CS/INFO 4154:
Analytics-driven Game Design

Lecture 11:

A/B Testing

Today

- Updates
- General presentation advice
- A/B Testing

Friends Release Tuesday 11/3

- Fixes discussed during team meetings
- Fifteen levels
- Tutorial messages
- Music
- Sound effects
- Logging

Not needed for Friends Release

- A/B test

Friends Postmortems Tuesday 11/10

Engage the audience

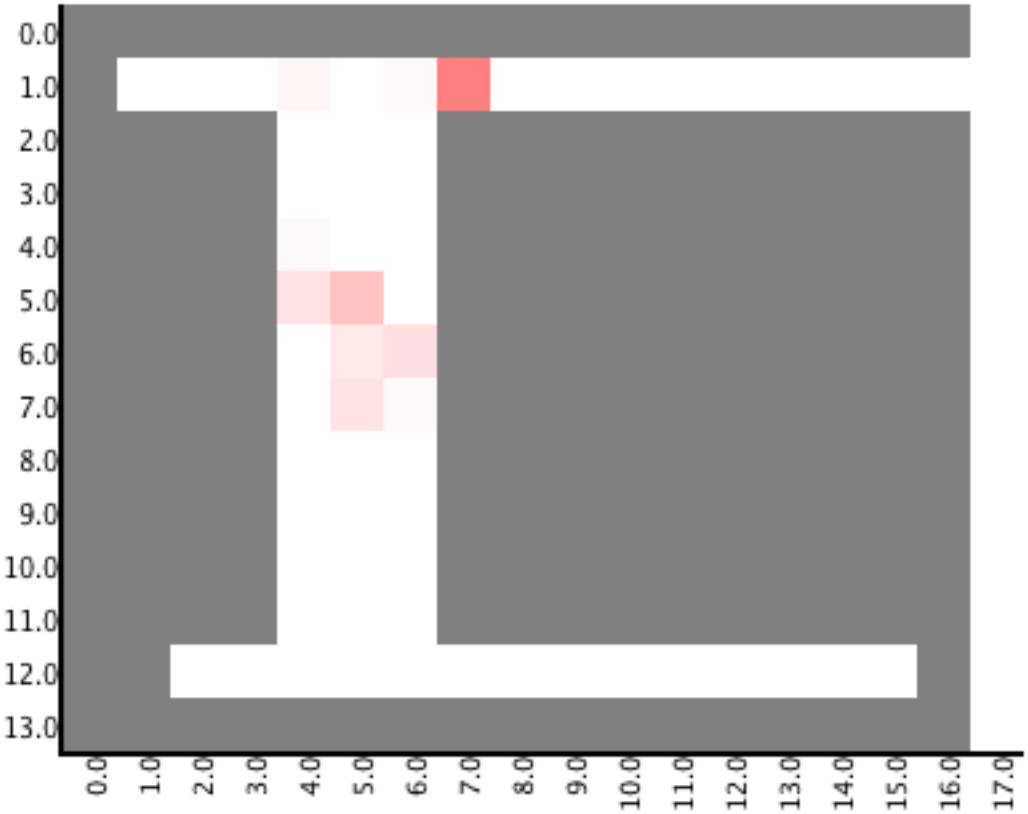
- Use sufficient volume

Use sufficient volume

Engage the audience

- Use sufficient volume
- Organize and motivate

Motivate


0.0
1.0
2.0
3.0
4.0
5.0
6.0
7.0
8.0
9.0
10.0
11.0
12.0
13.0
0.0
1.0
2.0
3.0
4.0
5.0
6.0
7.0
8.0
9.0
10.0
11.0
12.0
13.0
14.0
15.0
16.0
17.0

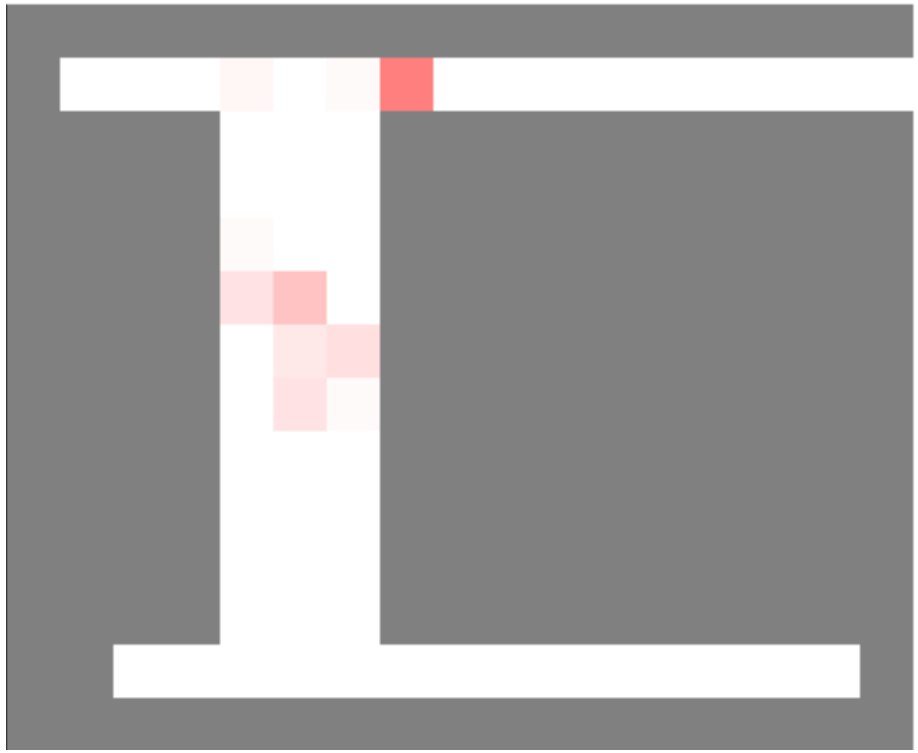
Engage the audience

- Use sufficient volume
- Organize and motivate
- Minimize cognitive load

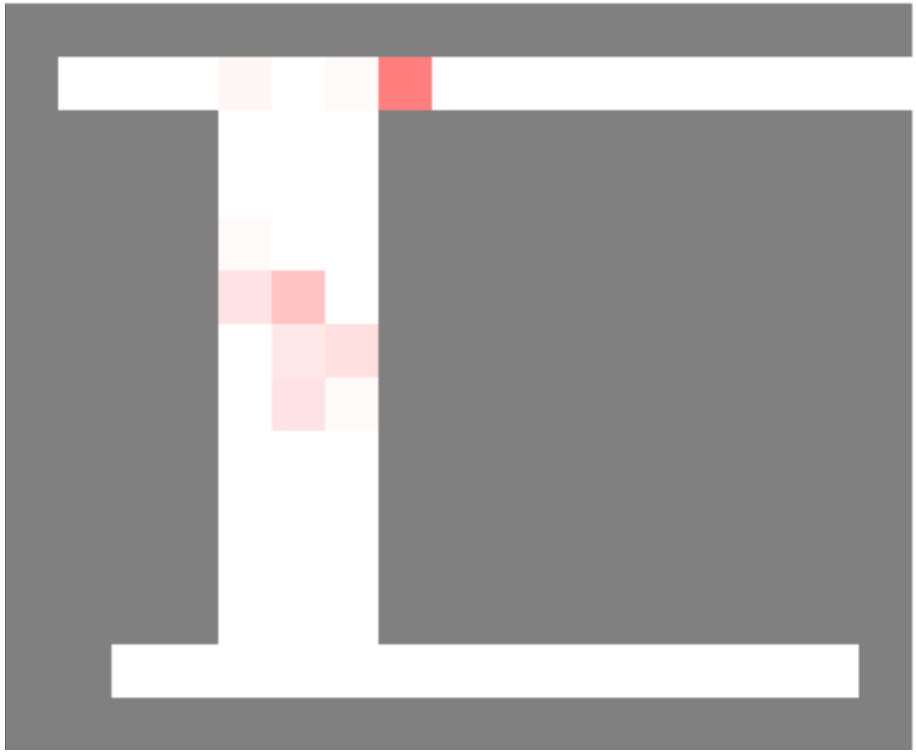
Avoid Death by Powerpoint

- People often put every word they are going to say on their slides
- It’s terrible
- Why?
  - No one can read this fast
  - It’s stressful to even have to look at this much text
  - No one can read and listen at the same time
    - If your audience is reading, ***they aren’t listening to you***

Remove unnecessary information



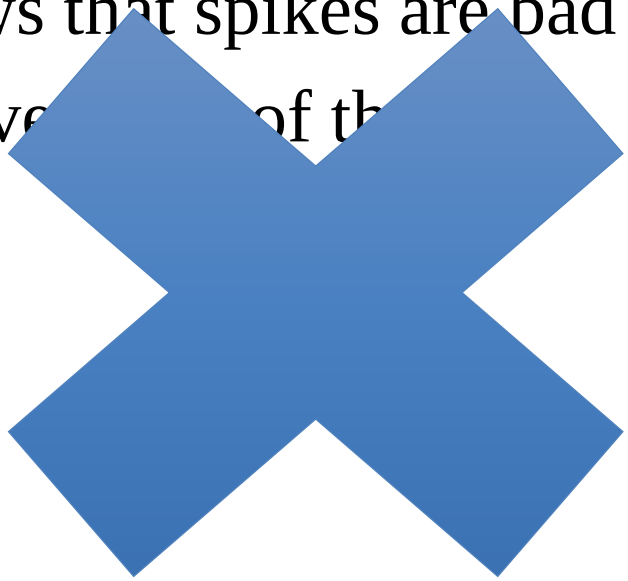
Include necessary information



Use pictures

Fixing the Spikes

- Our data shows that spikes are bad
- We will remove [illegible] of th



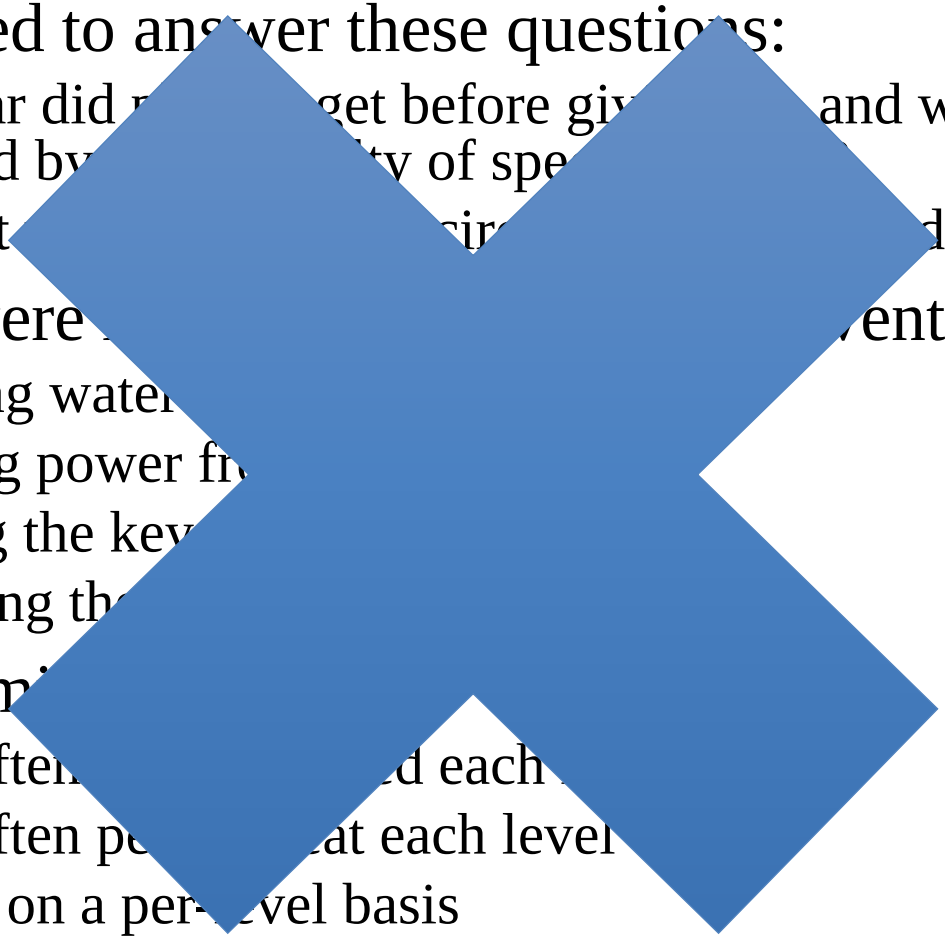
Plan: Reduce Spikes

For bulleted lists

- Simplify each point
- Bring in one-by-one
- This minimizes cognitive load

Logging Goals & Summary

- We wanted to answer these questions:
  - How far did
  - In what
- Entries were
  - Entering
  - Gaining power
  - Getting the
  - Finishing
- Data
  - How
  - How often
  - Events on a per-level basis



Logging Goals & Summary

- We wanted to answer these questions:
  - How far did players get before giving up, and was this affected by the difficulty of specific levels?
  - In what ways did players circumvent the intended solution?
- Entries were logged for the following events:
  - Entering water
  - Gaining power from a gate
  - Getting the key
  - Finishing the level
- Data examined:
  - How often people played each level
  - How often people beat each level
  - Events on a per-level basis

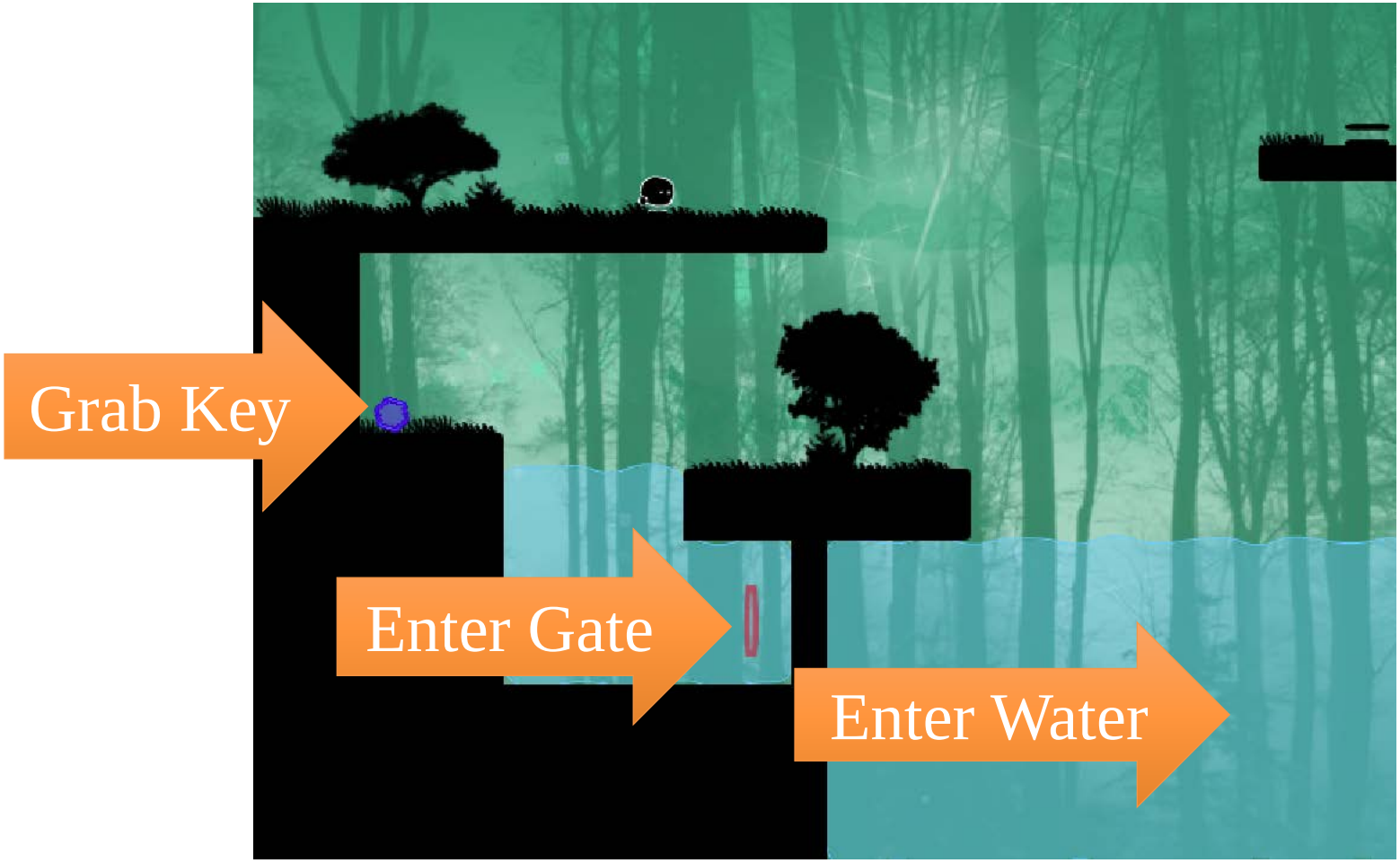
Design Questions

- How far did players get?
- Did they circumvent the intended solution?

Logging Goals & Summary

- We wanted to answer these questions:
  - How far did players get before giving up, and was this affected by the difficulty of specific levels?
  - In what ways did players circumvent the intended solution?
- Entries were logged for the following events:
  - Entering water
  - Gaining power from a gate
  - Getting the key
  - Finishing the level
- Data examined:
  - How often people played each level
  - How often people beat each level
  - Events on a per-level basis

Events

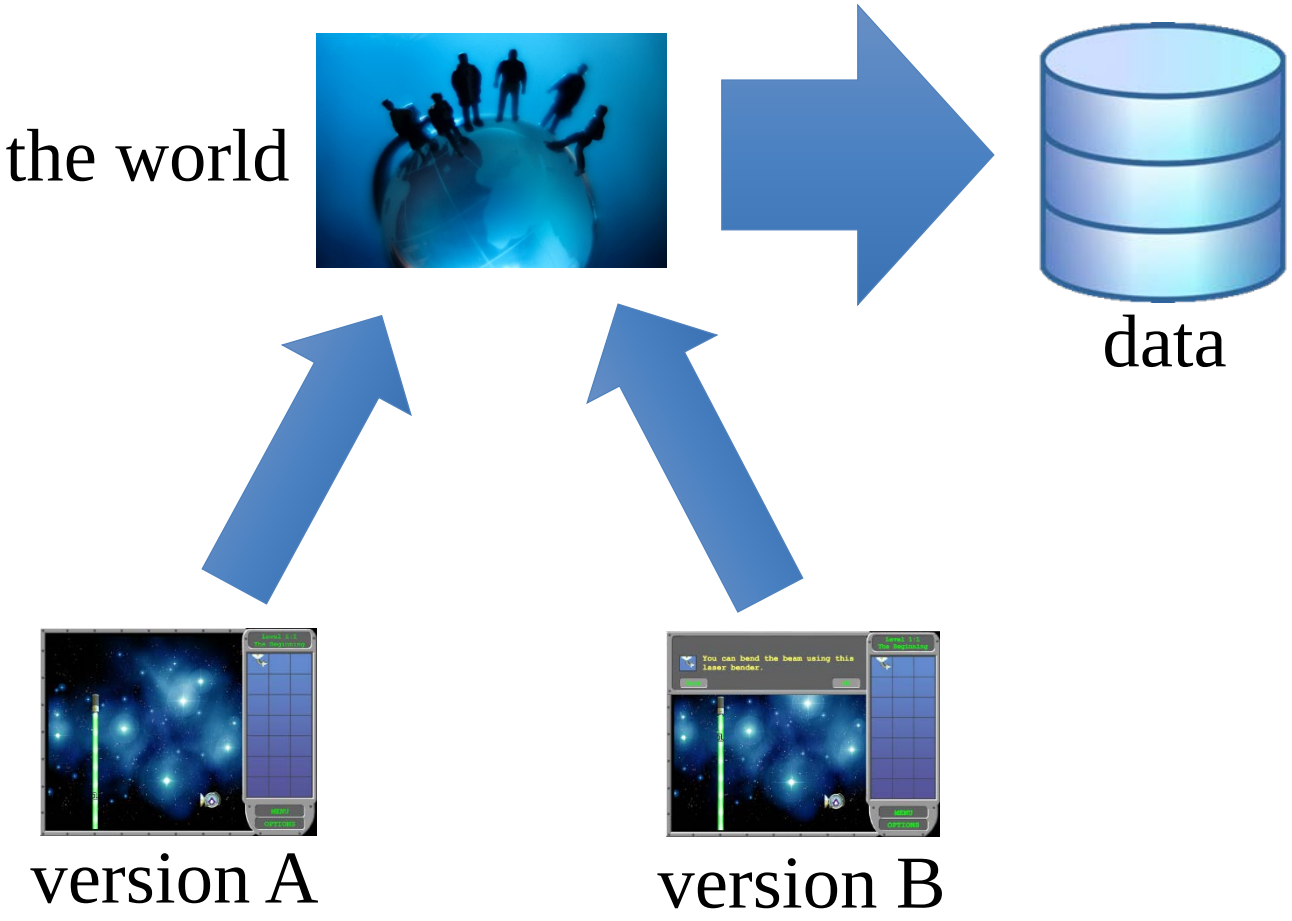


Don’t forget to summarize

- Use sufficient volume
- Organize and motivate
- Minimize cognitive load

Design decisions are a pain

A/B Testing



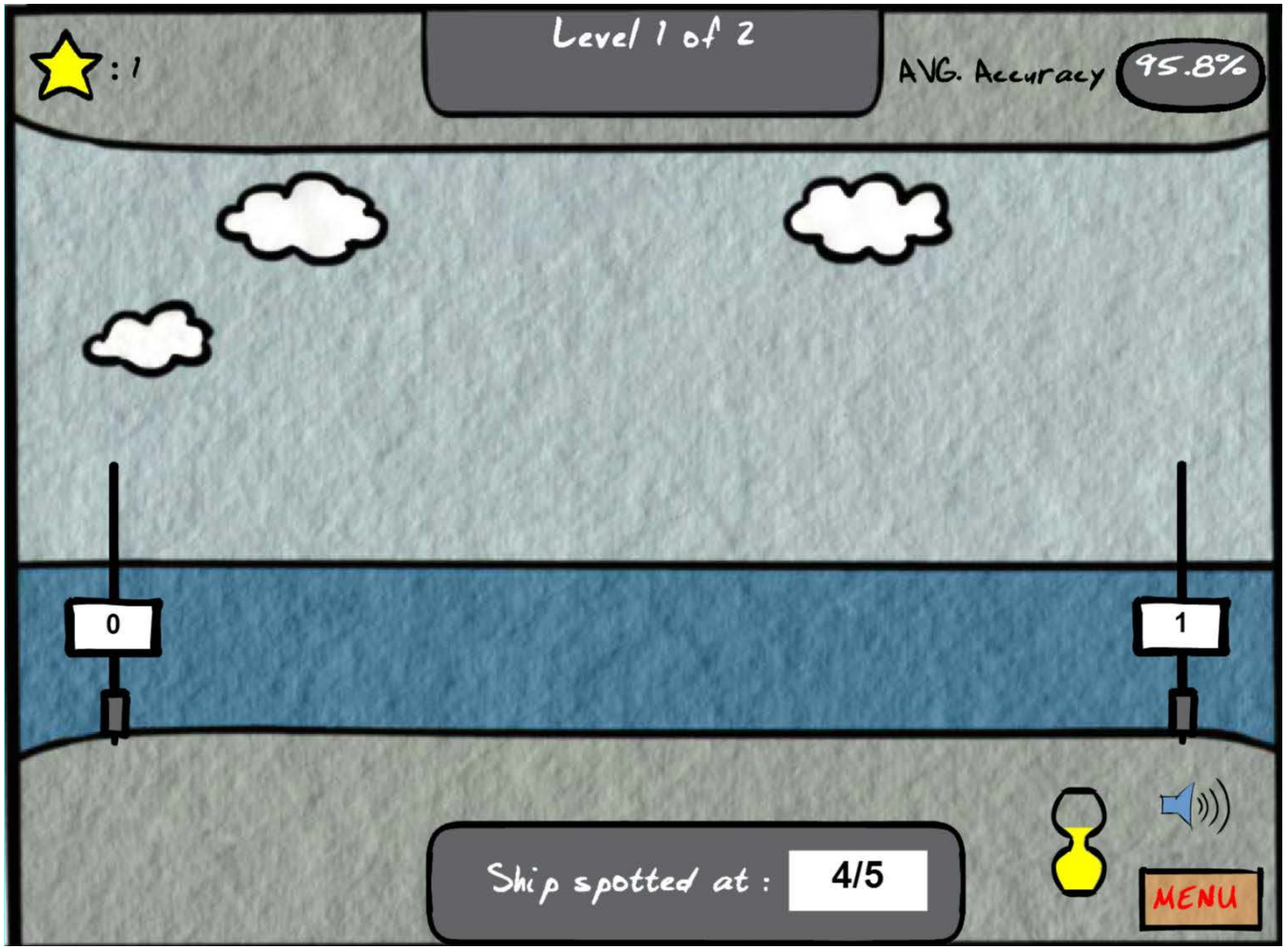
GSN Games A/B Testing

Revenue +12.3%

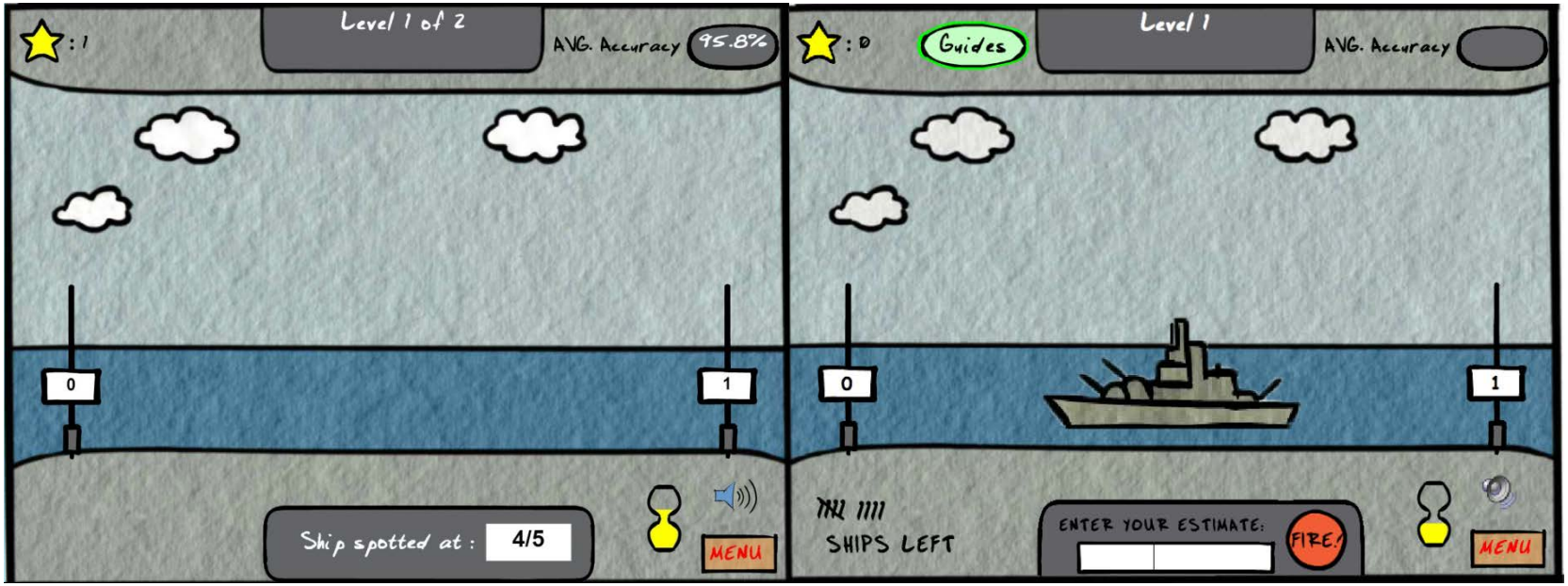
GSN Games A/B Testing

Revenue +11%

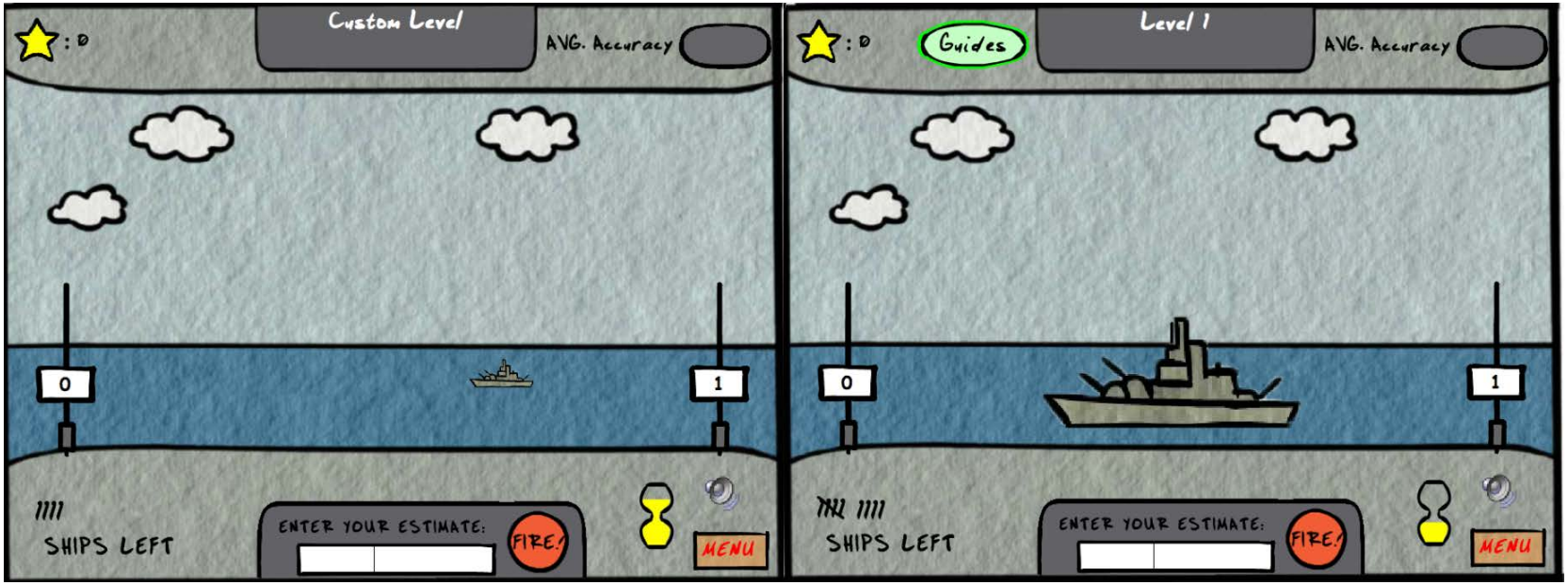
Battleship Numberline



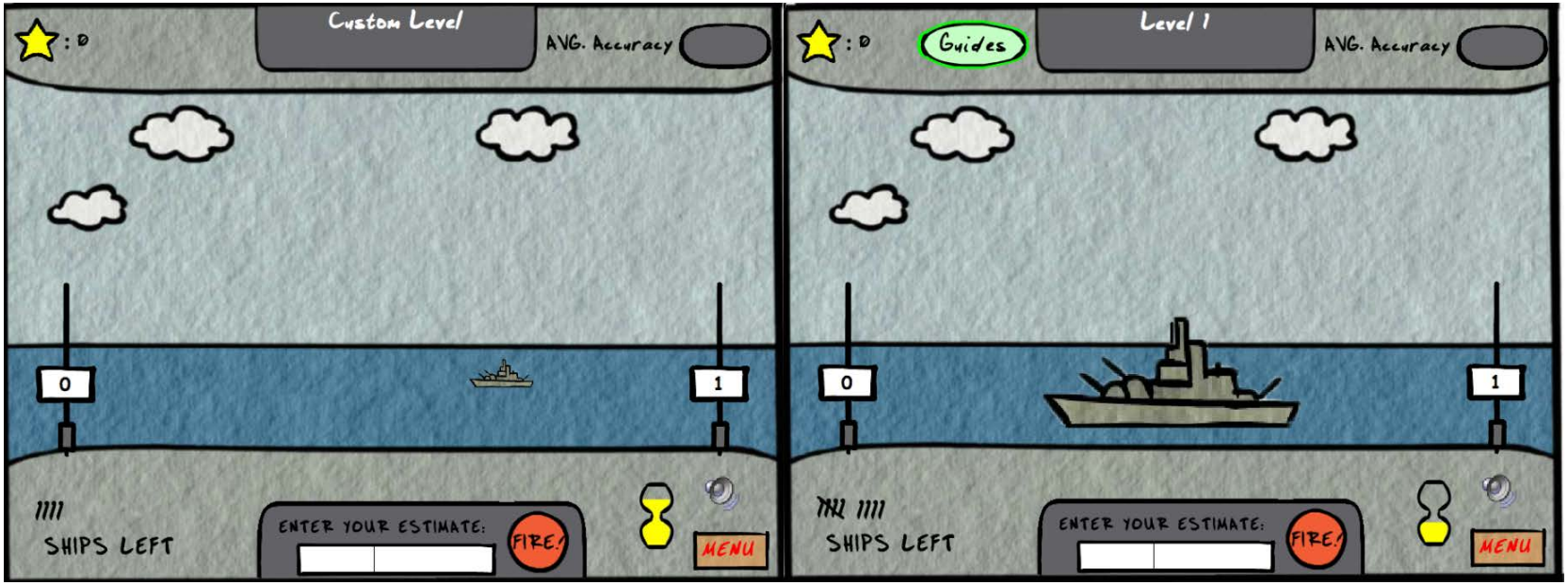
Optimal game type?



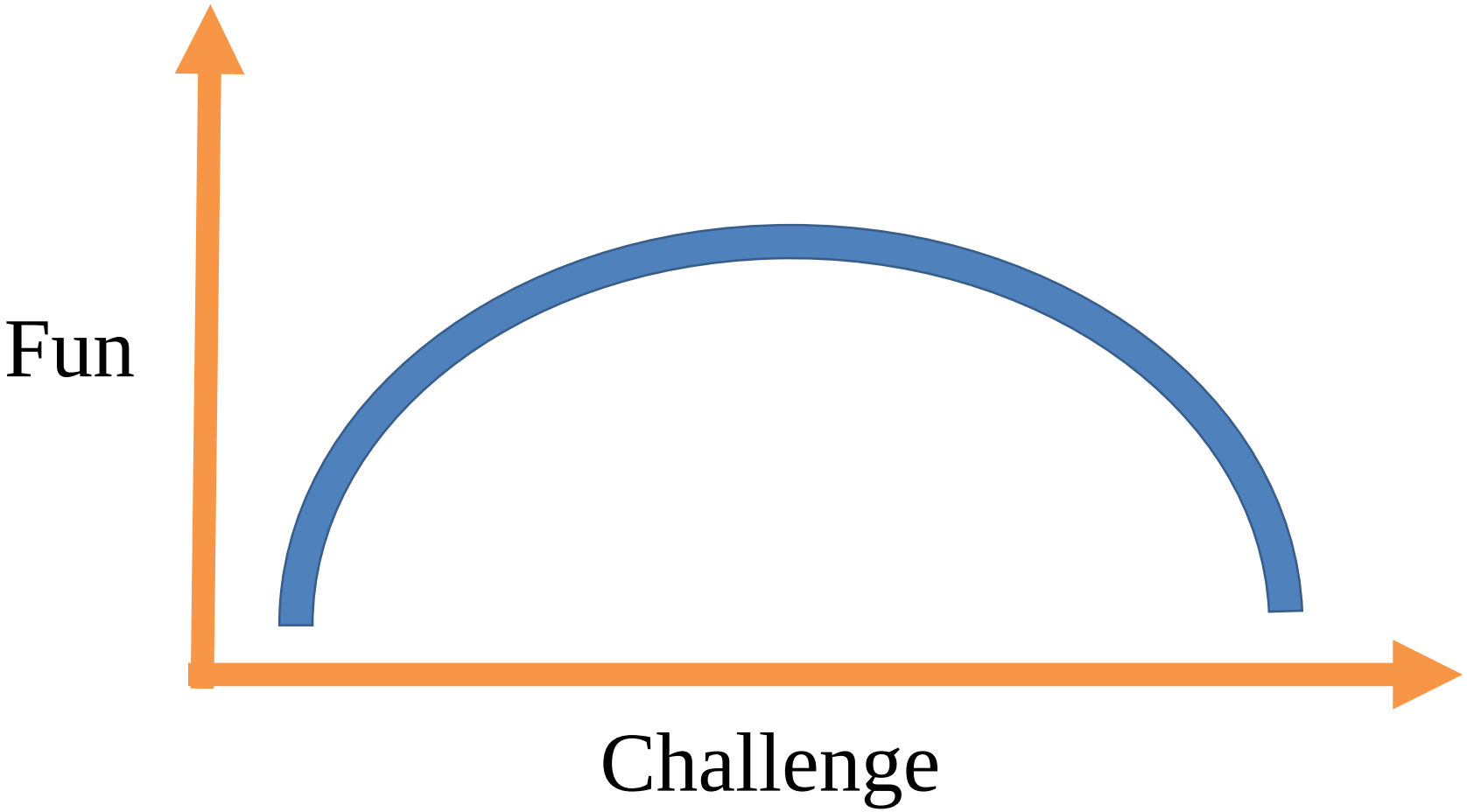
Optimal ship size?



Optimal time limit?



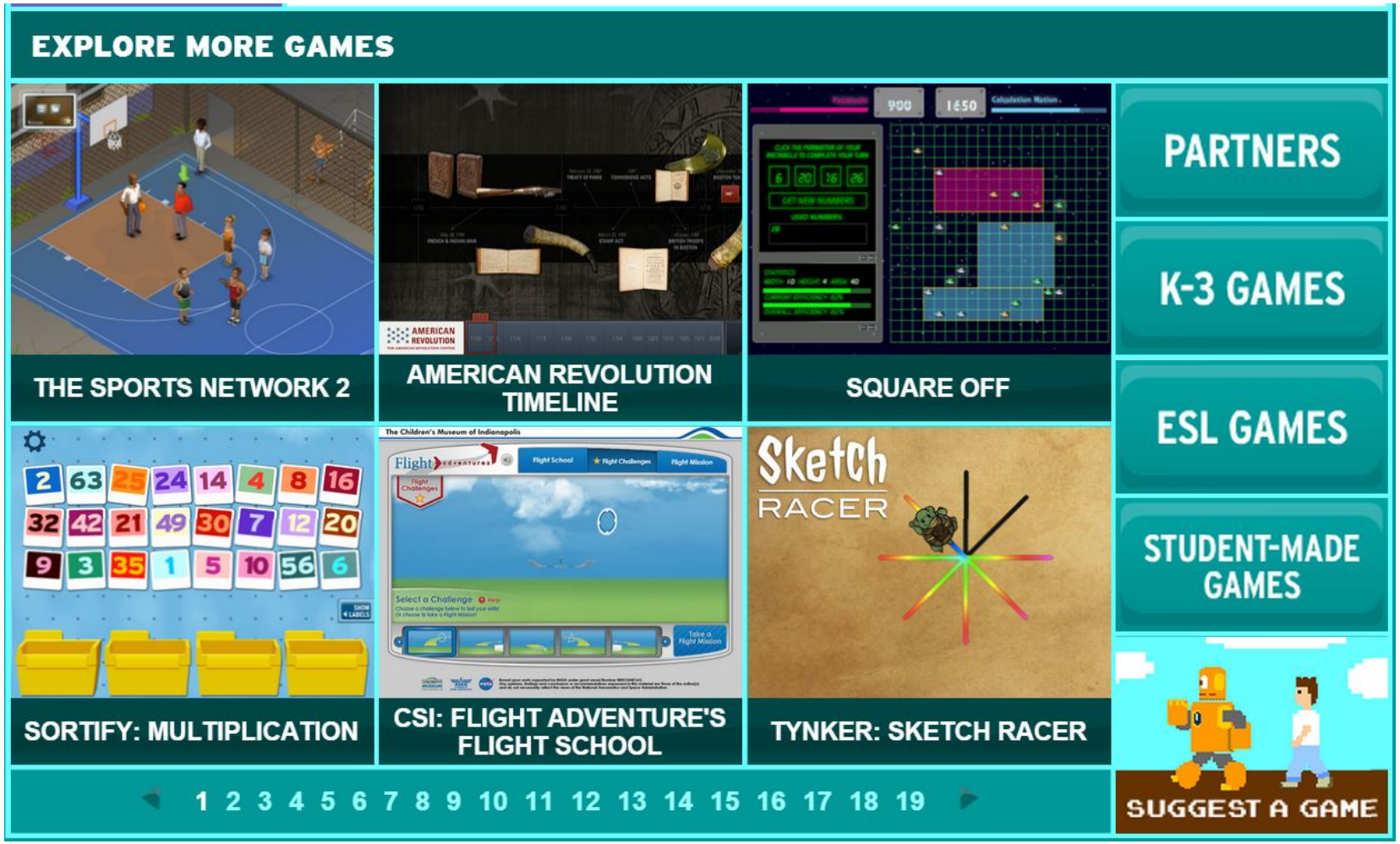
How much challenge?



Measures of engagement

- Engagement
  - Time played
  - Log(time * attempts)
- Challenge
  - Probability of success?

BrainPop

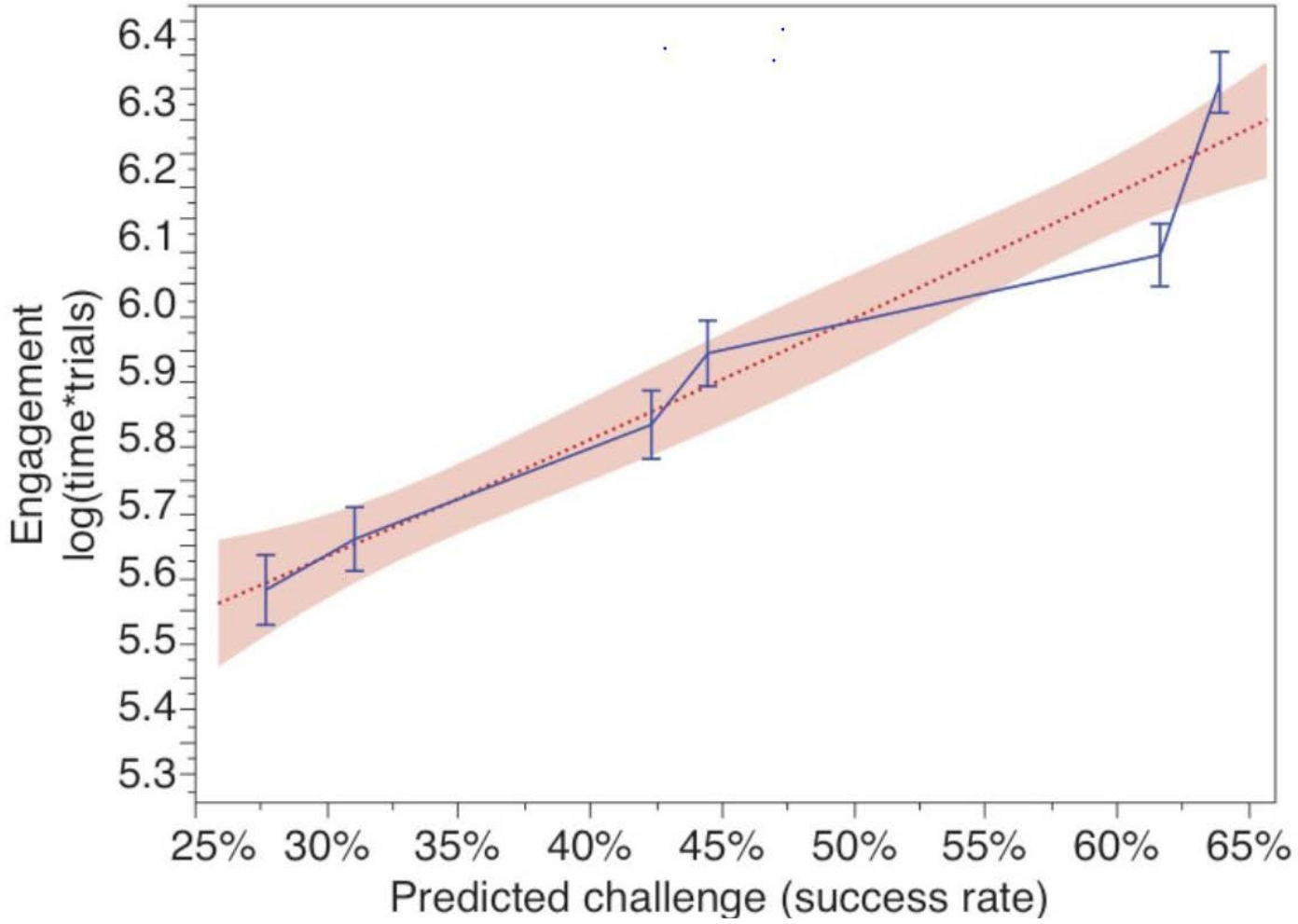

EXPLORE MORE GAMES
THE SPORTS NETWORK 2
AMERICAN REVOLUTION TIMELINE
SQUARE OFF
PARTNERS
K-3 GAMES
ESL GAMES
STUDENT-MADE GAMES
SORTIFY: MULTIPLICATION
CSI: FLIGHT ADVENTURE'S FLIGHT SCHOOL
TYNKER: SKETCH RACER
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19
SUGGEST A GAME

Experiment: 70,000 people

- Ship types: submarine and battleship
- Ship sizes: 4, 6, 8, 10, 16, 20, 24, 30, 40%
- Time limits: 2, 3, 4, 5, 8, 10, 15, 30 seconds

Results

- Clicking on target = more time played
- Bigger target = more time played
- Longer time limit = more time played


Engagement
log(time*trials)
6.4
6.3
6.2
6.1
6.0
5.9
5.8
5.7
5.6
5.5
5.4
5.3
25%
30%
35%
40%
45%
50%
55%
60%
65%
Predicted challenge (success rate)

A/B testing logistics

- What conditions?
- How to track players?
- Players per condition?

A/B testing in this class

1. Wait for logging update from Kelvin
2. Call reportPageLoad first
3. Then call recordABTestingValue
4. **Use the return value** of recordABTestingValue to set the condition

recordABTestingValue

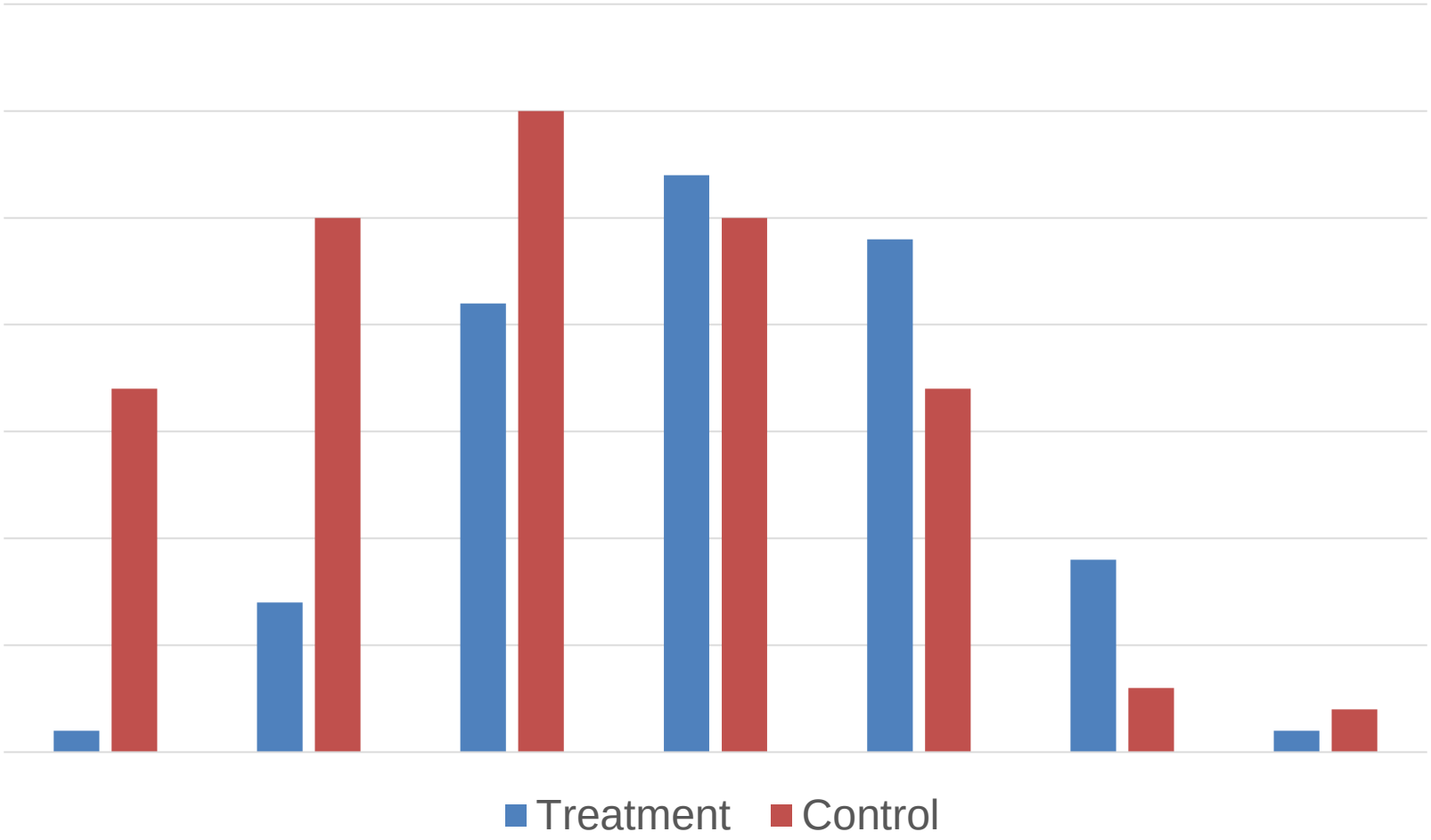
- Goals:
  - New condition for new players
  - Previous condition for returning players
- Takes a single parameter: ***proposed*** integer condition

```
var proposedCondition:int = Math.Floor(Math.Random() * 2) + 1;
var actualCondition:int = recordABTestingValue(proposedCondition);
if (actualConditon == 1) { … }
else if (actualCondition == 2) { … }
```

Assignment of players to condition

- 50% of players to each condition

Null hypothesis testing



Typical problem

- Version A
  - 100 players
  - Average time played: 120 seconds
  - Standard deviation: 5 points
- Version B
  - 100 players
  - Average time played: 105 seconds
  - Standard deviation: 5 points
- **What is the probability of obtaining a result at least this extreme?**

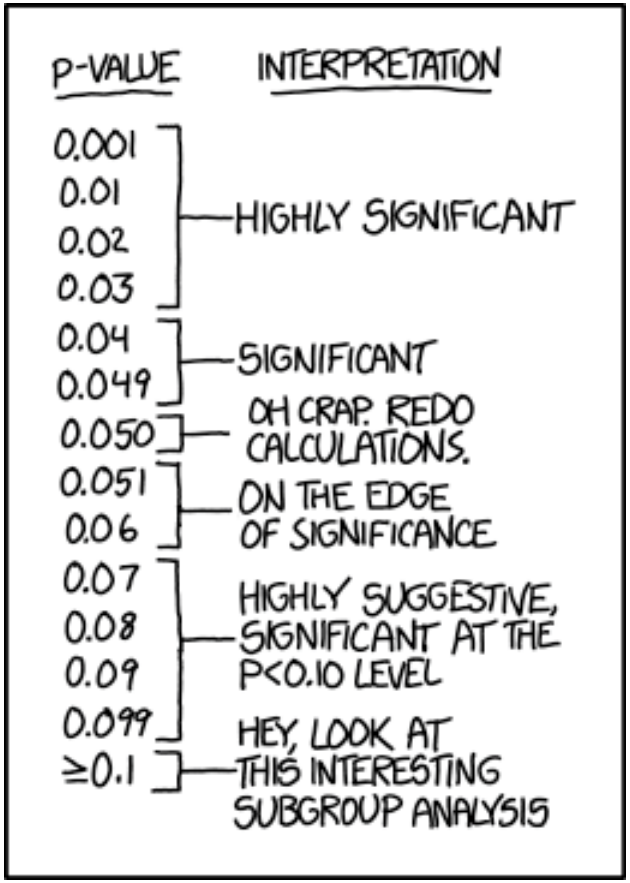
p-value

probability the null hypothesis is true

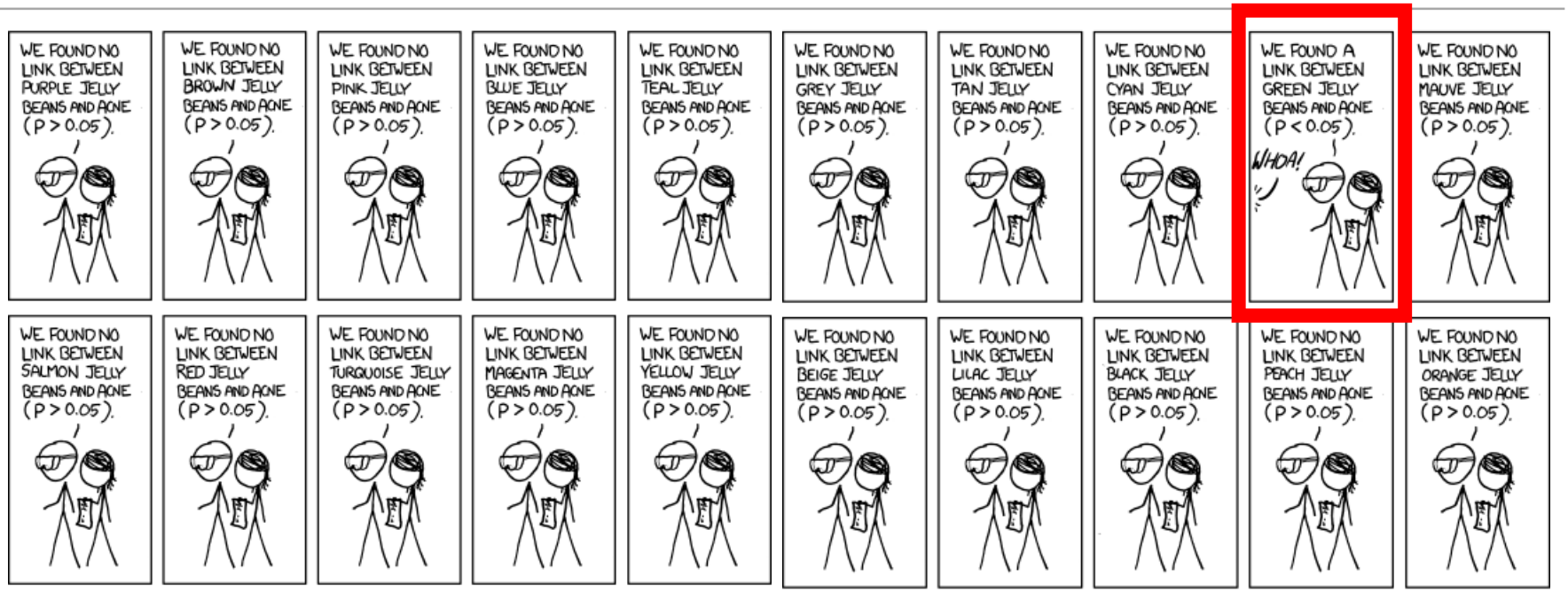
Common fallacies

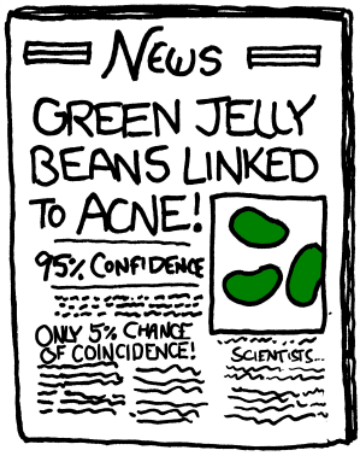
- lower p value means stronger effect
- p >= 0.05 means no effect

p-values (source: XKCD)



Fishing (source: XKCD)

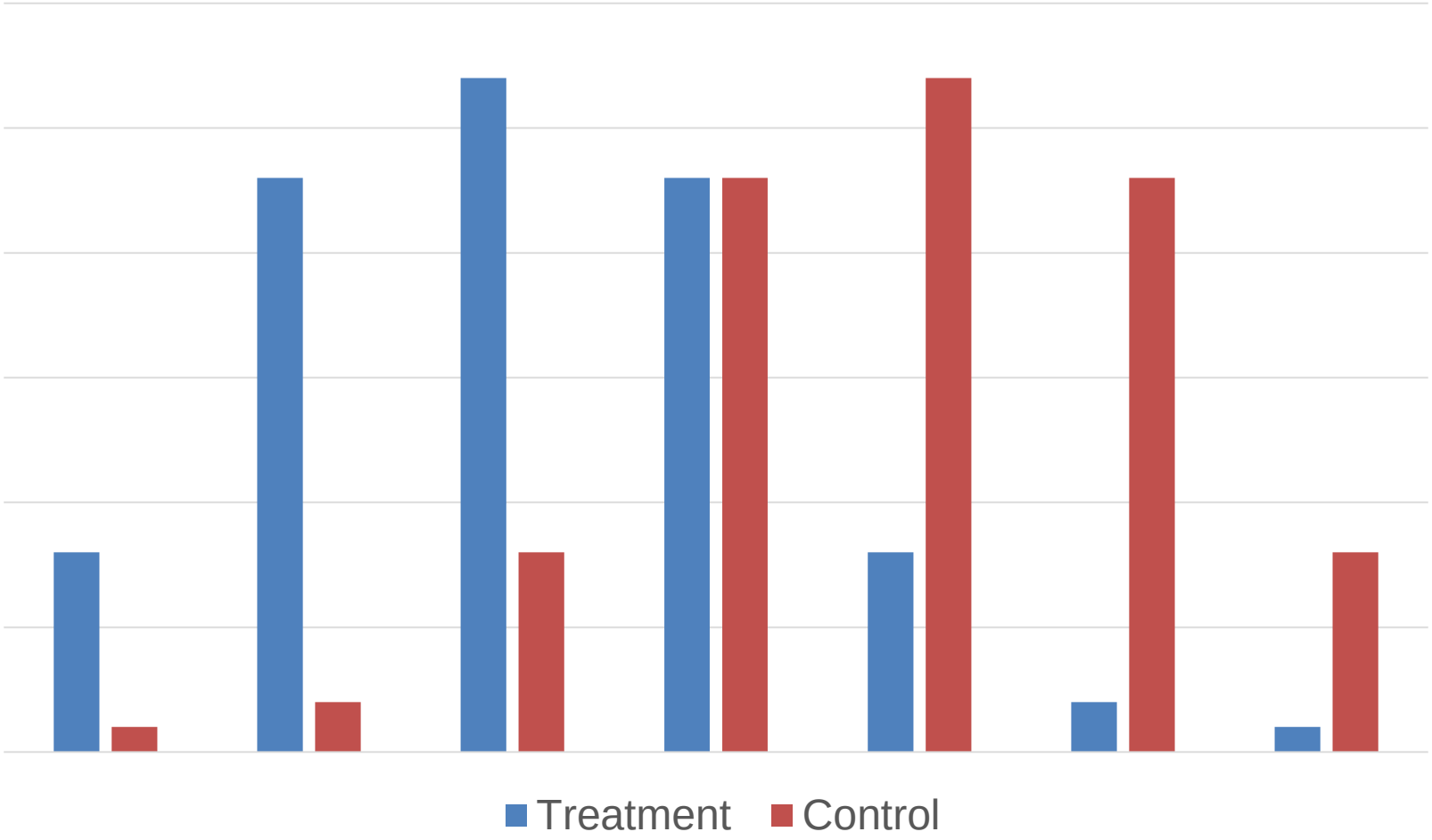




Student’s t-test

- Compares ***means***
- Assumes normal distribution

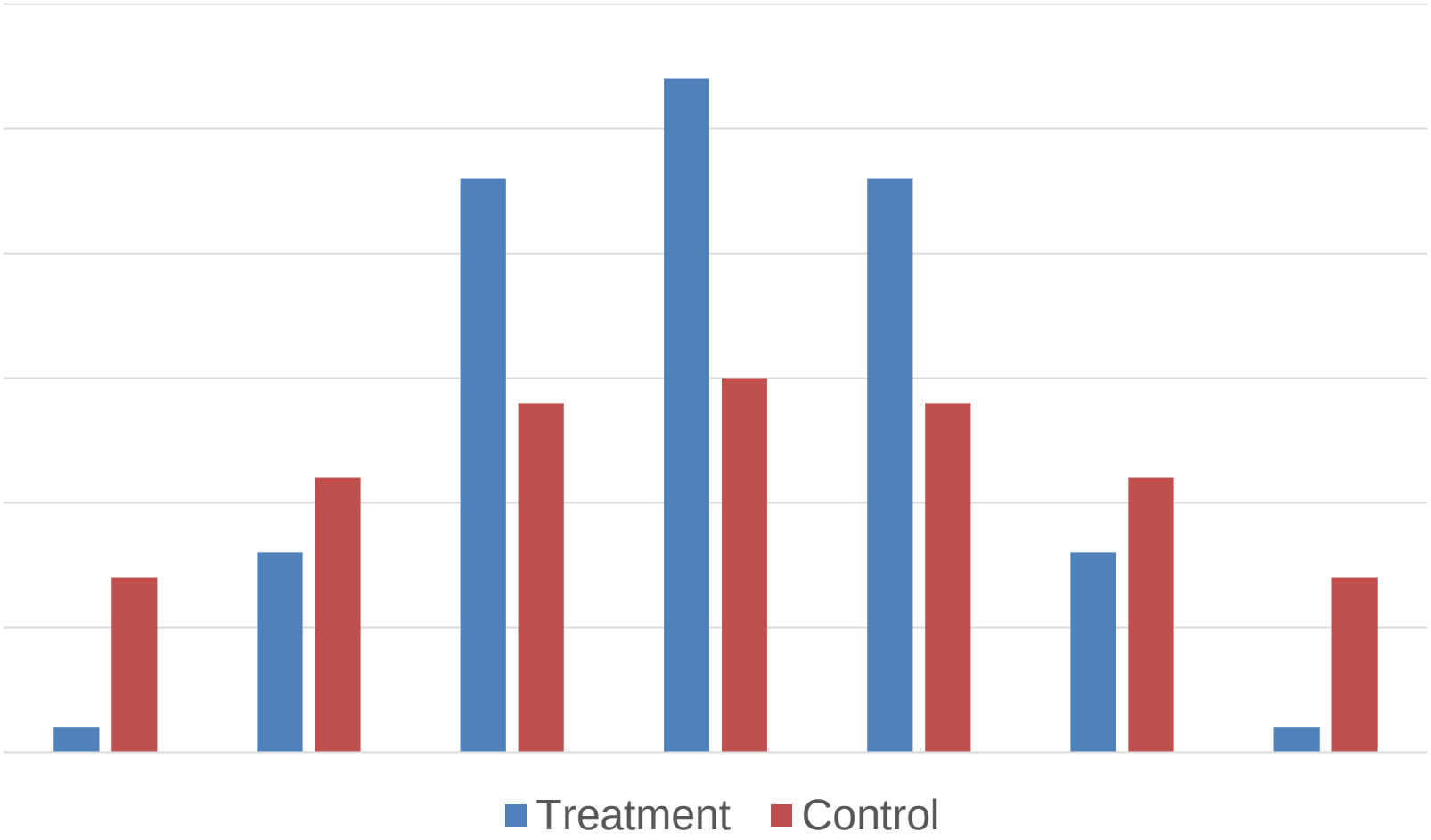
Student’s t-test



F-test

- Compares ***variance***
- Assumes normal distribution

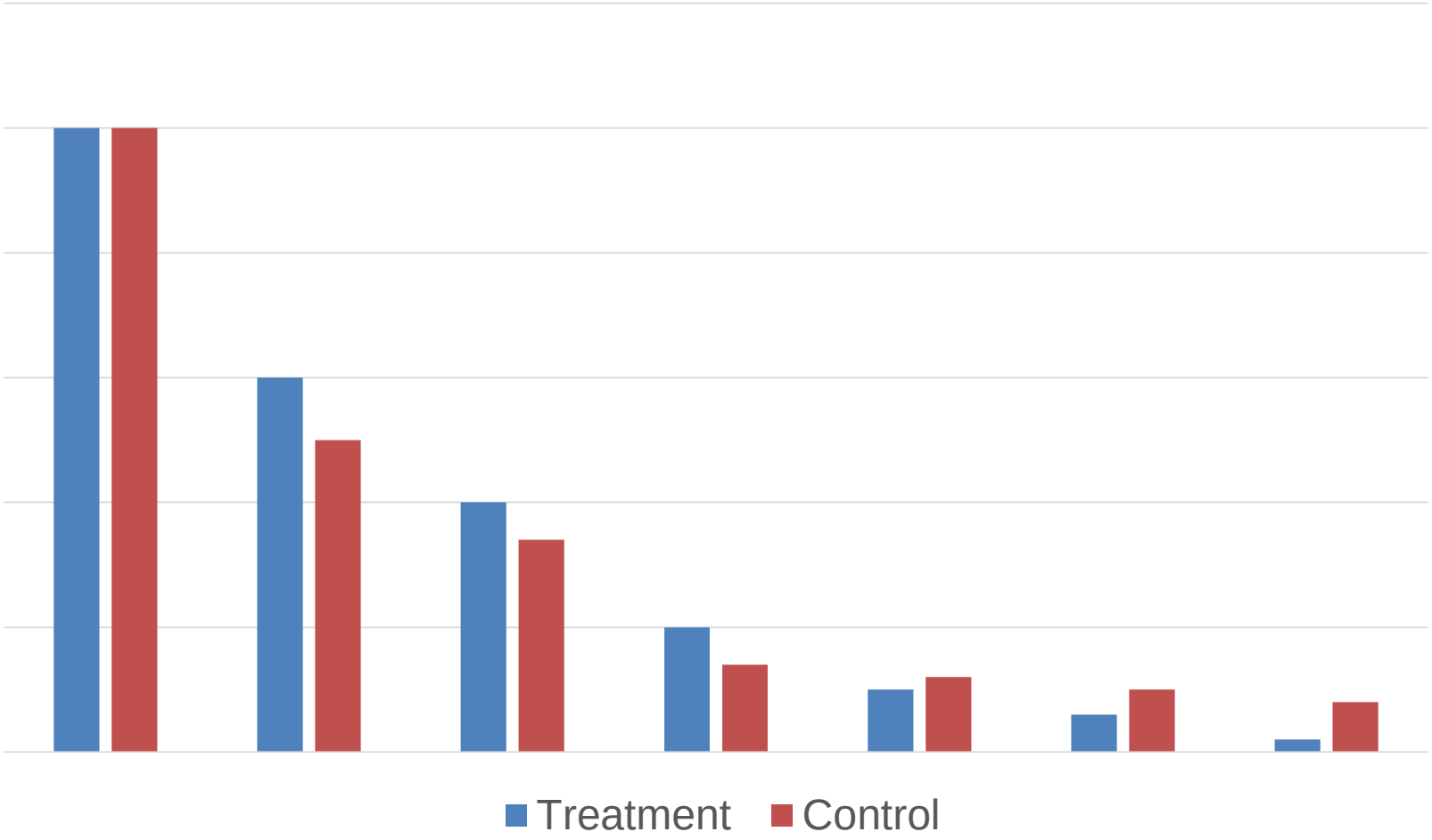
F-test



U-test

- Compares ***ranks***
- No assumption of normality

Mann-Whitney U-test
(aka Wilcoxon-Kruskal-Wallis two-tailed test)



How to run U-test

- Download 30-day free trial of JMP
  - Can also pay $15 for one year Cornell student license
- Can also use Excel or SPSS or R or …

Reporting U-test results

- Medians
- Z-statistic
- p-value

Group activity

- Brainstorm A/B tests!
  - What is your hypothesis?
  - What do you need to measure?
  - What do you need to log?