Unit 1: Lesson 4: Project 4: Evaluate an Experiment Analysis

Identify the flaws in each experiment and analysis, and describe what you would do to correct them.

Sith Lords: Recruiting Slogan

The Sith Lords are concerned that their recruiting slogan, "Give In to Your Anger," isn't very effective. Darth Vader develops an alternative slogan, "Together We Can Rule the Galaxy."

They compare the slogans on two groups of 50 captured droids each. In one group, Emperor Palpatine delivers the "Anger" slogan. In the other, Darth Vader presents the "Together" slogan.

20 droids convert to the Dark Side after hearing Palpatine's slogan, while only 5 droids convert after hearing Vader's. The Sith's data scientist concludes that "Anger" is a more effective slogan and should continue to be used.

Flaws:

1. Potential assignment bias: Technically, there are 5 different classes of droids. It's not clear if all 50 droids within and across groups were the same class/kind of droid. Or, if they were different, there is no mention of randomly assigning droids to group 1 and group 2. The different classes/kinds of droids who watched the slogans may have used different algorithms to compute their understanding of the slogan based on the kind of droid they were.
2. Condition bias: The conditions for each slogan's presentation were different. In one group, Emperor Palpatine delivered the slogan. In the other group, Darth Vader. It's possible the "Anger" slogan was less effective, but Emperor Palpatine was more persuasive as a speaker.

Adjustments:

1. Randomly assign the droids to each exposure group to minimize any potential differences between the droids.
2. Two possible ways to adjust for the bias introduced by having differing conditions between slogans (Emperor Palatine reading vs Darth Vader reading):
    a. Adjustment 1: Have both slogans read by either Emperor Palpatine or Darth Vader across both groups.

b. Adjustment 2: Capture another 100 droids that are randomly assigned to two groups and have Emperor Palpatine and Darth Vader switch the slogans they read.

Jedi: Public Relations

In the past, the Jedi have had difficulty with public relations. They send two envoys, Jar Jar Binks and Mace Windu, to four friendly and four unfriendly planets respectively, with the goal of promoting favorable feelings toward the Jedi.

Upon their return, the envoys learn that Jar Jar was much more effective than Windu: Over 75% of the people surveyed said their attitudes had become more favorable after speaking with Jar Jar, while only 65% said their attitudes had become more favorable after speaking with Windu.

This makes Windu angry, because he is sure that he had a better success rate than Jar Jar on every planet. The Jedi choose Jar Jar to be their representative in the future.

Flaws:

1. There is no clear definition or metric for “Friendly” vs “Unfriendly” planet. Either of these assignments exist on a continuum, so a planet that is “Very Unfriendly” will likely have less potential for having their attitudes swayed by a particular envoy vs a planet that is only “Slightly Unfriendly”.
2. There is a potential bias in assignment. If either of the envoys speaks the language of the “Unfriendly” assigned planet, or has family, or prior business relationships on that planet, it could affect his ability to “win over” the hearts and minds of that planet’s population.
3. There is no mention of the content of the envoy’s message. If Jar Jar promised each “Unfriendly” planet changes that they wanted, but Windu promised negotiation on the concerns his Unfriendly Planets had, there would be a different impact.
4. There is also no mention of the survey sampling and whether the methods used would be biased toward unfriendly vs. unfriendly responses to each of the envoy messages. Who does the surveys? Are the methods consistent?
5. Simpson’s Paradox: There is no discussion of planet population size - it is a lurking variable. If Jar Jar went to planets with smaller populations overall as compared to Windu, then any average would be misleading. Jar Jar’s mixed averages would look better than Windu’s averages purely because he had a fewer number of people whose attitudes he had to sway. Also, if there were a significant difference in population size between Friendly vs Unfriendly planets assigned to Jar Jar and Windu that would also be misleading. For example, if Jar Jar were able able to sway 75% of 100,000 people on his “Unfriendly” planets but Windu was able to sway 65% of 500,000 people on his

"Unfriendly" planets, then Windu would have been more effective at swaying more people than Jar Jar, even though Jar Jar's average would be higher.

Adjustments:

1. Define "Friendly" vs. "Unfriendly" and rate each planet on a scale. Use this scale to assign weights to survey results and adjust for difficulty in analysis. Assumption: the more "Unfriendly" the harder to sway.
2. Identify potential individual factors that might affect each envoy's ability to be successful in swaying people's attitudes. Use these factors to weight each individual's strengths/limitations and adjust for them in the statistical analysis. Assumption: the more favorable factors, the higher the difficulty in swaying the attitudes of "Unfriendly" planets
3. Standardize the envoy messages and have clear boundaries/limits to what either envoy can promise the people of the planets visited. If the messages represent the same accepted Jedi policies across planets, then swaying attitudes will have a higher likelihood of being a result of envoy approach vs. individual promises/content that may not necessarily represent Jedi policy
4. Use the Interplanetary Science Confederation to complete surveys and sampling for the study. Collection methods and surveys will be standardized across planets and will minimize bias due to planetary politics.
5. Adjust for population size in the statistical analysis of the planets visited.

IT and HR Job Satisfaction

A company with work sites in five different countries has sent you data on employee satisfaction rates for workers in Human Resources and workers in Information Technology.

Most HR workers are concentrated in three of the countries, while IT workers are equally distributed across worksites.

The company requests a report on satisfaction for each job type. You calculate average job satisfaction for HR and for IT and present the report.

Flaws:

1. Potential bias due to differences in sample groups. There is no mention if rates are presented "side-by-side" as though they are comparable or they can be compared against each other. Job satisfaction rates for each group can't be compared against each other or presented side-by-side by job type only because the groups also differ by their concentrations at each location. IT is distributed equally across 5 locations. HR is concentrated in 3 different locations. There are potentially multiple variables (cultural factors, language, local leadership style, city amenities, country of origin, living

accommodations, etc.) that change due to the location where each job type is placed and that can affect perceptions of job satisfaction.
2. There is no mention of numbers of employees by HR or IT job type. If the “n” of each group is significantly different, this would be another reason that it would be misleading to present job satisfaction ratings of both jobs side-by-side.

Adjustments:

1. Do A/A testing. Treat each job role separately and compare against itself by location. They are different jobs in different locations with potentially different numbers of respondents. By comparing IT against itself and HR against itself by location, you’ll be isolating location as the primary exposure factor that may affect differences in satisfaction. You’ll eliminate job type, potential differences in sample size, and impact of differences in concentrations of workers in some locations vs an even spread as potential variables affecting job satisfaction.

Happy Days Fitness Tracker App Opt In

When people install the Happy Days Fitness Tracker app, they are asked to "opt in" to a data collection scheme where their level of physical activity data is automatically sent to the company for product research purposes.

During your interview with the company, they tell you that the app is very effective because after installing the app, the data show that people's activity levels rise steadily.

Flaws:

1. There is no baseline (pre-app installation) against which Happy Days Fitness Tracker App can base their claims of effectiveness. There is no way to know pre-app installation exercise levels. It could be that people who get fitness tracker apps are already people who exercise regularly, and that they would’ve increased their exercise habits with or without the app. It could also be that the app itself has a flaw in it that calculates exercise/activity levels incorrectly. Since people’s lives tend to ebb and flow, with levels of exercise, it seems somewhat odd that activity levels rise steadily.
2. Also, it is possible that after “opting in” to the data collection subscribers are motivated to perform to make sure they “look good” - this could bias the results.
3. Effective at what? “Effective” is not defined in terms of the app’s purpose. Effective at tracking activity? If so, how is this supposed to help subscribers maintain or increase activity levels? Effective at assisting subscribers at maintaining exercise levels?

Adjustments:

1. Determine if rate of increase in activity is so regular and consistent across the population that a “steady increase” may be due to software related issue/programmatic issue that produces inaccurate data.
2. Define “effective” in terms of the purpose of the app. Difficult to design a way to test efficacy of app without this information.

Teacher Cheats Cheaters From Cheating

To prevent cheating, a teacher writes three versions of a test. She stacks the three versions together, first all copies of Version A, then all copies of Version B, then all copies of Version C.

As students arrive for the exam, each student takes a test.

When grading the test, the teacher finds that students who took Version B scored higher than students who took either Version A or Version C. She concludes from this that Version B is easier, and discards it.

Flaws:

1. There is no mention of how many of each version of the test the teacher created. If there are fewer Version B tests than the other versions, then it is possible that only “smart” students or students who studied harder for the exam picked up that version. If there were fewer Version B tests, then it would’ve been more likely for this scenario to occur.
2. There is a bias in selection for the Versions of the test. The versions of the test were not assigned to students randomly. Since students often walk into class with friends or people with whom they are closer, it’s possible that “smarter” students who are friends or students who studied more for the exam came in together and got the same version.
3. It is also possible that students who got Version B of the test arrived together and sat down adjacent to one another and were still able to cheat from each other.
4. It could’ve been a fluke that all students who took Version B of the test did well. Without using the same versions of the test on subsequent classes, it would be difficult to know if Version B test takers did well on the exam due to chance or if it was due to “ease” of the test or for some other reason.

Adjustments:

1. Lay out the exams on desks prior to students arriving so that students with the same version of the test are not sitting adjacent to one another.
2. Repeat method for subsequent classes to compare results