Now it's time to flex your critical evaluation skills. Read the following descriptions of an experiment and its analysis, identify the flaws in each, and describe what you would do to correct them.

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.

Two versions: The control group is the Palpatine group, while the "treatment" group is the Darth Vader group.

A sample, divided into two groups: 50 captured droids. These are 50 droids who were captured somehow, either in battle or overtaken by some other way. We assume that they are hostile to the Empire. Overall, I do not find fault with the two groups, in that I hope that both sets of sample populations were picked with randomization so that both sets of groups are as much alike as possible. It would be interesting to note if the 20 droids that converted to the Dark Side in the Palpatine group had been there longer than the droids in the Darth Vader side.

A hypothesis: The hypothesis here would be that using the new slogan will result in more droids converting to the Dark Side.

Outcome(s) of interest: The key metric here is the number of droids converting to the Dark Side.

Other measured variables: How was the slogan delivered? Were there armed Stormtroopers behind the Emperor pointing weapons at the droids? Is the Emperor more menacing than Darth Vader? Do the droids view the Emperor as more authoritative than Darth Vader.

What I would do would be to randomly pick 2 groups composed of 50 droids each and I would ensure that all of them are able to understand the language that is used and that all have been there an equal amount of time and have been treated more or less the same way and have been kept together the whole time they have been there. Next, after everyone has been fed breakfast and has not been treated maliciously or wrongly, I would lead the two groups, at the same time, to different chambers. There would no guards or menacing-looking people. I would next have a government

official deliver the slogans to each group at the same time. Then, I would see how many convert to the Dark Side.

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.

Two versions: Jar Jar Binks was sent to four friendly planets and Mace Windu was sent to four unfriendly planets. The goal is to promote favorable feelings toward the Jedi. This is comparable to two different test versions to compare to each other.

A sample, divided into 2 groups: One group would be the beings on the friendly planet that were visited by Jar Jar Binks and the other group would the beings on the unfriendly planets that were visited by Mace Windu.

A hypothesis: This is a testing of two different test versions and the hypothesis here is the outcome of the two different test versions.

Outcome of interest: The promotion of favorable feelings towards the Jedi is the outcome of interest. This is the key metric.

Other measured variables: Which planets were visited? How many people were seen by these respective ambassadors? What time of day did the ambassadors talk? Did the audience speak the same language? Was it a pro-Jedi planet or anti-Jedi planet?

The best way to fix this one would be for Jar Jar Binks and Mace Windu to each visit 2 pro-Jedi planets and 2 anti-Jedi planets. I would have them visit each planet in a pro-anti-pro-anti fashion. Also both would see an equal amount of people.

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.

Two versions: The company requests job satisfaction rates for people in HR and IT.

A sample, divided into 2 groups: One sample would be obtained from the HR employees and one sample from the IT employees. The company has employees in 5 different countries.

A hypothesis: This is a testing of two different test versions and the hypothesis here is the job satisfaction of each type of employees.

Outcome of interest: The average job satisfaction of HR and IT employees is the outcome of interest. This is the key metric.

Other measured variables: Job location, work hours, starting time, ending time, lunch time, age of the employee, gender of employee, type of employment (telecommuter, part-time, full-time, contract, salary, hourly wage), overall pay.

The best way to fix this one would be stratify each group according to location, job type, employment type, gender, and overall pay. Then I would get the average of each type of employee for each location and then get an equal number of employees from each location, based on stratification. Then for each location, compute how much overall % of employees are located in that location and then get a respective sample size from that location to represent that location, i.e. representative employee subtypes created then added together.

***Two groups - IT&HR. Within each group you compare lookalike employees with same/similar characteristics to achieve a job satisfaction score/rate and then take the whole by combining all job satisfaction/rates.

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.

***Find a way to neutralize the effect of the pro-workout group of people. The inherent problem here is that people who are going to be starting a

fitness plan or are avid fitness enthusiasts are the ones who will be downloading and using this app. So therefore, there is a huge selection bias when people are downloading this. Of course the company tells you that it is effective, that's because active people are the ones downloading it and using it. Also, people who have the mindset that they are going to be exercising are the ones who will be opting in to the program. People that are going to be lazy and not doing anything will not be opting in to the program. The key here is to somehow monitor the people who did not opt in, but that would be illegal.

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.

Three versions: A, B, C

A sample divided into 3 groups - students taking Version A, Version B, Version C.

Hypothesis and outcome of interest would be which version of the test most predicts that someone cheated?

Other measured variables: average gpa of the other students, number of students who took each version, average gpa of students by arrival time to the test center.

A better way to perform this would be to randomly intermix the versions of the test so that the students, as they come in, pick randomly what versions they are getting. Also, the teacher needs to make sure that students who are sitting next to one another or in close proximity do not have the same version.