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.

1. 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: The experiment doesn't whether the two groups of 50 captured droids are similar.

CORRECTIONS: Extract two groups of 50 captured droids from a random sample. Deliver both slogans to groups to understand recruiter/delivery impact.

1. 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: Self-reporting may introduce bias, and not disaggregating their individual effectiveness ratings by planet friendliness/unfriendliness leaves lurking variable(s) such as key talking points undetected - public relations is as much about altering the negative as it is about reinforcing the positive. CORRECTIONS: Give both envoys talking points that reinforce positive perception (for use with friendly planets) and talking points that address/alter negative perception (for use with unfriendly planets). Measure perception change indirectly (via social media reaction to the envoys visits for example) within an impact window for both envoys for friendly/unfriendly planets.

1. 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: Self-reporting may introduce bias and averaging doesn't account for the functional concentration differences by worksite, the resulting workload impacts, or other country-related factors that may be influencing employee satisfaction.

CORRECTIONS: Measure employee satisfaction indirectly (via attrition rates for example). Compare employee satisfaction within the IT function, focusing on worksite/country-related factors. Compare employee satisfaction within the HR function, focusing on concentration across worksites/countries. Create a heatmap that ordinally communicates employee satisfaction (as a numerical annotation - 1 through 10 for example), uses both axes for function-worksite/country combinations, and uses color to depict concentration.

1. 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: Self-selection bias means that the users opting into the data collection scheme may not be representative of the entire user base, as well as that their behavior may be driven less by the app's effectiveness and more by the fact that they are self-conscious about being monitored.

CORRECTIONS: Automatic data collection should be a precondition (rather than a postcondition) for app use/installation, so that the entire population's activity is measured and so that the likelihood that they're able to sustain behavior manipulation over an indefinite time horizon (defined as attempting to game the data collection system for ego/perception-driven reasons) is low.

1. 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: Assuming all three versions had unique questions and scrambled answer choices, the observed effect may be due to version clustering within the room.

CORRECTIONS: Randomize the seating arrangement as well the order in which the tests are handed out.