help performance most on the "average level" because what you learn on different levels is most transferable to a "similar level".

ex: say treatment levels are screens of different sizes:
  13 in 15 in 17 in 19 in 21 in
  even if you counterbalance, learning effects will benefit 17 in the most
Context Effects (affects both within-subjects and between-subjects)

People's past experiences affect results too.

25 people all with a 15-in monitor will do best on a 15-in treatment level.

Confounding variable of context
How many treatment levels?

You CANNOT infer ANYTHING about the effect of the indep. var. except at the treatment levels you actually test.

More treatment levels

→ more info about relationship between indep. and dep. var
→ more complicated stats
How many indep. variables?

If > 1, a more complicated experiment is needed.

Say indep. vars. are screen size and button size.

Design a factorial experiment

- m levels of indep. var. A
- n levels of indep. var. B

\[ \text{mn cells in your factorial experiment, put subjects in each cell.} \]

<table>
<thead>
<tr>
<th>12 in</th>
<th>15 in</th>
<th>17 in</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subj1</td>
<td>Subj3</td>
<td>Subj5</td>
</tr>
<tr>
<td>Subj2</td>
<td>Subj4</td>
<td>Subj6</td>
</tr>
<tr>
<td>0.25 in</td>
<td>0.5 in</td>
<td></td>
</tr>
<tr>
<td>Subj7</td>
<td>Subj9</td>
<td>Subj11</td>
</tr>
<tr>
<td>Subj8</td>
<td>Subj10</td>
<td>Subj12</td>
</tr>
</tbody>
</table>

\[ \text{between-subjects or within-subjects.} \]
Any difference between one factorial study and two studies with one indep. var.? 

<table>
<thead>
<tr>
<th>0.25 in</th>
<th>12 in</th>
<th>15 in</th>
<th>17 in</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>0.5 in</th>
<th>12 in</th>
<th>15 in</th>
<th>17 in</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>